The Casualty Actuarial Society *Forum*

Winter 2003 Edition

Including the Data Management Call Papers and Ratemaking Discussion Papers

To CAS Members:

This is the Winter 2003 Edition of the Casualty Actuarial Society *Forum*. It contains six Data Management, Quality, and Technology Call Papers, nine Ratemaking Discussion Papers, and seven additional papers.

The Casualty Actuarial Society *Forum* is a nonrefereed journal printed by the Casualty Actuarial Society. The viewpoints published herein do not necessarily reflect those of the Casualty Actuarial Society.

The CAS *Forum* is edited by the CAS Committee for the Casualty Actuarial Society *Forum*. Members of the committee invite all interested persons to submit papers on topics of interest to the actuarial community. Articles need not be written by a member of the CAS, but the paper's content must be relevant to the interests of the CAS membership. Members of the Committee for the Casualty Actuarial Society *Forum* request that the following procedures be followed when submitting an article for publication in the *Forum*:

1. Authors should submit a camera-ready original paper and two copies.
2. Authors should not number their pages.
3. All exhibits, tables, charts, and graphs should be in original format and camera-ready.
4. Authors should avoid using gray-shaded graphs, tables, or exhibits. Text and exhibits should be in solid black and white.
5. Authors should submit an electronic file of their paper using a popular word processing software (e.g., Microsoft Word and WordPerfect) for inclusion on the CAS Web Site.

The CAS *Forum* is printed periodically based on the number of call paper programs and articles submitted. The committee publishes two to four editions during each calendar year.

All comments or questions may be directed to the Committee for the Casualty Actuarial Society *Forum*.

Sincerely,

Joseph A. Smalley

Joseph A. Smalley, CAS *Forum* Chairperson

The 2003 CAS Data Management, Quality, and Technology Call Papers and Ratemaking Discussion Papers Presented at the 2003 Ratemaking Seminar March 27-28, 2003 San Antonio Marriott Rivercenter San Antonio, Texas

The Winter 2003 Edition of the CAS *Forum* is a cooperative effort between the CAS *Forum* Committee, the CAS Committee on Management Data and Information, and the CAS Committee on Ratemaking.

The CAS Committee on Management Data and Information presents for discussion six papers prepared in response to its Call for 2003 Data Management, Quality, and Technology Papers.

The CAS Committee on Ratemaking presents for discussion nine papers prepared in response to its Call for 2003 Ratemaking Discussion Papers.

This *Forum* includes papers that will be discussed by the authors at the 2003 CAS Ratemaking Seminar, March 27-28, 2003, in San Antonio, Texas.

Data Management Call Papers

Ratemaking Discussion Papers

Additional Papers

Additional Papers, continued

Applying Data Mining Techniques in Property/Casualty Insurance

Lijia Guo, Ph.D., ASA

Applying Data Mining Techniques in Property/Casualty Insurance

Lijia Guo, Ph.D., A.S.A.
University of Central Florida

Abstract

This paper addresses the issues and techniques for Property/Casualty actuaries using data mining techniques. Data mining means the efficient discovery of previously unknown patterns in large databases. It is an interactive information discovery process that includes data acquisition, data integration, data exploration, model building, and model validation. The paper provides an overview of the information discovery techniques and introduces some important data mining techniques for application to insurance including cluster discovery methods and decision tree analysis.

1. Introduction

Because of the rapid progress of information technology, the amount of information stored in insurance databases is rapidly increasing. These huge databases contain a wealth of data and constitute a potential goldmine of valuable business information. As new and evolving loss exposures emerge in the ever-changing insurance environment, the form and structure of insurance databases change. In addition, new applications such as dynamic financial analysis and catastrophe modeling require the storage, retrieval, and analysis of complex multimedia objects, which are often represented by high-dimensional feature vectors. Finding the valuable information hidden in those databases and identifying appropriate models is a difficult task.

Data mining (DM) is the process of exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns and rules (Berry and Linoff, 2000). A typical data mining process includes data acquisition, data integration, data exploration, model building, and model validation. Both expert opinion and data mining techniques play an important role at each step of this information discovery process.

This paper introduces two important data mining techniques for application to insurance: cluster discovery methods and decision tree analysis.

Cluster analysis is one of the basic techniques that are often applied in analyzing large data sets. Originating from the area of statistics, most cluster analysis algorithms have originally been developed for relatively small data sets. In recent years, the clustering algorithms have been extended to efficiently work on large data sets, and some of them even allow the clustering of high-dimensional feature vectors (see Ester, Kriegel, Sander, and Xu, and Hinneburg, and Keim, 1998, for example).

Decision tree analysis is another popular data mining technique that can be used in many areas of actuarial practice. We discuss how to use decision trees to make important design decisions and explain the interdependencies among the properties of insurance data. We will also provide examples of how data mining techniques can be used to improve the effectiveness and efficiency of the modeling process.

The paper is organized as follows. Section 2 provides an overview of data mining and a list of potential DM applications to insurance. Section 3 demonstrates the cluster analysis data mining techniques. Section 4 presents application of predictive data mining process. This section identifies factors that influence auto insurance claims using decision tree techniques and quantifies the effects and interactions of these risk factors using logistic regression. Model assessment is also discussed in this section. Section 5 concludes the paper.

2. Data Mining

In this section, we will provide an overview of the data mining process (2.1), data mining operations (2.2), data mining techniques and algorithms (2.3), and their potential applications in the insurance industry (2.4).

2.1 Data Mining Process

Data mining combines techniques from machine learning, pattern recognition, statistics, database theory, and visualization to extract concepts, concept interrelations, and interesting patterns automatically from large corporate databases. Its primary goal is to extract knowledge from data to support the decision-making process. Two primary functions of data mining are: *prediction,* which involves finding unknown values/relationships/patterns from known values; and *description,* which provides interpretation of a large database.

A data mining process generally includes the following four steps.

STEP 1: Data acquisition. The first step is to select the types of data to be used. Although a target data set has been created for discovery in some applications, DM can be performed on a subset of variables or data samples in a larger database.

STEP 2: Preprocessing data. Once the target data is selected, the data is then preprocessed for cleaning, scrubbing, and transforming to improve the effectiveness of discovery. During this preprocessing step, developers remove the noise or outliers if necessary and decide on strategies for dealing with missing data fields and accounting for time sequence information or known changes. In addition, the data is often transformed to reduce the effective number of variables under consideration by either converting one type of data to another (e.g., categorical values into numeric ones) or deriving new attributes (by applying mathematical or logical operators).

STEP 3: Data exploration and model building. The third step of DM refers to a series of activities such as deciding on the type of DM operation; selecting the DM technique; choosing the DM algorithm; and mining the data. First, the type of DM operation must be chosen. The DM operations can be classified as classification, regression, segmentation, link analysis, and deviation detection (see Section 2.2 for details). Based on the operation chosen for the application, an appropriate data-mining technique is then selected. Once a data-mining technique is chosen, the next step is to select a particular algorithm within the DM technique chosen. Choosing a data-mining algorithm includes a method to search for patterns in the data, such as deciding which models and parameters may be appropriate and matching a particular data-mining technique with the overall objective of data mining. After an appropriate algorithm is selected, the data is finally mined using the algorithm to extract novel patterns hidden in databases.

STEP 4: Interpretation and evaluation. The fourth step of the DM process is the interpretation and evaluation of discovered patterns. This task includes filtering the information to be presented by removing redundant or irrelevant patterns, visualizing graphically or logically the useful ones, and translating them into understandable terms by users. In the interpretation of results, we determine and resolve potential conflicts with previously found knowledge or decide to redo any of the previous steps. The extracted knowledge is also evaluated in terms of its usefulness to a decision maker and to a business goal. Then extracted knowledge is subsequently used to support human decision making such as prediction and to explain observed phenomena.

The four-step process of knowledge discovery should not be interpreted as linear, but as an interactive, iterative process through which discovery evolves.

2.2 Data Mining Operations

Assuming you have prepared a data set for mining, you then need to define the scope of your study and choose the subject of your study. This is referred as choosing a DM operation.

There are five types of DM operations: classification, regression, link analysis, segmentation, and deviation detection. Classification and regression are useful for prediction, whereas link analysis, segmentation, and deviation detection are for description of patterns in the data. A DM application typically requires the combination of two or more DM operations.

<u>Classification</u>
The goal of classification is to develop a model that maps a data item into one of several predefined classes. Once developed, the model is used to classify a new instance into one of the classes. Examples include the classification of bankruptcy patterns based on the financial ratios of a firm and of customer buying patterns based on demographic information to target the advertising and sales of a firm effectively toward the appropriate customer base.

<u>Regression</u>
This operation builds a model that maps data items into a real-valued prediction variable. Models have traditionally been developed using statistical methods such as linear and logistic regression. Both classification and regression are used for prediction. The distinction between these two models is that the output variable of classification is categorical, whereas that of

Table 1. DM Techniques for DM Operations

DM Technique	Induction	Neural Networks	Genetic Algorithms	Clustering	Logistic Regression	Association Discovery	Sequence Discovery	Visualization
DM Operation								
Classification	x	x	x					
Regression	x	x			x			
Link analysis		x	x			x	x	
Segmentation		x	x	x				
Deviation					x			x

Induction Techniques

Induction techniques develop a classification model from a set of records -- the training set of examples. The training set may be a sample database, a data mart, or an entire data warehouse. Each record in the training set belongs to one of many predefined classes, and an induction technique induces a general concept description that best represents the examples to develop a classification model. The induced model consists of patterns that distinguish each class. Once trained, a developed model can be used to predict the class of unclassified records automatically. Induction techniques represent a model in the form of either decision trees or decision rules. These representations are easier to understand, and their implementation is more efficient than those of neural network or genetic algorithms. A more detailed discussion on decision tree techniques and their applications will be presented in Section 4.

Neural Networks

Neural networks constitute the most widely used technique in data mining. They imitate the way the human brain learns and use rules inferred from data patterns to construct hidden layers of logic for analysis. Neural networks methods can be used to develop classification, regression, link analysis, and segmentation models. A neural net technique represents its model in the form of nodes arranged in layers with weighted links between the nodes. There are two general categories of neural net algorithms: supervised and unsupervised.

- Supervised neural net algorithms such as Back propagation (Rumelhart, Hinton, and Williams, 1986) and Perceptron require predefined output values to develop a classification model. Among the many algorithms, Back propagation is the most popular supervised neural net algorithm. Back propagation can be used to develop not only a classification model, but also a regression model.

- Unsupervised neural net algorithms such as ART (Carpenter and Grossberg, 1988) do not require predefined output values for input data in the training set and employ self-organizing learning schemes to segment the target data set. Such self-organizing networks divide input examples into clusters depending on similarity, each cluster representing an unlabeled category. Kohonen's Feature Map is a well-known method in self-organizing neural networks.

For organizations with a great depth of statistical information, neural networks are ideal because they can identify and analyze changes in patterns, situations, or tactics far more regression is numeric and continuous. Examples of regression are the prediction of change between the yen and the Government Bond Market and of the crime rate of a city based on the description of various input variables such as populations, average income level and education.

Link Analysis

Link analysis is used to establish relevant connections between database records. Its typical application is market-basket analysis, where the technique is applied to analyze point-of-sales transaction data to identify product affinities. A retail store is usually interested in what items sell together -- such as baby's diapers and formula -- so it can determine what items to display together for effective marketing. Another application could find relationships among medical procedures by analyzing claim forms submitted to an insurance firm. Link analysis is often applied in conjunction with database segmentation.

Segmentation

The goal is to identify clusters of records that exhibit similar behaviors or characteristics hidden in the data. The clusters may be mutually exclusive and exhaustive or may consist of a richer representation such as hierarchical or overlapping categories. Examples include discovering homogenous groups of consumers in marketing databases and segmenting the records that describe sales during "Mother's Day" and "Father's Day." Once the database is segmented, link analysis is often performed on each segment to identify the association among the records in each cluster.

Deviation Detection

This operation focuses on discovering interesting deviations. There are four types of deviation:

- Unusual patterns that do not fit into previously measured or normative classes,
- Significant changes in the data from one time period to the next,
- Outlying points in a dataset -- records that do not belong to any particular cluster, and
- Discrepancies between an observation and a reference.

Deviation Detection is usually performed after a database is segmented to determine whether the deviations represent noisy data or unusual casualty. Deviation detection is often the source of true discovery since deviations represent anomaly from some known expectation or norm.

2.3 Data Mining Techniques and Algorithms

At the heart of DM is the process of building a model to represent the data set and to carry out the DM operation. A variety of DM techniques (tools) are available to support the five types of DM operations presented in the previous section. The most popular data mining techniques include Bayesian analysis (Cheeseman et al., 1988), neural networks (Bishop, 1995; Ripley, 1996), genetic algorithms (Goldberg, 1989), decision trees (Breiman et al., 1984), and logistic regression (Hosmer and Lemeshow, 1989), among others.

Table 1 summarizes the DM techniques used for DM operations. For each of the DM techniques listed in Table 1, there are many algorithms (approaches) to choose from. In the following, some of the most popular technologies are discussed.

quickly than any human mind. Although the neural net technique has strong representational power, interpreting the information encapsulated in the weighted links can be very difficult. One important characteristic of neural networks is that they are opaque, which means there is not much explanation of how the results come about and what rules are used. Therefore, some doubt is cast on the results of the data mining. Francis (2001) gives a discussion on Neural Network applications to insurance problems.

Genetic Algorithms

Genetic algorithms are a method of combinatorial optimization based on processes in biological evolution. The basic idea is that over time, evolution has selected the "fittest species." For a genetic algorithm, one can start with a random group of data. A *fitness function* can be defined to optimizing a model of the data to obtain "fittest" models. For example, in clustering analysis, a fitness function could be a function to determine the level of similarity between data sets within a group.

Genetic algorithms have often been used in conjunction with neural networks to model data. They have been used to solve complex problems that other technologies have a difficult time with. Michaelewicz (1994) introduced the concept of genetic algorithms and applying them with data mining.

Logistic Regression

Logistic regression is a special case of generalized linear modeling. It has been used to study odds ratios (e^{β_j}, $j = 1, 2, \cdots, k$ as defined in the following), which compares the odds of the event of one category to the odds of the event in another category, for a very long time and its properties have been well studied by the statistical community. Ease of interpretation is one advantage of modeling with logistic regression. Assume that the data set consist of i = 1, 2, ..., n records. Let $p_i, i = 1, 2, \cdots, n$ be the corresponding mortality rate for each record and $x_i = (x_{1i}, x_{2i}, \cdots, x_{ki})$ be a set of k variables associated with each record. A linear-additive logistic regression model can be expressed as

$$\text{logit} = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ji}, \text{ where } i = 1, 2, \cdots, n.$$

If the model is correctly specified, each dependent variable affects logit linearly. Exponentiation of the parameter estimate of each slope, e^{β_j}, $j = 1, 2, \cdots, k$, can be interpreted as the odds ratio of the probability that p_i is associated with input variable x_{ji} (Kleinbaum, D., Kupper, L., and Muller, K., 1988). However, it poses several drawbacks especially with large data sets. The curse of dimensionality makes the detection of nonlinearities and interactions difficult. If the model is not correctly specified, the interpretation of the model parameter estimates becomes meaningless. In addition, the data might not be evenly distributed among the whole data space. It is very likely that some segments of the data space have more records than other segments. One model that fits the whole data space might not be the best choice depending on the intended application. Although there are many existing methods such as backward elimination and forward selection that can help data analyst to build logistic regression model, judgment should be exercised regardless of the method selected.

Clustering

Clustering techniques are employed to segment a database into clusters, each of which shares common and interesting properties. The purpose of segmenting a database is often to summarize the contents of the target database by considering the common characteristics shared in a cluster. Clusters are also created to support the other types of DM operations, e.g. link analysis within a cluster. Section 3 will introduce more details of clustering and its application to insurance.

Associated Discovery

Given a collection of items and a set of records containing some of these items, association discovery techniques discover the rules to identify affinities among the collection of items as reflected in the examined records. For example, 65 percent of records that contain item A also contain item B. An association rule uses measures called "support" and "confidence" to represent the strength of association. The percentage of occurrences, 65 percent in this case, is the confidence factor of the association. The algorithms find the affinity rules by sorting the data while counting occurrences to calculate confidence. The efficiency with which association discovery algorithms can organize the events that make up an association or transaction is one of the differentiators among the association discovery algorithms. There are a variety of algorithms to identify association rules such as Apriori algorithm and using random sampling. Bayesian Net can also be used to identify distinctions and relationships between variables (Fayyad et al., 1996).

Sequence Discovery

Sequence discovery is very similar to association discovery except that the collection of items occurs over a period of time. A sequence is treated as an association in which the items are linked by time. When customer names are available, their purchase patterns over time can be analyzed. For example, it could be found that, if a customer buys a tie, he will buy men's shoes within one month 25 percent of the time. A dynamic programming approach based on the dynamic time warping technique used in the speech recognition area is available to identify the patterns in temporal databases (Fayyad et al., 1996).

Visualization

A picture is worth thousands of numbers! Visual DM techniques have proven the value in exploratory data analysis, and they also have a good potential for mining large databases. Visualizations are particularly useful for detecting phenomena hidden in a relatively small subset of the data. This technique is often used in conjunction with other DM techniques: features that are difficult to detect by scanning numbers may become obvious when the summary of data is graphically presented. Visualization techniques can also guide users when they do not know what to look for to discover the feature. Also, this technique helps end users comprehend information extracted by other DM techniques. Specific visualization techniques include projection pursuit and parallel coordinates. Tufte (1983, 1990) provided many examples of visualization techniques that have been extended to work on large data sets and produce interactive displays.

2.4 Using Data Mining in the Insurance Industry

Data mining methodology can often improve existing actuarial models by finding additional important variables, by identifying interactions, and by detecting nonlinear relationships. DM can help insurance firms make crucial business decisions and turn the new found knowledge into actionable results in business practices such as product development, marketing, claim distribution analysis, asset liability management and solvency analysis. An example of how data mining has been used in health insurance can be found in Borok, 1997. To be more specific, data mining can perform the following tasks.

Identify Risk Factors that Predict Profits, Claims and Losses
One critical question in ratemaking is the following: "What are the risk factors or variables that are important for predicting the likelihood of claims and the size of a claim?" Although many risk factors that affect rates are obvious, subtle and non-intuitive relationships can exist among variables that are difficult if not impossible to identify without applying more sophisticated analyses. Modern data mining models such as decision trees and Neural Networks can more accurately predict risk than current actuarial models, therefore insurance companies can set rates more accurately, which in turn can result in more accurate pricing and hence a better competitive position.

Customer Level Analysis
Successfully retaining customers requires analyzing data at the most appropriate level, the customer level, instead of across aggregated collections of customers. Using the Associated Discovery DM technique, insurance firms can more accurately select which policies and services to offer to which customers. With this technique insurance companies can:

- Segment the customer database to create customer profiles.
- Conduct rate and claim analyses on a single customer segment for a single product. For example, companies can perform an in-depth analysis of a potential new product for a particular customer segment.
- Analyze customer segments for multiple products using group processing and multiple target variables. For example, how profitable are bundles of policies (auto, home, and life) for certain customer segments of interest?
- Perform sequential (over time) market basket analyses on customer segments. For example, what percentage of new policyholders of auto insurance also purchases a life insurance policy within five years?

Database segmentation and more advanced modeling techniques enable analysts to more accurately choose whom to target for retention campaigns. Current policyholders that are likely to switch can be identified through predictive modeling. A logistic regression model is a traditional approach to predict those policyholders who have larger probabilities of switching. Identifying the target group for retention campaigns may be improved by modeling the behavior of policyholders.

Developing New Product Lines
Insurance firms can increase profitability by identifying the most lucrative customer segments and then prioritize marketing campaigns accordingly. Problems with profitability can occur if

firms do not offer the "right" policy or the "right" rate to the "right" customer segment at the "right" time. For example, for an insurer or reinsurer to use the log normal distribution for rating when the Pareto distribution is the true distribution would likely prove to be an expensive blunder, which illustrates the importance of having the right tool to identify and estimate the underlying loss distribution. With DM operations such as segmentation or association analysis, insurance firms can now utilize all of their available information to better develop new products and marketing campaigns.

Reinsurance

DM can be used to structure reinsurance more effectively than the using traditional methods. Data mining technology is commonly used for segmentation clarity. In the case of reinsurance, a group of paid claims would be used to model the expected claims experience of another group of policies. With more granular segmentation, analysts can expect higher levels of confidence in the model's outcome. The selection of policies for reinsurance can be based upon the model of experienced risk and not just the generalization that it is a long tailed book of business.

Estimating Outstanding Claims Provision

The settlement of claims is often subject to delay, so an estimate of the claim severity is often used until the actual value of the settled claim is available. The estimate can depend on the following:

- Severity of the claim.
- Likely amount of time before settlement.
- Effects of financial variables such as inflation and interest rates.
- Effects of changing social mores. For example, the tobacco industry has been greatly affected by the changing views toward smoking.

DM operations such as Link Analysis and Deviation Detection can be used to improve the claim estimation.

The estimate of the claims provision generated from a predictive model is based on the assumption that the future will be much like the past. If the model is not updated, then over time, the assumption becomes that the future will be much like the distant past. However, as more data become available, the predictive DM model can be updated, and the assumption becomes that the future will be much like the recent past. Data mining technology enables insurance analysts to compare old and new models and to assess them based on their performance. When the newly updated model outperforms the old model, it is time to switch to the new model. Given the new technologies, analysts can now monitor predictive models and update as needed.

An important general difference in the focus between existing actuarial techniques and DM is that DM is more oriented towards applications than towards describing the basic nature of the underlying phenomena. For example, uncovering the nature of the underlying individual claim distribution or the specific relation between drivers' age and auto type are not the main goal of Data Mining. Instead, the focus is on producing a solution that can improve the predictions for future premiums. DM is very effective in determining how the premiums related to multidimensional risk factors such as drivers' age and type of automobile. Two examples of applying data mining techniques in insurance actuarial practice will be presented in the next two sections.

3. Clustering - Descriptive Data Mining

Clustering is one of the most useful tasks in data mining process for discovering groups and identifying interesting distributions and patterns in the underlying data. Clustering problem is about partitioning a given data set into groups (clusters) such that the data points in a cluster is more similar to each other than points in different clusters (Guha et al., 1998). For example, segmenting existing policyholders into groups and associating a distinct profile with each group can help future rate making strategies.

Clustering methods perform disjoint cluster analysis on the basis of Euclidean distances computed from one or more quantitative variables and seeds that are generated and updated by the algorithm. You can specify the clustering criterion that is used to measure the distance between data observations and seeds. The observations are divided into clusters such that every observation belongs to at most one cluster.

Clustering studies are also referred to as unsupervised learning and/or segmentation. Unsupervised learning is a process of classification with an unknown target, that is, the class of each case is unknown. The aim is to segment the cases into disjoint classes that are homogenous with respect to the inputs. Clustering studies have no dependent variables. You are not profiling a specific trait as in classification studies.

A database can be segmented by:

- Traditional methods of pattern recognition techniques,
- Unsupervised neural nets such as ART and Kohonen's Feature Map,
- Conceptual clustering techniques such as COBWEB (Fisher, Pazzani and Langley, 1991) and UNIMEM, or
- A Bayesian approach like AutoClass (Chessman, 1996).

Conceptual clustering algorithms consider all the attributes that characterize each record and identify the subset of the attributes that will describe each created cluster to form concepts. The concepts in a conceptual clustering algorithm can be represented as conjunctions of attributes and their values. Bayesian clustering algorithms automatically discover a clustering that is maximally probable with respect to the data using a Bayesian approach. The various clustering algorithms can be characterized by the type of acceptable attribute values such as continuous, discrete or qualitative; by the presentation methods of each cluster; and by the methods of organizing the set of clusters, either hierarchically or into flat files. K-mean clustering, a basic clustering algorithm is introduced in the following.

3.1 K-means clustering

Problem Description:

Given a data set with N n-dimensional data points x^n, the goal is to determine a natural partitioning of the data set into a number of clusters (*k*) and noise. We know there are *k* disjoint clusters containing N_j data points with representative vector μ_j where *j*=1, ..., *k*. The K-means algorithm attempts to minimize the sum-of-squares clustering function given by

$$J = \sum_{j=1}^{k} \sum_{n \in S_j} \left\| x^n - \mu_j \right\|^2$$

where μ_j is the mean of the data points in cluster S_j and is given by

$$\mu_j = \frac{1}{N_j} \sum_{n \in S_j} x^n .$$

The training is carried out by assigning the points at random to k clusters and then computing the mean vectors μ_j of the N_j points in each cluster. Each point is re-assigned to a new cluster according to which is the nearest mean vector. The mean vectors are then recomputed.

K-means clustering proceeds as follows:
1. Specify the number of clusters (classes) k.
2. Choose *k* initial cluster seeds.
3. Assign cases closest to seed *j* as belonging to cluster j, *j*=1,..., *k*.
4. Calculate the mean of the cases in each cluster, and move the *k* cluster seeds to the mean of their cluster.
5. Reassign cases closest to the new seed *j* as belonging to cluster *j*.
6. Take the mean of the cases in each cluster as the new cluster seed.

This procedure is repeated until there is no further change in clustering.

K-means clustering is an unsupervised classification method. It is computationally efficient provided the initial cluster seeds are intelligently placed. Clustering methods depend on a measure of distance or similarity between points. Different distance metrics used in k-means clustering can result in different clusters.

3.2. Example: Clustering Automobile Drivers

The ABC Insurance Company periodically purchases lists of drivers from outside sources. Actuaries at ABC want to evaluate the potential claim frequency for underwriting purposes. Based on their experience, they know that driver claim frequency depends on geographic and demographic factors. Consequently, they want to segment the drivers into groups that are similar to each other with respect to these attributes. After the drivers have been segmented, a random sample of prospects within each segment will be used to estimate the frequency. The results of this test estimate will allow the actuaries to evaluate the potential profit of prospects from the list, both overall as well as for specific segments.

The synthetic data that was obtained from the vendor is given in Table 2.

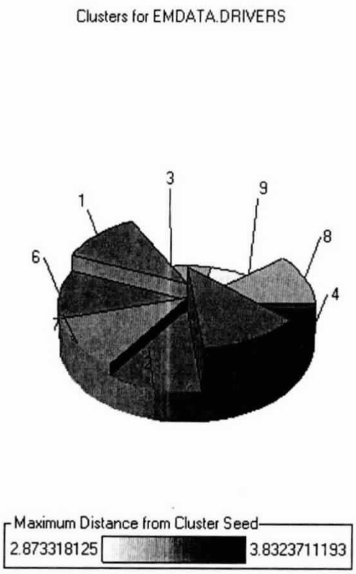
After preprocessing the data, which might include selecting a random sample of the data for initial analysis, filtering the outlying observations, and standardizing the variables in some way, we use the K-means clustering to form the clusters.

Table 2. Automobile Drivers Data

Variable	*Variable Type*	*Measurement Level*	*Description*
Age	Continuous	Interval	Driver's age in years
Car age	Continuous	Interval	Age of the car in years
Car type	Categorical	Nominal	Type of the car
Gender	Categorical	Binary	F=female, M=male
Coverage level	Categorical	Nominal	Policy coverage
Education	Categorical	Nominal	Education level of the drive
Location	Categorical	Nominal	Location of residence
Climate	Categorical	Nominal	Climate code for residence
Credit rating	Continuous	Interval	Credit score of the driver
ID	Input	Nominal	Driver's identification number

The following pie chart provides a graphical representation of key characteristics of the clusters.

Figure 1 Clusters Pie Chart

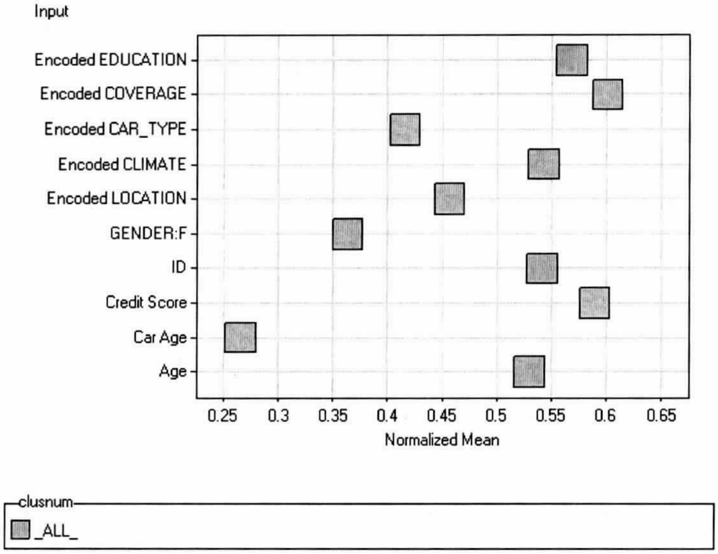


In the pie chart, slice width is the root-mean-square distance (root-mean-square standard deviation) between cases in the cluster; the height means the frequency and the color represents the distance of the farthest cluster member from the cluster. Cluster 5 contains the most cases while cluster 9 has the fewest.

Figure 2 below displays the input means for the entire data set over all of the clusters. The input means are normalized using a scale transformation

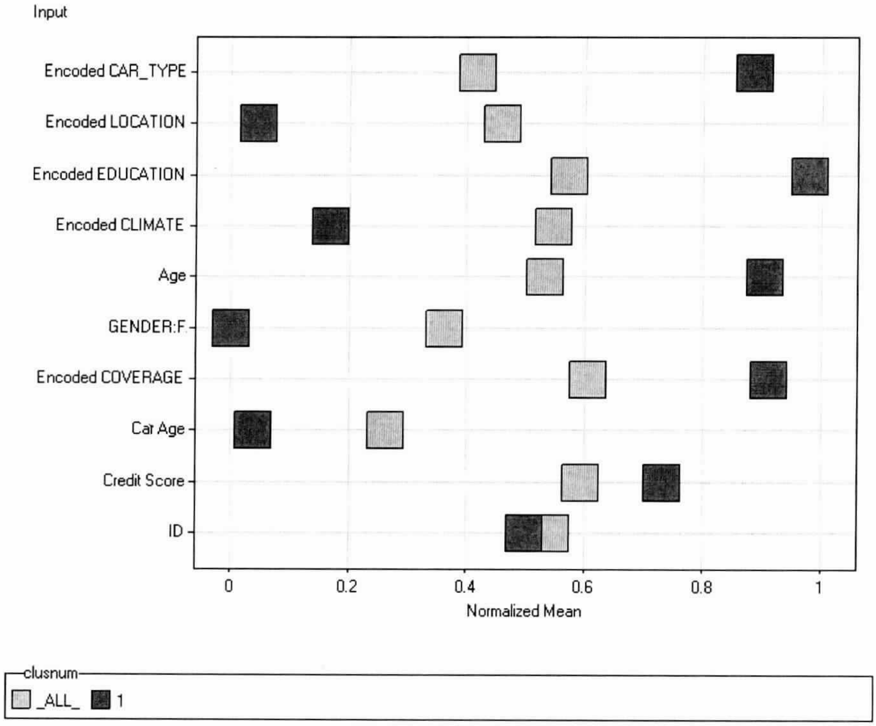
$$y = \frac{x - \min(x)}{\max(x) - \min(x)}.$$

Figure 2. Overall Input Means



The Normalized Mean Plot can be used to compare the overall normalized means with the normalized means in each cluster. Figure 3 compare the input means from cluster 1 (red blocks) to the overall input means (blue blocks). You want to identify the input means for clusters that differ substantially from the overall input means. The plot ranks the input based on how spread out the input means are for the selected cluster relative to the overall input means. The input that has the biggest spread is listed at the top and the input with the smallest spread is listed at the bottom. The input with the biggest spread typically best characterizes the selected cluster (Cluster 1 in Figure 3). Figure 3 shows that the variable "Car-Type" and "Location" are key inputs that help differentiate drivers in Cluster 1 from all of the drivers in the data set. Drivers in Cluster 1 tend to have higher than average education levels than average drivers in the data set.

Figure 3. Comparing the Input Means for Cluster 1 to the Overall Means



Cluster 5, as shown in Figure 4, has higher than average education and better than average credit scores. Most drivers in Cluster 5 live in location zone 4 and they drive newer car than average drivers. These characteristics can also be observed from Table 3.

Figure 4. Comparing the Input Means for Cluster 5 to the Overall Means

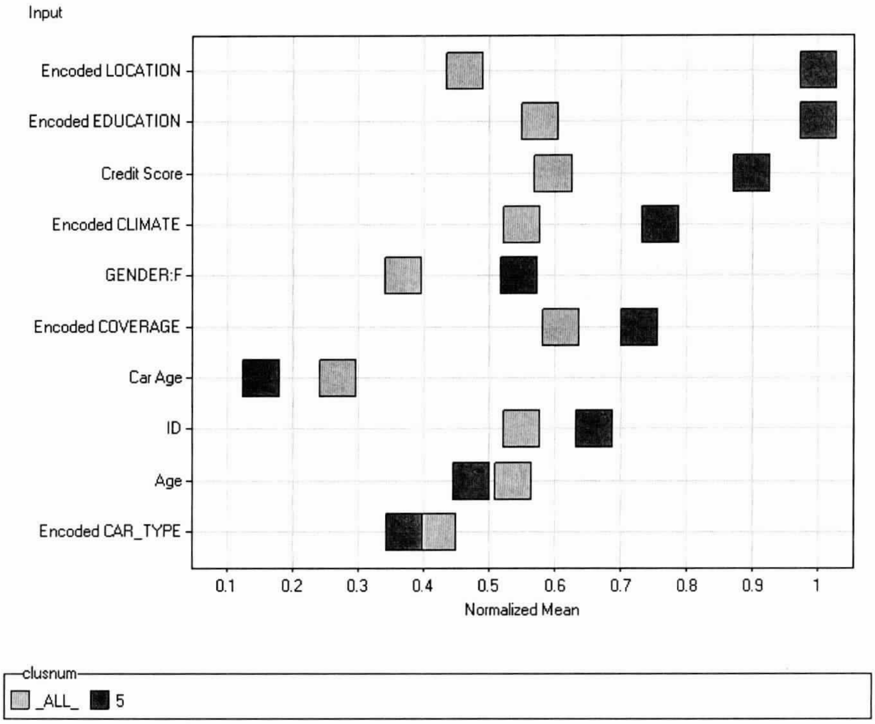


Table 3 displays information about each cluster. The statistics Root-Mean-Square Standard Deviation means the root-mean-square error across variables of the cluster standard deviations, which is equal to the root-mean-square distance between cases in the cluster.

Table 3. Clustering Summary Statistics

Cluster	Frequency of Cluster	Maximum Distance from Cluster Seed	Nearest Cluster	Distance to Nearest Cluster	Credit Score	Car Age	Age	Gender	Location	Climate	Car Type	Coverage	Education
9	7	2.87	5	2.82	0.86	3.29	35.57	1.00	3.43	1.29	3.57	2.43	1.86
8	20	3.22	7	2.40	0.62	2.15	46.65	0.65	2.80	2.55	2.25	2.85	1.85
7	22	3.25	2	2.25	0.65	2.73	24.59	0.27	1.95	2.09	1.45	2.36	2.27
6	21	3.38	4	2.41	0.81	6.52	35.19	0.43	2.00	1.48	1.67	1.19	1.76
5	33	3.41	4	2.37	0.82	3.00	32.79	0.58	3.82	2.33	2.03	2.39	3.03
4	18	3.83	5	2.37	0.59	5.17	34.44	0.39	3.50	1.83	2.72	1.44	2.56
3	7	3.21	7	3.14	0.46	8.00	20.57	0.43	3.57	2.43	1.14	1.43	2.00
2	18	3.38	7	2.25	0.56	3.56	26.00	0.28	2.89	2.67	1.28	2.28	1.39
1	27	3.40	5	2.55	0.75	2.37	44.15	0.07	2.04	1.52	3.30	2.70	3.00

During the clustering process, an importance value is computed as a value between 0 and 1 for each variable. Importance is a measure of worth of the given variable to the formation of the clusters. As shown in Table 4, variable "Gender" has an importance of 0, which means that the variable was not used as a splitting variable in developing the clusters. The measure of "importance" indicates how well the variable divides the data into classes. Variables with zero importance should not necessary be dropped.

Table 4. Variable Importance

Name	Importance
GENDER	0
ID	0
LOCATION	0
CLIMATE	0
CAR_TYPE	0.529939
COVERAGE	0.363972
CREDIT_SCORE	0.343488
CAR_AGE	0.941952
AGE	1
EDUCATION	0.751203

Clustering analysis can be used by the property/casualty insurance industry to improve predictive accuracy by segmenting databases into more homogeneous groups. Then the data of each group can be explored, analyzed, and modeled. Segments based on types of variables that associate with risk factors, profits, or behaviors often provide sharp contrasts, which can be interpreted more easily. As a result, actuaries can more accurately predict the likelihood of a claim and the amount of the claim. For example, one insurance company found that a segment of the 18- to 20-year old male drivers had a noticeably lower accident rate than the entire group of 18- to 20-year old males. What variable did this subgroup share that could explain the difference? Investigation of the data revealed that the members of the lower risk subgroup drove cars that were significantly older than the average and that the drivers of the older cars spent time customizing their "vintage autos." As a result, members of the subgroup were likely to be more cautious driving their customized automobiles than others in their age group.

Lastly, the cluster identifier for each observation can be passed to other nodes for use as an input, id, group, or target variable. For example, you could form clusters based on different age groups you want to target. Then you could build predictive models for each age group by passing the cluster variable as a group variable to a modeling node.

4. Predictive Data Mining

This section introduces data mining models for prediction (as opposed to description, such as in Section 3). Section 4.1 gives an overview of the Decision Tree DM algorithm. Section 4.2 presents a claim frequency model using Decision Trees and Logistic Regression.

4.1 Decision Trees

Decision trees are part of the Induction class of DM techniques. An empirical tree represents a segmentation of the data that is created by applying a series of simple rules. Each rule assigns an observation to a segment based on the value of one input. One rule is applied after another, resulting in a hierarchy of segments within segments. The hierarchy is called a tree, and each segment is called a node. The original segment contains the entire data set and is called the root node of the tree. A node with all its successors forms a branch of the node that created it. The final nodes are called leaves. For each leaf, a decision is made and applied to all observations in the leaf. The type of decision depends on the context. In predictive modeling, the decision is simply the predicted value.

The decision tree DM technique enables you to create decision trees that:

- Classify observations based on the values of nominal, binary, or ordinal targets,
- Predict outcomes for interval targets, or
- Predict the appropriate decision when you specify decision alternatives.

Specific decision tree methods include Classification and Regression Trees (CART; Breiman et. al., 1984) and the count or Chi-squared Automatic Interaction Detection (CHAID; Kass, 1980) algorithm. CART and CHAID are decision tree techniques used to classify a data set.

The following discussion provides a brief description of the CHAID algorithm for building decision trees. For CHAID, the inputs are either nominal or ordinal. Many software packages accept interval inputs and automatically group the values into ranges before growing the tree. For nodes with many observations, the algorithm uses a sample for the split search, for computing the worth (measure of worth indicates how well a variable divides the data into

each class), and for observing the limit on the minimum size of a branch. The samples in different nodes are taken independently. For binary splits on binary or interval targets, the optimal split is always found. For other situations, the data is first consolidated, and then either all possible splits are evaluated or else a heuristic search is used.
The consolidation phase searches for groups of values of the input that seem likely to be assigned the same branch in the best split. The split search regards observations in the same consolidation group as having the same input value. The split search is faster because fewer candidate splits need evaluating. A primary consideration when developing a tree for prediction is deciding how large to grow the tree or, what comes to the same end, what nodes to prune off the tree. The CHAID method of tree construction specifies a significance level of a Chi-square test to stop tree growth. The splitting criteria are based on p-values from the F-distribution (interval targets) or Chi-square distribution (nominal targets). For these criteria, the best split is the one with the smallest p-value. By default, the p-values are adjusted to take into account multiple testing.

A missing value may be treated as a separate value. For nominal inputs, a missing value constitutes a new category. For ordinal inputs, a missing value is free of any order restrictions.

The search for a split on an input proceeds stepwise. Initially, a branch is allocated for each value of the input. Branches are alternately merged and re-split as seems warranted by the p-values. The original CHAID algorithm by Kass stops when no merge or re-splitting operation creates an adequate p-value. The final split is adopted. A common alternative, sometimes called the exhaustive method, continues merging to a binary split and then adopts the split with the most favorable p-value among all splits the algorithm considered.

After a split is adopted for an input, its p-value is adjusted, and the input with the best-adjusted p-value is selected as the splitting variable. If the adjusted p-value is smaller than a threshold you specified, then the node is split. Tree construction ends when all the adjusted p-values of the splitting variables in the unsplit nodes are above the user-specified threshold.

Tree techniques provide insights into the decision-making process, which explains how the results come about. The decision tree is efficient and is thus suitable for large data sets. Decision trees are perhaps the most successful exploratory method for uncovering deviant data structure. Trees recursively partition the input data space in order to identify segments where the records are homogeneous. Although decision trees can split the data into several homogeneous segments and the rules produced by the tree can be used to detect interaction among variables, it is relatively unstable and it is difficult to detect linear or quadratic relationships between the response variable and the dependent variables.

4.2 Modeling claim frequency

We now start the modeling process by studying the relationship between claim frequency and underlying risk factors including age, gender, credit score, location, education level, coverage, …, and car age. Again, the synthetic data is used. A hybrid method is developed for this study – the modeling process is a combination of the decision tree techniques and logistic regression.

First, we use the decision tree algorithm to identify the factors that influence claim frequency. After the factors are identified, the logistic regression technique is used to quantify the claim frequency and the effect of each risk factor.

The data for the study has the following variables as shown in Table 5:

Table 5. Automobile Driver's Claim Information

Variable	*Variable Type*	*Measurement Level*	*Description*
Age	Continuous	Interval	Driver's age in years
Car age	Continuous	Interval	Age of the car
Car type	Categorical	Nominal	Type of the car
Gender	Categorical	Binary	F=female, M=male
Coverage level	Categorical	Nominal	Policy coverage
Education	Categorical	Nominal	Education level of the drive
Location	Categorical	Nominal	Location of residence
Climate	Categorical	Nominal	Climate code for residence
Credit rating	Continuous	Interval	Credit score of the driver
ID	Input	Nominal	Driver's identification number
No. of claims	Categorical	Nominal	Number of claims

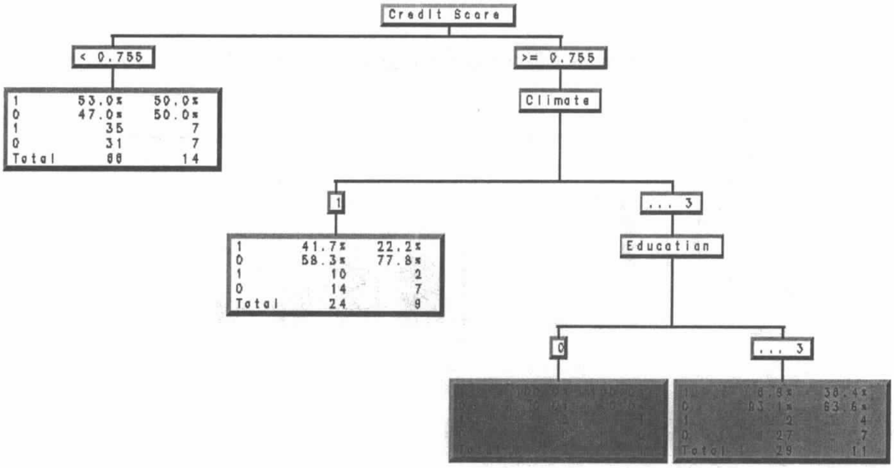
We now use the decision tree algorithm to analyze the influences and the importance of the claim frequency risk factors. The tree algorithm used in this research is SAS/Enterprise Miner Version 4.2 (2002). We built 100 binary regression trees and 100 CHAID-like trees for optimal decision tree. Our decision tree analysis reveals that the credit score has the greatest impact on the claim frequency. The claim frequency, and the interaction among different factors that affect the claim frequency, vary as the credit score status changes. Furthermore, there is a significant climate influence within the "higher credit score" status.

A Tree diagram contains the following items:

- Root node -- top node in the tree that contains all observations.
- Internal nodes – non-terminal nodes (including the root node) that contain the splitting rules.
- Leaf nodes -- terminal nodes that contain the final classification for a set of observations.

The tree diagram displays node (segment) statistics, the names of variables used to split the data into nodes, and the variable values for several levels of nodes in the tree. Figure 5 shows a partial profile of the tree diagram for our analysis:

Figure 5. Tree Diagram



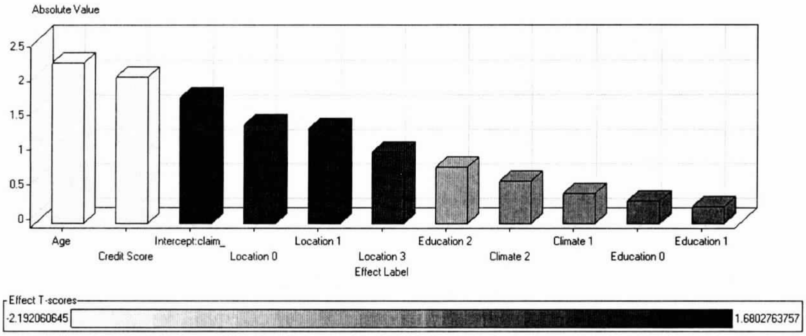
In Figure 5, each leaf node displays the percentage and n-count of the values that were used to determine the branching. The second column contains the learning from the training data including the percentage for each target level, the count for each target level, and the total count. The third column contains the learning from the validation data including the percentage for each target level, the count for each target level, and the total count. For example, among these drivers with credit score below 75.5%, 53% of them submitted a claim from the training data.

The assessment values are used to recursively partition the data in homogenous subgroups. The method is recursive because each subgroup results from splitting a subgroup from a previous split. The numeric labels directly above each node indicate at which point the tree algorithm found significant splits in interval level variable distributions or in categorical splits for nominal or ordinal level distributions. The character labels positioned central to each split are the variable names. You can trace the paths from the root to each leaf and express the results as a rule.

As shown in Figure 5, the claim frequency varies with the most important risk factor (the credit score status, in this study) among all the other variables. Based on tree analysis, the car age, coverage, and car-type are the irrelevant factors. They should not be included in the claim frequency model.

Based on the tree analysis, we now use logistic regression to estimate the probability of claim occurrence for each driver based on the factors under consideration. As discussed in Section 2, logistic regression attempts to predict the probability of a claim as a function of one or more independent inputs. Figure 6 shows a bar chart of the effect T-scores from the logistic regression analysis. An effect T-score is equal to the parameter estimate divided by its standard error.

Figure 6. Effect T-scores from the logistic regression

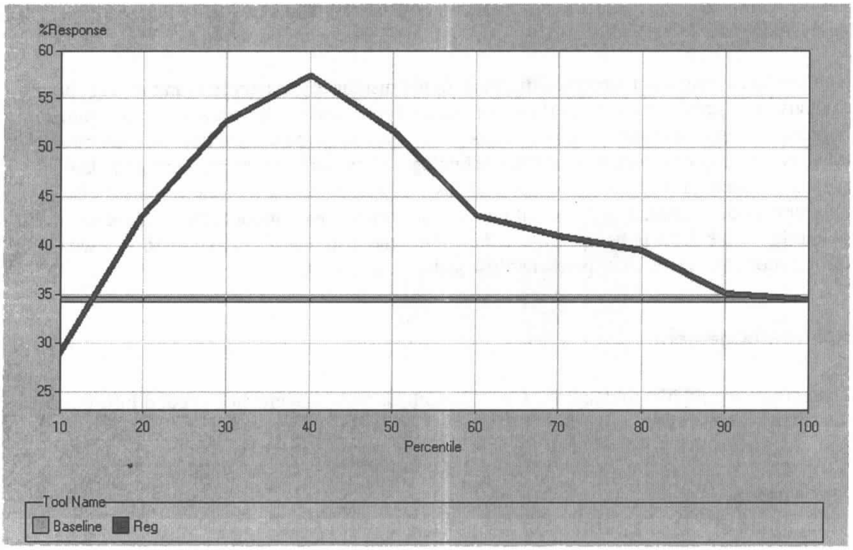


The scores are ordered by decreasing absolute value in the chart. The color density legend indicates the size of the score for a bar. The legend also displays the minimum and maximum score to the left and right of the legend, respectively. The vertical axis represents the absolute value for the effect. In this example, the first variable, Age has the largest absolute value, Credit Score has the second largest absolute value, and so on. The estimates for Location and Education are positive, so their bar values is colored a shade of orange. The estimates for Age and Credit Score have negative values, so their bars are displayed in yellow.

Assessment is the final part of the data mining process. The Assessment criterion is a comparison of the expected to actual profits or losses obtained from model results. This criterion enables you to make cross-model comparisons and assessments, independent of all other factors (such as sample size, modeling node, and so on).

Figure 7 is a cumulative % claim-occurrence lift chart for the logistic regression model. Lift charts show the percent of captured claim-occurrence (a.k.a. the lift value) on the vertical axis. In this chart the target drivers are sorted from left to right by individuals most likely to have an accident, as predicted by each model. The sorted group is lumped into ten percentiles along the X-axis; the left-most percentile is the 10% of the target predicted most likely to have an accident. The vertical axis represents the predicted cumulative % claim-occurrence if the driver from that percentile on down submitted a claim.

Figure 7. Lift Chart for Logistic Regression



The lift chart displays the cumulative % claim-occurrence value for a random baseline model, which represents the claim rate if you chose a driver at random, given the logistic regression model.

The performance quality of a model is demonstrated by the degree the lift chart curve pushes upward and to the left. For this example, the logistic regression model captured about 30% of the drivers in the 10th percentile. The logistic regression model does have better predictive power from about the 20th to the 80th percentiles. At about the 90th percentile, the cumulative % claim-occurrence values for the predictive model are about the same as the random baseline model.

5. Conclusions

This paper introduced the data mining approach to modeling insurance risk and some implementation of the approach. In this paper, we provide an overview of data mining operations and techniques and demonstrate two potential applications to property/casualty actuarial practice. In section 3.2, we used k-means clustering to better describe a group of drivers by segmentation. In section 4.2, we examined several risk factors for automobile drivers with the goal of predicting their claim frequency. The influences and the correlations of these factors on auto claim distribution were identified with exploratory data analysis and decision tree algorithm. Logistic regression is then applied to model claim frequency.

Due to our use of synthetic data, however, the examples show limited advantages of DM over traditional actuarial analysis. The great significance of the data mining, however, can only be shown with huge, messy databases. Issues on how to improve data quality through data acquisition, data integration, and data exploration will to be discussed in the future study.

The key to gaining a competitive advantage in the insurance industry is found in recognizing that customer databases, if properly managed, analyzed, and exploited, are unique, valuable corporate assets. Insurance firms can unlock the intelligence contained in their customer databases through modern data mining technology. Data mining uses predictive modeling, database segmentation, market basket analysis, and combinations thereof to more quickly answer crucial business questions with greater accuracy. New products can be developed and marketing strategies can be implemented enabling the insurance firm to transform a wealth of information into a wealth of predictability, stability, and profits.

Acknowledgement

The authors would like to thank the CAS Committee on Management Data and Information for reviewing this paper and providing many constructive suggestions.

References

Adya, M. 1998. "How Effective Are Neural Networks at Forecasting and Prediction? A Review and Evaluation." *Journal of Forecasting* 17(5-6, Sep-Nov): 481-495.

Berry, M. A., AND G. S. Linoff. 2000. *Mastering Data Mining*. New York, N.Y.: Wiley.

Bishop, C. M. 1995. *Neural Networks for Pattern Recognition*. New York: Oxford University Press.

Breiman, L., J. H. Friedman, R. A. Olshen, AND C. J. Stone. 1984. *Classification and Regression Trees*. New York, N.Y.: Chapman & Hall.

Borok, L.S. 1997. "Data mining: Sophisticated forms of managed care modeling through artificial intelligence." *Journal of Health Care Finance*. 23(3), 20-36.

Carpenter, G., AND S. Grossberg. 1988. "The ART of Adaptive Pattern Recognition by a Self-Organizing Neural Network." *IEEE Computer*, 21(3): 77-88.

Cheesman, P. 1996. "Bayesian Classification (AutoClass): Theory and Results," in *Advances in Knowledge Discovery and Data Mining*, ed. by Fayyad, U., G. Piatetsky-Shapiro, P. Smyth, AND R. Uthurusamy. Menlo Park, CA: AAAI Press/The MIT Press: 153-180.

Chessman, P., J. Kelly, M. Self, J. Stutz, W. Taylor, AND D. Freeman. 1988. "Auto Class: A Bayesian classification system." *5th Int'l Conf. on Machine Learning*, Morgan Kaufman.

Goldberg, D.E. 1989. *Genetic Algorithms in Search, Optimization and Machine Learning*. Morgan Kaufmann.

Guha, S., R. Rastogi, AND K. Shim K. 1998. "CURE: An Efficient Clustering Algorithm for Large Databases." *Proceedings of the ACM SIGMOD Conference*.

Ester, M., H. Kriegel, J. Sander, AND X. Xu. 1998. "Cluttering for Mining in Large Spatial Databases." *Special Issue on Data mining, KI-Journal*, 1. Scien Tec Publishing.

Fayyad, U. M., G. Piatetsky-Shapiro, P. Smyth, AND R. Uthurusamy(Eds.). 1996. *Advances in Knowledge Discovery and Data Mining*. Cambridge, MA: The MIT Press.

Fisher, D., M. Pazzani, AND P. Langley. 1991. *Concept Formation: Knowledge and Experience in Unsupervised Learning*. San Mateo, CA: Kaufmann.

Francis, L. 2001. "Neural Networks Demystified." *CAS Forum* 253-319.

Hand, D., H. Mannila, AND P. Smyth. 2001. *Principles of Data Mining*. Cambridge, Massachusetts: MIT Press.

Hinneburg, A., AND D.A. Keim. 1998. "An Efficient Approach to Clustering in Large Multimedia Databases with Noise." *Proceeding 4th Int. Conf. on Knowledge Discovery and Data Mining*.

Hosmer, D. W., AND S. Lemeshow. 1989. *Applied Logistic Regression*. New York, N. Y.: John Wiley & Sons.

Kass, G. V. 1980. *"An Exploratory Technique for Investigating Large Quantities of Categorical Data,"* Applied Statistics, 29:119-127.

Kleinbaum, D. G., L. Kupper, AND K. Muller. 1988. *Applied Regression Analysis and other Multivariable Methods, 2nd edition*. PWS-KENT Publishing Company, Boston.

Michaelewicz, Z. 1994. *Genetic Algorithms + Data Structures = Evolution Programs*. New York: Springer-Verlag.

Quinlin, J.R. 1983. "Induction of Decision Trees." *Machine Learning* 1: 81-106.

D.E. Rumelhart, G.E. Hinton, AND R.J. Williams. 1986. "Learning Internal Representation by Error Propagation." *Parallel Distributed Processing*, ed. by Rumelhart, D.E., J.L. McClelland, AND the PDP Research Group. Cambridge, MA: The MIT Press: 318-362.

SAS Institute. 2000. *Enterprise Miner*, Cary, N.C.: SAS Institute.

Tufte, E.R. 1983. *The Visual Display of Quantitative Information*, Graphics Press, Cheshire, CN.

Tufte, E.R. 1990. *Envisioning Information*, Graphics Press, Cheshire, CN.

Martian Chronicles: Is MARS Better than Neural Networks?

Louise A. Francis, FCAS, MAAA

Martian Chronicles: Is MARS better than Neural Networks?

by Louise Francis, FCAS, MAAA

Abstract:
A recently developed data mining technique, Multivariate Adaptive Regression Splines (MARS) has been hailed by some as a viable competitor to neural networks that does not suffer from some of the limitations of neural networks. Like neural networks, it is effective when analyzing complex structures which are commonly found in data, such as nonlinearities and interactions. However, unlike neural networks, MARS is not a "black box", but produces models that are explainable to management.

This paper will introduce MARS by showing its similarity to an already well-understood statistical technique: linear regression. It will illustrate MARS by applying it to insurance fraud data and will compare its performance to that of neural networks.

Acknowledgements:
The author thanks the Automobile Insurers Bureau of Massachusetts for supplying the data used in this paper and wishes to acknowledge the efforts of Richard Derrig in providing insight into the data and approaches to analyzing it. The author also acknowledges the following people who reviewed this paper and provided many constructive suggestions: Jane Taylor, Francois Morrisette, Christopher Yaure, Patricia Francis-Lyon and Virginia Lambert.

Martian Chronicles: Is MARS better than Neural Networks?

2 The discipline of artificial intelligence has contributed a number of promising techniques to the analyst's toolkit. The techniques have names such as "machine learning", "genetic algorithms" and "neural networks". These techniques are collectively known as data mining. Data mining uses computationally intensive techniques to find patterns in data. When data mining tools are applied to data containing complex relationships they can identify relationships not otherwise apparent. These complexities have been a challenge for traditional analytical procedures such as linear regression.

The casualty actuarial literature contains only a few papers about data mining techniques. Speights *et al.* (Speights *et al.*, 1999) and Francis (Francis, 2001) introduced the neural network procedure for modeling complex insurance data. Hayward (Hayward, 2002) described the use of data mining techniques in safety promotion and better matching of premium rates to risk. The methods discussed by Hayward included exploratory data analysis using pivot tables and stepwise regression.

In this paper, a new technique, MARS, which has been proposed as an alternative to neural networks (Steinberg, 2001), will be introduced. The name MARS, coined for this technique by its developer, Freidman, (Hastie, *et al.*, 2001), is an acronym for Multivariate Adaptive Regression Splines. The technique is a regression based technique which allows the analyst to use automated procedures to fit models to large complex databases. Because the technique is regression based, its output is a linear function that is readily understood by analysts and can be used to explain the model to management. Thus, the technique does not suffer from the "black box" limitation of neural networks. However, the technique addresses many of the same data complexities addressed by neural networks.

Neural networks are one of the more popular data mining approaches. These methods are among of the oldest data mining methods and are included in most data mining software packages. Neural networks have been shown to be particularly effective in handling some complexities commonly found in data. Neural networks are well known for their ability to model nonlinear functions. The research has shown that a neural network with a sufficient number of parameters can model any continuous nonlinear function accurately.[1] Francis (Francis, 2001) also showed that neural networks are valuable in fitting models to data containing interactions. Neural networks are often the tools of choice when predictive accuracy is required. Berry and Linoff (Berry and Linoff, 1997) suggest that neural networks are popular because of their proven track record.

Neural networks are not ideal for all data sets. Warner and Misra presented several examples where they compared neural networks to regression (Warner and Misra, 1996). Their research showed that regression outperformed neural networks when the functional relationship between independent and dependent variables was known. Francis (Francis,

---

[1] A more technical description of the property is that with a sufficient number of nodes in the neural network's hidden layer, the neural network can approximate any deterministic nonlinear continuous function.

2001) showed that when the relationship between independent and dependent variables was linear, classical techniques such as regression and factor analysis outperformed neural networks.

Perhaps the greatest disadvantage of neural networks is the inability of users to understand or explain them. Because the neural network is a very complex function, there is no way to summarize the relationships between independent and dependent variables with functions that can be interpreted by data analysts or management. Berry and Linoff (Berry and Linoff, 1997) state that "Neural networks are best approached as black boxes with mysterious inner workings, as mysterious as the origins of our own consciousness". More conventional techniques such as linear regression result in simple mathematical functions where the relationship between predictor and target variables is clearly described and can be understood by audiences with modest mathematical expertise. The "black box" aspect of neural networks is a serious impediment to more widespread use.

Francis (Francis, 2001) listed several complexities found in actual insurance data and then showed how neural networks were effective in dealing with these complexities. This paper will introduce MARS and will compare and contrast how MARS and neural networks deal with several common data challenges. Three challenges that will be addressed in this paper are:

1) Nonlinearity: Traditional actuarial and statistical techniques often assume that the functional relationship between the independent variables and the dependent variable is linear or some transformation of the data exists that can be treated as linear.
2) Interactions: The exact form of the relationship between a dependent and independent variable may depend on the value of one or more other variables.
3) Missing data: Frequently data has not been recorded on many records of many of the variables that are of interest to the researcher.

The Data

This paper features the application of two data mining techniques, neural networks and MARS, to the fraud problem. The data for the application was supplied by the Automobile Insurers Bureau of Massachusetts (AIB). The data consists of a random sample of 1400 closed claims that were collected from PIP (personal injury protection or no-fault coverage) claimants in Massachusetts in 1993. The database was assembled with the cooperation of ten large insurers. This data has been used by the AIB, the Insurance Fraud Bureau of Massachusetts (IFB) and other researchers to investigate fraudulent claims or probable fraudulent claims (Derrig *et al.*, 1994, Weisberg and Derrig, 1995, Viaene *et al.*, 2002). While the typical data mining application would use a much larger database, the AIB PIP data is well suited to illustrating the use of data mining techniques in insurance. Viaene *et al.* used the AIB data to compare the performance of a number of data mining and conventional classification techniques (Viaene *et al.*, 2002).

Two key fraud related dependent variables were collected in the study: an overall assessment (ASSESS) of the likelihood the claim is fraudulent or abusive and a suspicion score (SUSPICION). Each record in the data was assigned a value by an expert. The value indicates the expert's subjective assessment as to whether the claim was legitimate or whether fraud or abuse was suspected. Experts were asked to classify suspected fraud or abuse claims into the following categories: exaggerated damages, opportunistic fraud or planned fraud. As shown in Table 1, the assessment variable can take on 5 possible values. In addition, each claim was assigned a score from 0 (none) to 10 (very high) indicating the expert's degree of suspicion that the claim was abusive or fraudulent. Weisberg and Derrig (Weisberg and Derrig, 1993) found that more serious kinds of fraud, such as planned fraud were associated with higher suspicion scores than "softer" fraud such as exaggeration of damages. They suggest that the suspicion score was able to measure the range of "soft" versus "hard" fraud.

The database contains detailed objective claim information on each claim in the study. This includes information about the policy inception date, the date the accident occurred, the date it was reported, the paid and incurred loss dollars, the injury type, payments to health care providers and the provider type. The database also contains "red flag" or fraud indicator variables. These variables are subjective assessments of characteristics of the claim that are believed to be related to the likelihood of fraud or abuse. More information on the variables in the model is supplied below in the discussion of specific models.

Table 1

Assessment Variable

Value	Assessment	Percent of Data
1	Probably legitimate	64%
2	Excessive treatment only	20%
3	Suspected opportunistic fraud, no injury	3%
4	Suspected opportunistic fraud, exaggerated injury	12%
5	Suspected planned fraud	1%

We may use the more inclusive term "abuse" when referring to the softer kinds of fraudulent activity, as only a very small percentage of claims meet the strict standard of criminal fraud (Derrig, 2002). However, misrepresentation and exaggeration of the nature and extent of the damages, including padding of the medical bills so that the value of the claim exceeds the tort threshold, occur relatively frequently. While these activities are often thought of as fraud, they do not meet a legal definition of fraud. Therefore, they will be referred to as abuse. Overall, about one third of the claims were coded as probable abuse or fraud claims.

Nonlinear Functions

The relationships encountered in insurance data are often nonlinear. Classical statistical modeling methods such as linear regression have had a tremendous impact on the analysis and modeling of data. However, traditional statistical procedures often assume that the relationships between dependent and independent variables are linear. Traditional modeling also allows linear relationship that result from a transformation of dependent or independent variables, so some nonlinear relationships can be approximated. In addition, there are techniques specifically developed for fitting nonlinear functions such as nonlinear regression. However, these techniques require that theory or experience specify the "true" form of the nonlinear relationships. Data mining techniques such as neural networks and MARS do not require that the relationships between predictor and dependent variables be linear (whether or not the variables are transformed). Both neural networks and MARS are also considered nonparametric because they require no assumptions about the form of the relationship between dependent and independent variables.

For this illustration, a dependent variable that is not categorical (i.e. values have a meaningful order) was selected. The selected dependent variable was SUSPICION. Unlike the ASSESS variable, the values on the SUSPICION variable have a meaningful range, with higher values associated with suspicion of more serious fraud.

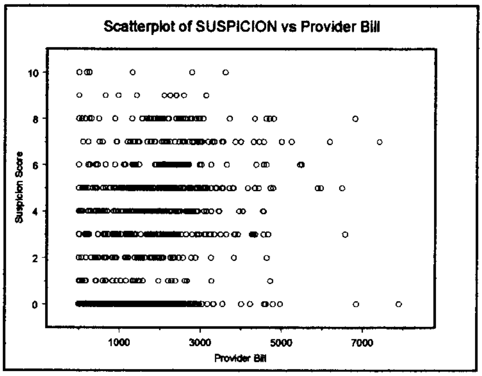
To illustrate methods of fitting models to nonlinear curves, a variable was selected which 1) had a significant correlation with the dependent variable, and 2) displayed a highly nonlinear relationship. Illustrating the techniques is the objective of this example. The data used may require significant time to collect and may therefore not be practical for an application where the objective is to predict abuse and fraud (which would require data that is available soon after the claim is reported). Later in the paper, models for prospectively predicting fraud will be presented. The variable selected was the first medical provider's bill[2]. A medical provider may be a doctor, a clinic, a chiropractor or a physical therapist. Prior published research has indicated that abusive medical treatment patterns are often key drivers of fraud (Derrig *et al.*, 1994, Weisberg and Derrig, 1995). Under no-fault laws, claimants will often deliberately run the medical bills up high enough to exceed tort thresholds. In this example the relationship between the first provider's medical bill and the value of the suspicion score will be investigated. The AIB fraud database contains the medical bills submitted from the top two health care providers. If more costly medicine is delivered to suspicious claims than non-suspicious claims, the provider bills should be higher for the suspicious claims.

Figure 1 presents a scatterplot of the relationship between SUSPICION and the provider bill. No relationship is evident from the graph. However, certain nonlinear relationships can be difficult to detect visually.

---

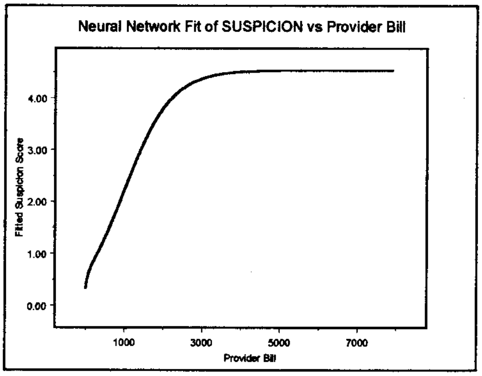
[2] Note that Massachusetts PIP covers only the first $8,000 of medical payments if the claimant has health insurance. Large bill amounts may represent data from claimants with no coverage. Bills may also exceed $8,000 even if payments are limited. However, the value of medical bills on some claims may be truncated because reimbursement is not expected.

Figure 1



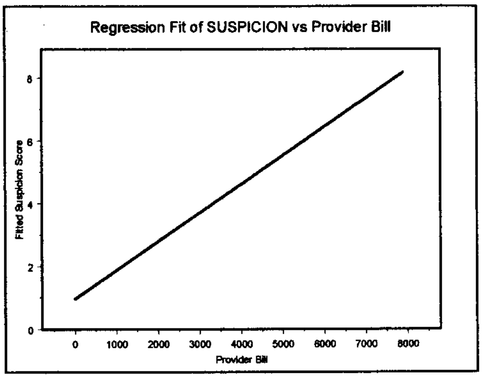
Neural networks will first be used to fit a curve to the data. A detailed description of how neural networks analyze data is beyond the scope of this paper. Several sources on this topic are Francis, Lawrence and Smith (Francis, 2001, Lawrence, 1994, Smith, 1996). Although based upon how neurons function in the brain, the neural network technique essentially fits a complex non-parametric nonlinear regression. A task at which neural networks are particularly effective is fitting nonlinear functions. The graph below displays the resulting function when the dependent variable SUSPICION is fit to the provider bill by a neural network. This graph displays a function that increases quickly at lower bill amounts and then levels off. Although the curve is flat over much of the range of medical bills, it should be noted that the majority of bills are below $2,000 (in 1993 dollars).

Figure 2



One of the most common statistical procedures for curve fitting is linear regression. Linear regression assumes the relationship between the dependent and independent variables is linear. Figure 3 displays the graph of a fitted regression line of SUSPICION on provider bill. The regression forces a linear fit to SUSPICION versus the payment amount Thus, rather than a curve with a rapidly increasing trend line that levels off, a line with a constant slope is fitted. If the relationship is in fact nonlinear, this procedure is not as accurate as that of the neural network.

Figure 3

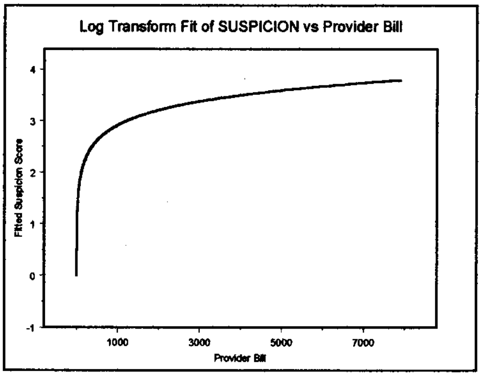


When the true relationship between a dependent and independent variable is nonlinear, various approaches are available when using traditional statistical procedures for fitting the curve. One approach is to apply a nonlinear transformation to the dependent or independent variable. A linear regression is then fit to the transformed variables. As an example, a log transform was applied to the provider bill variable in the AIB data. The regression fit was of the form:

$$Y = B_o + B_1 \ln(X)$$

That is, the dependent variable, the suspicion score, is assumed to be a linear function of the natural log of the independent variable, provider bill. Figure 4 displays the curve fit using the logarithmic transformation.

Figure 4

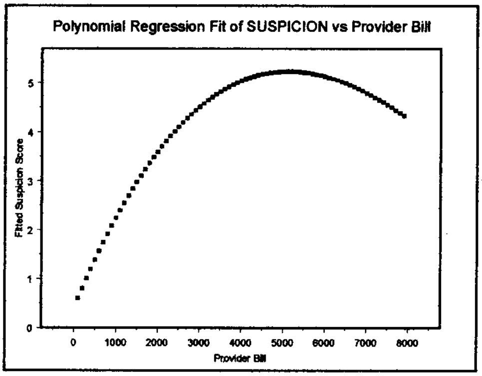


Another procedure which is used in classical linear regression to approximate nonlinear curves is polynomial regression. The curve is approximated by the function:

$$Y = B_o + B_1X + B_2X^2 + \ldots + B_nX^n$$

Generally, low order polynomials are used in the approximation. A cubic polynomial (including terms up to provider bill raised to the third power) was used in the fit. Figure 5 displays a graph of a fitted polynomial regression.

Figure 5



The use of polynomial regression to approximate functions is familiar to readers from its use in Taylor series expansions for this purpose. However, the Taylor series expansion is used to approximate a function near a point, rather than over a wide range. When evaluating a function over a range, the maximums and inflection points of the polynomial may not exactly match the curves of the function being approximated.

The neural network model had an R^2 (coefficient of determination) of 0.37 versus 0.25 for the linear model and 0.26 for the log transform. The R^2 of the polynomial model was comparable to that of the neural network model. However, the fit was influenced strongly by a small number of claims with large values. Though not shown in the graph, at high values for the independent variable the curve declines below zero and then increases again. This unusual behavior suggests that the fitted curve may not approximate the "true" relationship between provider bill and suspicion score well at the extremes of the data and may perform poorly on new claims with values outside the range of the data used for fitting.

Table 2 below shows the values of SUSPICION for ranges of the provider bill variable. The table indicates that SUSPICION increases rapidly at low bill amounts and then levels off at about $3,000.

Table 2

Suspicion Scores by Provider Bill

Provider Bill	Number of Claims	Mean Suspicion Score
$0	444	0.3
1 - 1,000	376	1.1
1,001 - 2,000	243	3.0
2,001 - 3,000	227	4.2
3,001 - 4,000	60	4.6
4,001 - 5000	33	4.2
5,001 - 6,000	5	5.8
6,001 - 7,000	12	4.3

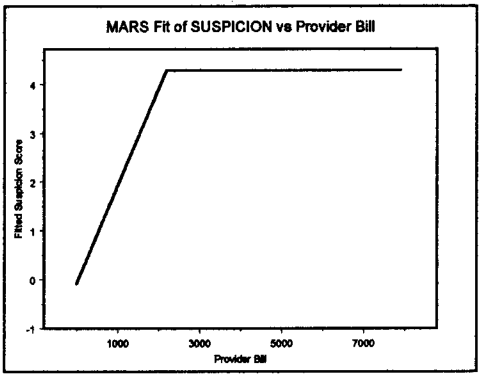
The examples illustrate that traditional techniques which require specific parametric assumptions about the relationship between dependent and independent variables may lack the flexibility to model nonlinear relationships. It should be noted, however, that Francis (Francis, 2001) presented examples where traditional techniques performed as well as neural networks in fitting nonlinear functions. Also, when the true relationship between the dependent and independent variables is linear, classical statistical methods are likely to outperform neural networks.

MARS and Nonlinear Functions

The MARS approach to fitting nonlinear functions has similarities to polynomial regression. In its simplest form MARS fits piecewise linear regressions to the data. That is, MARS breaks the data into ranges and allows the slope of the line to be different for the different ranges. MARS requires the function fit to be continuous, thus there are no jump points between contiguous ranges.

To continue the previous example, a function was fit by MARS. The graph below displays the MARS fitted function. It can be seen that the curve is broken into a steeply sloping line, which then levels off much the way the neural network fitted function did.

Figure 6



MARS uses an optimization procedure that fits the best piecewise regression. Simpler functions may adequately approximate the relationship between predictor and dependent variables and are favored over more complex functions. From the graph, it can be seen that the best MARS regression had two pieces:

1) The curve has a steep slope between bill amounts of $0 and $2,185
2) The curve levels off at bill amounts above $2,185

The fitted regression model can be written as follows:

BF1 = max(0, 2185 – X)
Y = 4.29 - 0.002 * BF1

where

Y is the dependent variable (Suspicion score)
X is the provider bill

The points in the data range where the curves change slope are known as knots. The impact of knots on the model is captured by basis functions. For instance BF1 is a basis function. Basis functions can be viewed as similar to dummy variables in linear regression. Dummy variables are generally used in regression analysis when the predictor variables are categorical. For instance, the Provider bill variable can be

MARS can perform regressions on binary variables. When the dependent variable is binary, MARS is run in binary mode. In binary mode, the dependent variable is converted into a 0 (legitimate) or a 1 (suspected fraud or abuse). Ordinary least squares regression is then performed regressing the binary variable on the predictor variables. Logistic regression is a more common procedure when the dependent variable is binary. Suppose that the true target variable is the probability that a given claim is abusive, and this probability is denoted *p(x)*. The model relating *p(x)* to the a vector of independent variables **x** is:

$$\ln(\frac{p}{1-p};\mathbf{x}) = B_0 + B_1X_1 + \ldots + B_nX_n$$

where the quantity $\ln(p(\mathbf{x})/(1-p(x)))$ is known as the logit function or log odds. Logistic regression can be used to produce scores that are between zero and one, consistent with viewing the score as a probability. Binary regressions can produce predicted values which can be less than zero and greater than one. One solution to this issue is to truncate the predicted values at zero and one. Another solution is to add the extra step of fitting a logistic regression to the data using the MARS predicted value as the independent variable and the binary assessment variable as the dependent variable. The fitted probabilities from the logistic regression can then be assigned as a score for the claim. The neural network model was also run in binary mode and also produced fitted values which were less than zero or greater than one. In this analysis, logistic regression was applied to the results of both the MARS and neural network fits to convert the predicted values into probabilities.

Variables in the Model

There are two categories of predictor variables that were incorporated into the models described in this section. The first category is red flag variables. These are primarily subjective variables that are intended to capture features of the accident, injury or claimant that are believed to be predictive of fraud or abuse. Many red flag variables represent accumulated industry wisdom about which indicators are likely to be associated with fraud or abuse. The information recorded in these variables represents an expert's subjective assessment of fraud indications, such as "the insured felt set up, denied fault" These variables are binary, that is, they are either true or false. Such red flag variables are often used to target certain claims for further investigation. The data for these red flag variables is not part of the claim file; it was collected as part of the special effort undertaken in assembling the AIB database for fraud research.

The red flag variables were supplemented with claim file variables deemed to be available early in the life of a claim and therefore of practical value in predicting fraud and abuse.

The variables selected for use in the full model are the same as those used by Viaene *et al.* (Viaene *et. al.*, 2002) in their comparison of statistical and data mining methods. While a much larger number of predictor variables is available in the AIB data for modeling fraud, the red flag and objective claim variables selected for incorporation into their models by Viaene *et al.* were chosen because of early availability. Therefore they are likely to be useful in predicting fraud and abuse soon enough in the claim's lifespan for effective mitigation efforts to lower the cost of the claim. Tables 6 and 7 present the red flag and claim file variables.

Table 6

Red Flag Variables

Subject	Indicator Variable	Description
Accident	ACC01	No report by police officer at scene
	ACC04	Single vehicle accident
	ACC09	No plausible explanation for accident
	ACC10	Claimant in old, low valued vehicle
	ACC11	Rental vehicle involved in accident
	ACC14	Property Damage was inconsistent with accident
	ACC15	Very minor impact collision
	ACC16	Claimant vehicle stopped short
	ACC19	Insured felt set up, denied fault
Claimant	CLT02	Had a history of previous claims
	CLT04	Was an out of state accident
	CLT07	Was one of three or more claimants in vehicle
Injury	*INJ01*	*Injury consisted of strain or sprain only*
	INJ02	No objective evidence of injury
	INJ03	Police report showed no injury or pain
	INJ05	No emergency treatment was given
	INJ06	Non-emergency treatment was delayed
	INJ11	Unusual injury for auto accident
Insured	INS01	Had history of previous claims
	INS03	Readily accepted fault for accident
	INS06	Was difficult to contact/uncooperative
	INS07	Accident occurred soon after effective date
Lost Wages	LW01	Claimant worked for self or a family member
	LW03	Claimant recently started employment

Table 7

Claim Variables Available Early in Life of Claim

Variable	Description
AGE	Age of claimant
POLLAG	Lag from policy inception to date of accident[8]
RPTLAG	Lag from date of accident to date reported
TREATLAG	Lag from date of accident to earliest treatment by service provider
AMBUL	Ambulance charges
PARTDIS	The claimant partially disabled
TOTDIS	The claimant totally disabled
LEGALREP	The claimant represented by an attorney

[8] POLLAG, RPTLAG and TRTLAG are continuous variables.

One of the objectives of this research is to investigate which variables are likely to be of value in predicting fraud and abuse. To do this, procedures are needed for evaluating the importance of variables in predicting the target variable. Below, we present some methods that can be used to evaluate the importance of the variables.

Evaluating Variable Importance

A procedure that can be used to evaluate the quality of the fit when fitting complex models is generalized cross-validation (GCV). This procedure can be used to determine which variables to keep in the model, as they produce the best fit, and which to eliminate. Generalized cross-validation can be viewed as an approximation to cross-validation, a more computationally intensive goodness of fit test described later in this paper.

$$GCV = \frac{1}{N}\sum_{i=1}^{N}[\frac{y_i - \hat{f}(x_i)}{1-k/N}]^2$$

where N is the number of observations
y is the dependent variable
x is the independent variable(s)
k is the effective number of parameters or degrees of freedom in the model.

The effective degrees of freedom is the means by which the GCV error functions puts a penalty on adding variables to the model. The effective degrees of freedom is chosen by the modeler. Since MARS tests many possible variables and possible basis functions, the effective degrees of freedom used in parameterizing the model is much higher than the actual number of basis function in the final model. Steinberg states that research indicates that k should be two to five times the number of basis functions in the model, although some research suggests it should be even higher (Steinberg, 2000).

The GCV can be used to rank the variables in importance. To rank the variables in importance, the GCV is computed with and without each variable in the model.

For neural networks, a statistic known as the sensitivity can be used to assess the relative importance of variables. The sensitivity is a measure of how much the predicted value's error increases when the variables are excluded from the model one at a time. Potts (Potts, 2000) and Francis (Francis, 2001) described a procedure for computing this statistic. Many of the major data mining packages used for fitting neural networks supply this statistic or a ranking of variables based on the statistic. Statistical procedures for testing the significance of variables are not well developed for neural networks. One approach is to drop the least important variables from the model, one at a time and evaluate whether the fit deteriorates on a sample of claims that have been held out for testing. On a large database this approach can be time consuming and inefficient, but it is feasible on small databases such as the AIB database.

Table 8 displays the ranking of variable importance from the MARS model. Table 9 displays the ranking of importance from the neural network model. The final model fitted by MARS uses only the top 12 variables in importance. These were the variables that were determined to have made a significant contribution to the final model. Only variables included in the model, i.e., found to be significant are included in the tables.

Table 8

MARS Ranking of Variables

Rank	Variable	Description
1	LEGALREP	Legal Representation
2	TRTMIS	Treatment lag missing
3	ACC04	Single vehicle accident
4	INJ01	Injury consisted of strain or sprain only
5	AGE	Claimant age
6	PARTDIS	Claimant partially disabled
7	ACC14	Property damage was inconsistent with accident
8	CLT02	Had a history of previous claims
9	POLLAG	Policy lag
10	RPTLAG	Report lag
11	AMBUL	Ambulance charges
12	ACC15	Very minor impact collision

The ranking of variables as determined by applying the sensitivity test to the neural network model is shown below.

Table 9

Neural Network Ranking of Variables

Rank	Variable	Description
1	LEGALREP	Legal Representation
2	TRTMIS	Treatment lag missing
3	AMBUL	Ambulance charges
4	AGE	Claimant age
5	PARTDIS	Claimant partially disabled
6	RPTLAG	Report lag
7	ACC04	Single vehicle accident
8	POLLAG	Policy lag
9	CLT02	Had a history of previous claims
10	INJ01	Injury consisted of strain or sprain only
11	ACC01	No report by police officer at scene
12	ACC14	Property damage was inconsistent with accident

Both the MARS and the neural network find the involvement of a lawyer to be the most important variable in predicting fraud and abuse. Both procedures also rank as second a missing value on treatment lag. The value on this variable is missing when the claimant has not been to an outpatient health care provider, although in over 95% of these cases, the claimant has visited an emergency room.[9] Note that both medical paid and total paid for this group is less than one third of the medical paid and total paid for claimants who visited a provider. Thus the TRTMIS (treatment lag missing) variable appears to be a surrogate for not using an outpatient provider. The actual lag in obtaining treatment is not an important variable in either the MARS or neural network models.

Explaining the Model

Below are the formulas for the model fit by MARS. Again note that some basis functions created by MARS were found not to be significant and are not shown. To assist with interpretation, Table 10 displays a description of the values of some of the variables in the model.

```
BF1 = (LEGALREP = 1)
BF2 = (LEGALREP = 2)
BF3 = ( TRTLAG = missing)
BF4 = ( TRTLAG ≠ missing)
BF5 = ( INJ01 = 1) * BF2
BF7 = ( ACC04 = 1) * BF4
BF9 = ( ACC14 = 1)
BF11 = ( PARTDIS = 1) * BF4
BF15 = max(0, AGE - 36) * BF4
BF16 = max(0, 36 - AGE ) * BF4
BF18 = max(0, 55 - AMBUL ) * BF15
BF20 = max(0, 10 - RPTLAG ) * BF4
BF21 = ( CLT02 = 1)
BF23 = POLLAG * BF21
BF24 = ( ACC15 = 1) * BF16

 Y = 0.580 - 0.174 * BF1 - 0.414 * BF3 + 0.196 * BF5 - 0.234 * BF7
+ 0.455 * BF9 + 0.131 * BF11 - 0.011 * BF15 - 0.006 * BF16 +
.135E-03 * BF18 - 0.013 * BF20 + .286E-03 * BF23 + 0.010 * BF24
```

[9] Because of the strong relationship between a missing value on treatment lag and the dependent variable, and the high percentage of claims in this category which had emergency room visits, an indicator variable for emergency room visits was tested as a surrogate. It was found not to be significant.

Table 10

Description of Categorical Variables

Variable	Value	Description
LEGALREP	1	No legal representation
	2	Has legal representation
INJ01	1	Injury consisted of strain or sprain only
	2	Injury did not consist of strain or sprain only
ACC04	1	Single vehicle accident
	2	Two or more vehicle accident
ACC14	1	Property damage was inconsistent with accident
	2	Property damage was consistent with accident
PARTDIS	1	Partially disabled
	2	Not partially disabled
CLT02	1	Had a history of previous claims
	2	No history of previous claims
ACC15	1	Was very minor impact collision
	2	Was not very minor impact collision

The basis functions and regression produced by MARS assist the analyst in understanding the impact of the predictor variables on the dependent variable. From the formulae above, it can be concluded that

1) when a lawyer is not involved (LEGALREP = 1), the probability of fraud or abuse declines by about 0.17
2) when the claimant has legal representation and the injury is consistent with a sprain or strain only, the probability of fraud or abuse increases by 0.2
3) when the claimant does not receive treatment from an outpatient health care provider (TRTLAG = missing), the probability of abuse declines by 0.41
4) a single vehicle accident where the claimant receives treatment from an outpatient health care provider (treatment lag not missing) decreases the probability of fraud by 0.23
5) if property damage is inconsistent with the accident, the probability of fraud or abuse increases by 0.46
6) if the claimant is partially disabled and receives treatment from an outpatient health care provider the probably of fraud or abuse is increased by 0.13

Of the red flag variables, small contributions were made by the claimant having a previous history of a claim[10] and the accident being a minor impact collision. Of the objective continuous variables obtained from the claim file, variables such as claimant age, report lag and policy lag have a small impact on predicting fraud or abuse.

Figures 11 and 12 display how MARS modeled the impact of selected continuous variables on the probability of fraud and abuse. For claims receiving outpatient health

[10] This variable only captures history of a prior claim if it was recorded by the insurance company. For some companies participating in the study, it was not recorded.

care, report lag has a positive impact on the probability of abuse, but its impact reaches its maximum value at about 10 days. Note the interaction between claimant age and ambulance costs displayed in Figure 12. For low ambulance costs, the probability of abuse rises steeply with claimant age and maintains a relatively high probability except for the very young and very old claimants. As ambulance costs increase, the probability of fraud or abuse decreases, and the decrease is more pronounced at lower and higher ages. Ambulance cost appears to be acting as a surrogate for injury severity.

Figure 11

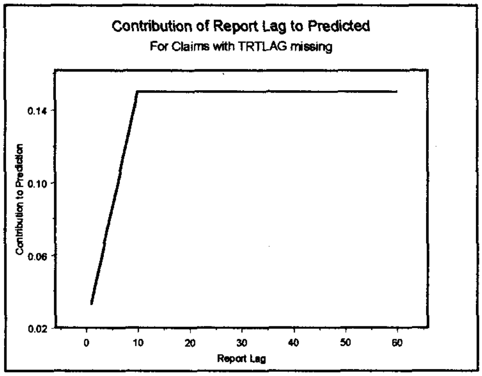
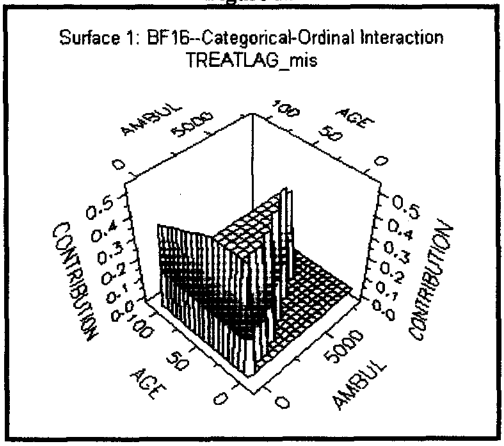


Figure 12



This section on explaining the model illustrates one of the very useful qualities of MARS ~~as compared to neural networks~~: the output of the model is a formula which describes the relationships between predictor and dependent variables and which can be used to explain the model to management. To some extent, the sensitivity measure assists us in understanding the relationships fit by the neural network model, as it provides a way to assess the importance of each of the variables to the prediction. However, the actual functional relationships between independent and dependent variables are not typically available and the model can be difficult to explain to management.[11]

Evaluating the Goodness of the Fit and Comparing the Accuracy
One approach for testing the accuracy of models that is commonly used in data mining applications is to have separate training and testing samples. This approach was used in the previous example. Typically one half to one third of the data is held out for testing. However, when the database used for modeling is small, the analyst may not want to lose a large portion of the data to testing. Moreover, as the testing is performed on a relatively small sample, the goodness of fit results may be sensitive to random variation in the subsets selected for training and testing. An alternative procedure that allows more of the data to be used for fitting and testing is cross-validation. Cross-validation involves iteratively holding out part of the sample, fitting the model to the remainder of the sample and testing the goodness of the fitted model on the held out portion. For instance, the sample may be divided into 4 groups. Three of the groups are used to fit the model and one is used for testing. The process is repeated four times, and the goodness of fit statistics for the four test samples are averaged. As the AIB database is relatively small for a data mining application, this is the procedure used. Testing was performed using four fold cross-validation.

Both a MARS model and a neural network model were fit to four samples of the data. Each time the fitted model was used to predict the probability of fraud or abuse for one quarter of the data that was held out. The predictions from the four test samples were then combined to allow comparison of the MARS and neural network procedures.

Table 11 presents some results of the analysis. This table presents the R^2 of the regression of ASSESS on the predicted value from the model. The table shows that the neural network R^2 was higher than that of MARS. The table also displays the percentage of observations whose values were correctly predicted by the model. The predictions are based only on the samples of test claims. The neural network model correctly predicted 79% of the test claims, while MARS correctly predicted 77% of the test claims.

Table 11

Four Fold Cross-validation

Technique	R^2	Percent Correct
MARS	0.35	0.77
Neural Network	0.39	0.79

[11] Plate (2000) and Francis (2001) present a method to visualize the relationships between independent and dependent variables, The technique is not usually available in data mining software.

Tables 12 and 13 display the accuracy of MARS and the neural network in classifying fraud and abuse claims.[12] A cutoff point of 50% was used for the classification. That is, if the model's predicted probability of a 1 on ASSESS exceeded 50%, the claim was deemed an abuse claim. Thus, those claims in cell Actual =1 and Predicted=1 are the claims assessed by experts as probably abusive which were predicted to be abusive. Those claims in cell Actual=1, Predicted =0, are the claims assessed as probable abuse claims which were predicted by the model to be legitimate.

Table 12

MARS Predicted * Actual

Predicted	**Actual**		
	0	1	Total
0	738	160	898
1	157	344	501
Total	895	505	

Table 13

Neural Network Predicted * Actual

Predicted	**Actual**		
	0	1	Total
0	746	127	873
1	149	377	526
Total	895	505	

Table 14 presents the sensitivity and specificity of each of the models. The sensitivity is the percentage of events (in this case suspected abuse claims) that were predicted to be events. The specificity is the percentage of nonevents (in this case claims believed to be legitimate) that were predicted to be nonevents. Both of these statistics should be high for a good model. The table indicates that both the MARS and neural network models were more accurate in predicting nonevent or legitimate claims. The neural network model had a higher sensitivity than the MARS model, but both were approximately equal in their specificities. The neural network's higher overall accuracy appears to be a result of its greater accuracy in predicting the suspected fraud and abuse claims. Note that the sensitivity and specificity measures are dependent on the choice of a cutoff value. Thus, if a cutoff lower than 50% were selected, more abuse claims would be accurately predicted and fewer legitimate claims would be accurately predicted.

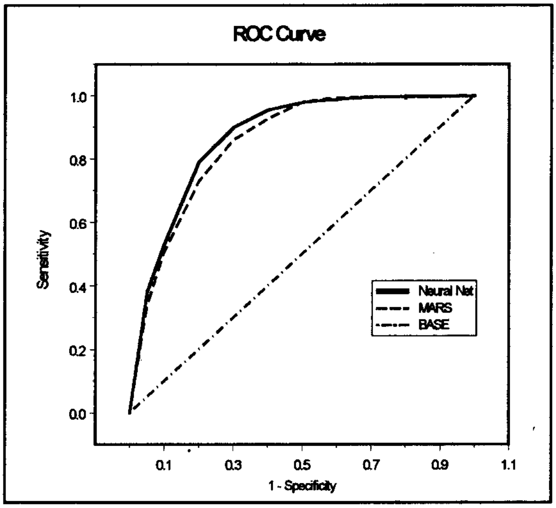
Table 14

Model	**Sensitivity**	**Specificity**
MARS	68.3	82.5
Neural Network	74.8	83.4

[12] These tables are often referred to as confusion matrices

A common procedure for visualizing the accuracy of models used for classification is the receiver operating characteristics (ROC) curve. This is a curve of sensitivity versus specificity (or more accurately 1.0 minus the specificity) over a range of cutoff points. When the cutoff point is very high (i.e. 1.0) all claims are classified as legitimate. The specificity is 100% (1.0 minus the specificity is 0), but the sensitivity is 0%. As the cutoff point is raised, the sensitivity increases, but so does 1.0 minus the specificity. Ultimately a point is reached where all claims are predicted to be events, and the specificity declines to zero. The baseline ROC curve (where no model is used) can be thought of as a straight line from the origin with a 45-degree angle. If the model's sensitivity increases faster than the specificity decreases, the curve "lifts" or rises above a 45-degree line quickly. The higher the "lift", the more accurate the model. It can be seen from the graph of the ROC curve that both the MARS and neural network models have significant "lift" but the neural network model has more "lift" than the MARS model.

Figure 13



A statistic that summarizes the predictive accuracy of a model as measured by an ROC curve is the area under the ROC curve (AUROC). A curve that rises quickly has more area under the ROC curve. Table 15 displays the AUROC for both models, along with their standard deviations and 95% confidence intervals. As the lower bound of the confidence interval for the neural network is below the higher bound of the confidence interval for MARS, it can be concluded that differences between the MARS model and the neural network model are not statistically significant.

Table 15

Statistics for Area Under the ROC Curve

Test Result Variables	Area	Std Error	Asymptotic Sig	Lower 95% Bound	Upper 95% Bound
MARS Probability	0.85	0.01	0.000	0.834	0.873
Neural Probability	0.88	0.01	0.000	0.857	0.893

Summary of Comparison

The ROC curve results suggest that in this analysis the neural network enjoyed a modest though not statistically significant advantage over MARS in predictive accuracy. It should be noted that the database used for this study was quite small for a data mining application and may produce results that do not generalize to larger applications. Steinberg (Steinberg, 2001) reports that on other applications MARS equaled or exceeded the performance of neural networks. It should also be noted that some of the key comparative strengths of MARS such as its ability to handle missing data were not a significant factor in the analysis, as all but one of the variables were fully populated. [13] In addition, MARS's capability of clustering levels of categorical variables together was not relevant to this analysis, as no categorical variable had more than two levels.

A practical advantage that MARS enjoys over neural networks is the ease with which results can be explained to management. Thus, one potential use for MARS is to fit a model using neural networks and then apply MARS to the fitted values to understand the functional relationships fitted by the neural network model. The results of such an exercise are shown below:

BF1 = (LEGALREP = 1)
BF2 = (LEGALREP = 2)
BF3 = (TRTLAG ≠ missing)
BF4 = (TRTLAG = missing)
BF5 = (INJ01 = 1)
BF7 = (ACC04 = 1) * BF3
BF8 = (ACC04 = 2) * BF3
BF9 = (PARTDIS = 1) * BF8
BF11 = max(0, AMBUL - 182) * BF2
BF12 = max(0, 182 - AMBUL) * BF2
BF13 = (ACC14 = 1) * BF3
BF15 = (CLT02 = 1) * BF3
BF17 = max(0, POLLAG - 21) * BF3
BF19 = max(0, AGE - 41) * BF3
BF20 = max(0, 41 - AGE) * BF3

[13] One of the claims was missing data on the AGE variable, and this claim was eliminated from the neural network analysis and from comparisons of MARS the neural network model. Had more claims been missing the AGE variable, we would have modeled it in the neural network.

BF21 = (INS06 = 1)
BF23 = max(0, RPTLAG - 24) * BF8
BF24 = max(0, 24 - RPTLAG) * BF8
BF25 = BF1 * BF4
BF27 = (ACC15 = 1) * BF8
BF29 = (INJ03 = 1) * BF2

Y = 0.098 - 0.272 * BF1 + 0.334 * BF3 + 0.123 * BF5 - 0.205 * BF7 + 0.145 * BF9 - .623E-04 * BF11 + .455E-03 * BF12 + 0.258 * BF13 + 0.100 * BF15 + .364E-03 * BF17- 0.004 * BF19 - 0.001 * BF20 + 0.152 * BF21 + .945E-03 * BF23 - 0.002 * BF24 + 0.135 * BF25 + 0.076 * BF27 - 0.073 * BF29

This model had an R^2 of 0.9. Thus, it was able to explain most of the variability in the neural network fitted model. Though the sensitivity test revealed that LEGALREP is the most significant variable in the neural network model, its functional relationship to the probability of fraud is unknown using standard neural network modeling techniques. As interpreted by MARS, the absence of legal representation reduces the probability of fraud by 0.272., even without interacting with other variables. LEGALREP also interacts with the ambulance cost variable, INJ03 (police report shows no injury) and no use of a health care provider (treatment lag missing). The sensitivity measure indicated that the presence or absence of a value for treatment lag was the second most important variable. As stated earlier, this variable can be viewed as a surrogate for use of an outpatient health care provider. The use of an outpatient health care provider (TRTLAG ≠ missing) adds 0.334 to the probability of fraud or abuse, but this variable also interacts with the policy lag, report lag, claimant age, partial disability, ACC04, (single vehicle accident), ACC14 (property damage inconsistent with accident) and CLT02 (history of prior claims).

The MARS model helps the user understand not only the nonlinear relationships uncovered by the neural network model, but also describes the interactions which were fit by the neural network.

A procedure frequently used by data mining practitioners when two or more approaches are considered appropriate for an application is to construct a hybrid model or average the results of the modeling procedures. This approach has been reported to reduce the variance of the prediction (Salford Systems, 1999). Table 16 displays the AUROC statistics resulting from averaging the results of the MARS and neural network models. The table indicates that the performance of the hybrid model is about equal to the performance of the neural network. (The graph including the ROC curve for the combined model is not shown, as the curve is identical to Figure 13 because the neural network and combined curves cannot be distinguished.) Salford Systems (Salford Systems, 1999) reports that the accuracy of hybrid models often exceeds that of its components, but usually at least equals that of the best model. Thus, hybrid models that combine the results of two techniques may be preferred to single technique models because uncertainty about the accuracy of the predicted values on non-sample data is reduced.

Table 16

Statistics for Area Under the ROC Curve

Test Result Variables	Area	Std Error	Asymptotic Sig	Lower 95% Bound	Upper 95% Bound
MARS Probability	0.853	0.01	0.000	0.834	0.873
Neural Probability	0.875	0.01	0.000	0.857	0.893
Combined Probability	0.874	0.01	0.000	0.857	0.892

Using Model Results

The examples in this paper have been used to explain the MARS technique and compare it to neural networks. The final example in this paper has been a fraud and abuse application that used information about the PIP claim that would typically be available shortly after the claim is reported to predict the likelihood that the claim is abusive or fraudulent. The results suggest that a small number of variables, say about a dozen, are effective in predicting fraud and abuse. Among the key variables in importance for both the neural network model and MARS are use of legal representation, use of an outpatient health care provider (as proxied by TRTLAG missing) and involvement in a single vehicle accident. Due to the importance of legal representation, it would appear useful for insurance companies to record information about legal representation in computer systems, as not all companies have this data available.

The results of both the MARS and neural network analysis suggest that both claim file variables (present in most claims databases) and red flag variables (common wisdom about which variables are associated with fraud) are useful predictors of fraud and abuse. However, this and other studies support the value of using analytical tools for identifying potentially abusive claims. As pointed out by Derrig (Derrig, 2002), fraud models can help insurers sort claims into categories related to the need for additional resources to settle the claim efficiently. For instance, claims assigned a low score by a fraud and abuse model, can be settled quickly with little investigative effort on the part of adjusters. Insurers may apply increasingly greater resources to claims with higher scores to acquire additional information about the claimant/policyholder/provider and mitigate the total cost of the claim. Thus, the use of a fraud model is not conceived as an all or nothing exercise that classifies a claim as fraudulent or legitimate, but a graduated effort of applying increasing resources to claims where there appears to be a higher likelihood of material financial benefit from the expenditures.

Conclusion

This paper has introduced the MARS technique and compared it to neural networks. Each technique has advantages and disadvantages and the needs of a particular application will determine which technique is most appropriate.

One of the strengths of neural networks is their ability to model highly nonlinear data. MARS was shown to produce results similar to neural networks in modeling a nonlinear function. MARS was also shown to be effective at modeling interactions, another strength of neural networks.

In dealing with nominal level variables, MARS is able to cluster together the categories of the variables that have similar effects on the dependent variable. This is a capability not possessed by neural networks that is extremely useful when the data contain categorical variables with many levels such as ICD9 code.

MARS has automated capabilities for handling missing data, a common feature of large databases. Though missing data can be modeled with neural networks using indicator variables, automated procedures for creating such variables are not available in most standard commercial software for fitting neural networks. Moreover, since MARS can create interaction variables from missing variable basis functions and other variables, it can create surrogates for the missing variables. Thus, on applications using data with missing values on many variables, or data where the categorical variables have many values, one may want to at least preprocess the data with MARS to create basis functions for the missing data and categorical variables which can be used in other procedures.

A significant disadvantage of neural networks is that they are a "black box". The functions fit by neural networks are difficult for the analyst to understand and difficult to explain to management. One of the very useful features of MARS is that it produces a regression like function that can be used to understand and explain the model; therefore it may be preferred to neural networks when ease of explanation rather than predictive accuracy is required. MARS can also be used to understand the relationships fit by other models. In one example in this paper MARS was applied to the values fit by a neural network to uncover the important functional relationships modeled by the neural network.

Neural networks are often selected for applications because of their predictive accuracy. In a fraud modeling application examined in this paper the neural network outperformed MARS, though the results were not statistically significant. The results were obtained on a relatively small database and may not generalize to other databases. In addition, the work of other researchers suggests that MARS performs well compared to neural networks. However, neural networks are highly regarded for their predictive capabilities. When predictive accuracy is a key concern, the analyst may choose neural networks rather than MARS when neural networks significantly outperform MARS. An alternative approach that has been shown to improve predictive accuracy is to combine the results of two techniques, such as MARS and neural networks, into a hybrid model.

This analysis and those of other researchers supports the use of intelligent techniques for modeling fraud and abuse. The use of an analytical approach can improve the performance of fraud detection procedures that utilize red flag variables or subjective claim department rules by 1) determining which variables are really important in predicting fraud, 2) assigning an appropriate weight to the variables when using them to predict fraud or abuse, and 3) using the claim file and red flag variables in a consistent manner across adjusters and claims.

References

Allison, Paul. *Missing Data*, Sage Publications, 2001

Berry, Michael J. A., and Linoff, Gordon, *Data Mining Techniques*, John Wiley and Sons, 1997

Brockett, Patrick L., Xiaohua Xia and Richard A. Derrig, 1998, "Using Kohonen's Self-Organizing Feature Map to Uncover Automobile Bodily Injury Claims Fraud", *Journal of Risk and Insurance*, June, 65:2.

Brockett, Patrick L., Richard A. Derrig, Linda L. Golden, Arnold Levine and Mark Alpert, "Fraud Classification Using Principal Component Analysis of RIDITs", *Journal of Risk and Insurance*, September, 2002, pp. 341-371.

Dhar, Vasant and Stein, Roger, *Seven Methods for Transforming Corporate Data Into Business Intelligence*, Prentice Hall, 1997

Derrig, Richard A., Herbert I. Weisberg and Xiu Chen, 1994, "Behavioral Factors and Lotteries Under No-Fault with a Monetary Threshold: A Study of Massachusetts Automobile Claims", *Journal of Risk and Insurance*, June, 1994, 61:2: 245-275.

Derrig, Richard A., and Krzysztof M. Ostaszewski, "Fuzzy Techniques of Pattern Recognition in Risk and Claim Classification", *Journal of Risk and Insurance*, September, 1995, 62:3: 447-482.

Derrig, Richard, "Patterns, Fighting Fraud With Data", *Contingencies*, Sept/Oct, 1999, pp. 40–49.

Derrig, Richard A., and Valerie Zicko, "Prosecuting Insurance Fraud: A Case Study of The Massachusetts Experience in the 1990s", Risk Management and Insurance Research, 2002.

Derrig, Richard, "Insurance Fraud", *Journal of Risk and Insurance*, September, 2002, pp. 271-287

Francis, Louise, "Neural Networks Demystified", *Casualty Actuarial Society Forum*, Winter 2001, *pp 253 - 320.*

Freedman, Roy S., Klein, Robert A. and Lederman, Jess, *Artificial Intelligence in the Capital Markets*, Probus Publishers 1995

Hastie, Trevor, Tibshirani, Robert, *Generalized Additive Models*, Chapman and Hall, 1990

Hastie, Trevor, Tibshirani, Robert and Freidman, Jerome, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer, 2001

Hayward, Gregory, "Mining Insurance Data to Promote Traffic Safety and Better Match Rates to Risk", *Casualty Actuarial Society Forum*, Winter 2002, *pp. 31 – 56.*

Hosmer, David W. and Lemshow, Stanley, *Applied Logistic Regression*, John Wiley and Sons, 1989

Keefer, James, "Finding Causal Relationships By Combining Knowledge and Data in Data Mining Applications", Paper presented at Seminar on Data Mining, University of Delaware, April, 2000.

Lawrence, Jeannette, *Introduction to Neural Networks: Design, Theory and Applications*, California Scientific Software, 1994

Little, Roderick and Rubin, Donald, *Statistical Analysis with Missing Data*, John Wiley and Sons, 1987

Marsh, Lawrence and Cormier, David, *Spline Regression Models*, Sage Publications, 2002

Martin, E. B. and Morris A. J., "Artificial Neural Networks and Multivariate Statistics", in *Statistics and Neural Networks: Advances at the Interface*, Oxford University Press, 1999, pp. 195 – 292

Miller, Robert and Wichern, Dean, *Intermediate Business Statistics*, Holt, Reinhart and Winston, 1977

Neal, Bradford, *Bayesian Learning for Neural Networks*, Springer-Verlag, 1996

Plate, Tony A., Bert, Joel, and Band, Pierre, "Visualizing the Function Computed by a Feedforward Neural Network", *Neural Computation*, June 2000, pp. 1337-1353.

Potts, William J.E., *Neural Network Modeling: Course Notes*, SAS Institute, 2000

Salford Systems, "Data mining with Decision Trees: Advanced CART Techniques", Notes from Course, 1999

Smith, Murry, *Neural Networks for Statistical Modeling*, International Thompson Computer Press, 1996

Speights, David B, Brodsky, Joel B., Chudova, Durya L., "Using Neural Networks to Predict Claim Duration in the Presence of Right Censoring and Covariates", *Casualty Actuarial Society Forum*, Winter 1999, pp. 255-278.

Steinberg, Dan, "An Introduction to MARS", Salford Systems, 1999

Steinberg, Dan, "An Alternative to Neural Networks: Multivariate Adaptive Regression Splines (MARS)", *PC AI*, January/February, 2001, pp. 38 – 41

Venebles, W.N. and Ripley, B.D., Modern *Applied Statistics with S-PLUS*, third edition, Springer, 1999

Viaene, Stijn, Derrig, Richard, Baesens, Bart and Dedene Guido, "A Comparison of State-of-the-Art Classification Techniques for Expert Automobile Insurance Fraud Detection", *Journal of Risk and Insurance*, September, 2002, pp. 373 – 421

Warner, Brad and Misra, Manavendra, "Understanding Neural Networks as Statistical Tools", *American Statistician*, November 1996, pp. 284 – 293

Weisberg, Herbert I., and Richard A. Derrig, "Fraud and Automobile Insurance: A Report on the Baseline Study of Bodily Injury Claims in Massachusetts", *Journal of Insurance Regulation*, June, 1991, 9:4: 497-541.

Weisberg, Herbert I., and Richard A. Derrig, "Quantitative Methods for Detecting Fraudulent Automobile Injury Claims", AIB Cost Containment/Fraud Filing (DOI Docket R95-12), Automobile Insurers Bureau of Massachusetts, July,1993, 49-82.

Weisberg, Herbert I., and Richard A. Derrig, "Identification and Investigation of Suspicious Claims", AIB Cost Containment/Fraud Filing (DOI Docket R95-12), Automobile Insurers Bureau of Massachusetts, July, 1995, 192-245, Boston.

Weisberg, Herbert I., and Richard A. Derrig, "Massachusetts Automobile Bodily Injury Tort Reform", *Journal of Insurance Regulation*, Spring, 1992, 10:3:384-440.

Zapranis, Achilleas and Refenes, Apostolos-Paul, *Principles of Neural Network Model Identification Selection and Adequacy*, Springer-Verlag, 1999

Rainy Day: Actuarial Software and Disaster Recovery

Aleksey S. Popelyukhin, Ph.D.

Rainy Day:

Actuarial Software and Disaster Recovery

Aleksey S. Popelyukhin, Ph.D.

Abstract

Tragic events with disastrous consequences that are happening all around the Globe made Disaster Recovery and Continuity Planning a **much higher** priority for every company. Scenarios, in which data centers, paper documents and even recovery specialists themselves may perish, became more probable.

Both, actuarial workflow and actuarial software design should be affected by disaster recovery strategy. Actuaries may simplify recovery task and insure higher rate of success if they properly modify their applications' architecture and their approaches to documenting algorithms and storing structured data.

The article attempts to direct actuaries to strategies that may increase chances of complete recovery: from separation of data and algorithms to effective storage of actuarial objects to automated version management and self-documenting techniques.

The matter of continuity of actuarial operations is in the hand of actuaries themselves.

Rainy Day:
Actuarial Software and Disaster Recovery

Aleksey S. Popelyukhin, Ph.D.

Failed Assumptions

Presumably, every insurance company has a backup system. Files, databases and documents are copied to tapes or CDs and stored offsite. It gives protection against hard disk failure, rogue viruses and provides an audit trail.

Many of the existing backup solutions, however, are built on the ***assumptions*** that after disaster strikes restoration will be performed by the same personnel to the same (compatible) hardware/software system. As the events of September 11th painfully demonstrated, these assumptions may not exactly hold true.

The following unfortunate scenarios became much more plausible:

- *One may have tapes (or other media), but not know what to do with them*
- *One may know what to do with the tape, but not have a compatible system to perform a restoration*
- *One may restore the files, but not have the software to read them*
- *One may get files restored and software working, but not have anyone around to explain how to use it.*

Consequently, disaster recovery and business continuity plans have to address them.

Personnel can perish

A company's tapes stored offsite may survive a disastrous event, but it does not mean they can be used effectively for the restoration. It may not be immediately clear how to perform a restoration: on what hardware with what backup software and in which order. It may also happen that the backup/restore software is so old it requires an older Operating System (OS) not available anymore. It may not be evident how to reinstall software without the manual and a license key. Moreover, there might not be anyone who remembers where to restore, what to restore and in which order.

Thus, it is *imperative* to escrow not only tapes, but also:

- *installation software (OS, backup/restore and other environment programs),*
- *manuals and documentation,*
- *licenses and support info,*
- *and restoration **instructions**.*

Sure, it is not up to actuaries to perform the restoration tasks, but it is in their best interests to make sure their software is part of the restoration effort (including installation disks, manuals and licenses) and that they do *everything* possible to simplify that effort.

Restoration Priorities

Any BIA (Business Impact Analysis) study will assign very low priority to the restoration of an Actuarial subsystem. Indeed, experience shows (see [1]) that the most important service for the business continuity is communications.

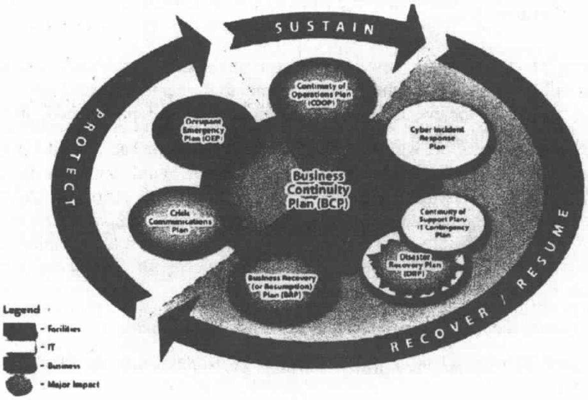


Fig 1 (see [2])

Experience shows that restoration priorities start with e-mail and end with the actuarial subsystem:

- *e-mail/communications*
- *accounting*
- *payroll*
- *trade/marketing*
- *underwriting*
- *claims*
- *actuarial applications*

There is nothing wrong with that picture: it just means that actuaries have to be ready to perform some or all restoration tasks by themselves and not wait for IT department help. It also implies that actuaries would be much better off if their applications were easy to restore or *reproduce* even if some knowledgeable personnel were not available.

Forced upgrading

The implicit assumption behind the majority of existing recovery plans is that restoration would be performed on the same version of hardware/OS combo that was used for backup. Or (given the long upgrade cycles of the recent past) quite similar and compatible versions. Not anymore. Every major OS upgrade may render backup/restore software useless; every advance in drive technology may make backup tapes unreadable. Skip a couple upgrade cycles and you may fail to find the appropriate drive to read your tapes and have no software to recognize recording format. And if the company's computers are destroyed, the company may be *forced* to upgrade.

Thus, downloading patches for backup software should be done as vigorously as downloading for anti-virus and security purposes. It is also crucial to monitor availability of the tape drives backward-compatible with existing tapes.

The same forced upgrading trap may occur with actuarial software. During restoration, one may discover that new computers come only with the newer OS, Utilities and Spreadsheet versions, which are not necessarily compatible with existing files. Imagine if one had to read VisiCalc or WordStar files today. Nobody guarantees that Oracle 7 will run on Windows XP or that Excel will properly interpret that old trustworthy *.wk3 file. It is even more of a problem for third-party proprietary software. It has to be maintained compatible with the latest OS, compiler, hardware key protection software and, possibly, a spreadsheet or a database: quite a formidable task.

Third-party actuarial applications

Sales data from the suppliers of actuarial software imply that actuaries heavily rely on "shrink-wrapped" applications from the third parties. Development, distribution and compatibility of theses applications are controlled by their vendors. Yet, disaster recovery cannot be completed without restoring full functionality of these programs. Actuaries cannot do much about these applications except to make sure that they can be restored.

Adequate code protection

Actuaries may require that their license agreement include a contractual obligations from the supplier for:

- *Adequate code base protection and*
- *Technology Assurance*

Adequate code base protection should include measures taken by the vendor to protect the application code with backups and offsite escrow. In addition, the vendor has to guarantee access to the code in case it goes out business or cannot longer support an application.

Technology Assurance is a fancy name for the continuous compatibility upgrades and patches that would guarantee application compatibility with the ever-changing software environment. Vendors should make sufficient effort to maintain their applications capable of running under the latest OS and interoperating with the latest spreadsheet or the underlying database.

Hardware keys

Actuaries also have to clarify a procedure for restoring third-party applications that utilize hardware-key protection schemes. In a plausible disaster scenario, hardware keys may cease to exist rendering an application useless. In that case

- *Does the License Agreement provide for* replacement *keys?*
- *Can the vendor deliver replacement keys from Australia (England, Connecticut)* fast*?*
- *Does the vendor provide a* downloadable *temporarily unprotected version?*

All these questions have to be answered before the disaster strikes: this way, actuaries can avoid a few unpleasant surprises during restoration.

In-house development

Programming cycle

Aside from analysis, actuaries perform some activities that closely resemble software development. Indeed, no matter what computer language they are using (Lotus, PL/1, APL, Mathematica, VBA), they are programming. Thus, as programmers, they have to conform to development cycle routines established in a programming world. Documentation, versioning, testing, debugging – these activities are well studied, and even automated.

Both actuarial workflow and software design should be affected by disaster the recovery strategy. Actuaries have to design their applications in such a way that somebody else other than the designers can *understand* the spreadsheet, the code and the logic.

Separating Data and Algorithms

A spreadsheet is a very popular actuarial tool. It is so versatile: it can be used as a database and as a calculation engine, as an exchange format and as report generator, as a programming environment and as a rich front-end to Internet. Actuaries use spreadsheets in all these aspects; the problem, though, arises when they use multiple features in one file. More precisely, when they use single spreadsheet file as the *engine* for calculations *and* as the *storage* for results of these calculations, creating multiple copies of the engine.

Actuaries do realize that input data like loss triangles, premiums vectors and industry factors come from outside and do not belong to their calculation template. What they rarely realize (or don't realize at all) is that output results such as predicted ultimates or fitted distribution parameters *do not belong* to the template either, and that they (results) have to be *stored outside* just like input data.

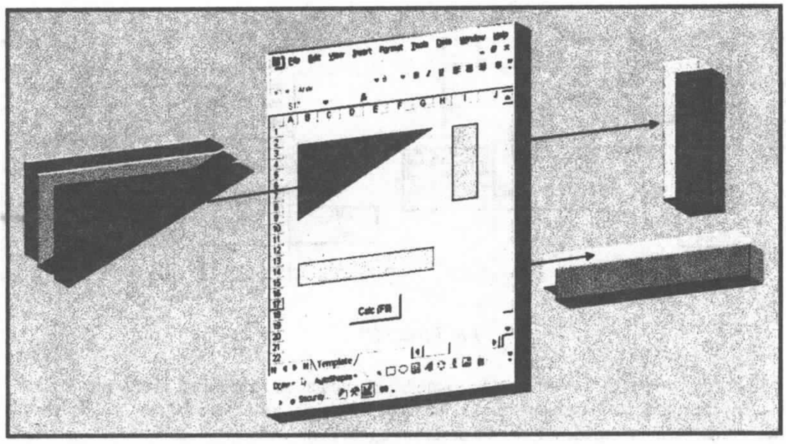


Fig 2

Usually, it is not the case: actuaries routinely create 72 files with the same algorithm for 72 lines of business rather than keep one file and storing 72 answers separately. This strategy creates an obstacle for the effective:

- *debugging/versioning,*
- *modifications/improvements,*
- *integrity/security,*
- *reporting and*
- *multi-user access.*

Indeed, correcting an error in one file takes 72 times less time than correcting the same error in 72 copies of that file. Extracting answers for reports from 72 files requires much more effort then summarizing 72 records in the database. And it is much harder to guarantee that nobody modified the 57th file incorrectly.

From a recovery standpoint, restoring a single file with formulas and separate data records is definitely easier then restoring 72 files with commingled data and formulae, especially given that the probability of a corrupted file is 72 times greater for 72 files.

It would be wise for actuaries to modify their workflow and spreadsheet design in order to separate data and algorithms. Rethinking their methodology in this light, actuaries *inevitably* will arrive at the idea to store data along with some kind of description, that is, to treat data as Actuarial Objects♣ (see [3]).

♣ See [5] to learn how to program custom objects in Excel VBA.

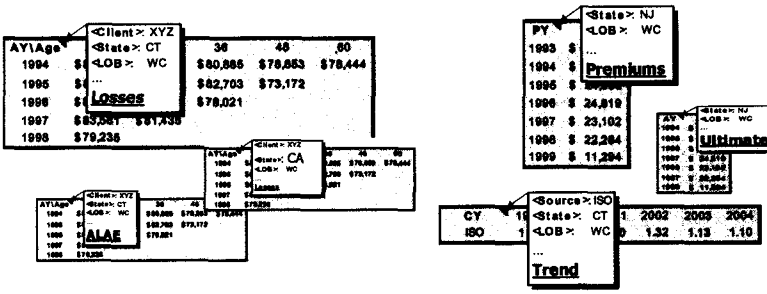


Fig 3 (see [4])

The logical extension of this idea would be to modify a calculation template in such a way that it understands these descriptions and acts upon them, serving like a traffic cop for the data objects. In other words, build an engine for objects processing (see [4]).

This architecture would streamline actuarial workflow, encourage debugging and modifications, simplify reporting and improve enormously recovery success chances.

Input and output to simpler formats

If (in addition to all its functionality) Microsoft Excel incorporated simulations, there would be only one actuarial software program: Excel. And, thus, no worries about file formats, conversions and availability of the file reader programs. Fortunately or unfortunately, this is not the case, and actuaries have to rely on applications with different file formats. The problem is that, as time passes, it will be harder and harder to find a reader program for some obscure and proprietary files. So, for the sake of disaster recovery actuaries should rely on the most ubiquitous file formats.

In the foreseeable future one can count on the ability to read ASCII (including XML and HTML), *.xls and *.doc files. Perhaps, *.dbf and *.pdf readers will be easily available too. Consequently, it is always a good idea to store an extra copy of the most crucial data in one of the aforementioned formats.

For example, SQL Server and Oracle tables can be dumped into ASCII*. Microsoft Access can import/export tables in *.xls or *.dbf formats. Excel files can be seamlessly converted to XML (structured ASCII). And, unless there are trade-secret considerations, it is always a good idea to export VBA modules to *.bas text files.

From a disaster recovery standpoint, it is important that third-party software has the capability to read and write to one of the ubiquitous formats (a side benefit of that capability would be possibility of data exchange and integration with other software programs).

---

* In fact, the whole database can be restored from ASCII files by running SQL scripts (ASCII) that recreate database structure and loading tables' content from the dump file (ASCII).

Version management

If a calculation algorithm is used (or is going to be used) *more than once*, it needs versioning. Indeed, if a "separation of data and algorithms" paradigm is embraced and implemented, it becomes quite practical and useful to maintain a version of the algorithm (in case of Excel: a version of template used for calculations).

The usefulness becomes obvious once one considers saving the version of the calculation engine along with results of calculation. Doing so helps immensely in audit trailing, debugging and, of course, recovery.

The practicality derives from the fact that (presumably) the number of calculation engines/templates is limited (usually the same algorithm can be reused for analysis of multiple contracts, LOBs and products). So maintaining version information for a few files is not an overburdening chore.

Microsoft Office applications provide adequate facilities for versioning: Word automatically updates "Revision number" (File/Properties/Statistics) and Excel allows custom properties to be linked to cells inside the spreadsheet.

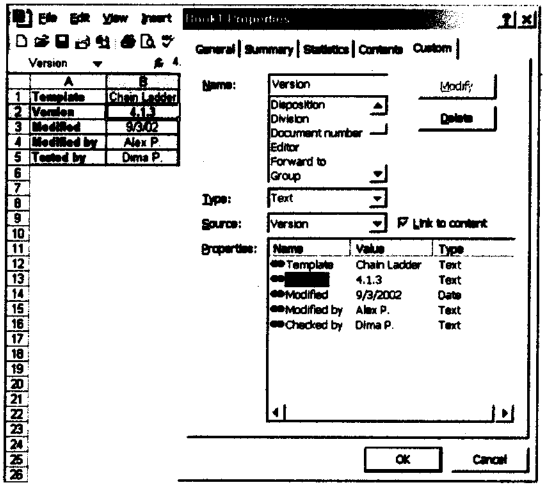


Fig 4

If the user would dedicate a cell in a spreadsheet to store version info and add one line of VBA code to the Workbook_BeforeSave event, he would get a "poor man" versioning mechanism for free.

```
Private Sub Workbook_BeforeSave(ByVal SaveAsUI As Boolean, Cancel As Boolean)
    Range("Version").Value = InputBox("Version number: ", "Properties", Range("Version"))
End Sub
```

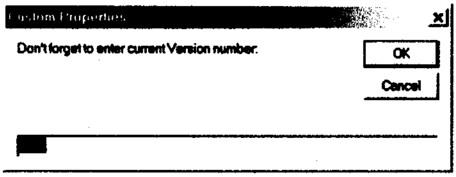


Fig 5

If there is a necessity to synchronize the version number through several files, the cell linked to the custom property can contain a formula referring to the information in the other (main) template.

Using files properties for versioning (and, possibly, other information♦) has some nice side benefits: one can use them for targeted file searches (File/Open/Tools/Search/Advanced).

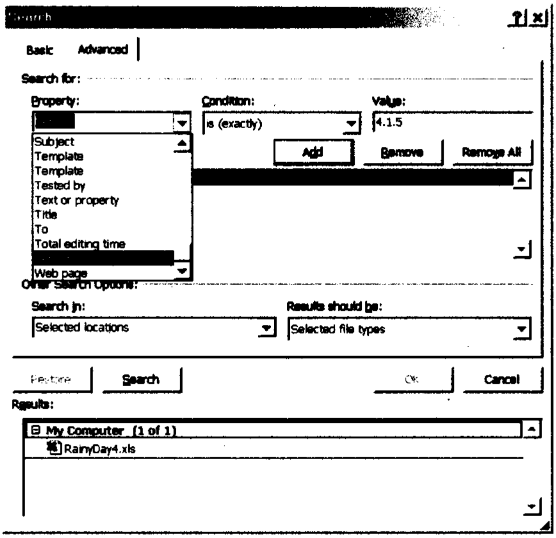


Fig 6

♦ It is always a good idea to dedicate an area (with named cells) in a spreadsheet to meta-information about it and link these cells to the Custom Properties like "Created by", "Last Modified by", "Verified by", etc…

More important, however, is the fact that Version info can be stored alongside with Results generated by the template, providing perfect means for the audit tracking.

Documentation

Knowledge about available documentation features and familiarity with restoration techniques may help actuaries to *design* their software in order to greatly simplify potential recovery efforts.

Usually, big nice printed manuals and an interactive online help system* are reserved for very large projects only. It is unreasonable to expect an actuary to write a manuscript for every Excel spreadsheet he creates in a hurry. Nevertheless, several simple approaches can be employed to greatly simplify restoration tasks:

- *Self-documenting,*
- *Excel Comments,*
- *Code Remarks.*

Self-documenting features

Microsoft Office programs provide adequate assistance for self-documentation attempts. If an author follows a few unobtrusive styling conventions, then Microsoft Word can easily generate an outline or table of content. Microsoft Access has an indispensable utility called Documenter (Tools/Analyze/Documenter):

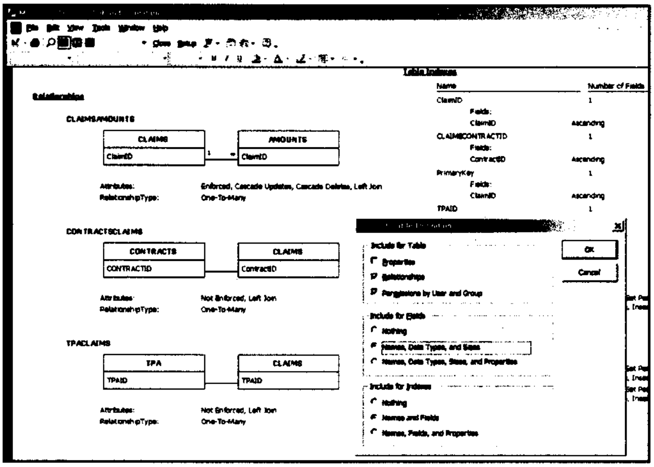


Fig 7

* Several packages on the market, most notably RoboHelp from eHelp, can convert Microsoft Word file(s) into a full-featured interactive help system, either Windows or HTML based.

Documenter generates Notepad or Excel files with the information about any object in a database with as many details as necessary.

Microsoft Access also provides facility for fields' descriptions (along with tables, queries, forms and reports descriptions): it would be unwise not to use it.

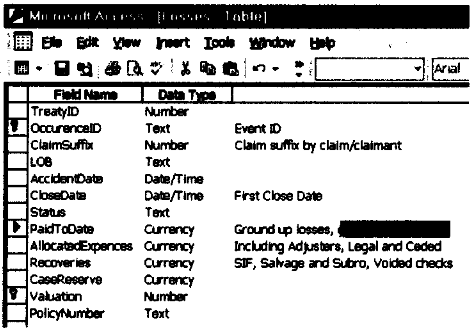


Fig 8

Excel is, by its very nature, self-documenting: by clicking on a cell the user can see the value in a cell and a formula behind it in a Formula Bar. To view formulae in multiple cells one can open a second window of the same spreadsheet (Window/new Window) and switch its mode to Formula View (Tools/Options/View/Formulas or just *press* CTRL - ~).

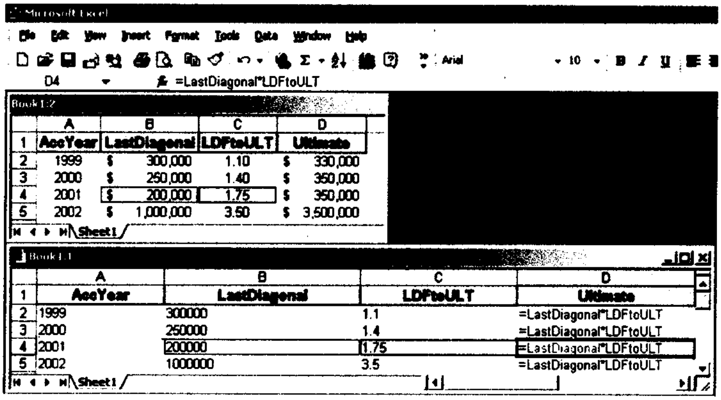


Fig 9

In recognition that building models in Excel is, essentially, some kind of programming, Microsoft added a quintessential debugging tool to Excel: Watch Window (Tools/Formula Auditing/Show Watch Window or *right click on cell*/Add Watch). Watch Window allows the user to track values and simultaneously see formulas of multiple cells located anywhere in a spreadsheet. The tool's value is not only in debugging, but also in ad-hoc goal seeking and audit trailing. Accompanying it is the step-by-step Formula Evaluation tool (Tools/Formula Auditing/Evaluate Formula), which used to be Excel4 Macro debugging instrument.

Sure, cells are not the only place for formulae and settings: PivotTables, Solver add-in, Links, External Data ranges, Web Queries and even Conditional Formatting contain important information which could be crucial for understanding the functionality of an algorithm.

To display Calculated Fields and Items in a PivotTable or PivotChart, click anywhere on a PivotTable and then in a PivotTable toolbar select Formulas/List Formulas.

To view the query behind the External Data range, right-click on it and select Edit Query item from the menu (or choose Data/Get External Data/Edit Query). The same procedure works for Web Queries. What's important is that in both cases Excel provides an option to save a query as a text file (*.dqy in case of Data Queries and *.iqy for Internet Queries). It is highly recommended to do so. The benefit is threefold: a) queries get documented, b) it is easier to modify them in this text form and c) it is so easy to execute them - just double-clicking on a *.dqy or *.iqy file.

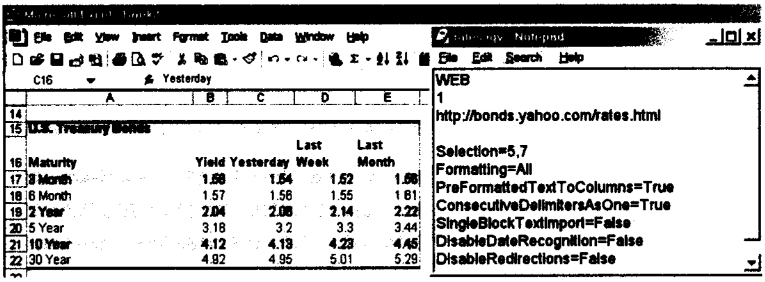


Fig 10

Solver* is a very popular goal-seeking tool and, thankfully, Excel preserves Solver settings, but only one per worksheet. This means that if Solver is used multiple times on the same sheet, it is a good practice to save its settings (Tools/Solver/Options/Save Model) in a descriptively labeled area.

Important settings are stored in Conditionally Formatted as well as Data Validated cells. To see validation settings and format conditions, navigate to these cells using Edit/Go To/Special dialog box.

* To access Solver select Tools/Solver from the Excel menu. If Solver is not listed in the menu, check whether it's installed (run Office install) and/or enabled (checkmark in Tools/Add-ins).

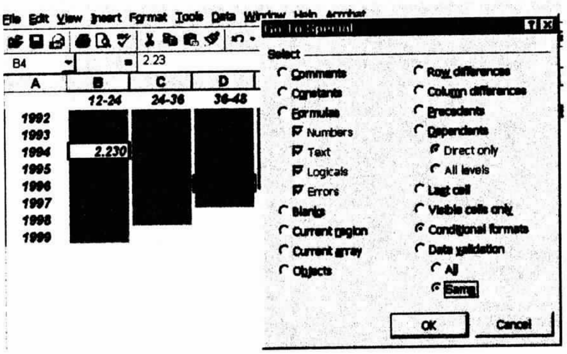
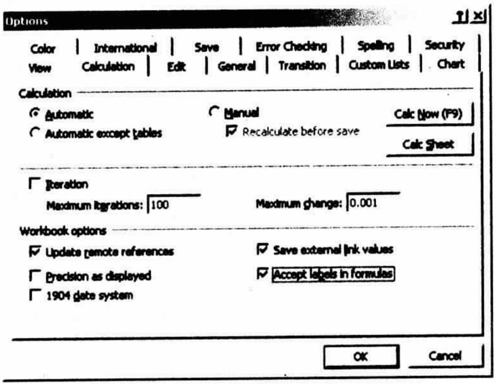


Fig 11

Most users use Names (Insert/Name/Define/Add) only for naming ranges. However, Excel allows giving a Name to any formula, even a User-Defined one. To display the list of Names with their definitions, press F3♦/Paste List. Names in Excel are too important -- and too convenient tool for documenting a spreadsheet -- to be ignored. In the ideal world, there should be no unnamed references in Excel formulas: every variable, input region and output location has to be named. Good naming conventions along with the habit of naming ranges and cells may prove invaluable not only for disaster recovery, but for debugging, modifications and education of the new employees.

Excel creators believe in named references so much they actually supply Names even if the user himself didn't define any. Since version 97, users can use column and row labels as if you created range names for rows and columns (since Office XP, the same syntax works for PivotTables♦♦). To enable this functionality, check the Tools/Options/Calculation/Accept Labels in Formulas option.



♦ Use F3 to paste a Name while typing formula text in a Formula Bar to avoid misprints.
♦♦ To get data from the PivotTable use GetPivotData function.

SUM ▾ ✗ ✓ fx =LastDiagonal*LDFtoULT

	A	B	C	D
1	AccYear	LastDiagonal	LDFtoULT	Ultimate
2	1999	$ 300,000	1.10	$ 330,000
3	2000	$ 250,000	1.40	$ 350,000
4	2001	$ 200,000	1.75	=LastDiagonal*LDFtoULT
5	2002	$ 100,000	3.50	$ 350,000
6				

Fig 12

Structured Comments

Excel comments, if used creatively, represent an amazingly powerful tool. Available through the Reviewing toolbar, comments can be toggled (by moving mouse over commented cell) or displayed permanently (Show Comment). They can be printed "in place" or as footnotes (File/Page Setup/Sheet/Comments *dropdown*). Moreover, comments (as in any programming environment) are invaluable for documenting designer's intentions and understanding algorithm's logic.

In addition to their special role in documentation of a spreadsheet, and consequently in any restoration effort, comments may play an even bigger role if used as Object's descriptors. Indeed, given that comments are

- *an ASCII text*
- *associated with a range and*
- *can be manipulated (created, used, modified and deleted) through VBA*

comments can be used for storing *structured attributes* of an object a la XML (see [4]):

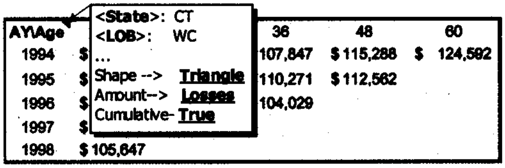


Fig 13 (from [4])

One can think of a comment as a "price" tag attached to an Actuarial Object. A user-defined VBA function that accepts such ranges as input may read that tag and decide what to do with associated ranges (if it is a triangle of losses the function may divide it by the corresponding claim counts; if it is a vector of loss development factors a function may multiply it by the last diagonal and if it is column of loss ratios the function will prohibit any attempt to add inflation factors to it).

Auto backup copy

To increase the chances of recovery of the most important Excel files, it is wise to enable a built-in facility for the automatic creation of backup copies. By launching "Save Options" dialog (File/Save As/Tools/General Options) and choosing "Always create backup" option, user can be assured that every time he saves the file an extra copy with the extension XLK is generated.

Still, for occurrences when files are corrupted or incompletely restored from the tapes, Excel 2002 has beefed up its file repair utility. Available through the (File/Open/Open *dropdown*/Open and Repair) menu item, the utility does a formidable job in recovering corrupted files.

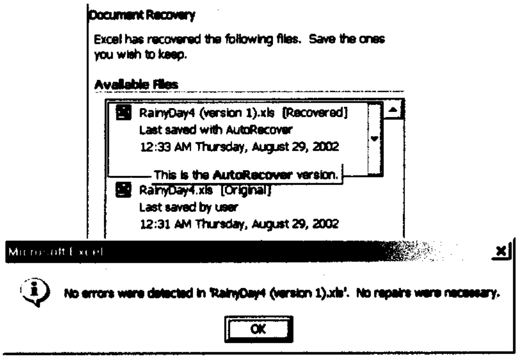


Fig 14

In a rare case, when an attempt to repair fails, as a last resort one can try to paste the content of the corrupted file into a new spreadsheet (see [6]). To do that, open two new files, select cell A1 and copy it to clipboard, switch to the second file and after pasting the link (Edit/Paste Special/Paste Link), change the link to the corrupted file. In most cases, Excel allows the user to access this way as much content as it could recover. The rest of the file (VBA modules, External Data queries and Pivot Table cubes) can be imported from the ASCII files.

Documenting Workflow

As crucial as preservation of files and documentation of algorithms is the process of diagramming actuarial workflow. The order of actions grouped by stages with the references to file locations and processes is an invaluable restoration asset.

There are many ways to document workflow. The most natural and powerful, though, is to use "smart diagramming" software like Microsoft Visio or Micrografx iGrafx by Corel. In addition to their ability to document, analyze and simulate workflow, these packages (empowered by VBA) may execute some actions automatically.

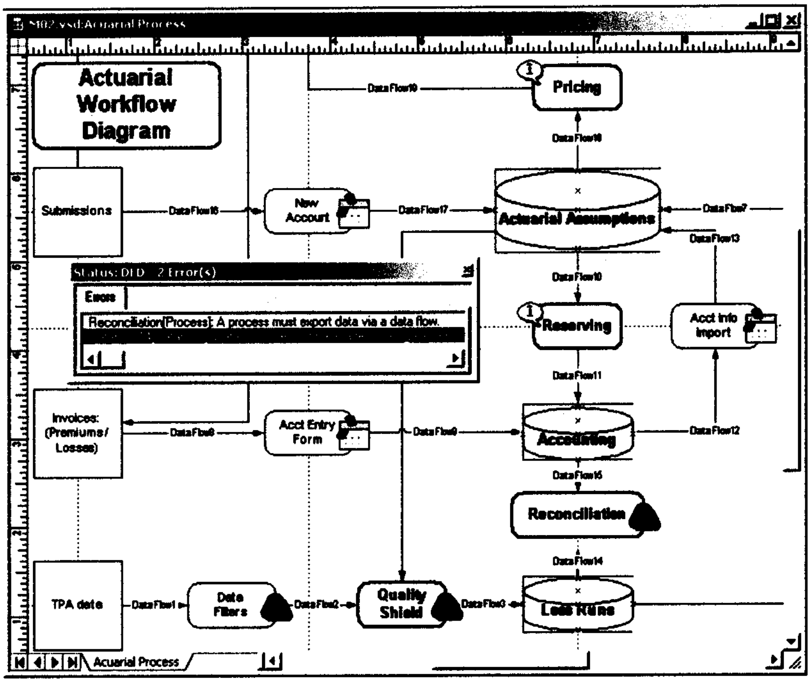


Fig 15

Workflow diagrams - as important as they are for disaster recovery – provide additional benefit as a way to look at the actuarial process as a whole, and possibly to streamline and simplify it.

Telecommute

Telecommuters present yet another challenge for flawless disaster recovery. A whole additional layer of network subsystems (terminal services, VPN access, firewall) has to be restored in order to enable their access to the company's applications. Home and mobile computers and devices represent an additional hazard for security and maintenance.

However, provided that security, connectivity, maintenance and support issues are solved, home computers will become a decentralized independent distributed file storage system: an additional chance to restore a copy of this most important lost file. Also, if configured accordingly, remote computers may serve as a temporary replacement system until restoration of the main system is complete. Indeed, many applications can be scaled to work on a standalone machine: major databases have compatible "personal" versions, while Office and many third-party actuarial applications are "personal" by nature. The synchronization with the "main" system can be possibly achieved via import/export to/from ubiquitous file formats.

Paperless Office

If the backup is set up and working smoothly, files are copied to tapes and stored offsite, then there is *no excuse **not to*** scan every paper document greatly reducing risk of losing it. Indeed, with advances in scanning quality and OCR (optical character recognition) accuracy, it makes perfect sense to convert all paper documents into computer readable files. The ubiquitous PDF (portable document format) file preserves the look of the original, while at the same time enabling index, catalog and search services to scan through its content as if it were simple text file. Even Internet Search Engines are now PDF-enabled, so Internet search queries are capable of looking for information inside PDF files. Thus, scanned paper documents can be organized into a useful searchable hierarchical "knowledge base" instead of lying in some storage boxes, being hard to find and, probably, unused.

Once again, an action geared toward better disaster protection may turn out to have a great side benefit, perhaps even greater than the initial purpose of the action.

Conclusion

Any type of action – from a big radical change of architecture in order to "separate data from algorithms" to a small conventional "enabling of auto-backups" in Excel - is better than no action.

Besides, all aforementioned recommendations help not only in the case of devastating disaster, but also in the event of a virus attack, malicious user actions, and staff rotation. In fact, benefits from such *preparation* measures as

- *clear documentation of actuarial procedures,*
- *streamlined algorithms and*
- *more effective workflow*

may far outweigh the potential payback from the original objective of disaster preparedness. These measures are more than worthy by themselves. Surely, the **cost of precautions should not exceed estimated damages**. However, side benefits such as audit trail capabilities, design discipline and improved understanding of calculations can easily justify disaster preparedness efforts.

Dedication

To Giya Aivazov and all friends and colleagues affected by The September 11.

Stamford, 2001

Bibliography

[1] Annlee Hines. *Planning for Survivable Networks: Ensuing Business Continuity.* 2002, Wiley Publishing

[2] http://csrc.nist.gov/publications/nistpubs/800-34/sp800-34.pdf

[3] Aleksey S. Popelyukhin. *The Big Picture: Actuarial Process from the Data Processing Point of View.* 1996, Library of Congress

[4] Aleksey S. Popelyukhin. *On Hierarchy of Actuarial Objects: Data Processing from the Actuarial Point of View.* Spring 1999, CAS Forum

[5] Michael Koflet. *Definite Guide to Excel VBA.* 2000, apress

[6] Mark Dodge. *Microsoft Excel Version 2002 Inside Out.* 2001, Penguin Books

Disaster Recovery websites:

[7] http://www.fema.gov

[8] http://www.disasterplan.com

[9] http://www.disasterrecoveryworld.com

Modeling Hidden Exposures in Claim Severity via the EM Algorithm

Grzegorz A. Rempala and Richard A. Derrig

Modeling Hidden Exposures in Claim Severity via the EM Algorithm

Grzegorz A. Rempala
Department of Mathematics, University of Louisville,
Richard A. Derrig
Automobile Insurers Bureau of Massachusetts

Abstract

We consider the issue of modeling the so-called hidden severity exposure occurring through either incomplete data or an unobserved underlying risk factor. We use the celebrated EM algorithm as a convenient tool in detecting latent (unobserved) risks in finite mixture models of claim severity and in problems where data imputation is needed. We provide examples of applicability of the methodology based on real-life auto injury claim data and compare, when possible, the accuracy of our methods with that of standard techniques.

1 Introduction

Actuarial analysis can be viewed as the process of studying profitability and solvency of an insurance firm under a realistic and integrated model of key input random variables such as loss frequency and severity, expenses, reinsurance, interest and inflation rates, and asset defaults. In a modern analysis of financial models of property-casualty companies, these input variables typically can be classified into financial market variables and underwriting variables (cf. e.g., D'Arcy et al. 1997). The financial variables generally refer to asset-side generated cash flows of the business, and the underwriting variables relate to the cash flows of the liabilities side. The process of developing any actuarial model begins with the creation of probability distributions of these input variables, including the establishment of the proper range of values of input parameters. The use of parameters is generally determined by the use of the parametric families of distributions, although the non-parametric techniques have a role to play as well (see, e.g., Derrig, et al. 2001). In this article we consider an issue of hidden or "lurking" risk factors or parameters and point out the possible use of the celebrated EM algorithm to uncover those factors. We begin by addressing the most basic questions concerning hidden loss distributions. To keep things in focus we will be concerned here only with two applications to modeling the severity of loss, but the methods discussed may be easily applied to other problems like loss frequencies, asset returns, asset defaults, and combining those into models of Risk Based Capital, Value at Risk, and general Dynamic

Financial Analysis, including Cash Flow Testing and Asset Adequacy Analysis. Our applications will illustrate the use of the EM algorithm (i) to impute missing values in an asset portfolio and (ii) to screen medical bills for possible fraud or abusive practices.

1.1 Hidden Exposures in Loss Severity Distributions

In many instances one would be interested in modeling hidden risk exposures as additional dimension(s) of the loss severity distribution. This in turn in many cases leads to considering mixtures of probability distributions as the model of choice for losses affected by hidden exposures; some parameters of the mixtures will be considered missing (i.e., unobservable in practice). During the last 20 years or so there has been a considerable advancement in statistical methodologies dealing with partially hidden or incomplete data models. Empirical data imputation has become more sophisticated and the availability of ever faster computing power have made it increasingly possible to solve these problems via iterative algorithms.

In our paper we shall illustrate a possible approach to two types of problems arising often in practical situations of modeling the severity of losses: (i) imputation of partially *missing* multivariate observations and (ii) identification of *latent* risks via fitting finite mixtures models.

The common feature of both of these issues is, generally speaking, the unavailability of complete information on the variables or parameters of interest. The statistical methodology which is especially well-suited for this type of circumstances is the so-called EM algorithm.

1.2 The EM Algorithm

In their seminal paper Dempster, Laird and Rubin (1977) have proposed the methodology which they have called the Expectation-Maximization (EM) algorithm as an iterative way of finding maximum likelihood estimates.[1] They demonstrated that the method was especially appropriate for finding the parameters of an underlying distribution from a given data set where the data was *incomplete* or had missing values. At present there are two basic applications of the EM methodology considered in the statistical literature. The first occurs when the data indeed has missing values, due to problems with or limitations of the data collection process. The second occurs when the original likelihood estimation problem is altered by assuming the existence of the *hidden* parameters or factors. It turns out that both these circumstances can be, at least initially, described in the following statistical setting. Let us consider a density function (possibly multivariate) $p(\cdot|\Theta)$ that is indexed by the set of parameters Θ. As a simple example we may take p to be a univariate Gaussian density and $\Theta = \{(\mu, \sigma)| -\infty < \mu < \infty, \sigma > 0\}$. Additionally, we have an observed data set $\mathcal{X}$ of size n, drawn from the distribution p. More precisely, we assume that the points of $\mathcal{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ are the realizations of some independent random variables distributed according to $p(\cdot|\Theta)$. We shall call $\mathcal{X}$ the *incomplete* data. In addition to $\mathcal{X}$, we also

[1] A full explanation of the role of the EM algorithm in missing data problems can be found in Little and Rubin, (1987) or in a somewhat more mathematically advanced monograph by McLachlan and Krishnan (1997).

consider a *complete* data set $\mathcal{Z} = (\mathcal{X}, \mathcal{Y})$ and specify the joint density

$$p(\mathbf{z}|\Theta) = p(\mathbf{x}, \mathbf{y}|\Theta) = p(\mathbf{y}|\mathbf{x}, \Theta)\, p(\mathbf{x}|\Theta). \tag{1}$$

As we can see from the last equality, this joint density $p(\mathbf{z}|\Theta)$ arises from considering the marginal density $p(\mathbf{x}|\Theta)$ and the specific assumptions on the relation between hidden (or missing) variables $\mathcal{Y} = (\mathbf{y}_1, \dots \mathbf{y}_n)$ and the observed incomplete data $\mathcal{X}$. Associated with the joint density is the joint likelihood function

$$\mathcal{L}(\Theta|\mathcal{Z}) = \mathcal{L}(\Theta|\mathcal{X}, \mathcal{Y}) = \prod_{i=1}^{n} p(\mathbf{x}_i, \mathbf{y}_i|\Theta)$$

which is often called the *complete* likelihood. For the sake of computational simplicity it is often more convenient to consider the logarithm of the complete likelihood

$$l(\Theta|\mathcal{Z}) = \log \mathcal{L}(\Theta|\mathcal{X}, \mathcal{Y}) = \sum_{i=1}^{n} \log p(\mathbf{x}_i, \mathbf{y}_i|\Theta). \tag{2}$$

Note that the function above may be thought of as a random variable since it depends on the unknown or missing information $\mathcal{Y}$ which by assumption is governed by an underlying probability distribution. Note also that in accordance with the likelihood principle, we now regard $\mathcal{X}$ as constant.

The EM algorithm as described in Dempster, Laird and Rubin (1977) consists of two steps repeated iteratively. In its *expectation* step or the E-step, the algorithm first finds the expected value of the complete log-likelihood function $\log p(\mathcal{X}, \mathcal{Y}|\Theta)$ with respect to the unknown data $\mathcal{Y}$ given the observed data $\mathcal{X}$ and the current parameter estimates. That is, instead of the complete log-likelihood (2) we consider the following

$$Q(\Theta, \Theta^{(i-1)}) = E\left[\log p(\mathcal{X}, \mathcal{Y}|\Theta)|\mathcal{X}, \Theta^{(i-1)}\right]. \tag{3}$$

Note the presence of the second argument in the function $Q(\Theta, \Theta^{(i-1)})$. Here $\Theta^{(i-1)}$ stands for the current value of the parameter Θ at the iteration $(i-1)$, that is, the value which is used to evaluate the conditional expectation.

After the completion of the E-step, the second step of the algorithm is to maximize the expectation computed in the first step. This is called the *maximization* or the M-step, at which time the value of Θ is updated by taking

$$\Theta^{(i)} = \underset{\Theta}{\operatorname{argmax}}\, Q(\Theta, \Theta^{(i-1)}) \tag{4}$$

The steps are repeated until convergence. It can be shown (via the relation (1) and Jensen's inequality) that if Θ^* maximizes $Q(\Theta, \Theta^{(i-1)})$ with respect to Θ for fixed $\Theta^{(i-1)}$ then

$$l(\Theta^*|Z) - l(\Theta^{(i-1)}|Z) \geq Q(\Theta^*, \Theta^{(i-1)}) - Q(\Theta^{(i-1)}, \Theta^{(i-1)}) \geq 0$$

and each iteration of the procedure indeed increases the value of complete log-likelihood (2). Let us note that from the above argument it follows that a full maximization in the M-step is not necessary: it suffices to find any value of $\Theta^{(i)}$ such that $Q(\Theta^{(i)}, \Theta^{(i-1)}) > Q(\Theta^{(i-1)}, \Theta^{(i-1)})$. Such procedures are called GEM (*generalized EM*) algorithms. For a complete set of references see, for instance, the monograph by McLachlan and Krishnan (1997) where also the issues of convergence rates for the EM and GEM algorithms are thoroughly discussed. For some additional references and examples see also Wu (1983) or the monographs by Little and Rubin (1987) and Hastie,Tibshirani, and Friedman (2001).

2 Modeling Hidden Risks via the EM Algorithm

As indicated in the previous section the primary application of the EM algorithm is in fitting the maximum likelihood models. Since this is accomplished by the M-step of the algorithm, the role of the E-step is, therefore, secondary – it is needed to facilitate the performance of the M-step in the presence of the missing or incomplete data. However, as in this paper we shall focus on the usefulness of the EM procedure in modeling hidden risks or variables, in our setup we shall be in fact more interested in the E-step of the algorithm, as it will provide us with the way to estimate or impute missing data and uncover hidden factors and variables. In our examples below we shall consider two types of hidden (latent) variables. The first one will arise when, due to some problems with the data collection, parts of the observations are missing from the observed dataset. We consider this problem via the EM method in the particular context of multivariate (loss) models.

2.1 Multivariate Severity Distributions. Data Imputation with EM

Although insurance has been traditionally build on the assumption of independence and the law of large numbers has governed the determination of premiums, the increasing complexity of insurance and reinsurance products has lead over past decade to increased actuarial interest in the modeling of dependent risks (see, e.g., Wang 1998 or Embrechts et al. 2000). Multivariate loss and risk models (and especially those based on elliptically contoured distributions) have been hence of interest in such areas as Capital Asset Pricing Model and the Arbitrage Pricing Theory (cf. e.g., Campbell, Lo, and MacKinlay 1996).

In some circumstances, however, parts of the observed multivariate data may be missing. Claim reporting systems depend heavily on the front-line adjusters to provide data elements beyond the simple payment amounts. In the absence of, or even in the presence of, system edits, daily work load pressures and the lack of interest in the coded data provide a deadly combination of disincentives for accurate and complete coding. Actuaries are quite familiar with missing data fields, which when

Table 1: 10 fictitious observed gains and losses from two risk portfolios in thousands.

0.914	2.088	2.644	0.477	-1.940	-0.245	0.362	1.147	?	?
3.855	4.025	2.092	3.400	1.520	2.626	?	?	5.473	6.235

essential to the analysis most often results in throwing the record out, thereby creating unknown 'hidden' biases. Likewise, financial time series data may be interrupted, unavailable, or simply lost for securities or portfolios that are not widely tracked.

As an illustration of an application of the EM algorithm in this setting let us consider a hypothetical example of 10 losses/gains from a two-dimensional vector of risk portfolios, which we have generated using a bivariate normal distribution. The data is presented in Table 1 (in thousands of dollars). As we can see parts of the last four observations are missing from the table. In fact, for the purpose of our example, they have been removed from the generated data. We shall illustrate the usefulness of the EM algorithm in estimating these missing values.

If we denote by $\mathcal{X}$ the observed (incomplete) data listed in Table 1 then following our notation from previous section we have the complete data vector $\mathcal{Z}$ given by

$$\mathcal{Z} = (\mathbf{z}_1 \dots \mathbf{z}_n) = \left(\mathbf{x}_1 \dots, \mathbf{x}_6, (x_{1,7}, y_{2,7})^T, (x_{1,8}, y_{2,8})^T, (y_{1,9}, x_{2,9})^T, (y_{1,10}, x_{2,10})^T\right)$$

where $\mathbf{x}_j = (x_{1,j}, x_{2,j})^T$ for $j = 1 \dots, 6$ is the set of pairwise complete observations. The missing data (corresponding to ? marks in Table 1) is, therefore,

$$\mathcal{Y} = (y_{2,7}, y_{2,8}, y_{1,9}, y_{1,10}). \tag{5}$$

Let us note that under our assumption of normality, the equation (2) now becomes

$$l(\Theta|\mathcal{Z}) = -n \log(2\pi) - \frac{1}{2} \log|\Sigma| - \frac{1}{2} \sum_{j=1}^{n} (\mathbf{z}_j - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{z}_j - \boldsymbol{\mu})$$

where $n = 10$, $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$ is a vector of means and

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}$$

is covariance matrix. The vector of unknown parameters, therefore, can be represented as

$$\Theta = (\mu_1, \mu_2, \sigma_{11}, \sigma_{12}, \sigma_{22}). \tag{6}$$

In order to describe the EM algorithm in this setting we need to find the particular form of $Q(\Theta, \Theta^{(i-1)})$ defined by (3). Due to the independence of the $\mathbf{z}_i$'s this is equivalent, in effect, to evaluating

$$E_{\Theta^{(i-1)}}(Y|\mathcal{X}) \qquad \text{and} \qquad E_{\Theta^{(i-1)}}(Y^2|\mathcal{X})$$

where Y is the underlying random variable for $\mathcal{Y}$, assumed to be normal. From the general formulae for conditional moments of a bivariate normal variable $X = (X_1, X_2)$ with the set of parameters Θ as above, we have that

$$\begin{aligned} E(X_2|X_1 = x_1) &= \mu_2 + \sigma_{12}/\sigma_{11}(x_1 - \mu_1) \\ Var(X_2|X_1 = x_1) &= \sigma_{22.1} = \sigma_{22}(1 - \rho^2) \end{aligned} \tag{7}$$

where ρ stands for the correlation coefficient. Interchanging the subscripts 1 and 2 in (7) gives the formulae for the conditional mean and variance of the distribution $X_1|X_2 = x_2$. Using the relations (7) and the usual formulae for ML estimators of the normal mean vector μ and the covariance matrix Σ, we may now state the EM algorithm for imputing missing data in Table 1 as follows.

Algorithm 1 *(EM version of Buck's algorithm)*

1. *Define the initial value $\Theta^{(0)}$ of the set of parameters (5). Typically, it can be obtained on the basis of the set of complete pairs of observations (i.e., $\mathbf{x}_1 \dots, \mathbf{x}_6$ in Table 1).*
2. *The E-step: given the value of $\Theta^{(i)}$ calculate via (7) the vector $\mathcal{Y}^{(i)}$ of the imputations of the missing data $\mathcal{Y}$ given by (6).*

$$y_{2k}^{(i)} = \mu_2^{(i)} + \frac{\sigma_{12}^{(i)}}{\sigma_{11}^{(i)}}\left(x_{1k} - \mu_1^{(i)}\right) \quad \textit{and} \quad y_{2k}^{2\,(i)} = \left(y_{2k}^{(i)}\right)^2 + \sigma_{22.1}^{(i)} \quad \textit{for } k = 7, 8$$

$$y_{1k}^{(i)} = \mu_1^{(i)} + \frac{\sigma_{12}^{(i)}}{\sigma_{22}^{(i)}}\left(x_{2k} - \mu_2^{(i)}\right) \quad \textit{and} \quad y_{1k}^{2\,(i)} = \left(y_{1k}^{(i)}\right)^2 + \sigma_{11.2}^{(i)} \quad \textit{for } k = 9, 10$$

3. *The M-step: given the current value of the imputed complete data vector $\mathcal{Z}^{(i)} = (\mathcal{X}, \mathcal{Y}^{(i)})$ set $M_k = \sum_{j=1}^n z_{k,j}^{(i)}/n$ and $M_{kl} = \sum_{j=1}^n z_{k,j}^{(i)} z_{l,j}^{(i)}/n$ for $k, l = 1, 2$, and calculate $\Theta^{(i+1)}$ as*

$$\begin{aligned} \left(\mu_1^{(i+1)}, \mu_2^{(i+1)}\right) &= (M_1, M_2) \\ \sigma_{kl}^{(i+1)} &= M_{kl} - M_k M_l \qquad \textit{for } k, l = 1, 2 \end{aligned}$$

4. *Repeat steps 2 and 3 until the relative difference of the subsequent values of $l(\Theta^{(i+1)}|\mathcal{Z}^{(i)})$ is sufficiently small.*

The above algorithm in its non-iterative version was first introduced by Buck (1960) who used the method of imputation via linear regression with subsequent covariance correction to estimate means and covariance matrices of p dimensional random vectors in case when some parts of the vector components were missing. For more details about Buck's imputation procedure, we refer to his original paper (Buck 1960) or to Chapter 3 of Little and Rubin (1987) or Chapter 2 of McLachlan and Krishnan (1997).

The numerical illustration of the algorithm is presented in Table 2. As we can see from the

Table 2: Selected iterations of the EM algorithm for data in Table 1.

Iteration	μ_1	μ_2	σ_{11}	σ_{12}	σ_{22}	$y_{2,7}$	$y_{2,8}$	$y_{1,9}$	$y_{1,10}$	$-2Q$
1	0.6764	3.5068	1.8170	0.3868	2.0671	3.4399	3.6069	1.0443	1.1867	65.7704
5	0.8779	3.6433	1.8618	0.8671	2.2030	3.4030	3.7685	1.5982	1.8978	64.7568
10	0.9279	3.6327	1.9463	0.9837	2.1724	3.3466	3.7433	1.7614	2.1061	64.5587
20	0.9426	3.6293	1.9757	1.0181	2.1639	3.3301	3.7345	1.8102	2.1683	64.5079
30	0.9435	3.6291	1.9775	1.0202	2.1634	3.3291	3.7339	1.8132	2.1722	64.5048
35	0.9436	3.6291	1.9776	1.0203	2.1633	3.3290	3.7339	1.8134	2.1724	64.5047
40	0.9436	3.6291	1.9777	1.0204	2.1633	3.3290	3.7339	1.8134	2.1724	64.5046
45	0.9436	3.6291	1.9777	1.0204	2.1633	3.3290	3.7339	1.8134	2.1725	64.5046

table with the accuracy of up to three significant digits, the algorithm seems to converge after about 30 steps or so and the estimated or imputed values of (5) are given by

$$\mathcal{Y}^{(em)} = (3.329, 3.734, 1.813, 2.173).$$

Let us note, for the sake of comparison, that if we were to employ the standard, "naive" linear or polynomial regression model based on 6 complete observations in order to fit the missing values in Table 1 we would have obtained in this case

$$\mathcal{Y}^{(reg)} = (2.834, 3.063, 2.700, 3.269).$$

Both $\mathcal{Y}^{(em)}$ and $\mathcal{Y}^{(reg)}$ can be now compared with the actual values removed from Table 1 which were

$$\mathcal{Y} = (3.362, 3.657, 1.484, 3.410).$$

As we can see, in our example the EM method did reasonably well in recovering the missing values.

2.2 Massachusetts Auto Bodily Injury Liability Data. Fraud and Build-up Screening via Mixture Models

By now it is fairly well known that fraud and build-up, exaggerated injuries and/or excessive treatment, are key components of the auto injury loss distributions (Derrig et al. 1994, Cummins and Tennyson 1996, Abrahamse and Carroll 1999). Indeed, injury loss distributions are prime candidates for mixture modeling, for at least the differing of payment patterns by injury type. Even within an injury type as predominant as strain and sprain,[2] there can be substantial differences in subpopulations arising from fraud and build-up. One common method of identifying these claims has been to gather additional features of the claim, the so-called fraud indicators, and to build models to identify those bogus claims (Brockett, et al. 1998). The acquisition of reliable indicators some of which may be highly subjective, is costly, and may not be efficient in uncovering abusive patterns in injury claims (Crocker and Tennyson 1999). The use of more flexible methods such as the fuzzy logic (see more below) may overcome the lack of this precision in subjective features in an economically efficient manner by running a background algorithm on adjusters' electronic files (see, for example, Derrig and Ostaszewski 1995, 1999).

Another approach to uncovering fraud and build up, perhaps grounded more in practical considerations, is to construct a filter, or screening algorithm, for medical provider bills (Derrig 2002). Routinely, excessive medical bills can be reduced to "reasonable and customary" levels by computer algorithms that compare incoming bills to right censored billing distributions with "excessive" being operationally defined to be above the censoring point. Less routine is the implementation of systematic analysis of the patterns of *a provider's billing practices* (Major and Riedinger 1992). Our second application of the EM algorithm is to build a first level screening device to uncover potential abusive billing practices and the appropriate set of claims to review. We perform the pattern analysis by uncovering abusive-like distributions within mixture models parametrized by the estimates obtained via the EM algorithm. An illustration of the method follows.

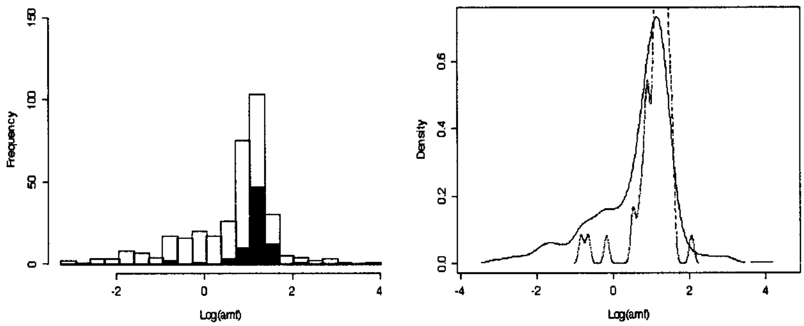
In the table provided in Appendix B we present a set of outpatient medical provider's total billings on the set of 348 auto bodily injury liability claims closed in Massachusetts during 2001. For illustration purposes, 76 claims with one "outlier" provider ("A") were chosen based on a pattern typical of abusive practice; namely, an empirical kurtosis more than five times the overall average. The "outlier" was then combined with medical bills in claims from a random sample of providers. The losses are recorded in thousands and are presented in column two. Column 4 identifies each medical billing amount as provider "A" or "other". We will use the EM algorithm applied to a normal (log) mixture model attempting to uncover provider A.

The relatively large volume of provider A's claims is clearly visible in the left panel of Figure 1, where it is presented as a portion of the overall claims

Whereas the volume of claims by itself never constitutes a basis for the suspicion of fraud or build-up, it certainly might warrant a closer look at the data at hand, especially via some type of

[2]Currently, Massachusetts insured bodily injury claims are upwards of 80 percent strain and sprain claims as the most costly part of the medical treatment. Of course, that may have a dependency on the $2,000 dollar threshold to file a tort claim.

Figure 1: Overall distribution of the 348 BI medical bill amounts from Appendix B compared with that submitted by provider A. Left panel: frequency histograms (provider A's histogram in filled bars). Right panel:density estimators (provider A's density in dashed line).



homogeneity analysis, since the second panel in Figure 1 clearly indicates the difference between the overall claims distribution and that of the provider A. Hence in this problem we shall be looking for a hidden exposure which could manifest itself as a non-homogenous component of the data, albeit we shall not be assuming that this component is necessarily due to provider A. In fact, as the initial inspection of the *overall* data distribution does not immediately indicate non-homogeneity we shall not make any prior assumptions about the nature or source of the possible non-homogeneity.

Since the standard analysis of the data by fitting a kernel density estimator (see the solid curve in the right panel of Figure 1) appears to give no definite indication of multimodality, it seems, that some more sophisticated methods are needed in order to identify any foreign components of the claims. Whereas many different approaches to this difficult problem are possible, we have chosen one that shall illustrate the applicability of the EM methodology in our setting. Namely, we shall attempt to fit a *log-mixture-normal* distribution to the data, that is, we shall model the logarithm of the claim outpatient medical billing distribution as a mixture of several normal variables. The use of normal distributions here is mostly due to convenience of the EM implementation and in more complicated real life problems can be inappropriate. However, the principle that we shall attempt to describe here is, in general, applicable to any mixture of distributions, even including non-parametric ones.[3]

[3]The notion of fitting non-parametric distributions via likelihood methods, which at first may seem contradiction in terms, has become very popular in statistics over the last decade. This is due to intensive research into the so called empirical likelihood methods (see for instance a recent monograph by Owen 2001 and references therein). In

In order to describe our method in the context of the EM algorithm we shall again relate the problem at hand to our EM methodology introduced in Section 1. In our current setting we shall consider the set of logarithms of the BI claim medical bills as the incomplete data $\mathcal{X}$. According to our model assumption we identify the underlying random variable X, of which $\mathcal{X}$ is a realization, as a mixture of several (say, $m \geq 2$) normal distributions[4]

$$X_i \sim N(\mu_j, \sigma_j) \quad \text{for} \quad j = 1, \ldots, m$$
$$X = \sum_{j=1}^{m} Y_j \cdot X_j, \tag{8}$$

where $Y_j \in \{0,1\}$ with $P(Y_j = 1) = \pi_j$ such that $\sum \pi_j = 1$ and the joint distribution of the vector $(Y_1, \ldots, Y_m)$ is multinomial with one trial, (i.e., $\sum Y_j = 1$). The right hand side of (8) is sometimes known as *generative* representation of a mixture. Indeed, if we generate a multinomial variable $(Y_1, \ldots, Y_m)$ with probabilities of $Y_j = 1$ equal to π_j, and depending on the index j for which outcome is a unity, deliver X_j, then it can be shown that the density of X is

$$\sum_{j=1}^{m} \pi_j \; p(x|\Theta_j) \tag{9}$$

where $p(\cdot|\Theta_j)$ is a normal density with the parameter

$$\Theta_j = (\mu_j, \sigma_j) \quad \text{for} \quad j = 1, 2 \ldots, m$$

Hence X is indeed a mixture of the X_j's. The density given by (9) is less helpful in our approach as it doesn't explicitly involve the variables Y_j's. Moreover, fitting the set of parameters[5]

$$\Theta = (\Theta_1, \ldots, \Theta_m, \pi_1, \ldots, \pi_{m-1}). \tag{10}$$

by considering log-likelihood of (9) is known to be numerically difficult as it involves evaluation of the sums under the logarithm. In contrast, the representation (8) provides for a simpler approach, which also suits better our purpose of illustrating the use of the EM methodology. In the spirit of the search for hidden exposure, we consider the (unobserved) realizations of random vector $(Y_1, \ldots, Y_m)$ in (8) as the missing data $\mathcal{Y}$. Let us note that unlike in the example discussed in Section 2 here we have in some sense artificially created the set $\mathcal{Y}$. In this setting the complete set of data is now $\mathcal{Z} = (\mathcal{X}, \mathcal{Y})$ or $\mathbf{z}_j = (x_j, y_{jk})$ for $j = 1, \ldots, n$, and $k = 1 \ldots, m$. Here $n = 348$ is the number of obervations, m is the number of components in the mixture, unspecified for now, x_j is (logarithm of) the observed medical bill value, and $y_{jk} \in \{0, 1\}$ is the auxiliary indicator variable

principle, with some modifications, the mixture approach discussed in this section and the associated EM algorithm can be applied to the empirical likelihood as well.

[4]Note that in our notation σ denotes the variance, not standard deviation.

[5]Note that we only need to estimate $m - 1$ proportions since $\sum \pi_i = 1$.

indicating whether or not x_j arrives from the distribution of X_k. In this setting the complete log-likelihood function (2) takes the form

$$l(\Theta|\mathcal{Z}) = \sum_{j=1}^{n}\sum_{k=1}^{m} y_{jk} \log p(x_j|\Theta_k), \tag{11}$$

and the conditional expectation (3) is given by

$$Q(\Theta, \Theta^{(i-1)}) = \sum_{j=1}^{n}\sum_{k=1}^{m} \delta_{jk} \ \log p(x_j|\Theta_k), \tag{12}$$

where

$$\delta_{jk} = E(Y_{jk}|\Theta^{(i-1)}, \mathcal{Z}) = P(Y_{jk} = 1|\Theta^{(i-1)}, \mathcal{X}) \quad \text{for} \quad j = 1 \dots, n; \quad k = 1, \dots, m. \tag{13}$$

As we can see from the above formulae, in this particular case $Q(\Theta, \Theta^{(i-1)})$ is obtained from the complete data likelihood by substituting for the unknown y_{jk}'s their conditional expectations δ_{jk}'s calculated under the current value of the estimates of Θ.[6] The quantity δ_{jk} is often referred to as the *responsibility* of the component X_k for the observation j. This terminology reflects the fact that we may think about final δ_{jk} as the conditional (posterior) probability of the j-th observation arriving from the distribution of X_k.

Once we have replaced the y_{jk}'s in (11) by the δ_{jk}'s, the maximization step of the EM algorithm is straightforward and applied to (12) gives the usual weighted ML estimates of the normal means, variances, and the mixing proportions (see below for the formulae). However, in order to proceed with the EM procedure we still need to construct the initial guesses for the set of parameters (10). A good way to do so (for a discussion, see, for instance, Chapter 8 of Hastie et al. 2001 or Xu and Jordan 1996) is to simply choose at random m of the observed claim values as the initial estimates of the means, and set all the estimates of the variances to the overall sample variance. The mixing proportion can be set uniformly over all components. This way of initiating the parameters ensures the relative robustness of the final estimates obtained via EM against any particular initial conditions. In fact, in our BI data example we have randomly selected several initial sets of values for the means and in all case have obtained convergence to the same set of estimates. Below we present the detailed EM algorithm we have used to analyze the Massachusetts auto BI data. In order to identify the number m of the mixture components in the model we have used the EM method to obtain the estimates of the complete log-likelihood function (as the final values of (12)) for $m = 2, 3, 4$ (we have had determined earlier that for $m > 4$ the BI mixture model becomes too cumbersome). The results are presented in Table 3. As can be seen from the last row of the table, $m = 3$ is the number of components minimizing the negative of the estimated log-likelihood (12). Henceforth we shall, therefore, take $m = 3$ for the BI mixture model.

[6]It may happen that some of the values y_{jk} are in fact available. In such cases, we would take $\delta_{jk} = y_{jk}$.

Table 3: Comparison of the mixture fit for the different values of m for the BI data

Parameter	$m=2$	$m=3$	$m=4$
μ_1	0.071	0.107	-0.01
μ_2	1.110	0.874	0.218
μ_3	-	1.248	0.911
μ_4	-	-	1.258
$\sigma_1^{1/2}$	1.265	1.271	1.201
$\sigma_2^{1/2}$	0.252	0.178	1.349
$\sigma_3^{1/2}$	-	0.146	0.214
$\sigma_4^{1/2}$	-	-	0.144
π_1	0.470	0.481	0.250
π_2	0.530	0.205	0.224
π_3	-	0.314	0.247
π_4	-	-	0.279
-2Q	819.909	811.381	811.655

Table 4: Selected iterations of the EM algorithm for the BI data with $m=3$.

Iteration	μ_1	μ_2	μ_3	$\sigma_1^{1/2}$	$\sigma_2^{1/2}$	$\sigma_3^{1/2}$	π_1	π_2	π_3	$-2Q$
1	0.229	0.785	0.885	1.172	0.89	0.843	0.35	0.329	0.321	973.115
5	-0.129	0.946	1.054	1.374	0.525	0.356	0.337	0.301	0.361	854.456
6	-0.131	0.953	1.083	1.357	0.499	0.300	0.349	0.281	0.370	839.384
10	-0.041	0.917	1.137	1.324	0.456	0.223	0.396	0.217	0.387	820.903
20	0.042	0.875	1.166	1.302	0.364	0.207	0.438	0.177	0.385	817.363
30	0.064	0.876	1.184	1.29	0.301	0.200	0.453	0.176	0.372	816.143
40	0.074	0.871	1.204	1.285	0.259	0.188	0.460	0.186	0.354	814.957
50	0.084	0.868	1.226	1.281	0.222	0.17	0.467	0.197	0.336	813.367
60	0.099	0.871	1.243	1.275	0.190	0.153	0.476	0.204	0.320	811.838
64	0.105	0.873	1.247	1.272	0.180	0.147	0.48	0.205	0.315	811.454
65	0.107	0.874	1.248	1.271	0.178	0.146	0.481	0.205	0.314	811.381

Algorithm 2 *(The EM algorithm for fitting m-component normal mixture)*

1. *Define the initial estimate $\Theta^{(0)}$ of the set of parameters* (10) *(see discussion above).*
2. *The E-step: given the current value of $\Theta^{(i)}$ compute the responsibilities δ_j as*

$$\delta_{jk} = \frac{\pi_k^{(i)}\, p(x_j|\Theta_k^{(i)})}{\sum_{l=1}^{m} \pi_l^{(i)}\, p(x_j|\Theta_l^{(i)})} \qquad j = 1, \ldots, n \quad \textit{and} \quad k = 1, \ldots, m.$$

3. *The M-step: compute the ML estimators of* (12) *as*

$$\pi_k^{(i+1)} = \frac{\sum_{j=1}^{n} \delta_{jk}}{n} \quad \textit{for} \quad k = 1, \ldots, m-1,$$

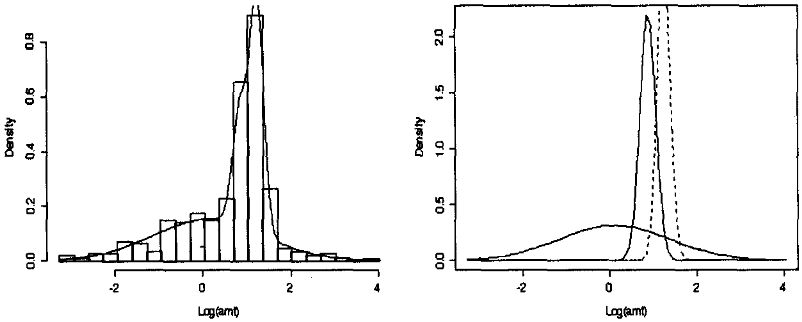
and

$$\mu_k^{(i+1)} = \frac{\sum_{j=1}^{n} \delta_{jk}\, x_j}{\sum_{j=1}^{n} \delta_{jk}}$$

$$\sigma_k^{(i+1)} = \frac{\sum_{j=1}^{n} \delta_{jk} \left(x_j - \mu_k^{(i+1)}\right)^2}{\sum_{j=1}^{n} \delta_{jk}} \quad \textit{for} \quad k = 1, \ldots, m.$$

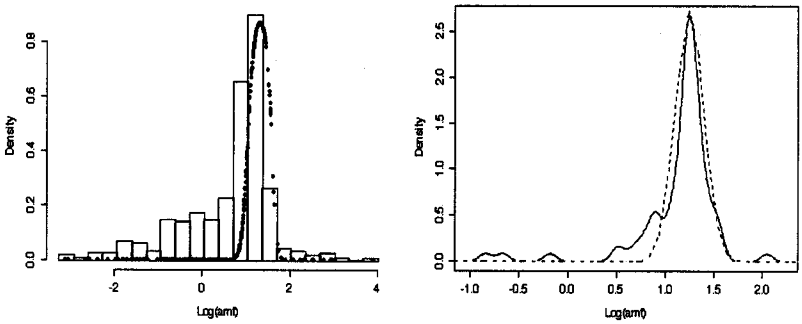
4. *Repeat steps 2 and 3 until the relative difference of the subsequent values of* (12) *is sufficiently small.*

Figure 2: EM Fit. Left panel: mixture of normal distributions fitted via the EM algorithm to BI data. Right panel: Three normal components of the mixture. The values of all the parameters are given in the last row of Table 4.



In Figures 2 and 3 we present graphically the results of the analysis of the BI data via the mixture model with $m = 3$ using the EM algorithm as described above. Some selected iterations of the EM algorithm for the three component normal mixture are presented in Table 4. In the left panel of Figure 2 we show the fit of the normal mixture fitted to the data using Algorithm 2 (with parameters values given by the last row of Table 4) . As we can see the fit looks reasonable and the fitted mixture distribution looks similar to the standard density estimator (solid curve in the right panel of Figure 1). The mixture components identified by the EM method are presented in the right panel of Figure 2 and clearly indicate non-homogeneity of the data which seems to consist of two (in fact, three) different types of claims. This is, obviously, related to a high volume of claims in the interval around 1.8–4.5 thousands (corresponding to the values .6–1.5 on the log scale). This feature of the data is modeled by the two tall and thin (i.e., with small dispersion) components of the mixture (corresponding in our notation to X_2 and X_3, marked as solid and dashed curves, respectively). Let us also note the very pronounced difference (over seven-fold) in the spread between the first and the two last components.

Figure 3: Latent risk in BI data modeled by the EM algorithm with $m = 3$. Left panel: set of the responsibilities δ_{j3}. Right panel: the third component of the normal mixture compared with the distribution of provider A's claims ("A" claims density estimator is a solid curve).



In the left panel of Figure 3 we present the set of responsibilities (δ_{j3}) of the model (or component) X_3 as calculated by the EM algorithm superimposed on the histogram of the BI data. The numerical values of the responsibilities for each data point are also listed in the last column of the table in Appendix B. The relationship between the set of responsibilities obtained via the EM procedure and the apparent lack of homogeneity of the data, demonstrated by Figure 2, is easy to see. The high responsibilities are clustered around the claim values within two standard deviations of the estimated mean (1.25) of the tallest distribution X_3. Hence the plot of responsibilities superimposed on the data distribution again uncovers the non-homogeneity or the risk factor which was initially hidden. As we can see from the right panel in Figure 3 the observed non-homogeneity may be attributed largely, as initially expected (and as the illustration intended), to the high kurtosis of "A" claims. Indeed, the superimposing of the distribution of "A" claims (solid curve) on the component X_3 (dashed curve) in the right panel of Figure 3 reveals a reasonably close match in the interval (.8, 1.7) or so. Outside this interval the normal approximation to the provider A's claims fails, mostly due to the fact that the normal mixture model employed is not sufficiently "fine tuned" in its tails to look for this particular type of distribution. The deficiency could be perhaps rectified in this particular case by incorporating some different (non-normal) components into the mixture model. However, our main task in this analysis was to merely uncover hidden factors (if any) and not necessarily to model them precisely, which should be done afterwards using some different, more sophisticated modeling approach depending on the type of problem at hand. See, for instance, Bilmes, (1998) who presents the extension of our Algorithm 2 to the so-called general hidden Markov model (HMM). For a full review of some possible approaches to fitting the finite mixtures models and the use of the EM methodology in this context, readers are referred to the recent monograph by McLachlan and Peel (2000) which also contains some descriptions of the currently available software for fitting a variety of non-normal mixtures via the EM method.

2.3 The EM Algorithm Output and Fuzzy Set Membership Function

As we have seen above, each run of the EM algorithm estimating an m-mixture model will produce responsibilities for each claim and for each one of the m mixture distributions. As mentioned earlier, they can be interpreted as the (posterior) probability that the claim "arises" from each of the components of the mixing distributions. They also can be interpreted as the membership functions for the fuzzy sets of "arising from the i-th mixture component". If for any claim the responsibility (membership) of a particular model component equals one, we say that the claim arises from that model component. When the responsibility is less than one, the claim arises partially from that component, and if the responsibility equals zero, we can say the claim does not arise from that component. In that context, every claim "belongs to" each of the mixing component with measurement value equal to the responsibility. Putting the EM algorithm within the fuzzy set context provides us with the well-known tools of fuzzy arithmetic to help interpret the EM output in a way that matches real-life actuarial choices (c.f., e.g., Derrig and Ostaszewski 1999).

Another advantage of portraying the responsibility probabilities as fuzzy sets relations is that the defuzzification operator known as the α-cut[7], can be used to illustrate the type I and II errors when the α-cut criterion is used to classify the claim as belonging to one of the mixture distributions. The α-cut classification table is presented in Table 5 below and shows the portions of "A" claims contained in each α-cut from 0.1 to 0.9 for each mixture component distribution. In particular, the α-cut analysis confirms our previous findings that "A" claims belong predominantly to the third

[7] For α equal to a number between zero and one, the α-cut of a fuzzy set consists of the (crisp) set of all elements that have a membership value greater than or equal to α (see, Derrig and Ostaszewski 1995).

mixing distribution (i.e., distribution of X_3). Indeed, the α-cut at about 0.5 provides us with a good indication that "A" arises from the third mixing distribution (corresponding to the value 75% in the table) but not from the first one (corresponding to 8% value only). These findings are consistent with those illustrated by Figure 3. In contrast, the second mixing distribution (distribution of X_2) does not allow us to classify correctly "A" and "other" in our three-mixture model. The low proportion of "A" claims assigned to the model X_2 indicates that they are generally unlikely to arrive from X_2 which may be an indication of some further non-homogeneity among claims, even after adjusting for the type "A". The X_2 component could be, therefore, the manifestation of some additional hidden factors, which again confirms the findings summarized in the previous section.[8]

Table 5: Fuzzy membership via responsibility probabilities

α	Resp. X_1		Resp. X_2		Resp. X_3	
	A	Other	A	Other	A	Other
0.9	0.05	0.42	0.00	0.00	0.00	0.00
0.8	0.05	0.42	0.00	0.00	0.54	0.13
0.7	0.07	0.45	0.09	0.11	0.62	0.19
0.6	0.08	0.46	0.13	0.20	0.70	0.22
0.5	**0.08**	**0.48**	**0.13**	**0.24**	**0.75**	**0.24**
0.4	0.11	0.49	0.16	0.28	0.78	0.26
0.3	0.16	0.54	0.21	0.30	0.79	0.30
0.2	0.22	0.64	0.24	0.35	0.79	0.34
0.1	0.95	1.00	0.33	0.41	0.82	0.38

2.4 Accuracy Assessment for the EM Output via Parametric Bootstrap

In our analysis of the BI data conducted in the previous sections we have used the numeric values of the estimated parameters (10) and the responsibilities (13). Since these values were estimated from the data via the EM algorithm, it is important to learn about their accuracy. In general, for the set of parameters (10) the usual approach to assessing accuracy based on the asymptotic normality of the maximum likelihood estimators can be applied here, as soon as we calculate the information matrix for Θ. This is slightly more complicated for the set of responsibilities (13) as they are the functions of Θ and hence require the appropriate transformation of the information matrix. However, a simpler method of obtaining, for instance, confidence intervals for the set of responsibilities and the model parameters can be also used, based on the so-called *parametric*

[8]An analysis of the mixture model applied only to 272 "other" claims shows that X_2 has a more pronounced representation (high α-cut proportions) of (i) chiropractic and physical therapy treatment, (ii) special investigations and independent medical examinations, and (iii) extended treatment delays.

bootstrap method outlined in Algorithm 3 below. The method can be shown to be asymptotically equivalent to the normal approximation approach and is known to be often more reliable for smaller sample sizes or for the heavily biased estimators (which will often be the case for the responsibilities (13)). The algorithm below describes how to obtain confidence intervals for the parameters given by (10) and (13) using bootstrap. For some more examples and further discussion see, for instance, McLachlan and Peel (2000) or the forthcoming paper by Rempala and Szatzschneider (2002) where also the issue of the hypothesis testing for the number of mixture components via the parametric bootstrap method is discussed.

Algorithm 3 *(Bootstrap confidence intervals)*

1 *Using the values of the model parameters* (10) *obtained from the EM algorithm generate the set of pseudo-data* $\mathcal{X}^*$ *(typically of the same length as the original data* $\mathcal{X}$*).*

2 *With* $\mathcal{X}^*$ *at hand, use Algorithm 2 in order to obtain a set of pseudo-values* Θ^*.

3 *Using the set of the original data values* $\mathcal{X}$ *and* Θ^* *from step 2 above, calculate the pseudo-responsibilities* δ^*_{jk} *as in Algorithm 2 step 2.*

4 *Repeat the steps 1–3 a large number of times, say,B.* [9]

5 *Use the empirical quantiles of the distributions of pseudo-values* Θ^* *and* δ^*_{jk} *to obtain confidence bounds for* Θ *and* δ_{jk}.

For illustration purpose we present the set of confidence intervals for the three-mixture-normal model parameters and the responsibilities (of X_3) obtained via the above algorithm for the BI data in Tables 6 and 7 below. The term "bootstrap estimate" in the tables refers to the average value of the B bootstrap pseudo-values obtained in steps 2 or 3.

3 Summary and Conclusion

This paper has introduced the statistical methodology for inference in the presence of missing data, known as the EM algorithm, into the actuarial settings. We have shown that this methodology is particularly appropriate for those practical situations which require consideration of the missing or incomplete data, the "lurking" variables, or the hidden factors. We believe that due to its conceptual simplicity, the EM method could become a standard tool of actuarial analysis in the future. Herein we have given only some example of its usefulness in modeling loss severity. Specifically, in modeling claim severities, the EM algorithm was used to impute missing values in a more sophisticated and statistically less biased way than simple regression methods as well as to uncover (hidden) patterns in the claim severity data. Actual auto bodily injury liability claims closed in Massachusetts in 2001 were used to illustrate a first stage screen for abusive medical providers, and

[9] In our setting B needs to be fairly large, typically at least a thousand. For a discussion see, for instance, McLachlan and Peel (2001).

Table 6: Accuracy of the parameter estimates for the BI data with B=1000

Parameter	Value	Bootstrap Estimate	95% CI
μ_1	0.107	0.104	(-0.115, 0.298)
μ_2	0.874	0.871	(0.809, 0.924)
μ_3	1.248	1.249	(1.216, 1.284)
$\sigma_1^{1/2}$	1.271	1.269	(1.132, 1.389)
$\sigma_2^{1/2}$	0.178	0.175	(0.125, 0.222)
$\sigma_3^{1/2}$	0.146	0.144	(0.117, 0.174)
π_2	0.205	0.207	(0.157, 0.253)
π_3	0.314	0.317	(0.268, 0.375)

Table 7: Accuracy of the selected responsibilities δ_{j3}

No (j)	Log Claim Value	δ_{j3} Value	Bootstrap Estimate	95% CI
100	0.380	0.000	0.000	(3.90e-12, 2.04e-06)
200	1.031	0.410	0.396	(0.243, 0.531)
300	1.353	0.854	0.863	(0.802 0.912)

their abusive claims, utilizing the EM algorithm. The usefulness of the EM output for classification purpose and its connections with fuzzy logic techniques were discussed. Namely, the EM algorithm output of posterior probabilities called responsibilities were reinterpreted as fuzzy set membership function in order to bring the machery of fuzzy logic to bear in the classification problem. The Monte-Carlo based method of assessing the accuracy of the model parameters fitted via the EM algorithm, known as the parametric bootstrap was also presented and the appropriate algorithm for its implementation was developed. The set of functions written in the statistical language R, implementing the EM algorithms discussed in the paper, have been included in Appendix A to allow readers to try different actuarial situations where missing data and hidden components might be found. A large variety of actuarial and financial applications of the presented methodology are possible, including its incorporation into models of Risk Based Capital, Value at Risk, and general Dynamic Financial Analysis. We hope that this paper shall promote enough interest in the EM methodology for further exploration of those opportunities.

Acknowledgement. The authors wish to acknowledge Dr Krzysztof M. Ostaszewski, FSA for his encouragements and helpful comments at the initial stages of this project.

References

Abrahamse, Alan F. and Stephan J. Carroll (1999). The Frequency of Excess Claims for Automobile Personal Injuries, in *Automobile Insurance: Road Safety, New Drivers, Risks, Insurance Fraud and Regulation,* Dionne, Georges and Claire Laberge-Nadeau, Eds., Kluwer Academic Publishers, pp. 131-150.

Bilmes, Jeff A. (1998). *Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models,* International Computer Science Institute. UC Berkeley.

Brockett, Patrick L., Xiaohua Xia and Richard A. Derrig (1998). Using Kohonen's Self-Organizing Feature Map to Uncover Automobile Bodily Injury Claims Fraud, *Journal of Risk and Insurance,* June, Volume 65, No. 2.

Campbell, John W., Andrew Y. Lo, and Archie Craig MacKinlay (1996). *The Econometrics of Financial Markets.* Princeton University Press.

Crocker, Keith J., and Sharon Tennyson (1999). Costly State Falsification or Verification? Theory and Evidence from Bodily Injury Liability Claims, in *Automobile Insurance: Road Safety, New Drivers, Risks, Insurance Fraud and Regulation,* Dionne, Georges and Claire Laberge-Nadeau, Eds., Kluwer Academic Publishers, pp.119-131.

Cummins, J. David, and Sharon Tennyson (1996). Controlling Automobile Insurance Costs, *Journal of Economic Perspectives,* Spring, Volume 6, No. 2. pp. 95–115.

D'Arcy, S. P.,Gorvett, R. W., Herbers, J. A., and Hettinger, T. E. (1997). Building a Dynamic Financial Analysis Model that Flies. *Contingencies* November/December, 40–45.

Dempster, Allan P., N. M. Laird and D. B. Rubin (1977). Maximum likelihood from incomplete data using EM algorithm, *Journal of Royal Statistical Society Series B,* Volume 39, No. 1. pp. 1–38.

Derrig, Richard A. (2002). Insurance Fraud. *Journal of Risk and Insurace,* Volume 69, No. 3. pp. 271–287.

Derrig, Richard A. and K.M. Ostaszewski (1999). Fuzzy Sets Methodologies in Actuarial Science, *Practical Applications of Fuzzy Technologies,* Hans-Jurgen Zimmerman Eds, Kluwer Academic Publishers, Boston, (November).

Derrig, Richard A., K.M. Ostaszewski, and G.A. Rempala (2001). Applications of Resampling Methods in Actuarial Practice, *Proceedings of the Casualty Actuarial Society,* Volume LXXXVII, pp. 322–364.

Derrig, Richard A., and K.M. Ostaszewski (1995). Fuzzy Techniques of Pattern Recognition in Risk and Claim Classification, *Journal of Risk and Insurance,* Volume 62, No. 3. pp. 447–482.

Derrig, Richard A., Herbert I. Weisberg and Xiu Chen (1994). Behavioral Factors and Lotteries Under No-Fault with a Monetary Threshold: A Study of Massachusetts Automobile Claims", *Journal of Risk and Insurance,* June, Volume 61, No. 2. pp. 245–275.

Embrechts, Paul, Alexander McNeil and Daniel Straumann (2000). Correlation and Dependence in Risk Management: Properties and Pitfalls, in *Extremes and Integrated Risk Management,* Paul Embrechts, Ed. pp 71–76. Risk Books. London.

Hastie, Trevor, Robert Tibshirani and Jerome Friedman (2001). *The Elements of Statistical Learning,* Springer-Verlag New York, Inc., New York.

Little, Roderick J.A., and Donald B. Rubin (1987). *Statistical Analysis with Missing Data,* John Wiley & Sons, Inc., Canada.

McLachlan, Geoffrey and David Peel (2000). *Finite mixture models,* Wiley-Interscience, New York.

McLachlan, Geoffrey and Thriyambakam Krishnan (1997). *The EM algorithm and extensions,* Wiley, New York.

Major, John A., and Dan R. Riedinger (1992). EFD: A Hybrid Knowledge/Statistical -Based System for the Detection of Fraud, International Journal of Intelligent Systems. Volume 7, pp. 687–703 (reprinted Journal of Risk and Insurance, Vol 69, No.3, pp 309–324 September, 2002).

Owen, Art, B. (2001). *Empirical Likelihood.* Chapman and Hall. New York.

Rempala, Grzegorz A. and Konrad Szatzschneider (2002). Bootstrapping Parametric Models of Mortality. Manuscript, to appear in *Scandinavian Actuarial Journal.*

Wang, Shaun (1998). Aggregation of Correlated Risk Portfolios: Models and Algorithms *Proceedings of the Casualty Actuarial Society,* Volume LXXXIII, pp 848–939

Wu, Jeff A (1983). On the convergence properties of the EM algorithm, *Annals of Statistics,* Volume 11, No. 1. pp. 95–103.

Xu, Li and Mike I Jordan (1996). On convergence properties of the EM algorithm for gaussian mixtures. *Neural computation,* Volume 8, pp. 129–151.

Appendix A. R Functions

We present here the implementation of Algorithms 1 and 2 in statistical software R which is a freeware version of the award winning statistical software S+ and is available from http://www.r-project.org. The functions below were used in the numerical examples discussed in the text.

```
#Algorithm 1: EM version of Buck's imputation procedure####################
#auxiliary function
inv<-function(m,...)  solve(m,diag(rep(1,length(m[,1]))),...);
#defining matrix inverse (for compatibility with older versions of R)
#input parameters
# d -dataframe of two columns containing complete observations
# d1-list of observations with missing second coordinate
# d2-list of observations with missing first coordinate
# B-maximal number of iterations (default value 500)
# eps- convergence criterion (default value .0001)
#########################################################################

em.buck <-function(d,d1,d2,B=500,eps=.0001) {
n<-length(d[,1]);
n1<-length(d1);
n2<-length(d2);
m<-apply(d,2,mean);
R<-cov(d);
rho<-cor(d)[1,2]; nLL.old<-eps; nLL.new<-100;
w<-rbind(d,cbind(d1,rep(m[2],n1)),cbind(rep(m[1],n2),d2));
i<-1; #mainloop#
while (abs(nLL.new-nLL.old)/nLL.old>eps && i<=B){
T1<-sum(w[,1]); T2<-sum(w[,2]); T12<-sum(w[,1]*w[,2]);
T11<-sum(w[(n+n1+1):(n+n1+n2),1]^2
+R[1,1]*(1-rho^2))+sum(w[-((n+n1+1):(n+n1+n2)),1]^2);
T22<-sum(w[(n+1):(n+n1),2]^2+R[2,2]*(1-rho^2))+sum(w[-((n+1):(n+n1)),2]^2);
R<-array(c(T11-T1^2/(n+n1+n2),T12-T1*T2/(n+n1+n2),T12-T1*T2/(n+n1+n2),
T22-T2^2/(n+n1+n2))/(n+n1+n2), c(2,2));
m<-c(T1/(n+n1+n2),T2/(n+n1+n2));
rho<-R[1,2]/sqrt(R[1,1]*R[2,2]);
w[(n+1):(n+n1),2]<-m[2]+R[1,2]*(w[(n+1):(n+n1),1]-m[1])/R[1,1];
w[(n+n1+1):(n+n1+n2),1]<-m[1]+R[1,2]*(w[(n+n1+1):(n+n1+n2),2]-m[2])/R[2,2];
nLL.old<-nLL.new;
s<-0; for (k in 1:(n+n1+n2)) s<-(w[k,]-m)%*%inv(R)%*%(w[k,]-m)+s;
nLL.new<-2*(n+n1+n2)*log(2*pi)+s+(n+n1+n2)*log(abs(det(R)));
i<-i+1; }; #end mainloop#
print(paste("n=", n, n1, n2, "Theta estimates=", m[1],m[2], R[1,1],
R[1,2], R[2,2],"iter=",i-1,"-2LL=",nLL.new,"rho=",rho))
return(list(m=m, R=R,iter=i-1,LL=nLL.new,w=w)); }
#output parameters:  list of objects (m,R,iter,LL,w)
```

```
# m -vector of estimated means
# R -estimated covariance matrix
# iter -number of iterations until convergence
# w-concatenated dataframe of d,d1,d2 along with imputed missing values
```

 # Algorithm 2: EM for normal mixtures ###############################

```
#auxiliary function
lsum<-function(a,p.new,m,s){ k<-length(m); ss<-0;
for (i in 1:k) ss<-ss+p.new[[i]]*dnorm(a,m[[i]],s[[i]]);
return(ss)}
# facilitates calculation of LL in the main procedure below

# input parameters:
# a -any list of numeric data
# pi -initial estimate of mixing proportions (default value:  uniform over three components)
# eps -desired convergence accuracy (default value .0001)
# B -maximal number of iterations allowed (default value 100)
# m -initial values of means estimates (default value:  random selection from a)
#####################################################################
em.multnorm<-function(a, pi=c(1/3,1/3,1/3),eps=.0001,B=100,m=sort(sample(a,3)))
{n<-length(a); k<-length(m); s<-rep(sd(a),k);
i<-1; p.new<-pi;
m0<-m;
logl.old<-1;
logl.new<-sum(log(lsum(a,p.new,m,s)));
#mainloop#
while (abs((logl.new-logl.old)/logl.old)>eps && i<=B)
{g<-NULL;
for (t in 1:k) g<-rbind(g, p.new[[t]]*dnorm(a,m[[t]],s[[t]])/lsum(a,p.new,m,s));
m<-g%*%a/g%*%rep(1,n);
s<-sqrt(g%*%a^2/g%*%rep(1,n)-m^2);
p.old<-p.new; p.new<-g%*%rep(1,n)/n; i<-i+1;
logl.old<-logl.new;
logl.new<-sum(log(lsum(a,p.new,m,s)));
};
#end mainloop#
print(paste("Theta estimates",m,s,"pi=",p.new,"iter=",i-1,"-2LL=",-2*logl.new));
return(list(m=m,s=s,pi=p.new,iter=i-1,start=m0,logl=-2*logl.new,resp=t(g),data=a)) }

# output parameters:  list of objects (m,s,pi,iter,logl,resp)
# m - vector of estimated values of means
# s - vector of estimated values of standard deviations
# pi -vector of estimated value of mixing proportions
# iter- number of iterations until convergence
# logl- final value of -2Q
# resp- matrix of responsibilities (columns correspond to mixture components)
```

Appendix B. Massachusetts Auto Insurance Bodily Injury Liability Data

Below we present the set of Auto Insurance Data discussed in the paper. Medical bill claim amounts are given in thousands. Responsibilities δ_{j3} are calculated according to Algorithm 2.

No	Claimed Amt	Log(Amt)	Provider	Resp. δ_{j3}	No	Claimed Amt	Log(Amt)	Provider	Resp. δ_{j3}
1	0.045	-3.101	Other	0.00	2	0.047	-3.058	Other	0.00
3	0.07	-2.659	Other	0.00	4	0.075	-2.590	Other	0.00
5	0.077	-2.564	Other	0.00	6	0.092	-2.386	Other	0.00
7	0.117	-2.146	Other	0.00	8	0.117	-2.146	Other	0.00
9	0.14	-1.966	Other	0.00	10	0.145	-1.931	Other	0.00
11	0.149	-1.904	Other	0.00	12	0.165	-1.802	Other	0.00
13	0.167	-1.790	Other	0.00	14	0.169	-1.778	Other	0.00
15	0.18	-1.715	Other	0.00	16	0.18	-1.715	Other	0.00
17	0.199	-1.614	Other	0.00	18	0.202	-1.599	Other	0.00
19	0.212	-1.551	Other	0.00	20	0.225	-1.492	Other	0.00
21	0.23	-1.470	Other	0.00	22	0.242	-1.419	Other	0.00
23	0.264	-1.332	Other	0.00	24	0.275	-1.291	Other	0.00
25	0.285	-1.255	Other	0.00	26	0.29	-1.238	Other	0.00
27	0.363	-1.013	Other	0.00	28	0.384	-0.957	Other	0.00
29	0.4	-0.916	Other	0.00	30	0.4	-0.916	Other	0.00
31	0.413	-0.884	Other	0.00	32	0.414	-0.882	Other	0.00
33	0.416	-0.877	Other	0.00	34	0.425	-0.856	Other	0.00
35	0.425	-0.856	Other	0.00	36	0.43	-0.844	Other	0.00
37	0.43	-0.844	A	0.00	38	0.431	-0.842	Other	0.00
39	0.45	-0.799	Other	0.00	40	0.46	-0.777	Other	0.00
41	0.486	-0.722	Other	0.00	42	0.5	-0.693	Other	0.00
43	0.5	-0.693	Other	0.00	44	0.514	-0.666	A	0.00
45	0.531	-0.633	Other	0.00	46	0.54	-0.616	Other	0.00
47	0.556	-0.587	Other	0.00	48	0.564	-0.573	Other	0.00
49	0.6	-0.511	Other	0.00	50	0.605	-0.503	Other	0.00
51	0.605	-0.503	Other	0.00	52	0.65	-0.431	Other	0.00
53	0.66	-0.416	Other	0.00	54	0.66	-0.416	Other	0.00
55	0.685	-0.378	Other	0.00	56	0.69	-0.371	Other	0.00
57	0.698	-0.360	Other	0.00	58	0.7	-0.357	Other	0.00
59	0.705	-0.350	Other	0.00	60	0.725	-0.322	Other	0.00
61	0.74	-0.301	Other	0.00	62	0.75	-0.288	Other	0.00
63	0.78	-0.248	Other	0.00	64	0.785	-0.242	Other	0.00
65	0.785	-0.242	Other	0.00	66	0.806	-0.216	Other	0.00
67	0.825	-0.192	Other	0.00	68	0.825	-0.192	Other	0.00
69	0.83	-0.186	Other	0.00	70	0.836	-0.179	A	0.00
71	0.87	-0.139	Other	0.00	72	0.9	-0.105	Other	0.00
73	0.934	-0.068	Other	0.00	74	0.95	-0.051	Other	0.00
75	0.954	-0.047	Other	0.00	76	0.956	-0.045	Other	0.00
77	0.962	-0.039	Other	0.00	78	0.97	-0.030	Other	0.00
79	0.975	-0.025	Other	0.00	80	0.988	-0.012	Other	0.00

No	Claimed Amt	Log(Amt)	Provider	Resp. δ_{j3}	No	Claimed Amt	Log(Amt)	Provider	Resp. δ_{j3}
81	1.015	0.015	Other	0.00	82	1.053	0.052	Other	0.00
83	1.058	0.056	Other	0.00	84	1.08	0.077	Other	0.00
85	1.161	0.149	Other	0.00	86	1.167	0.154	Other	0.00
87	1.195	0.178	Other	0.00	88	1.215	0.195	Other	0.00
89	1.242	0.217	Other	0.00	90	1.26	0.231	Other	0.00
91	1.295	0.259	Other	0.00	92	1.31	0.270	Other	0.00
93	1.319	0.277	Other	0.00	94	1.33	0.285	Other	0.00
95	1.34	0.293	Other	0.00	96	1.355	0.304	Other	0.00
97	1.39	0.329	Other	0.00	98	1.444	0.367	Other	0.00
99	1.455	0.375	Other	0.00	100	1.463	0.380	Other	0.00
101	1.49	0.399	Other	0.00	102	1.5	0.405	Other	0.00
103	1.542	0.433	Other	0.00	104	1.598	0.469	Other	0.00
105	1.616	0.480	Other	0.00	106	1.623	0.484	Other	0.00
107	1.64	0.495	Other	0.00	108	1.645	0.498	A	0.00
109	1.65	0.501	Other	0.00	110	1.66	0.507	Other	0.00
111	1.68	0.519	Other	0.00	112	1.695	0.528	Other	0.00
113	1.7	0.531	A	0.00	114	1.758	0.564	Other	0.00
115	1.759	0.565	Other	0.00	116	1.76	0.565	Other	0.00
117	1.896	0.640	Other	0.00	118	1.92	0.652	Other	0.00
119	1.923	0.654	Other	0.00	120	1.941	0.663	A	0.00
121	1.96	0.673	Other	0.00	122	1.972	0.679	Other	0.00
123	1.99	0.688	Other	0.00	124	2.005	0.696	Other	0.00
125	2.018	0.702	Other	0.00	126	2.02	0.703	Other	0.00
127	2.02	0.703	Other	0.00	128	2.03	0.708	Other	0.00
129	2.042	0.714	A	0.00	130	2.062	0.724	Other	0.00
131	2.063	0.724	Other	0.00	132	2.08	0.732	Other	0.00
133	2.087	0.736	Other	0.00	134	2.089	0.737	Other	0.00
135	2.1	0.742	Other	0.00	136	2.115	0.749	Other	0.01
137	2.12	0.751	Other	0.01	138	2.155	0.768	Other	0.01
139	2.159	0.770	Other	0.01	140	2.161	0.771	Other	0.01
141	2.184	0.781	A	0.01	142	2.188	0.783	Other	0.01
143	2.191	0.784	Other	0.01	144	2.196	0.787	Other	0.01
145	2.224	0.799	Other	0.02	146	2.237	0.805	Other	0.02
147	2.251	0.811	A	0.02	148	2.253	0.812	Other	0.02
149	2.288	0.828	Other	0.03	150	2.295	0.831	Other	0.03
151	2.318	0.841	Other	0.03	152	2.325	0.844	Other	0.03
153	2.325	0.844	A	0.03	154	2.335	0.848	Other	0.04
155	2.341	0.851	Other	0.04	156	2.35	0.854	Other	0.04
157	2.374	0.865	Other	0.05	158	2.39	0.871	Other	0.05
159	2.406	0.878	Other	0.06	160	2.434	0.890	Other	0.07
161	2.45	0.896	Other	0.08	162	2.453	0.897	A	0.08
163	2.468	0.903	Other	0.09	164	2.468	0.903	A	0.09
165	2.48	0.908	Other	0.10	166	2.48	0.908	A	0.10
167	2.49	0.912	A	0.10	168	2.498	0.915	Other	0.11
169	2.5	0.916	Other	0.11	170	2.5	0.916	Other	0.11
171	2.5	0.916	A	0.11	172	2.51	0.920	Other	0.11
173	2.532	0.929	Other	0.13	174	2.54	0.932	Other	0.14

No	Claimed Amt	Log(Amt)	Provider	Resp. δ_{j3}	No	Claimed Amt	Log(Amt)	Provider	Resp. δ_{j3}
175	2.543	0.933	Other	0.14	176	2.559	0.940	Other	0.15
177	2.572	0.945	Other	0.16	178	2.593	0.953	Other	0.18
179	2.601	0.956	Other	0.19	180	2.616	0.962	Other	0.20
181	2.619	0.963	Other	0.20	182	2.63	0.967	Other	0.21
183	2.635	0.969	Other	0.22	184	2.635	0.969	Other	0.22
185	2.653	0.976	Other	0.24	186	2.655	0.976	Other	0.24
187	2.675	0.984	Other	0.26	188	2.679	0.985	Other	0.26
189	2.697	0.992	Other	0.28	190	2.718	1.000	Other	0.31
191	2.73	1.004	Other	0.32	192	2.734	1.006	Other	0.32
193	2.755	1.013	Other	0.35	194	2.758	1.015	Other	0.35
195	2.773	1.020	Other	0.37	196	2.775	1.021	Other	0.37
197	2.78	1.022	Other	0.38	198	2.785	1.024	A	0.38
199	2.795	1.028	Other	0.40	200	2.805	1.031	Other	0.41
201	2.805	1.031	Other	0.41	202	2.808	1.032	A	0.41
203	2.88	1.058	Other	0.49	204	2.881	1.058	Other	0.50
205	2.881	1.058	A	0.50	206	2.924	1.073	A	0.54
207	2.93	1.075	Other	0.55	208	2.934	1.076	A	0.55
209	2.94	1.078	Other	0.56	210	2.972	1.089	Other	0.59
211	2.975	1.090	Other	0.59	212	3	1.099	Other	0.62
213	3	1.099	A	0.62	214	3.025	1.107	Other	0.64
215	3.058	1.118	Other	0.67	216	3.082	1.126	A	0.68
217	3.085	1.127	Other	0.69	218	3.095	1.130	Other	0.69
219	3.1	1.131	Other	0.70	220	3.102	1.132	A	0.70
221	3.106	1.133	Other	0.70	222	3.135	1.143	Other	0.72
223	3.17	1.154	Other	0.74	224	3.187	1.159	Other	0.75
225	3.192	1.161	A	0.75	226	3.193	1.161	Other	0.75
227	3.2	1.163	Other	0.76	228	3.21	1.166	Other	0.76
229	3.23	1.172	Other	0.77	230	3.23	1.172	Other	0.77
231	3.23	1.172	A	0.77	232	3.232	1.173	Other	0.77
233	3.235	1.174	Other	0.78	234	3.243	1.176	A	0.78
235	3.248	1.178	A	0.78	236	3.249	1.178	Other	0.78
237	3.26	1.182	Other	0.79	238	3.261	1.182	Other	0.79
239	3.272	1.185	A	0.79	240	3.29	1.191	Other	0.80
241	3.295	1.192	Other	0.80	242	3.304	1.195	Other	0.80
243	3.332	1.204	A	0.81	244	3.333	1.204	Other	0.81
245	3.338	1.205	Other	0.81	246	3.34	1.206	Other	0.82
247	3.341	1.206	A	0.82	248	3.349	1.209	A	0.82
249	3.349	1.209	A	0.82	250	3.349	1.209	A	0.82
251	3.353	1.210	A	0.82	252	3.36	1.212	Other	0.82
253	3.378	1.217	A	0.83	254	3.385	1.219	A	0.83
255	3.387	1.220	A	0.83	256	3.416	1.228	Other	0.84
257	3.429	1.232	A	0.84	258	3.438	1.235	A	0.84
259	3.444	1.237	A	0.84	260	3.469	1.244	A	0.85
261	3.473	1.245	A	0.85	262	3.473	1.245	A	0.85
263	3.475	1.246	A	0.85	264	3.477	1.246	A	0.85
265	3.505	1.254	Other	0.85	266	3.517	1.258	A	0.85
267	3.518	1.258	Other	0.85	268	3.527	1.260	A	0.85

No	Claimed Amt	Log(Amt)	Provider	Resp. δ_{j3}	No	Claimed Amt	Log(Amt)	Provider	Resp. δ_{j3}
269	3.535	1.263	A	0.86	270	3.547	1.266	A	0.86
271	3.55	1.267	Other	0.86	272	3.552	1.268	Other	0.86
273	3.567	1.272	A	0.86	274	3.57	1.273	Other	0.86
275	3.575	1.274	Other	0.86	276	3.58	1.275	Other	0.86
277	3.583	1.276	A	0.86	278	3.59	1.278	A	0.86
279	3.603	1.282	A	0.86	280	3.615	1.285	A	0.86
281	3.623	1.287	A	0.86	282	3.647	1.294	A	0.86
283	3.655	1.296	Other	0.86	284	3.655	1.296	A	0.86
285	3.658	1.297	Other	0.87	286	3.675	1.302	Other	0.87
287	3.675	1.302	Other	0.87	288	3.687	1.305	A	0.87
289	3.72	1.314	Other	0.87	290	3.72	1.314	Other	0.87
291	3.742	1.320	Other	0.87	292	3.757	1.324	A	0.87
293	3.765	1.326	Other	0.87	294	3.8	1.335	A	0.87
295	3.809	1.337	Other	0.86	296	3.848	1.348	A	0.86
297	3.857	1.350	A	0.86	298	3.867	1.352	Other	0.86
299	3.867	1.352	A	0.86	300	3.87	1.353	Other	0.86
301	3.883	1.357	Other	0.86	302	3.89	1.358	Other	0.86
303	3.905	1.362	A	0.86	304	3.907	1.363	A	0.86
305	4	1.386	Other	0.85	306	4.011	1.389	Other	0.85
307	4.039	1.396	A	0.84	308	4.065	1.402	A	0.84
309	4.095	1.410	Other	0.83	310	4.134	1.419	Other	0.82
311	4.147	1.422	Other	0.82	312	4.155	1.424	A	0.82
313	4.17	1.428	Other	0.81	314	4.179	1.430	A	0.81
315	4.2	1.435	Other	0.81	316	4.215	1.439	Other	0.80
317	4.257	1.449	A	0.79	318	4.3	1.459	Other	0.78
319	4.489	1.502	A	0.70	320	4.593	1.525	A	0.64
321	4.595	1.525	Other	0.64	322	4.63	1.533	A	0.62
323	4.653	1.538	Other	0.60	324	4.7	1.548	A	0.57
325	4.731	1.554	Other	0.55	326	4.741	1.556	A	0.55
327	4.75	1.558	Other	0.54	328	4.761	1.560	Other	0.53
329	4.81	1.571	Other	0.50	330	5.072	1.624	Other	0.31
331	5.161	1.641	Other	0.25	332	5.24	1.656	Other	0.20
333	5.64	1.730	Other	0.06	334	5.779	1.754	Other	0.03
335	6.166	1.819	Other	0.01	336	6.406	1.857	Other	0.00
337	6.725	1.906	Other	0.00	338	7.717	2.043	A	0.00
339	8	2.079	Other	0.00	340	9.5	2.251	Other	0.00
341	10.295	2.332	Other	0.00	342	12.533	2.528	Other	0.00
343	12.688	2.541	Other	0.00	344	16.043	2.775	Other	0.00
345	18.847	2.936	Other	0.00	346	19.5	2.970	Other	0.00
347	20.827	3.036	Other	0.00	348	50	3.912	Other	0.00

Where is My Market? How to Use Data to Find and Validate New Commercial Lines Market Niches

Lisa Sayegh, MBA, ARM

Where is My Market?
How–to Use Data to Find and Validate New Commercial Lines Market Niches

By Lisa Sayegh, MBA, ARM

Entering a new insurance market is not a decision to be taken lightly. Market segment analysis is a lengthy process, and finding the right data is just the beginning. Being able to make meaningful comparisons of data from various sources and across insurance lines is the key to identifying profitable markets. Fortunately, there are data sources and tools available that can help with the analysis, as well as provide quantifiable assessments of your niche-market recommendations. Here are some critical elements to keep in mind as you go through the process.

An approach to help insurance marketers identify profitable segments in an industry where costs can be unknown for a multiple year period:

1. Identify Your Company Strengths
2. Define Market Segments That Utilize Your Strengths
3. Validate Your Selections
   - Validate Using Insurance Data
   - Additional Considerations When Identifying Markets
   - What Data Is Available?
   - Using Multiple Data Sources
4. Provide High Quality Prospects

Why segment?

The segmentation process guides the marketer through defining market characteristics to enable comparison across slices of the market or segments and from which, attractive markets can be decided on. Consider two examples. A major insurer successfully used segmentation to identify optimum markets with established distribution channels by overlaying selected segments to agent territories. Another company learned that an adjacent state was not nearly as attractive for growth as expected, for reasons other than the regulatory environment of the state. Their initial expectation of the adjacent state was based solely on lower distribution costs. It turned out that by segmentation analysis, they found tremendous promise in a third region. This new segment was so attractive that it justified increased distribution costs to access the market.

Identifying target market segments is an iterative analytical process. The market researcher begins the analysis by identifying segments on a high level basis, then continues with further segmentation criteria to validate findings and to hone in on optimum market segments. Insurance market data can be used at all levels of the analysis to improve validations of the selected market segments. Use of data in the initial phase of the analysis should help your business canvas the markets and identify the most profitable segments. Once these initial segments are proposed, together with your business experience and knowledge of your business, data can be collected and monitored to help justify your selections. The data provides management with quantifiable assessments of markets and eases their decision process. It also should be emphasized that it is necessary to review current market penetrations on an ongoing basis to identify changing business environments.

All companies intuitively segment to a certain degree, but conducting the delineated segmentation analysis in this article can result in higher profitability. As per the example above, a company identified a profitable niche and decided to extend their program to the adjacent state. Via this analysis, the insurer found that expanding into a third region – not their first choice – provided a better opportunity. Particularly in these changing times, a company can easily miss profitable opportunities without segmentation analysis.

<u>Identify Your Company Strengths-</u>

The marketer would evaluate their business and company strengths by first identifying areas with performance success and then identifying the characteristics that make these areas successful. The marketer can look at functions within the company, such as underwriting, loss control and claims management - or industries, et al. Benchmarking can show that a company's book of funeral homes has fewer slips and falls than the funeral homes industry standard or that a closed-claim that is 3 days shorter than a comparable business.

If a company is unknown to a market, it appears that costly efforts would make it unreasonable to succeed in a new line of business where it's unknown. This analytical process can help the company to identify strengths that would be portable to what seemingly is a non-related business.

Company strengths can also be expertise in a pertinent industry segment. A company known as an auto coverage provider that decides to introduce coverage for high-price dwellings would need to assess costs associated with introducing their services to a marketplace that is unfamiliar with them. These costs can include advertising, new distribution channels, etc. Companies can have specialists on hand with skills that transfer to other industry segments. For example, inspectors for boiler and machinery can help to underwrite some types of manufacturing risks. Companies' sales distribution channels are also an asset not to be undervalued. Do your distributors have excellent people skills and show profitable returns on offering homeowner coverages? Can that be expanded to provide home-based business coverage, or businessowners coverages?

There are many sources of industry standard data that marketers can use to benchmark against, much of it not insurance-specific. Some brokers will provide comparative performance data of your company against their entire book. Many insurance associations may have pertinent information. One that comes to mind is the National Restaurant Association that conducts its own survey of insurance costs. (Note that only participants in the survey can access the information). The Risk and Insurance Management Society (RIMS) conducts a survey of the self-insured market that provides some industry slices.

Define market segments that utilize your strengths

Once you identify your strengths, you need parameters to define the market segments that utilize your strengths. This is a challenge for most of us but ultimately worthwhile. Effort expended to identify your own company strengths pays off when you can identify segments that correlate to those strengths.

Industry can be a critical segment definition as it enables you to compare results across line of insurance. Industry premium and loss information is typically categorized according to risk classification code that is specific to line of insurance. This risk classification-code-based data does not lend itself to easy market analysis. Comparison across individual lines of insurance with their different classifications is difficult and so, it is difficult to identify trends that can be applied to their industries. A prime example is the commercial auto line, which is categorized by weight of vehicle and distance of travel. If you identify a particular truck as very profitable, how can you apply that to other industries when you are not sure which industries use that type of truck?

Other parameters include:

- Demographics like geography play a major role in insurance costs as we are a regulated community and results can differ drastically in different geographies.
- Statistics such as sales or payroll can be helpful to identify risk potential.
- Business parameters such as size of risk can be critical to your analysis highlighting cases where the industry average differs greatly from a particular business-size result.

Validate Segment Selections with Premium Potential and Profitability-

The validation of the selected segments generally starts with non-insurance data as it is more abundant and cheaper. As things get refined, it moves to insurer data which is more scarce and more expensive.

Now that the segments have been identified, the marketer needs to support his/her findings. At this stage, it is important to use insurance specific information for the validation. Few industries have to calculate expenses that are paid over a period of years when determining profitability. Business demographics are valuable up to a point. A proposed market segment may show a large number of potential customers, but this may or may not be correlative to premium potential in the market. The number of establishments shows the number of potential customers. Compare the premium to number of establishments to ascertain the average premium and decide if that is a market segment that fits your business approach. Let us say your distribution channel supports companies with higher premium policies. Then you wouldn't be interested in a segment with a large number of establishments resulting in low average premiums. Make sure to check that the average premium per establishment is reflective of the market segment activity. A few outliers such as a few jumbo companies can skew the data and distort an average for a segment that may still be an option for your company.

It has always been a challenge for insurance marketers to obtain profitability data for analyzing markets. Pricing for policies is based on aggregated insurance information using classification codes.

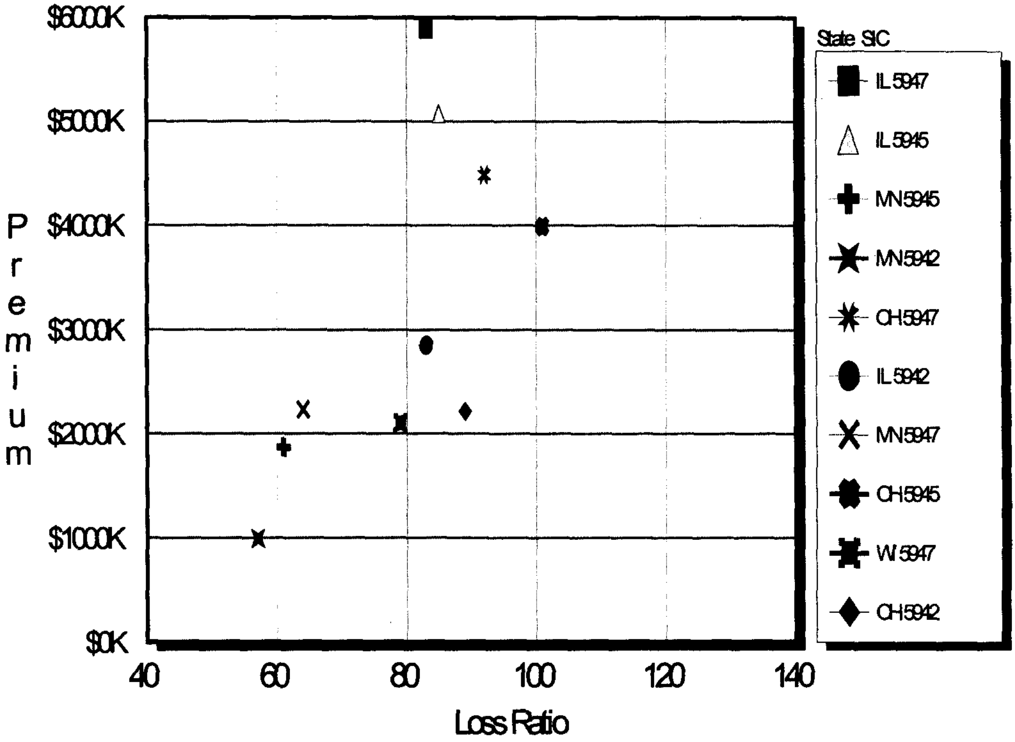
Classification codes reflect insurance risk and since risk characteristics vary by line of insurance, classification codes vary by line. As an example, ISO's classification scheme for General Liability has different class codes for restaurants with or without cooking, and/or with alcohol. The Commercial Fire and Allied line has a single code applying to all restaurants - including bars as well. For analyzing markets, class codes may not provide optimum information to enable comparison across niches. Even after identifying a classification code that provides good results, it would be difficult to apply these results to a particular industry and then apply to other industries. Therefore, it is helpful to be able to study the data by industry segment. To compensate for the lack of actual insurance statistics by industry, models were developed to "translate" insurance data collected by classification code to reflect an industry group.

One product that offers data for industry specific loss ratios by line of business for the three major lines is ISO's Market Profiler derived from modeling. This data can be used for profitability segment analysis.

An example of an analysis that can provide useful information quickly is to compare segment activity based on premium to loss ratio. The below graph compares activity for three 4-digit Standard Industrial Codes or SICs from the 5900 Miscellaneous Retail market segment - book stores (SIC 5942), hobby, toy and game stores (SIC 5945) and gift shops (SIC 5947). These selections are from the central region of the United States - Illinois, Wisconsin, Minnesota and Ohio.

The graph analyzes one year of activity. Other factors play a role in selecting optimum segments for your business, but at a glance, the segments for SIC 5947 in Illinois and again in Illinois for SIC 5845 show large premium opportunity with a comparatively reasonable loss ratio. Also attractive is SIC 5942 in Minnesota showing high profitability but with lower premium potential. The go/no go decision may depend on the ease of access to the Minnesota market.

PREMIUM VS. LOSS RATIO



Other analyses that can be conducted include evaluating trends – both historical and forecasted. Does the segment show reasonable sustained growth over a period of time or less reliable spiked growth? Forecasted data tells you if a trend is expected to continue. Some forecasters concentrate on industry, some on geography. Especially of interest to insurers when selecting markets would be expected premium growth or contraction.

Trending analysis can be conducted by SIC, by geographic selection, by line of business, etc. The below example shows the trend of the same data selection as above by 3 lines of insurance. (The forecasted years' premiums are based on history, economic and geographic forecast statistics, not price changes). The results are sums representing premiums of the 4 selected states. The example shows healthy growth in the general liability area, with commercial property also showing consistent growth. Annual growth in workers compensation seems to be questionable and bears further investigation.

Trend By Line

Conventional Premium- ($000's)

	1997	1998	1999	2000	2001	2002	2003
Line Of Business	**$PREM**	**$PREM**	**$PREM**	**$PREM**	**$PREM**	**$PREM**	**$PREM**
Workers' Compensation	5,742.88	6,196.22	6,689.67	6,527.55	6,580.21	6,727.10	6,822.86
Annual % increase	7.45%	7.89%	7.96%	-2.42%	0.81%	2.23%	1.42%
General Liability	6,606.47	6,759.85	7,328.01	7,827.93	8,204.38	8,652.13	9,106.23
Annual % increase	11.24%	2.32%	8.40%	6.82%	4.81%	5.46%	5.25%
Commercial Property	13,350.30	14,014.55	14,880.96	15,725.05	16,161.39	16,743.64	17,326.78
Annual % increase	1.96%	4.98%	6.18%	5.67%	2.77%	3.60%	3.48%

©ISO Market Profiler., 2001

An additional factor playing a role in segment validation is concentration. Do the number of players in the field leave room for additional penetration? Or do you have sufficient added value that you are confident to gain market share even against a major market holder?

These analyses raise "red flags" that warrant further investigation – the marketers should draw conclusions based on multiple data results and business savvy.

Additional considerations when identifying markets

Once the market segment is validated with supporting insurance information, there are additional checklist items.

- The cost to sell to the selected segment. Have you identified Georgia as an expansion state when your distribution is in California? As mentioned above, an insurer identified with personal auto would need to educate the marketplace when entering a new market. Costs to do so should be taken into any profitability measurements.
- As with any insurance decision, regulations should be examined to determine applicability of new program entry.
- New trends should be evaluated to determine the effect on your choices. The data source types listed below for marketing analysis include general news. If you were

conducting the analysis 10 years ago, would Internet companies have even showed up on the radar?

- An easy "sell" to management is cross selling. Can your identified segments or programs utilize the current customer base? The familiar is the most comfortable.

<u>What Data is Available?</u>

The below is a sample list of data sources. More importantly, it shows the types of data available for your analysis.

- Business information – provides demographic information such as location, sales, number of employees. Examples of sources are Dun & Bradstreet and Claritas / ABI. These sources are usually geared to specific company information and can be pricey options for initial market research when looking at industry slices at a high level.
- Insurance Statistics
  - AM Best- excellent source of data representing the aggregate of insurer-reported data on the Annual Statement - Page 15. However, AM Best provides only state totals, not industry specific information. This makes it difficult to use in market analysis.
  - Property & Casualty Statistical Agents – some of the statistical agents that collect information from insurers in accordance with regulatory requirements are the AAIS, ISO, NAII and NCCI[1]. NCCI aggregates workers compensation for about 80% of the states. ISO collects an estimated 70% of the commercial lines markets data (without workers compensation) for 11 lines of property & casualty lines. Other statistical agents may have strengths in one or more specific lines of insurance. The data is typically geared for pricing and actuarial analysis, is collected according to risk classification codes and may be difficult to apply to industry segment analysis.
  - State Insurance Departments- They may provide information on an individual company's rates which would be helpful when assessing new market penetration. The challenge is that each state has different rules on how to access the information. Using anecdotal sources, we're told that some states only enable access using a pen, not copying. Some states do not allow any access. Access differs, is not consistent and may be time consuming.
- Other insurance information includes access to Public Protection Classification used in fire rating, Fireline (brush fire data), and other geographic based information for fire rating and liability. Crime statistics can be valuable in assessing burglary and theft. Auto types according to industry are available as well as other information that is useful when assessing risk.
- Government information – The government offers a wide range of commercial information: the County Business Pattern from the Census; OSHA for lost workdays

[1] AAIS-American Association Insurance Services, Wheaton, Illinois
ISO- Insurance Services Office, Inc. Jersey City, New Jersey
NAII- National Association of Independent Insurers, Des Plaines, Illinois
NCCI- National Council of Compensation Insurance, Boca Raton, Florida

and other valuable workers compensation information; and the Bureau of Labor Statistics for a wide variety of other information.

- Association information – As mentioned above, RIMS is a source for self-insured information for those marketers targeting very large accounts. The National Restaurant Association is an example of an industry group that collects insurance information.
- Forecasting – Economic forecasting information can be used when assessing future trends. Even if not insurance specific, it can help identify areas of growth. For premium forecasting, ISO's Marketwatch tracks the change in renewal pricing which can be used for a trend analysis.
- News
- Your own data - Your own company data has tremendous value. If the information you need is not available in your current system, check to see if the information is on applications being received by the company but not coded.

Using multiple data sources

Items to question when comparing statistics across multiple data sources include:

- What constitutes a counted business? If a home-based business earns $1,000 a year, is it considered a business?
- How are leased employees handled?
- Is the information comparable i.e. can it be headquarter vs. branch location specific? Does it reflect the entire insurance market including the primary and self-insured portion?

When ready to integrate data across data sources, identify a denominator that applies to both sources. The easiest seems to be industry identification. Currently we use the government defined Standard Industrial Code known as SIC. There are over 900 4-digit SICs or industry slices available. SICs are being replaced by North American Industry Classification System (NAICs). The design of NAICs codes is expected to be easier to apply to insurance applications as it is more process oriented as well as more relevant to our current economy.

Integration across multiple data sources can be on an aggregate level and still provide very useful information. Other denominators for integration can be business size and geography. The Census Bureau - County Business Pattern defined over 10 sizes that are commonly accepted in the industry.

Analyzing with your own company data

Your own company data can become a benchmark tool if you can easily compare segments, however defined. To look at performance by industry, append SIC or NAICS to commercial account records. If you are not capturing this information at policy inception, you can submit your companies with their addresses to a business information provider that matches company by address. This will enable you to obtain SIC or NAICs. Address matching methodology typically identifies over 50% of your book. Business data providers can also offer a wide range of additional data sources. For future analysis, develop a way of appending additional business variables to your account records. Valuable data available include postal data which offers zip code-to-county mapping. And as mentioned above, you may be capturing valuable data already such as length in business, etc.

Provide High Quality prospects

Now that the analysis is done and findings have been validated, you would want to access the potential customers in the identified segments. To ensure that only leads that conform to your segment analysis are selected, focus on those customer characteristics that were identified. Some of the data available through lead generators that enable you to further hone in on your target market include:

- Business contact information
- Premium estimates by line
- Geographic areas (metropolitan statistical area, zip code, county options)
- Business size (# of employees, sales)
- Type of corporation
- Additional data sources for specialty programs
- Computer and telecommunications infrastructure
- Expiration dates
- Credit information

Conclusion

As lead time for successful returns from a new program or market decreases and more stress is placed on pre-qualifying new market choices, the benefit of accessing accurate data and insurance industry data about market segments is increasingly important. A marketer can utilize the data most effectively by following a plan that includes defining and analyzing the needs and strengths of the company, and then identifying additional market segments that would most benefit from those strengths.

Does Credit Score Really Explain Insurance Losses? Multivariate Analysis from a Data Mining Point of View

Cheng-Sheng Peter Wu, FCAS, MAAA, and
James C. Guszcza, ACAS

Does Credit Score Really Explain Insurance Losses? Multivariate Analysis from a Data Mining Point of View

by

Cheng-Sheng Peter Wu and James Guszcza

Abstract

One of the most significant developments in insurance ratemaking and underwriting in the past decades has been the use of credit history in personal lines of business. Since its introduction in late 80's and early 90's, the predictive power of credit score and its relevance to insurance pricing and underwriting have been the subject of debate [1-3]. The fact that personal credit is widely used by insurers strongly suggests its power to explain insurance losses and profitability. However, critics have questioned whether the apparently strong relationship between personal credit and insurance losses and profitability really exists. Surprisingly, even though this is a hot topic in the insurance industry and in regulatory circles, actuaries have not been actively participating in the debate. To date, there have been few actuarial studies published on the relationship of personal credit to insurance losses and profitability. We are aware of only two such studies: one published by Tillinghast, which was associated with the NAIC credit study [4], and the other by Monaghan [5]. A possible reason for the lack of published data is that many insurers view credit scores as a confidential and cutting-edge approach to help them win in the market place. Therefore, they might be reluctant to share their results with the public. In this paper, we will first review the two published studies and comment on their results. We will then share our own experience on this topic. We have conducted a number of comprehensive, large-scale data mining projects in the past that included credit information as well as an extensive set of traditional and non-traditional predictive variables. Because our projects have been true multivariate studies, conducted using rigorous statistical methodology on large quantities of data, our experience should add value to the debate. Our experience does suggest that such a relationship exists even after many other variables have been taken into account.

About the Authors
Cheng-Sheng Peter Wu, F.C.A.S, A.S.A., M.A.A.A., is a director in the Advanced Quantitative Services practice of Deloitte & Touche's Actuarial and Insurance Consulting Group. He is based in the Los Angeles, CA office. Mr. Wu received his masters degrees in chemical engineering and statistics from the Pennsylvania State University. Mr. Wu has published several papers in automotive engineering, tribology (lubrication engineering), statistics, and actuarial science.

James Guszcza, A.C.A.S., M.A.A.A., is a manager in the Advanced Quantitative Services practice of Deloitte & Touche's Actuarial and Insurance Consulting Group. He is based in the Los Angeles, CA office. Mr. Guszcza received his Ph.D. in Philosophy from the University of Chicago.

Messrs. Wu and Guszcza's address is: Deloitte & Touche LLP, 350 South Grand Avenue, Los Angeles, CA 90071

Acknowledgments
The authors would like to thank Matt Carrier, Ravi Kumar, Jan Lommele, Sara Schlenker, and an anonymous referee for their thoughtful comments.

Introduction

One of the more important recent developments in the U.S. insurance industry has been the rapidly growing use of credit scores to price and underwrite personal auto and homeowners insurance. But this development has not come without controversy. Perhaps the most important criticism raised is that there exists no convincing causal picture connecting poor credit history with high insurance loss potential [1-5]. Partly for this reason, many insurance regulators and consumer advocates have expressed doubts that the observed correlations between credit scores and insurance loss history truly reflect an underlying reality. Some critics have suggested that these correlations might be spurious relationships that would not survive more sophisticated (multivariate) statistical analyses.

Given the business significance and statistical nature of this topic, it is curious that actuaries have not participated more actively in the debate. We are aware of only two actuarial studies that have been published so far: one published by Tillinghast, which was associated with the NAIC credit study [4], and the other by Monaghan [5].

The aim of this paper is to review these studies and complement them with a qualitative description of our own experiences in this area. For reasons of confidentiality, we are not able to share detailed quantitative results in this forum. Our focus will be on the use of credit in the line of personal auto, but many of our comments will hold true for other lines of insurance. We will begin with several historical comments on the development of auto classification ratemaking in the United States, and with comments on the actuarial issues relating to the use of credit in auto ratemaking.

The Development of Auto Classification Ratemaking in the United States

Personal auto ratemaking came a long way in the 20th century [6]. Prior to World War II, auto ratemaking involved only three classes: adult, youthful operator, and business use. The three decades after the war saw a proliferation of new class categories such as vehicle characteristics (symbol, model year) and refined driver classifications.

Today, a typical personal auto rating plan contains hundreds, if not thousands of classes involving the following variables:

- *Territorial Characteristics*: insurers define intra-state rating territories that reflect such relevant aspects of the physical environment as population density and traffic conditions.
- *Vehicle Use*: examples include business use, pleasure use, and driving more or less than a certain number of miles per year.
- *Driver characteristics*: examples are age, gender, marital status, and good student status
- *Driving Record*: this is reflected by a point system based on accidents and violations.

- *Vehicle Characteristics*: this typically includes a vehicle symbol system as well as a model year rating structure.
- *Miscellaneous surcharges/discounts*: this is where rating plans vary the most from company to company. Special surcharges or discounts are used to reflect policy characteristics or advances in motor vehicle technology. Commonly seen discounts include multi-car discounts, homeowner discounts, safe driver discounts, anti-lock brake discounts, anti-theft discounts, affinity group factors, and so on.

In addition to the above class variables, a typical rating plan is not complete without a *tier rating structure*. A tier structure is designed to address rating inadequacies that an insurer believes exists in a class plan. For example, an insurer might create three companies for its preferred, standard, and high-risk books, and the rate differential for such companies can range from –20% to 20%. Such differentials are typically applied at the policy level, across all coverages. Tier rating factors can include characteristics that are not used in the class plan, such as how long an insured has been with the insurer. They can also include certain interactions of class factors, such as youthful drivers with poor driving records.

As class plan structures have become more complex, the problem of estimating rates for each combination of class variables has become more difficult. This is because many of the variables used to define rating factors are not statistically independent. For this reason, factors based on univariate analyses of the variables are not necessarily appropriate for a multi-dimensional rating structure. Some form of multivariate analysis is called for.

To take a concrete example, suppose that an existing rating plan charges youthful drivers 3 times that of mature drivers. Furthermore, we analyzed loss (pure premium) relativities by driver age group, and noticed that the youthful driver group has losses per exposure 4 times that of the mature driver group. But it does not follow that the youthful driver rating factor should be raised to 4. This is because other variables used in the class plan might be correlated with age group variable. For example, youthful drivers have more accidents and violations; they are more likely to drive sports cars; they are more likely to be unmarried, and so on. They are therefore likely to be surcharged along these other dimensions of the rating plan. To give them a driver age rating factor of 4 would possibly be to over-rate them.

This issue -- that non-orthogonal rating variables call for multivariate statistical analyses -- lies at the heart of the debate over credit. In addition, this issue is perhaps the key theme in the *methodological* development of classification ratemaking since the 1960's.

McClenahan's ratemaking chapter [7] in *The Foundations of Casualty Actuarial Science* outlines the univariate approach to ratemaking, an approach still employed by many insurance companies. Appealing to examples like the one just given, Bailey and Simon [8,9] pointed out that the univariate approach could lead to biased rates if the individual rating factors are non-orthogonal. Their proposed solution to this problem, the *minimum* *bias* procedure, involves assuming a mathematical relationship between the rating factors and pure premium.

The mathematics of minimum bias is pure algebra: Bailey and Simon derived their models without positing statistical models. In his 1988 paper, Robert Brown [10] showed that commonly used minimum bias formulas could be derived from statistical models via maximum likelihood. Stephen Mildenhall's 1999 paper [11] is the most rigorous examination to date of the statistical underpinnings of the minimum bias method. Thanks to Brown, Mildenhall, and others [12, 13], it is now abundantly clear that Bailey-type actuarial analyses are in fact special cases of Generalized Linear Models. Multi-dimensional classification ratemaking projects should therefore be viewed as exercises in multivariate statistical modeling.

The lesson is obvious: a multivariate statistical analysis is necessary to establish the importance of credit for personal auto ratemaking.

How Credit is Currently Used in Personal Auto Ratemaking

During 1970's and 1980's, when classification ratemaking was undergoing its methodological development, no major rating *variables* were introduced. This changed in the late 1980's and 1990's when credit scores were introduced to personal lines insurance [1].

Raw credit information is supplied by several major credit bureaus, including Choice Point, TransUnion, and Experian. These companies collect individuals' credit data and in turn sell this data in the form of credit reports. Credit reports contain a wealth of information that can be grouped into four classifications:

- General information
- Trade line information
- Inquiries
- Public Records and Collections

The raw fields on these reports can be combined in many ways to create a plethora of random variables. Examples include number of trades, months since oldest trade, amount past due, trade line balance-to-limit ratio, number of inquiries, number of collections, and number of lawsuits. Using various statistical techniques (such as multiple regression, principal components analysis, clustering, Classification and Regression Trees) these random variables can in turn be combined to create *credit scores.*

Using credit scores to segment risks is hardly a new idea. For many years the lending industry has used such scores to underwrite loan applications. The Fair, Isaac Company is a leading vendor of one such score, called the FICO score.

Linking credit scores to personal auto and homeowners profitability, however, *was* a new idea, when they were introduced to the insurance industry approximately15 years ago. A typical credit score used in personal lines insurance might be calculated based on 10 to 30 variables. Conning's latest report [1] indicates that today more than 90% of insurance companies use credit scores or credit information in one way or another.

As noted above, the growing use of credit scores in insurance underwriting and ratemaking has garnered controversy along many fronts. We will set aside the political and social aspects of the debate and focus on the more purely actuarial issue: *do credit scores really help explain insurance profitability?* As we will discuss further, answering this question in the affirmative involves more than simply demonstrating a correlation between credit and loss ratio.

In the remainder of this paper, we will review the answers given to this question by the Tillinghast [4] and Monaghan [5] studies, and then add our own perspective. But first, it would be good to briefly discuss some general actuarial and statistical issues.

Background Actuarial and Statistical Considerations

Loss (Pure Premium) Relativity vs. Loss *Ratio* (Profitability) Relativity: The distinction between these concepts might not be clear to a non-actuarial audience, but it is absolutely critical. Because premium reflects all of the components of a rating plan, a correlation between a new variable (say, credit score) and loss ratio indicates the degree to which this variable can explain losses not already explained by the existing rating plan. For example, a critic might question the power of credit scores by claiming that credit is correlated with driver age. Since driver age is already in the class plan, there is no need to include credit as well. This argument would have some validity if it were in response to a pure premium relativity analysis. However, it would have much less validity if the relativity were based on loss ratios. Returning to the above example, the premium for youthful drivers is already 3 times that of mature drivers. Therefore a correlation between credit and loss ratio indicates the extent to which credit explains losses not already explained by the youthful driver surcharge.

Non-Independent Rating Variables: We believe that this is the key issue of the debate over the explanatory power of credit score. Intuitively, independence means that knowing the probability distribution of one variable tells you absolutely nothing about the other variable. Non-independence is common in insurance data. For example, youthful drivers have more accidents and violations than do mature drivers; mature drivers have more cars on their policies than do youthful drivers; number of drivers are correlated with number of vehicles. We can therefore expect that credit score will exhibit dependences with other insurance variables, such as driver age, gender, rating territory, auto symbol, and so on.

Univariate v. Multivariate Analyses: In the case of independent random variables, univariate analyses of each variable are entirely sufficient -- a multivariate analysis would add nothing in this case. Failure of independence, on the other hand, demands multivariate analysis. Furthermore, the results of multivariate analyses can be surprising. Below, we will give a hypothetical example in which an apparently strong relationship between credit and loss disappears entirely in a multivariate context.

Credibility vs. Homogeneity: paying attention to the credibility and homogeneity of one's data is important when we review any actuarial study and is essential in this debate for the usefulness of credit scores. Sparse data present the danger that one's model will fit noise rather than signal, leading to non-credible results. Non-homogenous data present the danger that extrapolating from one sub-population to another will lead to inaccurate predictions.

With these general remarks in hand, let us turn to the Tillinghast [4] and Monaghan [5] studies.

Tillinghast's Study

Tillinghast's credit study was undertaken on behalf of the Fair, Isaac Company for use in its discussions with the National Association of Insurance Commissioners (NAIC). The purpose of the study was to establish a relationship between Insurance Bureau credit scores with personal auto and homeowners insurance. Tillinghast received the following information for each of nine personal lines insurance companies:

- Credit score interval
- Interval midpoint
- Earned premium
- Loss ratio relativity

For the most part, the credit score intervals were constructed to contain roughly equal amounts of premium. The results for these 9 companies are given in Exhibit 1.

Clearly, the information provided to Tillinghast only allowed for a univariate study, and this is all Tillinghast set out to perform. Tillinghast's report displays tables containing each interval's loss ratio relativity alongside the interval's midpoint. These numbers are also displayed graphically. The report comments, "From simply viewing the graphs... it seems clear that higher loss ratio relativities are associated with lower Insurance Bureau Scores."

No detailed information is provided on the data used, or about the 9 companies that provided the data. Therefore we cannot comment on how credible the results are. The loss ratio relativity curves are somewhat bumpy for certain of the 9 companies; and the loss ratio spreads varies somewhat from company to company. But the patterns are clear enough to strongly suggests that the relativity spreads are robust, and not merely company-specific fluctuations in the data.

Furthermore, the relativities produced by credit are fairly large. The 10% of the companies' books with the best credit have anywhere from –20% to –40% loss ratio relativities. The worst 10% have relativities ranging from +30% to +75%. These loss ratio spreads compare favorably with those resulting from traditional rating variables. For example, based on our experience, about 20% to 30% of a standard auto book will have point surcharges for accidents or violations. The average surcharge might range from 15% to 40%. Therefore, the loss ratio spread indicated in the study is no less than the accident and violation point surcharge. In addition, the credit loss ratio spread can largely support the commonly seen rate differentiation for the tier rating. Examples such as this make it clear why insurers are embracing the use of credit scores.

In additional to displaying tabular/graphical evidence, Tillinghast computed regression slope parameters and their associated p-values. The p-values were all below 0.1, and often well below 0.05. (The p-value is defined as the probability of observing the actual slope parameter -- or a greater slope parameter -- given that the "true" slope parameter is zero.) The Tillinghast report concluded: "from the data and P-Values, we conclude that the indication of a relationship between Insurance Bureau Scores and loss ratio relativities is highly statistically significant."

Simpson's Paradox and the Perils of Univariate Analyses

This is reasonable as far as it goes. Unfortunately, univariate statistical studies such as Tillinghast's do not always tell the whole story. A statistical phenomenon knows as *Simpson's Paradox* [14,15] illustrates what can go wrong. A famous example of Simpson's Paradox is the 1973 study of possible gender bias in graduate school admissions at the University of California at Berkeley [16]. We will stylize the numbers for ease of presentation, but the point will remain the same.

Suppose it was reported 1100 men and 1100 women applied for admission to Berkeley in 1973. Of these people, 210 men were accepted for admission, while only 120 women were accepted. Based on this data, 19% of the men were accepted, while only 11% of the women were accepted. This is a univariate analysis (somewhat) analogous to Tillinghast's, and it seems to prove decisively that there was serious gender bias in Berkeley's 1973 graduate admissions.

But in fact this univariate analysis does not tell the whole story. When the admissions were broken down by division (suppose for simplicity that there were only two divisions: Arts & Sciences and Engineering) the data looked more like this:

	Applicants			# Accepted			% Accepted		
	Arts	**Eng.**	**Total**	**Arts**	**Eng.**	**Total**	**Arts**	**Eng.**	**Total**
Women	1000	100	1100	100	20	120	10%	20%	11%
Men	100	1000	1100	10	200	210	10%	20%	19%

Now our analysis is multivariate, by virtue of the fact that we are including *division applied to*, in addition to gender. The multivariate analysis quite clearly shows that the acceptance rate for men and women *within each division* was identical. But because a greater proportion of women applied to the division with the lower admission rate (Arts & Sciences), fewer women overall were accepted.

This is a very simple example of what can go wrong when one's data does not contain all relevant variables: *an apparent correlation between two variables can disappear when a third variable is introduced.*

In order to make the link to regression analysis, let us analyze this data at the un-grouped level. The reader can reproduce the following results with a simple spreadsheet exercise. Create 2200 data points with a {0,1}-valued target variable (ACCEPTED) and two {0,1}-valued predictive variables (MALE, ENGINEERING). 1000 of the points are males who applied to engineering {MALE=1, ENGINEERING=1}. For 200 of these points ACCEPTED=1, for the remaining 800 ACCEPTED=0, and so on.

If we regress ACCEPTED on MALE, we get the following results:

	Beta	**t-statistic**
Intercept	.1091	10.1953
MALE	.0818	5.0689

As expected, this univariate regression analysis indicates that gender is highly predictive of acceptance into graduate school, and indeed it is: a greater proportion of males *were* accepted! However this analysis is potentially misleading because it does not help *explain why* males are accepted at a higher rate.

When we regress ACCEPTED on MALE *and* ENGINEERING, we get quite different results:

	Beta	**t-statistic**
Intercept	.1	9.1485
MALE	0	0
ENGINEERING	.1	3.8112

When the truly relevant variable is introduced, the spurious association between gender and acceptance goes away (the beta and t-statistics for MALE are both 0). This multiple regression approach on un-grouped data is illustrative of our data mining work involving credit and other predictive variables.

(Of course logistic regression is usually a more appropriate way to model a binary target variable such as application acceptance or auto claim incidence. But such an analysis could not easily be replicated in a spreadsheet. Because ordinary multiple regression gives the same results in this simple case, it is sufficient for our illustrative purpose.

Readers are encouraged to try logistic regression, from which precisely the same conclusion will be reached.)

Returning to the Tillinghast study, consider the following scenario: suppose our credit variable has two levels (good/bad). Rather than academic division, suppose that the "true" confounding variable is urban/rural (territory). Thus good/bad correspond to male/female in the Berkeley example, and urban/rural corresponds to arts/engineering. Rather than acceptance into school, the target variable is now having a personal auto claim. Now our data is:

	Exposures			# Claims			Claim Freq		
	Rural	Urban	Total	Rural	Urban	Total	Rural	Urban	Total
Good credit	1000	100	1100	100	20	120	10%	20%	11%
Poor credit	100	1000	1100	10	200	210	10%	20%	19%

If we similarly re-label the terms of our regressions, we will again see that (in this purely hypothetical example) the GOOD_CREDIT indicator loses its apparent significance once the URBAN indicator is introduced.

These considerations make it clear that a multivariate analysis is needed to assess whether credit history bears a *true* relation with insurance loss experience. A univariate analysis might produce a statistical illusion, not true insight.

Of course, given our discussion of the difference between a pure premium study and a loss ratio study, it is not entirely fair to call the Tillinghast study "univariate". Recall that Tillinghast's target variable was *loss ratio relativity*, not claim frequency. In the above example, suppose all claims have a uniform size of $1000, and further suppose that the territorial rates are $2000 for urban territories, and $1000 for rural territories. Now the loss ratio relativity in each cell will be exactly 1.0. In this (again, purely hypothetical) case, Tillinghast's methodology would (correctly) show no relationship between credit and loss ratio relativity.

In other words, to the extent that all possible confounding variables are perfectly accounted for in premium, Tillinghast's "univariate" analysis *is* implicitly a multivariate analysis, and is therefore convincing. But realistically, this may not be the case. For example, in our work we regularly regress loss ratio on such zip code-based variables as population density and median population age. If territory were entirely accounted for in premium, such variables would never appear statistically significant. But in fact they sometimes do. Therefore a true multivariate study is desirable even if loss ratio is used as the target variable.

Monaghan's Study

James Monaghan's paper on "The Impact of Personal Credit History on Loss Performance in Personal Lines" is an advance over the Tillinghast study partly because he addresses the multivariate issue. Monaghan asks: if the correlation between credit and loss ratio exists, "is it merely a proxy, i.e., is the correlation actually due to other characteristics (which may already be underwritten for or against, or rated for)?" And, "are there dependencies between the impact of credit history on loss performance and other policyholder characteristics or rating variables?"

Monaghan's study for auto is based on three calendar years of data (1993-95). Each record in his database contains premiums and losses accumulated over this entire three-year period. So each record may have different length for the term. Losses are evaluated at 6/30/1995. For this reason, losses on different records might be evaluated at varying states of maturity. Losses include reserves, salvage and subrogation recoveries, and allocated loss adjustment expenses. The credit information used in this study was a "snapshot view" taken at the policy inception date. Approximately 170,000 records were used in the analysis. The total premium and loss in these records were $393 million and $300 million, respectively.

The amount of data in Monaghan's study is very large. While we don't know all the details about the data, the large amount of premium indicates that it is probably based on a countrywide population. Our experience on auto data indicates that on average there will be 150 to 400 claims per $1 million in premium, depending on the geographic concentration, program type, and policy type (liability only vs. full coverage) represented in the data. This suggests that there will be on the order of a hundred thousand claims in Monaghan's study. According to actuarial credibility theory [17], Monaghan's data should provide very credible results.

Monaghan discusses credit variables from a number of angles. First, he performs a number of univariate studies comparing *individual* credit variables (such as Amounts Past Due, Derogatory Public Records, Collection Records, Age of Oldest Trade, Number of Inquiries, Account Limits, and Balance-to-Limit Ratios) with fitted loss ratio relativity. In each case, there exists a positive correlation. This part of Monaghan's study is much like Tillinghast's study. The difference is that Monaghan analyses individual credit variables, whereas Tillinghast analyses a composite credit *score*.

While not conclusive for the reasons given above, this part of Monaghan's study is helpful in that it unpacks credit score into its component variables. The relationship between credit score is not entirely the result of some mysterious or proprietary interaction of the components credit variables. Rather, each of these component variables is individually somewhat predictive of insurance losses. For the record, the results Monaghan reports in this section are consistent with our experience working with credit data.

Note that these univariate results -- as well as Monaghan's multivariates to be described below -- are in terms of loss *ratio* relativity. Therefore, Monaghan's work (like the Tillinghast study) indicates the degree to which credit is able to capture loss variation not captured by the existing rating plan.

Next, Monaghan studies credit in conjunction with several traditional underwriting characteristics. Monaghan uses the above credit variables to profile policies into four groups, A, B, C, and D. For example, group A (the profile with the worst loss ratio) is characterized by one or more derogatory public records, high amounts past due, and so on. Group D (the profile with the lowest loss ratios) is characterized by long credit histories, low balance-to-limit ratios, and so on. Consistent with Monaghan's earlier results and Tillinghast's study, Monaghan shows that group A has a loss ratio relativity of 1.33; and group D has a relativity of 0.75.

Monaghan displays several two-way tables showing loss ratio relativity by credit group and an underwriting variable. The auto underwriting variables he displays in conjunction with credit include past driving record, driver age, territory, and classical underwriting profile. The last variable is a composite variable combining marital status, multicar, homeowner, and clean driving record. (Monaghan supplies similar tables for homeowners rating variables. We will not review the specifics of these tables here.)

In no case did Monaghan's inclusion of the rating factor cause the relationship of credit with loss ratio to disappear (as in the Simpson illustration above). Indeed, Monaghan's tables contain some very telling relationships. For example, the loss ratio relativity of drivers with clean driving histories and poor credit was 1.36. In contrast the relativity for drivers with good credit and poor driving records was only 0.70!

It is possible to reinforce Monaghan's conclusions by performing multivariate calculations on his data. Rather than use Bailey's iterative minimum bias equations, we performed equivalent Generalized Linear Model calculations using the PROC GENMOD facility in SAS. Recall [11,12] that the multiplicative Bailey model is equivalent to a GLM with the Poisson distribution and log link function; the additive Bailey model is equivalent to a GLM with the normal distribution and the identity link function. Note also that this latter model is simply a classical multiple regression model. Exhibits 2-4 contain GLM analyses of credit group by Driver Record, Driver Age, and Classical Underwriting Profile.

The results of the GLM analyses are striking, and they buttress Monaghan's claims. For example, the multiplicative Bailey factors arising from the credit/driving record analysis are 1.709, 1.339, 1.192, and 1.0 for credit groups A-D. These are quite close to the univariate loss ratios relativities that can be calculated from Monaghan's data (1.757, 1.362, 1.204, 1.0). This is excellent confirmation that credit is largely uncorrelated with driving record: the multiplicative Bailey factors are almost the same as the factors that would arise from a univariate analysis!

Furthermore, the GLM parameter estimates are quite large relative to their standard errors. Also, the Chi-squared statistics for the four credit groups are high, and the associated *p*-values are very low. These observations add statistical rigor to the claim that the loss ratio "lift" resulting from credit score is "real". These observations hold equally well for the other two variables as well. Finally, performing an additive Bailey analysis (normal/identity GLM – not shown) produces qualitatively similar results.

Monaghan reports that he produced such two-way tables for a large number of other traditional underwriting characteristics. He says, "there were no variables that produced even roughly uniform results across the credit characteristics."

Applying Data Mining Methodology to Credit Data

For several years, we have applied data mining methodology and a range of predictive modeling techniques to build insurance profitability and underwriting models for writers of both commercial and personal lines insurance. Credit variables and credit scores are typically included along with a comprehensive set of other traditional and non-traditional insurance variables. Because of the truly multivariate context in which we employ credit information, our findings lend further support to the conclusions reached in the Tillinghast and Monaghan studies. For reasons of confidentiality, we are not at liberty to share quantitative results in this paper. However, we shall describe our methodology and modeling results in a qualitative way.

We follow a standardized, disciplined methodology when embarking upon a data mining project. The first several steps involve studying internal and external data sources and generating predictive variables. Typical internal data sources include statistical records for premiums and losses, "snapshot" data for policyholder characteristics from legacy systems or a data warehouse, driver data, vehicle data, billing data, claims data, an agent database, and so on. Typically, several years of the company's relevant data sources will be utilized in the study. Commonly used external data sources include credit reports of the kind used by Monaghan, MVR (Moving Violation Records) data and CLUE (Claims Loss Underwriting Exchange) data. But other external data is available. For example, useful predictive variables at the zip-code level can be generated from data available from the US Census Department and the US Weather Bureau.

By the end of this process, literally *hundreds* of predictive variables will have been created from the internal and external data sources. The goal is to create upfront as many variables as possible that might be related to insurance loss and profitability. These variables represent a wide range of characteristics about each policyholder.

Typically we design our analysis files in such a way that each data record is at a policy-term level. For example, personal auto policies usually have a six-month term. If a policy has two years of experience in our study, we will generate four 6-month term data points in the study. This design, which is different from that of Monaghan's study, will give each record equal weight for the term in the analysis process. All of the predictive variables, including the credit variables, are evaluated as of the beginning of the term-effective date.

Target variables, including loss ratio, frequency, and severity, are created in parallel with the predictive variables. Losses are usually evaluated a fixed number of months from the term effective date. The reason for this is to minimize any chance of bias appearing in the target variables due to varying loss maturities. In addition, we will incorporate various actuarial techniques that we deem necessary to adjust the target information. Such adjustments include loss trending, premium on-leveling, re-rating, loss capping, cat loss exclusion, and so on.

Once the generation of target and predictive variables has been accomplished, we will merge all the information together to produce a policy-term level database. This database contains all of the predictive variables, as well as such target information as claim frequency, claim severity, loss ratio, capped loss ratio, and so forth. The database is then used to produce univariate reports showing the relationship of each predictive variable with the target information. This is essentially a collection of reports containing one Tillinghast-type study for each of the hundreds of predictive variables. This database is a useful exploratory data analysis (EDA) prelude to the multivariate modeling phase of our projects.

This database of univariate results also provides invaluable information for multivariate modeling regarding (1) whether to discard the variable right away because it has no/little distribution or because there is any business or other reason to do so; (2) how to cap the variable either above or below; (3) what to do with missing values; and (4) whether to treat the variable as a continuous or categorical random variable. Other needed transformations might be suggested by this univariate study.

Once the Exploratory Data Analysis stage is completed, we are ready to begin the modeling process. The first sub-phase of this process is to search for an optimal multiple regression model. Criteria used to judge "optimality" include (but are not limited to) strong t-statistics, parameter estimates that agree with business intuition, and not overfitting data used to estimate the parameters. This model serves as a useful benchmark for comparison purposes. In addition, the parameter estimates, and the t- and F-statistics generated by regression models are useful for such interpretive issues as the topic of this paper.

Once the optimal regression model has been selected, we turn to more advanced model building techniques such as Neural Networks [18-20], Generalized Linear Models [8-13], Classification and Regression Trees (CART) [21] and Multivariate Adaptive Regression Splines (MARS) [22]. These more advanced techniques can potentially provide more accurate predictions than a multiple regression model, but this additional predictive power often comes at a cost: more complex models can be harder to interpret and explain to underwriters, upper management, and insurance regulators.

We use a *train/test* methodology to build and evaluate models. This means that the modeling dataset is randomly divided into two samples, called the training and test samples. A number of models are fit on the training sample, and these models are used to "score" the test sample. The test sample therefore contains both the *actual* loss ratio (or any other target variable) as well as the *predicted* loss ratio, despite the fact that it was not used to fit the model. The policies in the test sample are then sorted by the score, and then broken into (for example) ten equal-sized pieces, called *deciles*. Loss ratio, frequency, and capped loss ratio are computed for each decile. These numbers constitute *lift curves*. A model with a low loss ratio for the "best" decile and a very high loss ratio for the "worst" decile is said to have "large lift". We believe that the lift curves are as meaningful for measuring the business value of models as such traditional statistical measures as mean absolute deviation or R^2. The purpose of setting aside a test set for model evaluation is to avoid "*overfit*". (Of course a lift curve can also be computed on the training dataset. Naturally, this lift will be unrealistically large.) A third sample, called a *validation* sample, sometimes will also be set aside to produce an unbiased estimate of the future performance of the final selected model.

We have performed several large data mining projects that included credit variables and credit scores. Similar to the Tillinghast study and Monaghan's study, we have studied data from various sources, different distribution channels, and different geographic concentrations. Our studies are very large in size, similar to Monaghan's study, usually with several hundred thousand data points that contain a total of hundreds of millions of dollars of premium. Our approach is tailored to the use of large datasets, the use of train/test methodology, the use of lift curves to evaluate models, and the exploratory use of a variety of modeling techniques. These are all hallmarks of the data mining approach to statistical problems. We believe that our analyses are true multivariate analyses that yield very robust and credible results. It is precisely this kind analysis that makes it possible to decisively answer the question: does credit *really* help explain insurance losses and profitability?

Our Findings: the Importance of Credit Variables in a Data Mining Context

First, through our univariate databases we note that composite credit score and many of its associated credit variables invariably show strong univariate relationships with frequency, severity, and loss ratio. Our univariate experience is entirely consistent with that of Tillinghast and Monaghan.

Turning to our multivariate modeling work, the estimates and statistics coming from our multiple regression models are useful for evaluating the importance of credit relative to the other variables considered in our model building process. Several points are worth making. First, credit variables consistently show up as among the most important variables *at each step* of the modeling process. As noted by Tillinghast and Monaghan, they dependably show strong univariate relationships with loss ratio. Furthermore, they are typically among the first variables to come out of a stepwise regression analysis.

Second, the parameter estimates for credit variables are consistently among the strongest of the parameters in our regression models. As illustrated in the Simpson's paradox example, credit score would have a small beta estimate and t-statistic were it a mere proxy for another variables or some combination of other variables. But this is not the case. Rather, we have repeatedly seen that credit adds predictive power *even in the presence of a comprehensive universe of traditional and non-traditional predictive variables, all used in conjunction with one another, on a large dataset.*

We are basing our conclusion in part on the t-statistics of the credit variables in our underwriting/pricing regression models. To this one might object: "but one of the assumptions of regression analysis is a normally distributed target variable. It is obvious that loss ratio is not normally distributed, therefore your t-statistics are meaningless." In response, it is true that loss ratios are not normally distributed. Nevertheless, the models we build using regression analysis reliably produce strong lift curves on test and validation data. Therefore, our models do "work" (in the sense of making useful predictions) in spite of the lack of normality.

It is also true that because of the lack of normality, we cannot use our models' t-statistics to set up traditional hypothesis tests. But neither our analyses nor our conclusions are based on hypothesis tests. We interpret t-statistics as measures of the *relative importance* of the variables in a model. Consider ranking the variables in a regression model by the absolute value of their t-statistics. The resulting order of the variables is the same as the order that would result from ranking the variables by their marginal contribution to the model's R^2 (in other words the additional R^2 that is produced by adding the variable after all of the other variables have been included in the model). This interpretation of t-statistics does *not* depend on the normality assumption.

To summarize, our reasoning is as follows:

- Our models effectively predict insurance losses. The evidence for this is repeated, unambiguous empirical observations: these models dependably distinguish profitable from unprofitable policies on out-of-sample data. In other words, they produce strong lift curves on test and validation datasets.

- Furthermore, credit variables are among the more important variables in these models. This is evidenced by the following observations: (i) the univariate relationship between credit and loss ratio is as strong or stronger than that of the other variables in the model. (ii) Credit variables reliably appear in a stepwise regression performed using all of the available variables. (iii) Credit variables typically have among the largest t-statistics of any of the variables in the model.

- Supporting the above observations, removing the credit variable(s) from a model generally results in a somewhat dampened lift curve.

- The implication of the above two bullets is that credit variables add measurable and non-redundant predictive power to the other variables in the model. Therefore, we believe that the observed correlation between credit and loss ratio cannot be explained away as a multivariate effect that would go away with the addition of other available variables.

Furthermore, this is true not just of the final selected regression model, but of most or all of the models produced along the way. In addition, we have noticed this result applies in all different lines of insurance, in both personal lines and commercial lines. For this reason, we feel comfortable saying that credit bears an unambiguous relationship to insurance loss, and is not a mere proxy for other *available* kinds of information.

But *Why* is Credit Related to Insurance Losses?

It is important to emphasize the word *available* because poor credit is obviously not in itself a *cause* of poor loss experience. In this sense, it is analogous to territory. Presumably credit is predictive because it reflects varying levels of "stress", planning and organization, and/or degrees of risk-taking that cannot be directly measured by insurers. These specific conjectures have been offered many times and they are intuitively plausible. However it is less conjectural to say that whatever credit might be a proxy for, it is not a proxy for any other variable (or combination of variables) practically available to insurers. In our data mining projects we explicitly set out to generate the most comprehensive universe of predictive variables possible. In this sense, we therefore use credit in the "ultimate" kind of multivariate analysis. Even in this truly multivariate setting, credit is indicated to have significant predictive power in our models.

It is beyond the scope of this paper to comment on the societal fairness of using credit for insurance pricing and underwriting. From a statistical and actuarial point of view, it seems to us that the matter is settled: credit *does* bear a real relationship to insurance losses.

Conversely: Can We Predict Insurance Losses without Credit? Can We Go beyond Credit?

Our experience does indicate that credit score is a powerful variable when it is used *alone* for a standard rating plan. In addition, our large-scale data mining results suggest that just about any model developed to predict insurance profitability will be somewhat stronger with credit than without credit. Typically credit score, when added to an existing set of non-credit predictive variables, will be associated with a relatively large beta estimate and t-statistic. Consistent with this, the resulting model will have higher "lift" than its counterpart without credit.

The results we have described might create an impression that credit variables are an essential part of any insurance predictive modeling project. But this would be an exaggeration. *Our experience also shows that pricing and underwriting models created without credit variables can still be extremely good.* The key to building a non-credit predictive model is to fully utilize as many available internal data sources as possible, incorporate other types of external information, use large amount of data, and apply multivariate modeling methodologies. Given all the regulatory and public policy issues surrounding insurers' use of credit, such non-credit models provide the insurance industry with a valuable alternative to using credit scores for pricing and underwriting.

Conclusion: Predicting the Future

Our data mining projects are multivariate predictive modeling projects that involve hundreds of variables being used to analyze many thousands of records. Many of these variables are credit variables, which play an important role even in this broad context. Our experience using credit scores and credit variables in a truly multivariate statistical setting has allowed us to add a new perspective to the debate over credit.

The use of credit in insurance underwriting and ratemaking might seem like a rather specialized topic. But we believe the issue reflects two important trends in the development of actuarial science. First, credit scores come from a non-traditional data source. The advent of the Internet makes it likely that other new data sources will become relevant to actuarial practice. Credit information is probably just the beginning.

The second issue is the increasingly multivariate nature of actuarial work. Credit scores themselves are inherently "multivariate" creatures in that they are composites built from several underlying credit variables. In addition, recall that we have reviewed and discussed three ways of studying the relationship between credit scores and insurance losses and profitability. Each study has been progressively more multivariate than its predecessor. This reflects the methodological development of classification ratemaking from univariate to multivariate statistical analyses (Generalized Linear Modeling).

In our opinion, the adoption of modern data mining and predictive modeling methodologies in actuarial practice is the next logical step in this development. Bailey's minimum bias method might seem like actuarial science's in-house answer to multivariate statistics. On the contrary, Mildenhall's paper makes it clear that conceptually, nothing separates minimum bias from work done by mainstream statisticians in any number of other contexts. But why stop at Generalized Linear Modeling?

We live in an information age. The availability of new data sources and cheap computing power, together with the recent innovations in predictive modeling techniques allow actuaries to analyze data in ways that were unimaginable a generation ago. To paraphrase a famous logician, actuaries inhabit "a paradise of data". This, together with our insurance savvy and inherently multivariate perspective, puts us in an excellent position to benefit from the data mining revolution.

Given the success of credit scores and predictive modeling, we expect actuaries to be enlisted to push this type of work even further. Here are examples of future questions we anticipate being asked of actuaries:

- Are we currently getting the most predictive power out of the internal and external information/data sources that we are currently using? Are we really analyzing data in a rigorous multivariate fashion?

- What other powerful variables and data sources are "out there" that we are not aware of? How do we go beyond credit?

- Are there other ways insurance companies (and indeed other kinds of companies) can leverage predictive modeling? For example, predictive modeling has a proven record of success in such applications as target marketing, customer retention/defection analysis, predicting cross-sales, customer profiling, and customer lifetime value. These are all important projects at which actuaries can excel. Furthermore, they are not insurance-specific. An actuary with expertise in these areas could transfer his or her skills to other industries.

To conclude, our multivariate predictive modeling work supports the widely held belief that credit scores help explain insurance losses, and that they go beyond other sources of information available to insurers. However it is unclear to what extent insurers will be permitted to used credit for future pricing and underwriting. For this reason insurers might want to consider non-credit scoring models as an alternative to traditional credit scores. For actuaries, the use of credit scores and predictive modeling is the beginning of a new era in insurance pricing and underwriting.

References

1. "Insurance Scoring in Personal Automobile Insurance – Breaking the Silence", *Conning Report,* Conning, (2001).

2. "Insurers Battling Credit-Scoring", *National Underwriter*, March 5th Issue, (2002).

3. "Insurers Lose a Credit Scoring Battle", *National Underwriter*, February 21st Issue, (2002).

4. "Credit Reports and Insurance Underwriting", *NAIC White Papers*, National Association of Insurance Commissioners, (1997).

5. Monaghan, J. E., "The Impact of Personal Credit History on Loss Performance in Personal Lines", *CAS Forum*, Casualty Actuarial Society, (2000).

6. Finger, R. J., "Risk Classification", *Foundations of Casualty Actuarial Science*, Chapter 5, Casualty Actuarial Society, (1990).

7. McClenahan, C.L., "Ratemaking", *Foundations of Casualty Actuarial Science*, Casualty Actuarial Society, Chapter 2, (1990).

8. Bailey, R. A., "Insurance Rates with Minimum Bias", *Proceedings of Casualty Actuarial Society,* Vol. L, Casualty Actuarial Society, (1960).

9. Bailey, R. A. and Simon, L. J., "Two Studies in Auto Insurance Ratemaking", *Proceedings of Casualty Actuarial Society,* Vol. XLVII, Casualty Actuarial Society, (1963).

10. Brown, R. L., "Minimum Bias with Generalized Linear Models", *Proceedings of Casualty Actuarial Society,* Vol. LXXV, Casualty Actuarial Society, (1988).

11. Mildenhall, S.J., "A Systematic Relationship Between Minimum Bias and Generalized Linear Models," *Proceedings of Casualty Actuarial Society,* Vol. LXXXVI, Casualty Actuarial Society, (1999).

12. Holler, K.D., Sommer, D.B.; and Trahair, G., "Something Old, Something New in Classification Ratemaking with a Novel Use of GLMS for Credit Insurance," *CAS Forum*, Casualty Actuarial Society, (1999).

13. Brockman, M. J., Wright, T. S., "Statistical Motor Rating: Making Effective Use of Your Data," *Journal of the Institute of Actuaries*, Vol. 119, Part III.

14. Simpson, E. H., "The Interpretation of Interaction in Contingency Tables," *Journal of the Royal Statistical Society, Series B*, 13: 238-241, (1951).

15. Pearson, K., Lee, A., Bramley-Moore, L., "Genetic (Reproductive) Selection: Inheritance of Fertility in Man", *Philosophical Transactions of the Royal Society A*, 73: 534-539, (1899).

16. Bickel, P. J., Hamel, P.A., O'Connell, J.W., "Sex Bias in Graduate Admissions: Data from Berkeley", Science, (1975).

17. Mahler, H.C., and Dean, C.G., "Credibility", *Foundations of Casualty Actuarial Science,* Casualty Actuarial Society, Chapter 8, (2001).

18. Francis, L., "Neural Networks Demystified," *CAS Forum*, Casualty Actuarial Society, (2001)

19. Wu, C. P., "Artificial Neural Networks – the Next Generation of Regression" *Actuarial Review,* Casualty Actuarial Society, Vol. 22, No. 4, (1995).

20. Zizzamia, F., Wu, C. P., "Driver by Data: Making Sense Out of Neural Networks", *Contingencies*, American Academy of Actuaries, May/June, (1998).

21. Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J., "Classification and Regression Trees", Monterrey: Waldsworth and Brooks/Cole, (1984).

22. Friedman, J. H., "Multivariate Adaptive Regression Splines," *Annals of Statistics*, Vol. 19, (1991).

Exhibit 1
Tillinghast -NAIC Study of Credit Score [4]

Company 1
Scores & Loss Ratio Relativity Summary

Score Interval	Midpoint	Earned Premium	Loss Ratio Relativity
813 or more	850.0	10.2%	0.657
768-812	790.0	9.9%	0.584
732-767	749.5	11.0%	0.692
701-731	716.0	10.9%	0.683
675-700	687.5	10.4%	1.184
651-674	662.5	9.8%	0.793
626-650	638.0	9.9%	1.332
601-625	613.0	10.0%	1.280
560-600	580.0	9.4%	1.214
559 or less	525.0	8.6%	1.752

Company 2
Scores & Loss Ratio Relativity Summary

Score Interval	Midpoint	Earned Premium	Loss Ratio Relativity
840 or more	854.0	10.0%	0.607
823-839	831.0	10.0%	0.813
806-822	814.0	10.0%	0.626
789-805	797.0	10.0%	1.342
771-788	779.5	10.0%	1.059
748-770	759.0	10.0%	1.019
721-747	734.0	10.0%	1.322
686-720	703.0	10.0%	0.810
635-685	660.0	10.0%	0.986
635 or less	592.0	9.9%	1.417

Company 3
Scores & Loss Ratio Relativity Summary

Score Interval	Midpoint	Earned Premium	Loss Ratio Relativity
826 or more	845.0	10.0%	0.723
803-826	814.5	10.0%	0.903
782-803	792.5	10.0%	0.895
759-782	770.5	10.0%	0.795
737-759	748.0	10.0%	1.073
710-737	723.5	10.0%	0.941
680-710	695.0	10.0%	0.912
640-680	660.0	10.0%	1.115
583-640	611.5	10.0%	1.221
583 or less	535.0	10.0%	1.421

Company 4
Scores & Loss Ratio Relativity Summary

Score Interval	Midpoint	Earned Premium	Loss Ratio Relativity
832 or more	859.0	10.0%	0.672
803-832	817.5	10.0%	1.027
767-803	785.0	10.0%	0.823
739-767	753.0	10.0%	1.036
720-739	729.5	10.0%	0.775
691-720	705.5	10.0%	1.000
668-691	679.5	10.0%	1.041
637-668	652.5	10.0%	1.023
602-637	619.5	10.0%	1.251
602 or less	571.0	10.0%	0.135

Company 5
Scores & Loss Ratio Relativity Summary

Score Interval	Midpoint	Earned Premium	Loss Ratio Relativity
845 or more	857.0	10.0%	0.800
830-845	837.5	10.0%	0.919
814-830	822.0	10.0%	0.740
798-814	806.0	10.0%	0.733
779-798	788.5	10.0%	0.855
757-779	768.0	10.0%	0.889
730-757	743.5	10.0%	0.993
695-730	712.5	10.0%	1.143
643-695	669.0	10.0%	1.300
643 or less	600.0	10.0%	1.628

Company 6
Scores & Loss Ratio Relativity Summary

Score Interval	Midpoint	Earned Premium	Loss Ratio Relativity
810 and up	837.5	19.7%	0.656
765-809	777.0	20.1%	0.795
715-764	739.5	20.8%	0.911
645-714	679.5	20.2%	1.066
Below 645	600.0	19.2%	1.593

Company 7
Scores & Loss Ratio Relativity Summary

Score Interval	Midpoint	Earned Premium	Loss Ratio Relativity
750 and up	795.0	21.3%	0.783
685-749	717.0	25.8%	0.900
630-684	657.0	19.6%	1.083
560-629	594.5	19.3%	1.150
Below 560	520.0	13.9%	1.200

Company 8
Scores & Loss Ratio Relativity Summary

Score Interval	Midpoint	Earned Premium	Loss Ratio Relativity
755 or more	775.0	8.9%	0.767
732-754	743.0	9.3%	0.798
714-731	722.5	9.6%	0.859
698-713	705.5	9.9%	0.969
682-697	689.5	10.3%	0.922
666-681	673.5	9.7%	0.978
647-665	656.0	10.5%	1.070
625-646	635.5	10.2%	1.107
592-624	608.0	10.7%	1.122
591 or less	562.0	10.8%	1.324

Company 9
Scores & Loss Ratio Relativity Summary

Score Interval	Midpoint	Earned Premium	Loss Ratio Relativity
780 and up	815.0	16.8%	0.637
745-779	762.0	13.7%	0.715
710-744	727.0	13.9%	0.734
670-709	689.5	15.0%	0.807
635-669	652.0	12.1%	0.909
590-634	612.0	11.2%	1.241
530-589	559.5	9.8%	1.357
Below 530	495.0	7.5%	2.533

Exhibit 2
Bailey Analysis of Monaghan's Two-Way Study
Credit Score vs. Driving Record

	Credit Group A		Credit Group B		Credit Group C		Credit Group D		Overall			**Bailey**
Prior Driving Record	Prem	LR	Prem	LR	Prem	LR	Prem	LR	Prem	LR	LR Rel	**Factor**
No incidents	28.4	93%	66.0	71%	30.70	64%	45.80	53%	170.90	68.6%	1.000	**1.000**
1 minor	8.0	94%	17.3	68%	7.50	68%	8.40	50%	41.20	69.4%	1.012	**0.987**
1 at-fault accident	3.7	101%	7.7	74%	4.10	68%	5.90	65%	21.40	75.0%	1.094	**1.096**
1 non-fault accident	6.6	109%	14.8	81%	7.30	70%	9.90	70%	38.60	80.9%	1.180	**1.176**
2 minors	2.5	86%	6.0	59%	1.90	41%	2.40	43%	12.80	58.6%	0.855	**0.827**
2 incidents (any)	6.5	108%	13.5	96%	6.60	82%	7.90	64%	34.50	88.3%	1.287	**1.268**
All other (> 2 incid.)	18.6	114%	33.7	95%	10.80	83%	11.50	66%	74.60	93.5%	1.364	**1.289**
Overall	74.3	101%	159	79%	68.9	69%	91.8	58%				
LR Rel		1.757		1.362		1.204		1.000				
Bailey Factor		**1.709**		**1.339**		**1.192**		**1.000**				

Generalized Linear Model Details

		estimate	exp* estimate	Bailey factor	s.e.	Chi Squared	*p* -value
Credit Group	A	0.1631	1.177	1.709	0.026	40.60	0.000
	B	-0.0807	0.922	1.339	0.023	12.13	0.001
	C	-0.1970	0.821	1.192	0.031	40.37	0.000
	D	-0.3727	0.689	1.000	0.031	148.23	0.000
Prior Driving Record	No incidents	-0.2540	0.776	1.000	0.026	96.80	0.000
	1 minor	-0.2667	0.766	0.987	0.038	50.12	0.000
	1 at-fault accident	-0.1624	0.850	1.096	0.047	11.93	0.001
	1 non-fault accident	-0.0922	0.912	1.176	0.037	6.34	0.012
	2 minors	-0.4438	0.642	0.827	0.065	46.44	0.000
	2 incidents (any)	-0.0169	0.983	1.268	0.037	0.21	0.647
	All other (> 2 incid.)	0	1.000	1.289	0	--	--

* Because the log link function was used, the GLM parameter estimate must be exponentiated

Exhibit 3
Bailey Analysis of Monaghan's Two-Way Study
Credit Score vs. Driver Age

	Credit Group A		Credit Group B		Credit Group C		Credit Group D		Overall			**Bailey**
Age of Driver	Prem	LR	Prem	LR	Prem	LR	Prem	LR	Prem	LR	LR Rel	**Factor**
<25	3.8	121%	23.6	75%	1.40	51%	1.90	53%	30.70	78.2%	1.000	**1.000**
25-34	21.1	103%	55.8	79%	22.60	66%	8.90	63%	108.40	79.6%	1.018	**1.023**
35-39	13.0	100%	21.8	81%	12.90	65%	13.00	54%	60.70	75.9%	0.970	**1.007**
40-44	12.4	109%	18.5	82%	10.40	76%	15.60	52%	56.90	78.6%	1.004	**1.055**
45-49	9.8	93%	14.6	83%	8.20	76%	14.80	58%	47.40	76.1%	0.972	**1.036**
50-59	9.2	97%	14.4	78%	7.90	68%	16.50	53%	48.00	71.4%	0.913	**0.985**
60+	3.8	110%	8.3	75%	4.90	81%	20.00	67%	37.00	75.1%	0.959	**1.129**
Overall	73.1	103%	157	79%	68.3	70%	90.7	58%				
LR Rel		1.775		1.367		1.202		1.000				
Bailey Factor		**1.805**		**1.394**		**1.220**		**1.000**				

Generalized Linear Model Details

		estimate	exp* estimate	Bailey factor	s.e.	Chi Squared	*p* -value
Credit Group	A	0.1214	1.129	1.805	0.052	5.44	0.020
	B	-0.1372	0.872	1.394	0.050	7.66	0.006
	C	-0.2707	0.763	1.220	0.055	24.11	0.000
	D	-0.4692	0.626	1.000	0.048	94.19	0.000
Age	<25	-0.1214	0.886	1.000	0.067	3.30	0.069
	25-34	-0.0985	0.906	1.023	0.053	3.52	0.061
	35-39	-0.1146	0.892	1.007	0.057	4.10	0.043
	40-44	-0.0674	0.935	1.055	0.057	1.42	0.234
	45-49	-0.0865	0.917	1.036	0.059	2.15	0.142
	50-59	-0.1366	0.872	0.985	0.059	5.28	0.022
	60+	0	1.000	1.129	0	--	--

* Because the log link function was used, the GLM parameter estimate must be exponentiated

Exhibit 4
Bailey Analysis of Monaghan's Two-Way Study
Credit Score vs. Classical Underwriting Profile

	Credit Group A		Credit Group B		Credit Group C		Credit Group D			Overall		**Bailey**
Prior Driving Record	Prem	LR	Prem	LR	Prem	LR	Prem	LR	Prem	LR	LR Rel	**Factor**
MMH, Clean**	10.2	97%	22.3	77%	14.50	76%	20.20	57%	67.20	73.8%	1.000	**1.000**
MMH, Other	10.6	102%	20.2	85%	13.50	76%	16.00	58%	60.30	78.8%	1.068	**1.051**
not MMH, Clean	27.8	92%	62.9	69%	24.40	58%	34.40	50%	149.50	67.1%	0.909	**0.877**
not MMH, Other	25.6	113%	53.4	88%	16.70	74%	21.20	70%	116.90	88.2%	1.195	**1.125**
Overall	74.2	101%	158.8	79%	69.1	69%	91.8	58%				
LR Rel		1.761		1.365		1.202		1.000				
Bailey Factor		**1.739**		**1.354**		**1.196**		**1.000**				

Generalized Linear Model Details

		estimate	exp* estimate	Bailey factor	s.e.	Chi Squared	*p*-value
Credit Group	A	0.1268	1.135	1.739	0.028	20.56	0.000
	B	-0.1237	0.884	1.354	0.024	26.83	0.001
	C	-0.2479	0.780	1.196	0.035	51.46	0.000
	D	-0.4265	0.653	1.000	0.034	162.37	0.000
Prior Driving Record	MMH, Clean	-0.1175	0.889	1.000	0.035	11.02	0.000
	MMH, Other	-0.0679	0.934	1.051	0.036	3.59	0.000
	not MMH, Clean	-0.2485	0.780	0.877	0.029	75.81	0.001
	not MMH, Other	0	1.000	1.125	0.000	0.00	0.012

* Because the log link function was used, the GLM parameter estimate must be exponentiated

**MMH = Married Multicar Homeowner
Clean = Clean driving record

Credit & Surety Pricing and the Effects of Financial Market Convergence

Athula Alwis, ACAS, and
Christopher M. Steinbach, FCAS, MAAA

Credit & Surety Pricing and the Effects of Financial Market Convergence

By: Athula Alwis, ACAS, American Re and Chris Steinbach, FCAS, Swiss Re

Abstract:

This paper describes how the convergence of the insurance and financial markets is affecting Credit & Surety insurance. It explains why prior experience has become an unreliable measure of exposure and how this paradigm shift affects the pricing of Credit & Surety products. It proposes a new exposure based method for analyzing Credit & Surety that combines the best practices of insurance and financial market pricing theory. Discussions about its implementation as well as sample calculations for, both primary and reinsurance pricing are included. This paper also discusses the new breed of Commercial Surety bonds that have been recently developed to compete with traditional financial products. Finally, the paper addresses the need for better and more sophisticated risk management techniques for the industry.

1 Introduction

There is a revolution occurring in Credit & Surety. The convergence of the insurance and financial markets is resulting in dramatic changes to these insurance products. There has been an explosion of new forms and some new coverages as insurers attempt to compete with financial institutions. There is new creativity to coverage structures as insurers rethink traditional practices and their applicability in today's environment. Increased competition by financial institutions for business that was traditionally considered insurance is the end result. All of these changes present new opportunities and new risks for the industry. The final outcome must be a revolution in our practices, which affects the actuarial profession in two ways.

First, as our products become increasingly sophisticated, our risk management practices must keep pace. We cannot rely on naïve diversification as much as we have in the past. This became apparent in the past year as unprecedented credit events educated us as to the true nature of our exposures and the weaknesses of our current risk management systems. Most Credit & Surety insurers have since made a concerted effort to improve their credit risk management systems to suit the new environment.

Second, convergence has resulted in competition between the insurance and financial industries, creating arbitrage opportunities between insurance and financial markets pricing theories. Insurance and financial markets pricing theories are very different and can produce completely different results for the same risk. Recent experience has shown that insurers, more often than not, are the losers when arbitrage occurs. Many insurers have witnessed entire segments of their portfolio perform poorly, particularly with regards to their new products. This has caused some Credit & Surety insurers to reconsider what they write and how, and for others to reconsider whether they want to be in this business at all.

The challenges that actuaries currently face in both risk evaluation and risk management are problems that the financial markets have already conquered. So, financial markets theory is the natural place for actuaries to turn for solutions. Over the past few years, financial markets theory has been finding its way into Credit & Surety insurers and reinsurers alike. This paper describes the financial market theories that can be applied to Credit and Surety, the benefits they bring, and how they could be implemented.

2 History & Current Events

Surety is unique in the insurance industry in that it is the only three-party insurance instrument. It is a performance obligation, meaning it is a joint undertaking between the principal and the surety to fulfill the performance of a contractual obligation. The principal is primarily responsible for the obligation and the surety guarantees fulfillment. If the principal fails to fulfil the obligation, then the surety steps into the shoes of the principal to complete the obligation. Surety obligations are divided into two general categories. Contract Surety guarantees the completion of a construction project, such as a road or building. Contract Surety is the largest segment of the Surety market because all government construction must be bonded. All other Surety products are called Commercial or Miscellaneous Surety. This covers a wide assortment of obligations, such as Bail bonds, the delivery of natural gas paid for in advance, the environmental reclamation of a strip mine, or the proper administration of a self-insured Worker's Compensation program. This is a smaller, but rapidly growing, segment of the Surety market.

Credit Insurance is a demand obligation, meaning it indemnifies the insured for un-collectable receivables if there is default. It is commonly used in retail, since many stores do not pay for the merchandise on their shelves until they themselves sell it to consumers. Another common use is in shipping, since merchandise is typically not paid for until it is delivered and inspected. Note that the majority of the Credit Insurance market is outside the United States. However, the US credit insurance market is growing rapidly. Inside the United States, companies typically use banking products such as loans and letters of credit instead of credit insurance.

Financial Guarantee Insurance is a demand obligation that consists of two distinct categories. The first involves policies that insure against defaults of financial obligations, such as insurance guaranteeing payment of the principal and interest of a municipal bond. The second involves insurance against certain fluctuations of the financial markets, such as insurance ensuring a minimum performance for an investment. Note that in many states, this second category is not permissible because it lacks a valid insurable interest. Financial Guarantee insurance was regulated in the 1980s because of New York State's concerns that the line's rising popularity and enormous policy limits could result in insurer defaults that would swamp the state's insolvency fund. So, New York regulated the line, requiring that Financial Guarantee writers be well-capitalized mono-line insurers that are not eligible for insolvency fund protection. Today, New York's strict regulations effectively control how Financial Guarantees are written. However, the other states and the rating agencies also work to exert their influence over the line, significantly complicating the regulatory landscape.

Closely related to Financial Guarantees is a collection of minor lines that often are treated separately. These include Residual Value, Mortgage Guarantee, Credit Unemployment, Student Loan Guarantees and many life insurance schemes. These are often considered separate from Financial Guarantee simply because they were already regulated when the Financial Guarantee regulations were written. However, it pays to do research when working in these lines because Financial Guarantee regulations are still evolving and different states have different opinions.

Credit Derivatives are financial instruments that pay when default occurs, whether or not the default results in a loss. Credit Derivatives are financial products, and as such do not require that a valid insurable interest exist. The most common Credit Derivative pays out on default the notional amount of a bond in exchange for receipt of the actual bond, so the loss is the difference between the notional of the bond and the market value of the underlying security. Credit Derivatives can be quite complex. They do not require underlying securities, so they are ideal hedging instruments for credit insurance risk. They also can be constructed to have additional triggers, such as a rise in the price of gas or a fluctuation in currency rates.

Traditionally, Surety Bonds, Credit Insurance, Financial Guarantees and Credit Derivatives have been distinct products. In the past few years, the boundaries between these products have blurred considerably. They are now part of a continuous spectrum of products that insure financially related obligations. They start with Surety, where the insurer is entitled to be very active in managing the insured risk, and end with Credit Derivatives, where the insurer is entirely passive. This blurring has enabled products from different financial sectors to compete with each other. It also permits insurers to tailor make products of varying insurer supervision, fiduciary duty, and regulatory control.

The biggest development in the past few years has been the explosion of the Commercial Surety market. Commercial Surety products traditionally have been relatively simple bonds with modest limits, such as Bail bonds and License & Permit bonds. But, recently they have evolved into sophisticated financial products with complex triggers and limits of hundreds of millions of dollars. One area that has generated increasing activity in recent years has been the use of Commercial Surety to mimic other types of financial instruments. In many cases, the Commercial Surety obligations tread very close to Financial Guarantees as defined by the New York State Department of Insurance. For this reason, the insurance industry has begun to call them "Synthetic Financial Guarantees."

The simplest financial market application of Commercial Surety involves using bonds to secretly credit enhance financial products. For example, banks frequently are involved in the short-term leasing industry. Banks typically securitize their lease portfolios, paying hefty rates if the portfolio contains many poor or mediocre credits. But, in this scenario, the bank requires the lessors to purchase Surety bonds that guarantee that all lease payments will be made. This enhances the credit quality of the portfolio, dramatically reducing the risk of loss. The bank now pays significantly lower rates for the securitization. This ultimately is cheaper for the lessees because insurers charge less than the capital markets to assume this risk, allegedly because insurers are able to wield influence over the risk. The insurance products would appear to be standard Lease Bonds, except for the fact that the ulterior motive is to credit enhance a financial instrument.

The applications can get significantly more complicated. But, part of the additional complication is due to the fact that knowledge of the intricacies of insurance law is very important. Whether the policy is enforceable in the manner for which it is intended depends on much more than just the wording of the policy. For example, suppose Eastern Power Company sells one year of electricity to Western Power Company for $100mm, to be paid in advance on January 1, 2003. Simultaneously, Western Power sells the same one year of electricity to Eastern Power for $105mm, to be paid in arrears on January 1, 2004. The two contracts cancel each other out, resulting in no effect other than the difference in payment terms. Western Power also purchases a Surety bond that guarantees that Eastern Power will pay the $105mm they owe. What this deal effectively reduces to is that Western Power has loaned Eastern Power $100mm at 5% and convinced an insurer to take the default risk. Here, a lack of full disclosure to the insurer would be very material in the advent of a loss. Surety is a performance obligation and the Surety could argue that the performance of the underlying obligation was never intended, so they owe nothing.

Commercial Surety has dramatically grown in popularity. Commercial Surety is now being used to secure letters of credit, to secure bank lines, and to enhance credit. Favorable historic loss ratios and limited abilities to grow other parts of the book has created the incentive for most major Surety writers to grow their Commercial Surety books. Furthermore, clients have flocked to Commercial Surety because they offer straightforward financial protection in a favorable regulatory environment. The ability of the market to arbitrage the various rating methodologies has also been a key factor.

The ability to use Commercial Surety for arbitrage purposes has revealed striking differences in how the various markets price their products. For example, it is not unusual for insurers to see Commercial Surety bonds that sell for a quarter of what Credit Derivatives sell for, with nearly identical terms. The problem that causes this discrepancy is that insurers generally do not differentiate risks as well as the financial markets do. Insurance pricing focuses primarily on making sure that the overall rate is adequate while financial market pricing focuses more on risk differentiation. This difference can be best demonstrated via the data each industry uses for establishing rate relativities. Insurance rate relativities are generally based on the company's own limited experience while financial market rate relativities are based on long periods of rating agency (industry) data. The larger data set enables the financial industry to calculate relativities that have a greater resolution than what the insurance industry calculates, creating the arbitrage opportunity. The result is the fact that insurers typically overprice short term / good credits and underprice long term / poor credits, compared to the financial markets.

Recent events, particularly those involving the use of Commercial Surety bonds to mimic Financial Guarantees, have gone a long way toward dulling the popularity of these new products for both insurer and insured alike. Several insurance companies are currently addressing severe anti-selection problems in their portfolio. Several others are in court dealing with the fact that insurers and financial institutions have different customs and practices for their products – a fact that has significantly confused their customers. Insurers have reacted to these problems by either pulling out of the Credit & Surety market entirely or by significantly curtailing their Commercial Surety writings. But, these problems have not killed the demand for Commercial Surety bonds, and the reduced supply probably will not last long.

3 Credit Risk Management

Credit risk management practices of the financial markets have always been more advanced than that of the insurance industry. Until recently, insurance credit risk management has been largely limited to purchasing reinsurance and managing the book to a targeted loss ratio. In contrast, credit risk management in the financial markets is a wide collection of tools. The similarity of Credit & Surety insurance to financial markets products permits the insurance industry to borrow financial market's risk management techniques. Several techniques transfer particularly well.

The first requirement is to make sure that the portfolio is not excessively exposed to any single credit event. The most cost-effective way to do this is by implementing concentration limits by counter party, industry sector and country. These concentration limits must take into consideration the quality of the credit risks. Since poorer credits have higher frequencies and severities during economic downturns, it is important to have the concentration limits be lower for poorer quality risks. In this way it is possible to keep the expected loss for an event relatively constant throughout the portfolio. It is also important to take correlation into consideration when establishing concentration limits. Setting all of the concentration limits lower than an independence analysis would suggest, or establishing a tiered system of concentration limits can achieve this.

Since the portfolio and the economy are both always changing, it is also important to have mechanisms for repositioning the portfolio over time. Three methods are commonly used. First, it is important to manage bond durations by counter party, industry sector and country. An insurer who manages their durations well can progressively reduce their exposure to a deteriorating segment of the business by not writing new bonds. This is known as an "orderly exit." Second, covenants can be placed in the contracts that require the insureds to post collateral if certain thresholds are breached. These thresholds can be established to generally coincide with the deterioration of that part of the portfolio, keeping the total exposure under the concentration limits and withdrawing the exposure while the risks are still solvent. However, covenants are losing their effectiveness because they are becoming too popular of a solution, contributing to the trend of marginally solvent companies crashing dramatically into bankruptcy. Third, the partial derivatives of the expected loss relative to changes in various economic indices measure the sensitivity of the portfolio to macroeconomic events. In financial market risk management theory, the partial derivatives are known as the "Greeks" because a particular Greek letter typically represents each distinct partial derivative. Analysis of the Greeks assists insurers in managing their risk to macroeconomic events, such as a rise in interest rates.

Insurers are also getting better at actively managing their credit risk profile with reinsurance, retrocessions, credit default swaps, and other financial instruments. This is a powerful technique for managing the portfolio because it is able to change the risk profile after the fact. But, it is more difficult to implement in practice than it appears. Reinsurance is becoming increasingly expensive and credit markets are not that liquid. Thus, the required credit protection is often not affordable or available. This is because the names that have exhausted the credit capacity of the insurance company have also generally exhausted the credit capacity of the other credit markets as well. Furthermore, the insurer always runs the risk of having the reinsurance/swap cost more than the premium the insurer collected. For this reason it is important that the pricing of the insurance product specifically incorporate the cost of any risk transfer or hedging activity. Also, those using hedges must note that they are often inefficient. The trigger for the insurance policy and the trigger for the hedge are usually slightly different. This inefficiency must be factored into the hedge and the premium charged for the insurance product if hedges are being used. Finally, reinsurance and hedge transactions are usually conducted with reinsurance & financial institutions. This is a problem because most reinsurance & financial institutions are themselves peak credit risks in the insurance portfolio. The insurance company must also manage what would happen if the counter party goes bankrupt, causing the reinsurance or hedge to fail.

While substantial similarities between insurance companies and financial institutions enable the insurance industry to borrow liberally from financial risk management theory, there is also an important difference to note. For financial institutions, credit risk is highly correlated and dominates the portfolio. Credit risk management practices focus on fencing in this risk. For insurance companies, credit risk diversifies with the other risks that the insurer writes, such as CAT (Catastrophe) covers. Insurance credit risk management can take advantage of this diversification.

Credit risk management in financial institutions focuses on fencing in the potential damage from highly correlated losses. Periodically, defaults occur in a highly correlated way and this is known as the credit cycle. The cycle begins when bad economic news causes a large amount of money to flee the credit markets, resulting in the cost of credit to suddenly and dramatically increase. This causes many companies that were only barely surviving to fail simultaneously. Their failure in turn adds additional financial stress to their creditors, customers and suppliers, causing more failures. The failures ripple through the market, taking out many of the financially weak and some of the financially strong. Credit cycles often center about the specific countries and industry sectors that generated the initial bad economic news. Financial risk management is heavily focused on quantifying the amount of loss the institution is potentially exposed to when credit becomes scarce, causing counter-parties to go bankrupt. It is managed in a manner very similar to the way insurers manage earthquakes.

When insurers manage the credit cycle, they have the added luxury that the credit cycle and the underwriting cycle are natural hedges; that is, they anti-correlate. Both are driven by the availability of capital. When capital is scarce and credit dries up, counter parties go bankrupt and financial markets suffer catastrophic losses. However, when capital is scarce and capacity dries up, insurance premiums rise and insurance markets are at their most profitable. The opposite relationship holds when capital is plentiful. As a result, it is possible for insurers to implement a risk management strategy that integrates credit-related products, other insurance products and the investment portfolio results. This strategy aims to immunize the portfolio by balancing the effect of the credit and underwriting cycles. Currently this idea is more theory than practice, although several companies are implementing risk aggregation models that would permit them to implement such a risk management strategy. These models are effectively detailed DFA (Dynamic Financial Analysis) models of the corporation and all of its parts.

It is often thought that the only goal of risk management is in making sure that the company survives to see tomorrow. But, an equally important goal is to be able to determine which products add value. Traditionally, "value" has been measured via profit & loss reports. But this approach is really only able to identify which products are unprofitable or under-performing relative to historic norms. It is not able to reliably identify the products that create a drag on the stock price. In order to identify these products, a system that measures a product's contribution to the ROE (Return on Equity) is required. Several companies are experimenting with such a measurement system.

The reason most insurance companies today are not able to identify the products that decrease their ROE is because most companies do not have the risk measurement systems capable of quantifying the amount of capital each risk requires. The question becomes particularly complicated for Credit & Surety products since these have many risk characteristics that the other lines can often downplay, such as correlation and hedging activities. The most immediate obstacle is the fact that some companies do not yet capture the information required for such an analysis. The new risk management techniques are extremely data intensive, and sophisticated inventory systems are required to implement them.

The risk profile for Credit & Surety is extremely complicated. It is easy for the insurer to accidentally take on an unacceptable amount of risk, requiring the utilization of an unacceptable amount of capital. It is also easy to make simple mistakes, such as paying more for reinsurance than the amount you collected, quickly dooming the insurer to certain loss. Measuring profitability requires understanding how much capital the risk requires relative to the profit that the risk generates and including the effects of all of the reinsurance and hedging purchased. This comparison can only be done within the framework of an advanced risk management system.

4. Pricing

4.1 Traditional Approach to Pricing

Credit & Surety has always been viewed as a form of property insurance because it shares the defining characteristics of property business. Most important is the fact that the severity distribution is relative to the limit of insurance. But, it also shares numerous other characteristics: there is a very wide variation in the limits commonly purchased (several orders of magnitude), Credit & Surety is subject to large shock losses, as well as catastrophes (known as the "loss cycle").

However, Credit & Surety does have distinguishing characteristics of its own. First is the fact that the loss cycle (the catastrophe) is not random, but appears at eight to twelve year intervals. This means that the way the loss cycle is incorporated into the pricing is different than regular property catastrophe pricing. Second is the fact that Credit & Surety underwriting requires more judgment than other types of property underwriting because insurers understand the causes of fire much better than the causes of insolvency. As a result, there can be enormous variations in experience from one insurer to another as underwriting practices differ. Third is the fact that Credit & Surety is "underwritten to a zero loss ratio." This does not mean that insurers have years without losses. What it implies is that the goal of Credit & Surety is to actively monitor the risks and to proactively respond to problems in order to prevent losses from happening. As a result, Credit & Surety underwriting focuses primarily on reducing loss frequency (i.e. default risk). Therefore, when the loss experience of two primary companies differs, most of the difference is in their frequencies.

Because of the similarities between Credit & Surety and property, most actuaries approach Credit & Surety pricing in the same manner as the other property lines. Both experience rating and exposure rating methods are commonly used. The benefit of having two different methodologies is that when the assumptions underlying one methodology fail, then the other methodology can generally be relied upon. But for Credit & Surety, the assumptions underlying both methodologies are equally questionable. Thus, Credit & Surety pricing has always been somewhat of an art form. The fact that Credit & Surety pricing requires this judgment is a particular weakness during a soft market, because it is not unusual for market pressures to compromise actuarial judgement.

Experience rating is theoretically appealing because it calculates the correct rate for a portfolio based on its own experience. This means that we do not need to make many assumptions about the applicability of the data when we price. But experience rating does have weaknesses. Primarily, it is a demanding methodology with regards to data quantity and quality. It requires reasonably extensive data, restricting its applicability to larger volumes of business and longer time intervals. It also requires reasonably good quality data. Shock losses need to be massaged to match long-term expectations and the loss cycle needs to be carefully built in. But shock losses and loss cycles are rare, so actuaries must chose between long time periods full of ancient data or short time periods that lack credible experience. The adjustments required to get around these problems are judgmental and threaten the credibility of the analysis. It is unfortunate that experience rating's major weakness is Credit & Surety's major characteristic.

Exposure rating is theoretically appealing because it permits the use of industry experience. This permits the experience of shorter time periods to be more credible. Furthermore, the way Credit & Surety is underwritten means that industry severity data should not need to be manipulated when applied in an exposure rating. Only the frequency estimates should require judgement. However, both are difficult to calculate in practice. First, sharing of data is not common in the Credit & Surety industry. There is not a lot of experience available, and those who do have books large enough to have credible experience want to use this as a competitive advantage over those who do not. Furthermore, when experience is shared (for ex. Surety Association of America (SAA) or reinsurers), it usually is without the corresponding exposure values. So, it is difficult to compile industry data. The industry also does not have a uniform standard for recording data. There are a wide variety of definitions for "loss" and an even wider variety of definitions for "exposure." So, if and when one is able to compile a collection of industry experience, it does not have quite as much meaning as we would like. This reduces the selection of exposure rating parameters to an act of judgement.

Historically, experience rating has been the approach used by insurers when reviewing the rate adequacy of their book and by reinsurers when pricing reinsurance. This is because the lack of credible exposure rating parameters is generally a greater problem than the judgement required for experience rating. Primary companies have always used some form of exposure rating for pricing individual insureds, but they seldom even look at this data when reviewing the profitability of their entire portfolio. This is partly due to the fact that exposure rating systems typically require so many soft factors that the results are unsuitable for the purpose of portfolio analysis. Portfolio reviews are almost exclusively performed via experience rating. This is in stark contrast to the financial markets that rely heavily on the exposure rating when performing portfolio analyses.

4.2 Introducing Financial Market Theory

Combining traditional exposure rating with modern financial markets pricing theory results in a Credit & Surety pricing methodology that is considerably more flexible than traditional insurance pricing methodologies. This development is made possible by the fact that insurers and reinsurers are now adopting financial markets risk management methodologies, making new data available for pricing. The mixed approach combines the best practices of both theories.

The characteristic that most distinguishes financial markets pricing theory from insurance pricing theory is the way exposure is measured. Credit & Surety insurance currently follows the property tradition by using the policy limit or a PML (probable maximum loss) as the exposure base. The financial markets use an exposure base that is significantly more sophisticated. As with insurance, it starts with the policy limit modified to reflect the value realistically exposed. This effectively gives us a PML. The financial markets then further modify the quantity to reflect the credit rating of the counter party. Better credit ratings have lower losses with respect to the amount exposed. Finally, correlation is introduced to get the correct measure of aggregate risk.

The goal of the financial markets approach to exposure measurement is to precisely quantify the expected loss of a risk with as little subjective judgement as possible. This would appear to be impossible when you consider all of the qualitative risk assessments that must go into the analysis. For this reason, the financial markets have established the use of public credit ratings as a way of validating the judgement of the analysts. The credit ratings contain all of the judgmental factors so that the other components can be entirely objective. Making the credit ratings public knowledge permits analysts to be able to compare their assessments with those of other analysts and ensure that their assessments are not wildly different from the rest of the market. The consistency in approach and public application compensates for the necessary subjectivity of financial markets pricing. Credit & Surety pricing could benefit from this approach towards pricing. The change would also enable insurers to incorporate their Credit & Surety exposures into their credit risk management framework, giving them a more complete picture of the risk their company has to stress from the financial markets.

However, there are differences between the two markets that hinder the combining of their theories. Two important differences are the fact that insurance often has triggers that differ materially from simple financial default and insurers have significantly more control over the risk. For example, if a construction company defaults, then the Surety will look for ways to keep projects going forward either by loaning money to the contractor or by finding a replacement contractor. The loss will emerge over time according to decisions the Surety makes about how to handle it. It is even possible for default to ultimately result in no loss at all. On the other hand, in the financial markets, default results in payment according to the obligation. Therefore, while the risk profiles of insurers are strongly correlated with the risk profile of the financial market, they are also markedly different.

Another important difference between insurance and financial markets theory that must be reflected is the fact that insurance companies regularly review their base rate while the financial markets do not. Financial markets price each risk separately, but they do not review the portfolio in total and calculate base rate changes. Financial markets pricing theory focuses more on differentiating risks than on making sure that the aggregate return is adequate. In financial markets theory, the company does not attempt to set the average return but rather lets market forces dictate what that return should be. It is assumed that the rate is adequate because market forces will push all unprofitable business into the lower credit ratings. The goal then is to provide the risk differentiation information required to make the market operate efficiently. In order for this theory to be useful to insurers, the financial market's ability to differentiate risks must be married with the insurance market's ability to measure whether the correct aggregate premium is collected.

The differences between insurance and the financial market can be incorporated into the pricing by using the financial markets pricing as a benchmark and adding a deviation factor for the insurance differences. In its most general form, the expected loss as an insurance product can be represented as follows:

Insurance Policy E[L] = f(Financial Instrument E[L], α)

Here, α represents a factor measuring the differences in the loss triggers and other advantages of writing insurance. The value for α should vary with the type of product being modeled. The function used to apply α could model expected loss in total, frequency and severity separately, or some variation thereof. Refer to the Appendix 1, Page 2 for a simple example.

Establishing a value for α that is appropriate for the insurance product is critical to this exercise. Two main approaches are possible. The first is to use the historical data of that product and to back into the α's that reconcile the experience and benchmark. In other words, estimate the expected loss as a financial product (using historical exposures and ratings at a given point in time) and compare that number to historical surety losses developed to ultimate for the same time period. The second is to establish α judgmentally by comparing that product to others for which α is known. For example, if $f(E[L],\alpha) = \alpha E[L]$, then a general rule of thumb is: For high risk Commercial Surety (bonds that act as financial instruments), α is one. For very low risk Commercial Surety, α is close to zero. For Contract Surety and Credit Insurance, α is somewhere in the middle.

Credit Default Swaps are the preferred benchmark because Credit Default Swaps are standardized and actively traded on the open market. This permits the insurance company to see what the market's consensus opinion of that credit's risk is without having to adjust the data for specialized terms and conditions. This is called the "price discovery process." The riskiness implied by the market price can then be compared to the riskiness as measured by the commercial credit ratings and the riskiness as measured by the insurer's own credit models. Another benefit is that the market reacts to information faster than any other credit rating process. This makes the pricing more responsive to current events and the ability to keep up with the market prevents arbitrage opportunities. Finally, Credit Default Swaps are becoming an increasingly popular tool for hedging credit risk exposures. Using Credit Default Swaps as the basis for the pricing and risk management process makes the hedging calculations easier.

The calculations can be accomplished with varying sophistication. This paper presents a simplistic approach that can be applied to any Credit & Surety product. The calculations in this paper will be based on the following definitions:

Notional Amount = Exposure

This is the sum of all of the bond limits for that principal. (Contract Surety sometimes uses Work on Hand.) This amount goes by many names, including: Aggregate Bonded Liability, Aggregate Penal Sum, and Un-exonerated Bond Amount.

Default Rate = Probability of Default as a Financial Instrument

Default means that the company was not able to continue servicing its debt. The default rates for securities are based on their credit ratings from credit rating agencies such as Moodys and Standard & Poor. Companies that are un-rated by the rating agencies can be rated using computer programs such as KMV rating and Moody's Risc Calc. Moody's Idealized Defaults Rate table is presented in Appendix 1, Page 3 of this paper.

Please note that we used a somewhat narrower definition of default rates in this paper. Moodys and S&P, at times, use a broader definition of default rates depending on the purpose of the exercise.

α = Probability of Loss as a Surety Product / Probability of Default as a Financial Instrument

The loss triggers for surety are stricter than those assumed in the definition of default for a pure financial instrument. For example, a missed interest payment or a restructuring of debt would trigger a financial default. On the other hand, the contractor or the commercial surety principal has to be bankrupt for a surety default to take place. Therefore, a relativity factor needs to be applied to get the correct frequency for the product being priced.

1 - Recovery Rate = Severity

One minus the recovery rate equals the expected severity. It is stated as a percentage of the exposure. The recovery rate should be based on the type of bond (Contract, Low Risk Commercial, Workers Compensation, etc). Note that recovery rates are correlated with default rates – poor quality credits have both higher frequencies and higher severities.

Expected Loss = Notional Amount x Default Rate x α x (1- Recovery Rate)

This is the expected loss for the principal. Including α permits us to reflect the bond's unique characteristics in the pricing. Omitting α gives us the expected loss for a comparable credit default swap.

When applying these formulae for risk management purposes, it is important to take into account the correlation within and between industry sectors. Correlation also exists between regions/countries and between Credit & Surety insurance and the company investment portfolio. It is important to include all sources of credit risk in this calculation, including all corporate bonds that your investment department has purchased. There are several methodologies to perform this calculation. Two are frequently used:

a. Downgrade the credit ratings of securities in sectors that have exceeded specific concentration thresholds. For ex: 10% concentration → one-notch down grade, 15% concentration → two-notch down grade, et cetera. This gives correlation a cost (the cost of the required hedge) enabling underwriters to manage correlation within the pricing formulae and creating a disincentive for adding more of this risk.

b. Create a simulation model that accurately reflects the characteristics of the original portfolio, including correlation. Note that many different options exist for the design of the correlation engine.

Finally, it is important to note that pricing is not independent of risk management. The outputs of the pricing exercise are the inputs for the risk management system. For this reason, it is important to design any pricing system so that both needs are met. In general, the output from the pricing exercise should contain the following:

- Average portfolio default rate and rating
- A distribution of default rates (or ratings)
- Average notional amount
- Expected loss
- A distribution of losses
- Expected excess loss
- A distribution of excess losses

4.3 Primary Insurance Application

Most primary companies use industry based rating tables for small risks, such as the Surety Association of America's Surety Loss Cost tables, and their own proprietary rating systems for large risks. Increasingly, these proprietary systems refer to the credit rating of the risk being insured.

One can use a modified credit default swap pricing methodology as the approach for pricing bonds. Consider the example of an insured that wants to insure a $25mm receivable from a power company, payable in 5 years. (Appendix 1, Page 1) It is a high-risk bond that behaves very much like a financial guarantee. Suppose the power company has a Moody's credit rating of Baa3. Referring to Moody's Default Rate table, the five-year default rate for Baa3 credits is 3.05%. Since the insurance policy behaves similarly to a financial guarantee, α is chosen to be one. The expected loss is thus $686K. Reflecting five years of investment income gives us a discounted expected loss of $538K.

4.4 Proportional Reinsurance Application

Both quota share and surplus share reinsurance are common in Credit & Surety. Quota share reinsurance is the easiest to price since cedants are able to provide all of the pricing information listed in section 4.2. The most difficult part of the quota share pricing exercise is in modeling the commission terms, since they generally are a function of the treaty results. When computing the appropriate aggregate loss distribution, it is critical to accurately reflect the correlation within the portfolio. Surplus share reinsurance is more difficult to price because the ceded amount varies with the bond limit. A standard way to approach this is to restate the exposure profile to reflect the surplus share terms and to then price the treaty as if it were a 100% quota share.

It is increasingly common for reinsurers to request a complete listing of all of the credits in the portfolios so that the reinsurer can incorporate the information in their credit risk management system. This detail of data also enables the reinsurer to independently assess the adequacy of the primary company risk evaluation and management process. Lead reinsurers typically review the historical accuracy of the cedant's pricing relative to the results that insurer experienced. The reinsurer calculates a cedant specific α in addition to the α' **s** it uses for the products to adjust for the primary's underwriting quality. The pricing then proceeds as described above.

4.5 Excess Reinsurance Application

The credit default swap approach can also be an effective way to approach the pricing of excess reinsurance. Consider the example of a portfolio presented in Appendix 1, Page 2. Excess reinsurance covers losses occurring this year, so the term is always one no matter what the length of the underlying obligations are. (Beware of the optional tail coverage!) We look up this product in the pricing tables to get the default values for α and the recovery rate. This information is then enough to compute the expected loss for the excess reinsurance layers.

A major benefit of this methodology is that the relationship between pricing and risk management is considerably clearer. It is now obvious to the cedant how different risk management rules will affect their reinsurance costs. For example, two observations are immediately apparent in the sample exhibit on Appendix 1, Page 2.

First, credit A could potentially cause a loss that greatly exceeds the amount of excess reinsurance purchased. The insurer did not purchase enough excess reinsurance to adequately protect itself. But, high layer reinsurance is expensive, especially if it is not well used. Just like lines of credit issued by banks, reinsurers typically charge capacity fees for excess layers that have low activity because they might be used. A more cost effective way to manage the portfolio is to not let any one risk get that large in the first place.

Second, credits A, B and C all have inadequate credit quality for their size. Notice how the expected loss for the 5x5 layer is almost entirely due to these three risks. It is unlikely that these three risks are able to support all of a reinsurer's capital and frictional costs by themselves. So, the 5x5 layer is probably uneconomical for the cedant. A more cost effective way to manage the portfolio is to place lower maximum limits on poor credits and to establish an orderly exit process to address deteriorating credits.

Incorporating additional information into the exhibit enables us to perform even more analyses. For example, comparing the direct premiums collected for the bonds with the ground up expected loss calculation gives us a diagnostic for reviewing α. Other information that is potentially useful includes bond type, industry group, collateral, hedges, retrocessions, and the Greeks.

4.6 Agg-Stop Application

Aggregate stop loss reinsurance (Agg-stops) is the insurance version of a collateralized debt obligation (CDO). Both involve collecting a large portfolio of risks and then the slicing the portfolio into horizontal tranches. Since correlated risk exacerbates aggregate loss, most of the correlated risk ends up residing in the upper tranches while most of the uncorrelated risk ends up residing in the lower tranches. Therefore, the function of Agg-stops for Credit & Surety portfolios is to strip the correlation from the portfolio. And since the correlated risk is the largest consumer of credit capacity, Agg-stops release significant amounts of credit capacity for the primary insurer, but at the cost of consuming significant amounts of credit capacity from the reinsurer.

Since the risk in an Agg-stop treaty is almost entirely correlated risk, it is critical that the model used for pricing Agg-stops has a sophisticated treatment of correlation. Typically a simulation model is used such that industry group could accurately model each necessary term, such as sub limits. Simulations also permit the reinsurer to analyze the effect of including the treaty in their portfolio. This permits the reinsurer to more accurately assess their cost of capital loads for the treaty.

A simulation model involves stochastically generating frequency, exposures and severities. For homogenous portfolios, the simulation model can be relatively simple. If we assume that all of the correlation is the result of frequency, then the key to the model is the frequency simulation. A commonly used approach in this situation is the Binomial Model. It is best described by example: We simulate the frequency of loss for 100 correlated risks as 50 uncorrelated risks, adjusting the expected severities so that the expected total loss is correct. We established number 50 by applying the Method of Moments to an estimated aggregate variance. After the frequency is simulated, then for each simulated event, the sizes of the exposures are modeled, usually from a Lognormal distribution fitted to the exposure profile and scaled as required by the Binomial Model. Finally, the loss severities are modeled as a percentage of the exposure, usually from a Gamma or a Pareto.

5 Specific Issues

5.1 Frequency

The ability to get improved frequency estimates is a key reason why many insurers have begun to adopt the credit scoring algorithms of the rating agencies. One major advantage rating agency algorithms have over other pricing methodologies is that the rating agencies have the most complete and longest running histories publicly available. A second major advantage is that the rating agencies are relatively quick to reflect any changes in probabilities of default in their credit ratings. This allows financial companies to continually revise their assessment of the quality of their portfolio without having to continually re-rate all of the credits themselves. Alternatively, an even more responsive indicator of change is the credit spreads in the market. A credit spread is the difference between the rate for treasury notes and the rate for a similar bond issued by the company. Since credit spreads rise monotonically as credit ratings fall, the market's spreads can be used to establish the market's consensus credit rating. The credit derivative market is a common place for financial institutions to get consistent information about credit spreads. It is estimated that up to 90% of the activity on the credit derivative market is solely for "price discovery" purposes.

While the credit scoring approach has its benefits, it also has its limitations. For example, it is important to remember that credit ratings are designed for assessing the pricing of debt instruments, not insurance. Also, rating agencies have also been known to approach the same calculations in different ways for different publications, depending on what the information is intended for. Rating agency information must be used with caution.

If pulling default statistics from a publication, it is important to note precisely what the statistics measure. This is not always clear. Sometimes the statistics are pure frequency statistics and sometimes they measure expected loss costs. Furthermore, since frequency and severity strongly correlate, the different default statistics are not always easily distinguished. A detailed knowledge of how the statistics were calculated and the assumptions underlying them is necessary before attempting to use them in pricing.

Credit ratings have a fair amount of subjectivity to them. Rating agencies judgmentally segregate the credits into rating categories and then calculate statistics on the categories. The subjectivity of the data means that there are trends that must be identified and compensated for. For example, from 1984 to 1991, the annual default rate for Moodys B1 rated securities always stayed within the range of 4.36% to 8.54%. From 1992 to 2000, the annual default rate for Moodys rated B1 securities always stayed within the range of 0.00% to 4.57%. Was this the result of a changing economic environment or a change in the definition of a B1 security? A review of the aggregate default rate for all corporate bond issuers demonstrates that the two periods were not significantly different. Therefore, we can conclude that the change in experience is due, at least in part, to a change in the definition of a B1 security.

While credit-scoring models can be used to improve credit default rate predictions, they cannot always produce accurate frequency predictions for the insurance products we are pricing. This is because the insurance industry's definitions of credit default can differ markedly. To a rating agency, default means the failure to service debt. To an insurer it can mean many things, such as the failure to pay a bill or the failure to fulfil a bonded obligation. Insurance default rates can be greater or less than commercial debt default rates, depending on the nature of the insured obligation. For this reason, it is best to use credit-scoring models with great care in order to be useful in surety pricing.

5.2 Severity

Severity (recovery rates) can be analyzed using data and models similar to those used for frequency. Recovery rates vary with both credit rating and debt seniority, thus they are specific to the insured and the instrument being priced. Severity distributions are harder to fit than frequency distributions because they are more complex. For frequency, we only need to be concerned about the average probability of default. While for severity, we need the full distribution. However, the increased complexity of fitting severity curves is partly mitigated by the fact that Credit & Surety underwriting places an overwhelming focus on frequency.
When pricing for retentions or excess layers, it is important to put a distribution around the average recovery rate. Then, the expected loss cost for each exposure in the portfolio is calculated using the Limited Expected Values of the recovery rate distribution. Typically, a Beta distribution is used if it is impossible for the loss to exceed the notional amount. If it is possible for the loss to exceed the notional amount (i.e.: Contract Surety), then property distributions typically are used. If the list of exposures is not known, then a LogNormal distribution is typically used for representing the distribution of potential exposure sizes.

For primary insurers, an accurate representation of the recovery rate distribution is essential if the insured has a significant retention or posts significant collateral. The recovery rate curves will determine how much credit to give to the collateral and retention. Inaccurate curves increase the risk of over/under pricing the business. A review of the insurer's hit ratios by retention would indicate whether inaccuracies exist.

Establishing accurate recovery rate distributions for new products pose a particularly difficult challenge. This is generally accomplished by borrowing distributions from other related products. Insurers can improve the variety of their severity distributions in the following manner: First, fit a recovery rate distribution for each category that has sufficient experience. Use the same form of distribution for each fit so that the equations are identical and only the parameters change. Plot the parameters onto a grid, labeling each point with the product that generated it. Then, recovery rate curves for new products can be selected judgmentally from the grid by placing a point onto the grid that makes sense relative to the existing portfolio of products.

For reinsurers, the exact shape of the recovery rate distribution is often not that important. For most reinsurance applications, only the mean and variance of the recovery rate distributions are significant. This is because we are applying the same curve to a large number of exposures and the Law of Large Numbers smoothes out the inaccuracies of the distribution. It is important that the first two moments are correct, but the higher moments are often smoothed out. However, note that the Law of Large Numbers breaks down in the high excess layers. If pricing these layers, it is important that the tail of the recovery distribution be adequately represented otherwise the layers will be under-priced.

Reinsurers must also pay attention to whether the distribution of exposures is changing or can change. A trend in average exposure size will materially effect the excess severity distribution. Furthermore, the existence of excess reinsurance often provides the incentive for cedants to put the coverage to greater use. An excess layer that does not currently have many exposures in it may have significantly more by the end of the term. Therefore, it is common for reinsurers to charge a capacity fee (similar to the fee banks charge for keeping lines of credit open) for excess layers that are lightly exposed. This pays for the potential for the cedant to write more bonds that expose the layer. Note that such a fee is not required if the layer is written on a cessions basis.

5.3 Loss Cycle

The loss cycle is when Credit & Surety loss activity dramatically increases. During a loss cycle, loss ratios typically are double or treble their historic levels. Loss cycles generally are caused by credit cycles but may also be caused by other contagious events, such as a rapid contraction in the amount of government spending on capital projects. Loss cycles typically focus around a particular industry and region, meaning that there are multiple overlapping cycles that could potentially affect an insurer's results. Preparing the insurance company for future loss cycles is one of the most difficult tasks a Credit & Surety actuary must perform.

Surety & Credit has two main loss cycles. First, large contractors and most non-construction companies finance their operations through credit. A contraction of the credit market causes the less stable corporations to fail. However, small contractors tend to finance their operations by kiting funds from one job to the next. A reduction in the amount of new construction has the same effect on this market as a contraction of credit has on the market as a whole. Typically, the availability of credit drives the loss cycle for the large Contract Surety, Commercial Surety and Credit insurance markets while the amount of new construction drives the loss cycle for the small Contract Surety market.

Financial Guarantees and Credit Derivatives also have loss cycles that are largely driven by the credit cycle, and thus are strongly correlated with large contractors and Commercial Surety. However, a large part of the Financial Guarantee market is municipals and these behave very differently. The risk for municipals is that the politicians do not want subject the public to the pain required for them to maintain their financial obligations.

The existence of the loss cycle complicates the pricing of these products. The actuary must keep both the long term and short-term horizons in mind when pricing. For example, if the loss ratio averages 30% in normal years and 80% during a loss cycle, and if the loss cycle comes once every decade and lasts for two years, then the long-term loss ratio is 40% (= 0.8 x 30% + 0.2 x 80%). Therefore, in order to make money over the long term, an insurance company must charge between 10% and 33% more for its products than it would if it was taking purely a myopic view towards pricing (depending on whether expenses are loaded as fixed or variable). The market does not easily support such pricing. Thus, strict discipline by actuaries and underwriters is required.

In reinsurance, managing the horizons also consists of paying attention to the "banks" that insurers have developed with the reinsurer. The bank is the amount of excess funds that the reinsurer has collected over the good years in order to pay for the bad. Without the building up of banks, reinsurers cannot be profitable over the long term. Reinsurance rates should reflect the size of the bank that the insurer has. Returning to the above example, if the insurer has a fully funded bank, then the reinsurer can charge a rate contemplating a 30% loss ratio. But, if the insurer has no bank at all, then the reinsurer should charge a rate contemplating a 40% loss ratio (or higher).

Even if the insurer/reinsurer intends to withdraw from the market when the loss cycle begins, they generally do not get a chance to withdraw until their contracts end. That means that the insurers & reinsurers must first witness the beginning of the loss cycle before knowing it is time to withdraw, and by then it is generally too late to avoid the bulk of it. The loss cycle is relatively short – it is over before much action can be taken.

Loss cycles also have another insidious side that make identifying them particularly difficult. Loss cycles have the tendency to be devastating to insureds that already have open credit-related claims, meaning that these claims are severely exacerbated and the resulting extraordinary loss is recorded with the date of the original claim. Therefore, the loss cycle is actually much shorter than the actuarial loss experience suggests. For example, in the most recent loss cycle, losses grew in 2000, peaked in 2001 and may have begun to decline in 2002. But, the loss cycle was not apparent to the market until late 2001. Most of the losses in 2001 and all of the losses in 2000 are due to the aggravation of losses that already were in claim. In the context of the actuarial loss history, the loss cycle was not identifiable until it was already half over.

The loss cycle has often been compared to hurricanes and other natural catastrophes and they both are managed in similar ways. But, they are very different to a pricing actuary. Unlike most other catastrophes, the loss cycle is not a Poisson process. If we have a loss cycle this year, then it will be a few years before we have another one. Loss cycles require weak companies and excessive competition. It takes time for these economic conditions to redevelop once a loss cycle occurs. However, the fact that there are many different types of loss cycles makes the time between loss cycles very difficult to predict. For example, the time between the last two Contract Surety cycles was about 13 years (1987 to 2000). But, if we include Commercial Surety, the period drastically shortens. The last Commercial Surety loss cycle (credit cycle) was in 1992. The fact that we have multiple different types of loss cycles does add some Poisson-style risk to the pricing, but does not make it a full Poisson process.

Understanding the loss cycle is a vital part of the pricing process. It ultimately determines whether the insurer makes money or not. It is a particularly difficult component to price because the long time periods which separate loss cycles limits the usefulness of loss histories. Predictions are as much art as science and crystal balls invariably find their way into the process. Some actuaries have expressed their confidence in the new economy and that the durations between loss cycles are increasing. Others point to the increasing reliance businesses have on the credit markets as a fundamental destabilizing force, which should shorten the durations between loss cycles and increase their severities. Today, there is no consensus. The only general conclusion that can be drawn is the fact that insurers tend to be too optimistic. Historically, too many have found themselves with inadequate banks when the loss cycle begins.

6 Conclusion

There are many new products at the intersection of the insurance and financial markets, and some of the traditional insurance products now have financial flavors. The traditional insurance methods for evaluating and managing these risks have become out-dated. The goal of this paper is not to give a definitive proposal, but to invite actuaries, underwriters and senior managers to look at these products from a new perspective. The biggest danger to insurance is in not changing. This was made very evident by the enormous exposures insurers had to Enron and by the fact that many of the resulting claims by Enron's obligees were entirely unanticipated. In conclusion, we strongly believe that following the lead of financial markets could help the insurance industry quantify and manage Credit & Surety risks more effectively and more efficiently. This will ensure the long-term availability of sufficient capital, and thus capacity, for this line of business.

Acknowledgements

The authors would like to gratefully acknowledge Scott Orr, Frank Bonner, Marianne Wilbert, Scott MacColl, John Boulton, Kathleen Miller-Boreman, Roger Amrein for reviewing early drafts of this work and offering their feedback. We would like to thank Nick Pastor, David Smith and Chris Walker of the CAS Ratemaking Committee for providing constructive feedback. Finally, we thank our spouses Dr. Latha Cherath and Ms. Akiko Ito for their support and encouragement. The authors remain solely responsible for errors that still exist in the paper.

Credit & Surety Pricing and the Effects of Financial Market Convergence

The Calculation of Expected Loss

(1)	Principal	XYZ Inc.
(2)	Credit Rating	Baa3
(3)	Industry	Power
(4)	Exposure	25,000,000
(5)	Duration [A single payment of $25mm is due in 5 years.]	5
(6)	Moody's Default Rate	3.05%
(7)	Average Recovery Rate [Default value for high risk bonds]	10%
(8)	α [Bond is a no-recourse demand obligaton.]	100%
(9)	Expected Loss = (4) x (6) x [1 - (7)] x (8)	686,250
(10)	Discount (@ 5%)	0.784
(11)	PV(Expected Loss) = (9) x (10)	537,695

Credit & Surety Pricing and the Effects of Financial Market Convergence

The Calculation of Expected Loss for XOL Reinsurance

(0)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
												Reinsurance Loss	
Index	Name of Credit	Notional Amount	Term	S&P Rating	Moody Rating	Selected Rating	Default Rate	α	Recovery Rate	Ground Up Expected Loss	1M X 1M	3M X 2M	5M X 5M
1	A	20,000,000	1.000	B+	B2	B2	7.160%	70%	10%	902,160	50,120	150,360	250,600
2	B	12,000,000	1.000	B+	B1	B1	4.680%	70%	10%	353,808	32,760	98,280	163,800
3	C	9,000,000	1.000	B	B3	B3	11.620%	70%	10%	658,854	81,340	244,020	252,154
4	D	8,000,000	1.000	BB+	Ba1	Ba1	0.870%	70%	10%	43,848	6,090	18,270	13,398
5	F	7,500,000	1.000	BBB	Baa2	Baa2	0.170%	70%	10%	8,033	1,190	3,570	2,083
6	G	7,000,000	1.000	BB+	Ba1	Ba1	0.870%	70%	10%	38,367	6,090	18,270	7,917
7	H	7,000,000	1.000	A	A2	A2	0.011%	70%	10%	479	76	228	99
8	I	6,800,000	1.000	BB+	Ba1	Ba1	0.870%	70%	10%	37,271	6,090	18,270	6,821
9	J	6,500,000	1.000	BB+	Ba1	Ba1	0.870%	70%	10%	35,627	6,090	18,270	5,177
10	K	6,000,000	1.000	A	A2	A2	0.011%	70%	10%	411	76	228	30
11	L	5,000,000	1.000	BB	Ba1	Ba1	0.870%	70%	10%	27,405	6,090	15,225	-
12	M	5,000,000	1.000	BBB	Baa2	Baa2	0.170%	70%	10%	5,355	1,190	2,975	-
13	N	5,000,000	1.000	BBB-	Baa3	Baa3	0.420%	70%	10%	13,230	2,940	7,350	-
14	O	4,000,000	1.000	BB-	Ba1	Ba1	0.870%	70%	10%	21,924	6,090	9,744	-
15	P	3,500,000	1.000	B+	B1	B1	4.680%	70%	10%	103,194	32,760	37,674	-
16	Q	2,000,000	1.000	AA	Aa2	Aa2	0.001%	70%	10%	17	8	-	-
17	R	2,000,000	1.000	BB+	Ba1	Ba1	0.870%	70%	10%	10,962	4,872	-	-
18	S	1,500,000	1.000	BB	Ba2	Ba2	1.560%	70%	10%	14,742	3,822	-	-
19	T	1,500,000	1.000	B	B2	B2	7.160%	70%	10%	67,662	17,542	-	-
20	U	1,500,000	1.000	BBB-	Baa3	Baa3	0.420%	70%	10%	3,969	1,029	-	-
										Expected Loss in Layer:	266,265	642,735	702,078

Notes: (10) = (2) x (7) x (8) x [1-(9)]
(11) = (7) x (8) x Min[Max[(2)x[1-(9)]-1M,0],1M]
(11) = (7) x (8) x Min[Max[(2)x[1-(9)]-2M,0],3M]
(12) = (7) x (8) x Min[Max[(2)x[1-(9)]-5M,0],5M]

Credit & Surety Pricing and the Effects of Financial Market Convergence

Moodys Idealized Corporate Default Table

	Life of Asset										
	0	1	2	3	4	5	6	7	8	9	10
NoDef	0.000%	0.000%	0.000%	0.000%	0.000%	0.000%	0.000%	0.000%	0.000%	0.000%	0.000%
Aaa	0.000%	0.000%	0.000%	0.001%	0.002%	0.003%	0.004%	0.005%	0.007%	0.008%	0.010%
Aa1	0.000%	0.001%	0.003%	0.010%	0.021%	0.031%	0.042%	0.054%	0.067%	0.082%	0.100%
Aa2	0.000%	0.001%	0.008%	0.026%	0.047%	0.068%	0.089%	0.111%	0.135%	0.164%	0.200%
Aa3	0.000%	0.003%	0.019%	0.059%	0.101%	0.142%	0.183%	0.227%	0.272%	0.327%	0.400%
A1	0.000%	0.006%	0.037%	0.117%	0.189%	0.261%	0.330%	0.406%	0.480%	0.573%	0.700%
A2	0.000%	0.011%	0.070%	0.222%	0.345%	0.467%	0.583%	0.710%	0.829%	0.982%	1.200%
A3	0.000%	0.039%	0.150%	0.360%	0.540%	0.730%	0.910%	1.110%	1.300%	1.520%	1.800%
Baa1	0.000%	0.090%	0.280%	0.560%	0.830%	1.100%	1.370%	1.670%	1.970%	2.270%	2.600%
Baa2	0.000%	0.170%	0.470%	0.830%	1.200%	1.580%	1.970%	2.410%	2.850%	3.240%	3.600%
Baa3	0.000%	0.420%	1.050%	1.710%	2.380%	3.050%	3.700%	4.330%	4.970%	5.570%	6.100%
Ba1	0.000%	0.870%	2.020%	3.130%	4.200%	5.280%	6.250%	7.060%	7.890%	8.690%	9.400%
Ba2	0.000%	1.560%	3.470%	5.180%	6.800%	8.410%	9.770%	10.700%	11.660%	12.650%	13.500%
Ba3	0.000%	2.810%	5.510%	7.870%	9.790%	11.860%	13.490%	14.620%	15.710%	16.710%	17.660%
B1	0.000%	4.680%	8.380%	11.580%	13.850%	16.120%	17.890%	19.130%	20.230%	21.240%	22.200%
B2	0.000%	7.160%	11.670%	15.550%	18.130%	20.710%	22.650%	24.010%	25.150%	26.220%	27.200%
B3	0.000%	11.620%	16.610%	21.030%	24.040%	27.050%	29.200%	31.000%	32.580%	33.780%	34.900%

Dynamic Pricing Analysis

Charles H. Boucek, FCAS, MAAA,
and Thomas P. Conway, ACAS, MAAA

Abstract

This paper presents a methodology that represents a significant enhancement to current pricing practices. The goal of this methodology is to estimate the impact that a rate change will have on a company's policyholder retention and the resulting profitability of this transformed book of business. The paper will present the basics of this methodology as well as where future work will need to be done to bring this methodology into mainstream pricing. The work that the authors have done in this area has focused on Private Passenger Auto Insurance but these techniques could be applied to other lines of business.

Introduction

There is a wealth of actuarial literature regarding appropriate methodologies for using exposure and claims data in order to calculate indicated rates. Techniques have been developed to address difficult issues such as small volumes of data, years that are particularly immature and high excess layers of coverage. All of these techniques ultimately produce a set of actuarially indicated rates and rating factors. When it comes to deciding on the rates and rating factors that will actually be used in the marketplace, however, a new dynamic begins to enter the picture.

A revised set of rates will impact the profitability of the companies' book of business in a number of different ways. There is the obvious impact that the revised rates will have on the premiums that policyholders are paying. There is also the more intangible impact of the policyholder reaction to the rate change. A rate change exceeding a certain threshold will likely send a customer shopping for an alternate insurer. Depending on the alternative premiums that are available in the market, that customer may decide to insure with another company. If a rate change produces a large number of such non-renewals within the company's book of business, the revised rates could impair the intended benefits of the rate change. Alternatively, if the non-renewals that occur are in classes of business that are particularly unprofitable for the company, its profitability could actually be enhanced by the non-renewal activity.

Companies often have a number of ad hoc "rules of thumb" for determining the amount of a rate change that the market will bear, but very few rigorous models exist that attempt to estimate the likely customer reaction to a rate change. An approach to pricing that considers not only the impact of the new rates on the average premium charged, but also on the renewal behavior of policyholders can thus be a significant step forward for determining appropriate prices and likely future profitability. The question that must be asked is "Are there a family of models that can model the renewal behavior of policyholders?"

Actuarially Sound Rates

This paper will present methodologies that will allow the consideration of the impact of policyholder retention in the pricing process. As such, the rates that are being considered in such an approach may not be the same as the actuarially indicated rate. However, since no actuarial method produces an indicated rate that is precisely correct in all situations, there will always be a reasonable range of actuarially sound rates. The methodologies presented in this paper demonstrate how the decision regarding which rate to implement can be made with more rigor than is possible with the current approaches used in the industry.

I. Agent Based Modeling

A. What is It?

A family of techniques that has been successfully applied to model similar behavior in the past is called Agent Based Modeling. Simply put, in using these techniques, models are built which contain factors, agents and rules. Factors are the quantitative measures of the system that is being modeled. In the example of modeling customer reaction to rate changes, the factors would encompass the rates and rating factors for a company and its competitors. It would also include the loss potential of various classes of business that would be used to determine profitability of those business classes. The agents in the model are the units between which interactions take place. In the modeling of rate change reactions, agents would consist of customers, competitors, insurance agents, etc. The rules describe how the different agents in the model will interact.

One of the problems encountered in applying agent based modeling to insurance is with nomenclature. We have agents in the model and agents who are selling policies. To complicate matters even further, the insurance agents are one of the agents in the model. Throughout this paper, in order to assure that the terminology is succinct, an agent in the model will be referred to as an economic agent while an agent selling insurance will be referred to as an insurance agent.

Economic agent behavior is assigned rules, based on a combination of historical data, surveys, focus groups, and analysis. The models are run under various scenarios and the results can be used to help determine a strategic direction with insights that cannot be discerned with the current "rules of thumb" type approach.

An example of a successful application of economic agent based modeling is the modeling of changes in retirement behavior, measured by retirement ages, in response to law changes. In an article in *Behavioral Dimensions of Retirement Economics* by Robert Axtell and Joshua Epstein[1], the changes in retirement behavior since 1961 were successfully modeled. In 1961, the minimum age at which a worker could receive Social Security benefits was reduced from 65 to 62. It was expected that the average retirement age would reduce somewhat quickly to the lower age as a result of this change in benefits.

The actual experience however, was somewhat different. Peoples' average retirement ages did move towards younger ages but the transformation took nearly three decades, which was much longer than expected. What was missing from the original estimations that caused the actual experience to differ so greatly?

[1] Robert L. Axtell and Joshua M. Epstein, 1999, "Coordination in Transient Social Networks: An Agent-Based Computational Model of the Timing of Retirement." *Behavioral Dimensions of Retirement Economics*: 161-186

The original models were predicated on the assumption that the primary factor that influenced retirement age was the laws surrounding Social Security benefits. Reality, however, is more complicated than that. After researching the factors that influenced retirement ages, it was determined that a major factor in the decision to retire is the retirement decisions of other people in an individual's social network. By constructing models that consider these social interactions, a set of economic agents and rules were developed that accurately predicted the retirement age decisions of a population of individuals. This seemingly complex decision making of individuals could thus be accurately modeled, in the aggregate, with the proper alignment of economic agents and straightforward decision rules.

B. Constructing a Model of Insurance Retention

If models can be constructed that accurately predict the retirement decisions of a population, it is not difficult to imagine the construction of models that accurately predict the decisions of a group of policyholders to remain with their current insurer or to switch from their current insurer to another. In order to construct such a model, the first step is to describe the process that an insured will utilize in deciding to renew his policy or switch to another insurer.

1. The insured receives his renewal notice approximately 45 days prior to policy expiration.
2. If the premium decreases or increases modestly, the insured will likely renew the policy with his current insurer.
3. If the premium increases significantly, the insured will likely begin shopping around.
4. The insured will do some market research by calling other insurance agents or getting quotes over the phone or internet.
5. Depending upon the savings that can be realized, the insured will either stay or move.

This is a rather simple model as it relies solely on price as the factor upon which the decision is made. In reality, the process is more complex as other factors, such as quality of service, brand name recognition and financial stability enter into the decision as to where to buy insurance. However, many recent studies have shown that price is the most significant factor. Thus, once models can be constructed that accurately model behavior based on price, more complex models can subsequently be constructed that would consider elements other than price. Methods used to modify the basic model in consideration of these other elements will be discussed further in section II.F.

Throughout this paper, private passenger auto insurance will be used as an example line of business. These techniques could apply to other lines as well.

C. Economic Agent Based Approach versus Current Approach

The current "rules of thumb" approach may have been good enough at one time. It may also be true that this approach will be acceptable today in a situation where the rate change is simple. An example would be a rate change that applies only to the base rates. However, one of the trends for virtually all lines of business is that rate structures have become more refined over time. Using automobile insurance as an example, the number of different possible combinations of rate classes is so great that it is not possible to assess all of the changes that individual policyholders will experience in a rate change where base rates, territorial factors, driver classification factors and accident surcharges all change at the same time.

The economic agent based approach requires a model that analyzes the impact of a rate change at the individual insured level, taking into account class, territory, etc. The rate impact on detailed classifications can be assessed and thus the likely behavior of members of each of the classifications can also be assessed. By combining this retention information with information regarding the profitability of each of these individual classes, a powerful tool is built. This tool can be used to test a number of different rate scenarios in order to determine an optimal combination of profitability and retention.

For the application we created, the ABM modeling approach has advantages over traditional economic approaches to estimating buyer elasticity of demand. Traditional approaches would require empirical studies of policyholder reaction to rate changes and then the construction of elasticity curves from this analysis. While traditional approaches are useful during both a stable economic and competitive environment these conditions rarely exist for an extended period of time. The ABM approach allows for the ability to separate the impacts of the economy on a policyholder's propensity to shop and the level of price competition on the policyholder's ability to find an alternate policy at a lower price.

Another advantage of the ABM approach is that it allows for the modeling of emergent behavior. These are behavioral impacts, which may seem irrational at an individual level but are exhibited when the behavior of a group is analyzed as a whole. An example of this phenomenon is the observed behavior of groups of insured to leave when they are presented with a rate decrease. This seems irrational at an individual level but this phenomenon is accepted as regularly occurring

An additional key issue to note is that, while the model operates at the individual insured level, the goal of the model is to project the aggregate behavior of an entire book of policyholders. Thus, precise modeling of the behavior of each individual insured is not required in order to accurately model the overall behavior of a book of policyholders.

II. An Actual Model in Operation

A. The Economic agents in the Model

In constructing such a model, the first decision that must be made is "what are the appropriate economic agents to include in the model?" In the case of the retention/profitability model, there are four economic agents in the model. These economic agents are

1. The policyholders
2. The company that is considering changing its rates
3. The other companies that form the competition in the state
4. The insurance agents who are selling policies

B. The Factors in the Model

As previously mentioned, factors are the quantitative measures of the system that is being modeled. For a retention model, factors would comprise the companies new and old rate sets as well as the rates of market competitors. In addition, the claims frequencies and severities by major risk class will also need to be entered into the model. The methodology used to process this information will be described more fully in section II D.

C. The Rules for Interaction

Once the economic agents in the model are determined, the rules for interaction must then be determined. Using the structure of the model described above, the rules required for the model can be developed.

1. Policyholder/Company Interaction

   When the policyholder receives his renewal notice, a number of factors will determine his likelihood of shopping for an alternate insurer. These include the amount of a rate increase that he sees, his satisfaction with the handling of a claim (if this occurred during the most recent policy period), his satisfaction with policyholder service that he may have received throughout the policy period (e.g. for a change in vehicle), past rate changes that the policyholder experienced and position in the underwriting cycle. The focus of this paper is the amount of rate increase that the policyholder experiences and the impact that the change has on the policyholders propensity to shop and switch his policy.

   The likelihood to shop is related to the concept of the price elasticity of demand. Since auto insurance is a mandatory product in most states, a significant increase in price does not normally cause a driver to forego purchasing insurance, but instead causes him to shop. At what amount of rate change does the decision to shop occur? The decision of whether to seek an alternate insurer can be expressed as a probability function describing the relationship between the dollar (or percentage) change in an individual insured's premium and the likelihood of that insured researching the premiums of alternate insurers. We will refer to this function as the shopping function. The shopping function could be expressed either as a discrete function or a continuous function. The following would be a simple example of such a function

Premium Increase	Likelihood of Shopping
$0 and below	2%
$1 to $50	5%
$51 to $100	25%
$101 to $200	70%
$201 to $300	85%
$301 and above	100%

2. Policyholder/Competitor Interaction

If the rate increase is significant enough, and the policyholder decides to shop for coverage from another insurer, the rates of those insurers will come into play. If the policyholder finds that the price he has with his current insurer is less expensive than the prices charged by other companies, then the policyholder is unlikely to move coverage to another company. If, however, the price charged by the other companies is significantly less than that charged by the current insurer, the likelihood of the policyholder moving coverage to one of the other companies would be significant. Similar to the shopping function, the likelihood of a policyholder moving from one company to another can be described by a probability function. This function would describe the likelihood that a policyholder would move to another company given the amount of savings that could be realized. We will refer to this function as the switching function.

Similar to the shopping function, the switching function could be expressed either as a discrete function or a continuous function. The following would be a simple example of such a function

Premium Savings	Likelihood of Switching
$0 and below	2%
$1 to $50	15%
$51 to $100	40%
$101 to $200	95%
$201 to $300	98%
$301 and above	100%

The shopping and switching functions shown above are for illustration purposes only and are not based on independent research performed by the authors. The functions also show the probability based on the dollar amount of change. Work done by the authors has shown that both the dollar amount and percentage change are important predictors of shopping and switching behavior. The process required to develop shopping and switching functions is described later in this paper

3. Policyholder/Agency interaction

Whether the policyholder uses an insurance agent or buys coverage directly from the company is likely to have an impact on the likelihood of switching from one company to another. For policyholders using an independent insurance agent, alternative quotes can be obtained from the policyholders' own insurance agent. Thus, the likelihood of switching is probably greater with an independent insurance agent than with a direct insurance agent since it is easier for the customer to obtain alternative quotes from the independent agent. This would be addressed in the model by having one shopping function used for shopping for alternative quotes via the policyholders own insurance agent and a separate shopping function, with higher threshold amounts, that would be used for determining whether the policyholder will seek alternative quotes from other insurance agents or through other direct writers.

Insurance agents may also influence the behavior of their customers. For example, an insurance agent who considers a particular insurer to be a good business partner due to strong policyholder and claim service or a favorable commission structure may try to keep policies with that particular company, regardless of rates being offered by competitors.

The direct sale of insurance could have a significant impact on these functions. One of the reasons that direct sale of insurance is becoming more prevalent is that policyholders have better access to competitive information and indeed have the ability to purchase a policy via the Internet. As purchasing coverage directly from a company becomes more popular, the likelihood of switching from one company to another should increase as the effort required to comparison shop a policy will be reduced. The direct sale of insurance increasing the amount of price shopping that occurs, increases the need to perform the type of modeling that is described in this paper.

In order to keep the presentation in the paper more straightforward, the illustrations in this paper do not consider policyholder/insurance agent interaction.

D. Model in Operation

In order to develop and utilize such a model, the following information must be input to the model.

1. Current rates and rating factors of the company used to determine premiums
2. Proposed rates and rating factors of the company
3. Rates and rating factors of key competitors - These should be the largest companies in the state as these are the companies from which policyholders are most likely to get quotes. In addition, as policyholders of companies using independent insurance agents are likely to get an initial quote from the insurance agent, the most common competitors in insurance agents offices should also be used in the model.
4. The profitability of the different classes of business for the company
5. The current in force distribution of policyholders in the various rate classes

With this information put into the model, it can be run to test alternative proposed rating structures. This is done by creating a virtual marketplace and applying a Monte-Carlo simulation to the individual policyholders in the marketplace. The model will first generate a group of policyholders consistent with the companies' policyholder distribution across key classes (age, gender, marital status, etc.). The current and proposed rates of the company will then be used to determine the amount of rate change that the individual policyholders will experience. The premium change, in combination with the shopping function, will determine the probability of each individual policyholder shopping. The simulation is then run which results in certain individual policyholders deciding to shop. The policyholders that decide to shop will go into the market to seek alternative quotes and will thus determine the possible savings by switching to another company. The savings, in combination with the switching function will determine the probability that an individual policyholder will switch. A second simulation is then run which results in certain individual policyholders deciding to switch to another company. In addition, certain policyholders of the competitors will shop their policies as well. While there are a number of factors driving this behavior, our model assumes that all policyholders of the competitors are likely to shop with equal probability. The model will be run for multiple iterations until the results converge to an equilibrium level of retention and profitability.

Essentially the model tracks the distribution of policyholders across various rate classes before and after the rate change. The model then combines this information with the profitability by class in order to produce an estimate of the total profitability that will be realized under each rate scenario as well as the volume of business that will be written under each scenario.

E. Example of Determining Customer Retention

An example will help to clarify the operation of the model. This example describes how the model will work for an individual policyholder. Consider an example driver with the following characteristics

Age	35-44
Gender	Female
Marital Status	Married
Single Car/Multi-Car	Multi-Car
Driving Record	Clean
Number of Years Loss Free	5
Vehicle usage	Drive to Work < 10 miles
Rating Territory	12
Liability Limit	100/300/50
Comprehensive Deductible	100
Collision Deductible	250
Vehicle Model Year	1996

In addition, the following premiums apply for this policyholder.

Current Premium	$500
New Premium	$555
Premiums of Competitors	
Competitor #1	$619
Competitor #2	$452
Competitor #3	$544
Competitor #4	$592

Since the policyholder experiences a $55 increase in premium, the shopping function would predict that 25% of these policyholders will shop. When they shop, they learn that a savings of $103 is possible and 95% of the policyholders will switch for this amount of savings. Via the simulation, the rate increase will cause 24%(=25%x95%) of policyholders to leave the company. If there were 250 policyholders with the above characteristics, the simulation would be expected to result in the company losing 60 of these policyholders due to the rate change. If these policyholders were expected to earn an annual profit of 15% (i.e. this assumes that total losses and expenses are $472 per policy) at the higher rate level, the following table describes the expected premium and profit at the higher rate level

Current Premium	$125,000 (=$500x250)
Premium after rate change	$105,450 (=$555x190)
Profit after rate change	$15,818 (=$555x190x15%)

These results can be compared to the results of no rate change and a 9% rate increase to a premium of $545. Note that with a 0% change, the profit would be 5.6% of premium and with a 9% increase, the profit would be 13.4 % of premium (again assuming that total losses and expenses are $472 per policy). In addition, the 9% increase would result in losing only 2% of policyholders as 5% would shop for alternatives and 40% of those would switch companies for the $93 savings that could be achieved.

Premium with 0% change	$125,000 (=$500x250)
Profit with 0% change	$7,000 (=$500x250x5.6%)
Premium after rate change	$132,300 (=$540x245)
Profit after rate change	$17,728 (=$540x245x13.4%)

Obviously, the 9% rate increase is preferable as it produces a larger total profit. These results for each individual class will be accumulated in order to determine the total projected premium and profit for a specific rate scenario. The volume/profit tradeoff under each scenario can thus be reviewed in order to determine the rate structure that is considered best for the company.

F. Aggregation of Results

The example in the previous section was simplistic from a number of different perspectives. First, it dealt with a single combination of risk classes while a company's book of business is comprised of numerous different risk classes. Obviously, each such combination will have its own premium. Second, the shopping and switching functions in the example are also simplistic. The example functions shown describe the behavior of one combination of risk classes, however, the shopping and switching functions need to be more specific to the behavior of different risk classes in order for the model to accurately estimate the behavior of an entire book of business.

Constructing a more comprehensive model requires that the rate change, competitive position and shopping and switching propensities of different risk classes are known. The complexities of current rate structures requires that a thorough modeling framework be developed in order to model results at the individual class level and then aggregate the results across all of the possible combinations of driver class, territory, driving record, number of times renewed, etc. However, at its basis, the process involves aggregating results at the individual class level as described in section E.

By constructing models for different rate change scenarios, the retention and profitability of these different scenarios can be examined. The following three examples show how different rate scenarios can be analyzed in order to provide greater insight into their impact on retention and profitability.

1. Different scenarios with same overall rate change
   A company is targeting a 5% overall rate increase, but there are three specific sets of changes that are being considered for implementation.

   Scenario 1
   Only changing territorial base rates.
   This option would have rate changes vary by territory but not across other rating variables.

   Scenario 2
   Changing territorial base rates so that the territorial relativities are the same as Scenario 1. Changing a policy renewal discount to give a greater discount to policyholders that have been insured by the company for six or more years since the loss data of the company indicates that a greater discount is justified. Currently, all policyholders receive a 5% discount after three years. The proposed rate structure adds a 10% discount after six years. In this scenario, the territorial base rates would be offset in order to make up for the increase in the renewal discount and thus maintain the targeted 5% increase.

   Scenario 3
   Changing territorial base rates so that the territorial relativities are the same as Scenario 1.

Changing a policy renewal discount to give a greater discount to policyholders that have been insured by the company for six or more years since the loss data of the company indicates that a greater discount is justified.
Changing the Multi Car discount to give a greater discount to multi car policies since the loss data of the company indicates that a greater discount is justified. Again, the territorial base rates would be offset in order to maintain the overall 5% increase

As you move from scenario 1 to scenario 3, policyholders will experience more extreme rate changes. For instance, consider a territory in scenario #1 in which all coverage base rates are increased by 5%. All policyholders will experience a 5% increase unless there is a change in policy characteristics such as a change in vehicle or driving record. However in scenario #2, base rates need to increase by 7% in order to make up for the greater discount for policyholders insured for six or more years. Policyholders insured for five or fewer years will experience a 7% increase while policyholders insured for six or more years will experience a 2% increase (7% base rate increase combined wit a 5% greater discount). This demonstrates the more extreme rate changes experienced in scenario 2 versus scenario 1.

Thus, policyholders will experience the most extreme rate changes in scenario #3 and thus more policyholders will tend to shop their policies in this scenario. What impact will this have on the book of business in terms of retention and profitability? By applying the techniques mentioned previously, the following table can be produced to compare the modeled results of these different scenarios.

Scenario	Premium	Retention	Loss Ratio	Operating Result
1	149,134	88.5%	66.5%	$3,672
2	151,263	88.6%	65.5%	$5,318
3	154,412	88.7%	64.4%	$7,031

As this table demonstrates, Scenario 3 yields the best profit since the company is retaining more of the more profitable multi-car, renewal policies.

2. Different overall rate changes
A company is considering different base rate changes. The three options being considered are for a 3% overall rate increase, a 5% overall rate increase and a 7% overall rate increase. Which scenario will produce the best profit? Will the highest rate change produce unacceptably low retention values? A similar table to the previous table can be presented that compares these different scenarios.

Scenario	Premium	Retention	Loss Ratio	Operating Result
+3%	151,885	88.5%	65.92%	$4,678
+5%	153,606	87.5%	64.46%	$6,980
+7%	155,336	86.9%	63.02%	$9,283

The above table shows that the operating result projected under the +7% scenario produces the greatest profit. This is because the reduction in the retention is more than offset by the increase in premium and so the total premium is greatest under the +7% scenario. The question then remains as to whether the 86.9% retention is acceptable to the company. This decision will need to be based on the growth goals of the company. The decision process will vary from company to company.

3. Effect of rating plan changes

Most companies offer a discount for renewing a policy with one company. Is a company better off taking a deeper discount on the older more profitable policies in order to retain them for a longer period of time at a less profitable level? Or is the company better off by offering less of a discount in order to maintain a higher profitability but suffer somewhat in policyholder retention? Retention modeling allows for these types of questions to be answered. For example, consider a situation in which a company is concerned that its current rate structure is too heavily discounted for policies with longer renewal persistency. It is considering two rate changes of +5% overall. In one scenario, it is making no changes to its renewal discount. In another scenario, it is reducing the discount for policies at the sixth and greater renewals.

Scenario 1

# Times Renewed	Rate Change	Premium (000's)	Retention	Loss Ratio	Operating Result
0-2	+5%	57,064	84.8%	67.5%	856
3-5	+5%	34,050	90.4%	67.0%	681
6-8	+5%	25,153	90.5%	63.2%	1,459
9+	+5%	36,437	91.8%	55.3%	4,992
Total	+5%	152,704			7,988

Scenario 2

# Times Renewed	Rate Change	Premium (000's)	Retention	Loss Ratio	Operating Result
0-2	+3%	56,383	85.2%	68.6%	230
3-5	+3%	34,791	90.6%	67.3%	591
6-8	+7%	25,902	90.3%	61.7%	1,880
9+	+7%	36,830	91.2%	53.8%	5,598
Total	+5%	153,906			8,299

The above examples show that the improvement in profitability more than outweighed the loss of premium due to lower retention on policies at six and subsequent renewals and thus the company would be in a better financial position with the lower discount. These calculations could be run over a number of years in order to determine what the lifetime value of this group of insureds will be. The discussion of lifetime customer value is outside the scope of this paper but readers are referred to the paper by Feldblum[2] on the subject.

These three examples demonstrate the power of these modeling techniques. The decision making process that companies have historically utilized require that judgments be made regarding the impact of the rate change on retention, however, these judgments were largely based on anecdotal information regarding customer reaction to a rate change. In addition, these decisions are often based on review of a limited number of risks reviewed as part of a competitive analysis. The modeling techniques described in this paper combine a quantitative specification of how customers do react to a rate change with a review of thousands of different risks (potentially, a company's entire book of business) in order to model customer retention and the resulting impact on profitability. These techniques also allow for modeling over multiple time periods. Modeling over multiple time periods has the advantage of projecting results over more than a one-year time horizon in order to see if there will be any negative consequences of taking less than the indicated rate change in the current year. However, it also requires estimates of future competitive position and this adds more uncertainty to the model.

G. Parameterization of the Switching and Shopping Functions

The switching and shopping functions are the most difficult elements of the model to parameterize. Possible sources for this information are the following:

1. Industry studies - Industry studies of the relationship between customer loyalty and price have been done and these could be used to determine appropriate probabilities for these functions.
2. Surveys of policyholders - A company could survey its own policyholders to determine the likelihood of shopping and switching policies at various price levels. Work done by the authors has shown the following to be significant predictors of retention

   - Amount of Rate Change
   - Competitive position
   - Driver Age
   - Multi Car/Single Car
   - Existence of other policies (e.g. existence of a homeowners policy)
   - Number of Times Renewed
   - Channel (Agency company versus a direct writing company)

3. Actual company experience - by matching rate change histories with renewal data, a company could determine the likelihood of a policyholder switching based on actual company data.

---

[2] Sholom Feldblum, 1996, "Personal Automobile Premiums: An Asset Share Pricing Approach for Property/Casualty Insurance", *Proceedings of the Casualty Actuarial Society*: 190-296.

The industry data could be used as a starting point for determining these functions, but this approach has two main disadvantages. First, the industry studies will, by definition, describe average policyholder behavior for the entire industry. As mentioned earlier, the likelihood of a policyholder moving is different for a company selling policies directly as compared to one that is selling via insurance agents. Other dynamics will also cause differences in the shopping behavior of the policyholders of different companies. Thus, once a model is developed using industry data, it should be enhanced with company specific switching and shopping functions. This could be accomplished via the surveys and company experience discussed above to determine the appropriate functions. The second disadvantage of using industry studies as a starting point is that these studies were not designed for the specific purpose of developing shopping and switching functions and thus will not have all of the required data.

Once the basic shopping and switching functions have been developed, they need to be tested. This can be accomplished by running a historical rate change through the model and comparing the modeled results to the actual results to test model performance. The testing may be done at a fairly detailed level or at a more aggregate level depending on the intended use of the model. It is the experience of the authors that survey data will predict actual retention behavior of a group of policyholders quite well for certain dimensions of a company's rate structure but will under performs in other dimensions. Thus the process of back-testing the model in order to tune the shopping and switching functions to a company's actual experience is critical to the accurate performance of the model.

H. Future Enhancements

As is the case with any modeling exercise, it generally starts with a simple example and then more complexity is added in order to improve the accuracy of the basic model as the technique matures. The basic model presented here could be enhanced in the following ways.

1. Refinement of shopping and switching functions

The sample shopping and switching functions presented in this paper are very simplistic and are merely designed to demonstrate the concept of how such models could work. In reality, modeling of behavior is much more complicated than the functions presented in this paper. Proper design of these functions is critical to model development.

In order to assess the shopping function, policyholder surveys will probably be required. Analysis of a company's retention data will not give information about all of the policyholders that decide to "test the waters" for another insurer. Such a survey should be directed at a proper cross section of the policyholder base in order to assess the effect of age, gender, current premium level, etc on the propensities to shop and switch. This survey could arm the company with a detailed shopping function that would be fully indicative of its policyholder base.

The switching function could be determined through a combination of data analysis and surveys. The advent of insurance market websites provides a single source of information regarding the premium of the prior company and the premium of the new company to which the policyholder switched. Data from such market websites probably holds the most promise for determining the switching function. A survey could also be conducted on policyholders of a company that cancel voluntarily to determine the premium savings that were required in order for them to switch to a different insurer.

Alternatively, one overall survey could be conducted that would attempt to determine both the shopping and the switching functions. Similar surveys have shown that a policyholder generally underestimates the amount of premium that he actually pays, unless he has a copy of his last renewal notice in front of him,. Thus, while such surveys would provide the quickest method of determining the shopping and switching functions, they should be verified with actual data analysis for a company.

The specificity of the shopping and switching functions is also an area that is of key concern. Different policyholders will have different shopping and switching propensities. For instance, the more often a policyholder renews with his current insurer, the less likely he is to shop his policy, regardless of price changes. In addition, if a policyholder has more than one type of policy with a given company, he is less likely to shop his policy. Regional differences in shopping behavior also affect the propensity to shop. The work done by the authors has shown these to be some of the more critical factors in predicting the shopping and switching behavior. Increasing the specificity of these functions will improve the performance of the model especially when the rate change is less uniformly distributed by rate class.

2. Build brand name into the model

Individual companies' policyholder retention rates can differ quite dramatically. The reasons for these differences are varied and have an impact on an individual company's shopping and switching functions. The different levels of policyholder service offered by insurers also have an impact on retention. Companies operating within a particular market niche will have different shopping and switching functions than companies operating across the entire market. For instance, a company whose marketing strategy is to emphasize lower prices is probably attracting a more price-sensitive customer than a company that is emphasizing policyholder service. These types of different marketing strategies should be reflected in the model via shopping and switching functions that are geared towards a company's own marketing strategy. Thus, the shopping and switching functions should be company-specific. The techniques mentioned previously in this paper should be applied to develop company-specific shopping and switching functions.

3. Changing data structures to support retention modeling

Data structures in company databases could be modified in order to better track the impact of rate changes on a policyholder and thus better assess the shopping and switching functions.

One example of additional data that should be captured is premium data on policyholders that did not renew policies. Usually, these policyholders are eliminated from most "actuarial" databases and thus information allowing the assessment of the likelihood of policyholder non-renewal is lost.

A second example of an additional data element is an indicator to tell whether there was a change in policy characteristics that caused a change in premium. For instance, the policyholder may have replaced a car or had an accident in the previous year causing an increase in the premium. It is likely that the policyholder will be less likely to shop their policy if a premium increase is due to one of these reasons, however, the authors are not aware of any published studies examining this phenomenon. Perhaps a $50 increase in premium due to a rate increase is just as likely to cause someone to shop as is a $50 increase due to the policyholder receiving a moving violation. By assessing the different causes of premium increases and the associated changes in retention rates, this issue can be better understood.

4. Factors other than price producing shopping behavior

As previously mentioned, the shopping and switching functions predict customer movement based solely on rate activity. There are other factors that could cause the policyholders to shop for an alternative insurer. These include the policyholder not being satisfied with claim service or policy service throughout the year. This could be included in the retention model by estimating the number of policyholders that will require claim service or policy service throughout the year and estimating the percentage of these policyholders that will be unsatisfied with this service and thus shop for alternatives. The number of policyholders that will require claim service and policy service can be based on the company's internal data. The percentage of policyholders dissatisfied with service can be estimated through survey data.

An alternative would be to predict some amount of random shopping based on the same type of company data as mentioned in the previous paragraph. This random shopping could likely form an accurate estimate of the impact of non-price factors on retention.

5. Impact of Internet/Direct Advertising

The internet has created a new era of consumerism. The level of insurance price comparison-shopping that can be accomplished via the internet was previously unheard of. Consumers can now receive several comparable quotes through a single internet quoting service. Online comparisons are also available through individual company websites such as Progressive. Also contributing to the increase in comparison-shopping is the dramatic increase in TV and direct mail advertising urging policyholders to shop their policies. These advertising campaigns could lead to a rise in spontaneous shopping by policyholders

Retention modeling is still in a nascent stage of development. The five areas mentioned here are areas where some of the more significant work is required in order to bring retention modeling into the mainstream of pricing practice. However, as work in these areas progresses, retention modeling techniques will become more accepted and accurate.

III. Conclusion

An old adage is that "You can't stop progress". One of the ways that rate structures have progressed is that they have become more and more refined over time. The advances in computing power have allowed analysis of new data elements and provide the ability to discern patterns in the data that simply could not be recognized using single dimensional cross cuts of data that, at one time, were the norm.

While improved technology has produced tremendous advances in determining proper premiums for individual policyholders, it has made the assessment of the impact of rate activity a very difficult matter. It is now time to use the progress in computing power to address this problem as well. By using information on the elasticity of demand to estimate policyholder retention, rate structures can be determined that will produce the optimal combination of policyholder growth and profitability.

This paper has presented the framework for a modeling methodology that can be used to find this optimal combination. However, much more work in this area needs to be done. Future papers to be presented by the authors will focus on more specific shopping and switching functions and a case study of modeled results versus actual results from an actual rate change.

Statistical Learning Algorithms Applied to Automobile Insurance Ratemaking

Charles Dugas, Yoshua Bengio,
Nicolas Chapados, Pascal Vincent,
Germain Denoncourt, FCAS, FCIA, MAAA,
and Christian Fournier, FCAS, FCIA

Statistical Learning Algorithms Applied to Automobile Insurance Ratemaking

Charles Dugas, Yoshua Bengio, Nicolas Chapados and Pascal Vincent
{DUGAS,BENGIO,CHAPADOS,VINCENT}@APSTAT.COM
Apstat Technologies Inc.

Germain Denoncourt, F.C.A.S., F.C.I.A.
GDENONCOURT@ASSURANCE-ALPHA.COM
L'Alpha Compagnie d'Assurance Inc.

Christian Fournier, F.C.A.S., F.C.I.A.
CHRISTIAN.FOURNIER@ICBC.COM
Insurance Corporation of British Columbia

Abstract

We recently conducted a research project for a large North American automobile insurer. This study was the most exhaustive ever undertaken by this particular insurer and lasted over an entire year. We analyzed the discriminating power of each variable used for ratemaking. We analyzed the performance of several models within five broad categories: linear regressions, generalized linear models, decision trees, neural networks and support vector machines. In this paper, we present the main results of this study. We qualitatively compare models and show how neural networks can represent high-order nonlinear dependencies with a small number of parameters, each of which is estimated on a large proportion of the data, thus yielding low variance. We thoroughly explain the purpose of the nonlinear sigmoidal transforms which are at the very heart of neural networks' performances. The main numerical result is a statistically significant reduction in the out-of-sample mean-squared error using the neural network model and our ability to substantially reduce the median premium by charging more to the highest risks. This in turn can translate into substantial savings and financial benefits for an insurer. We hope this paper goes a long way towards convincing actuaries to include neural networks within their set of modeling tools for ratemaking.

1. Introduction

Ratemaking is one of the main mathematical problems faced by actuaries. They must first estimate how much each insurance contract is expected to cost. This conditional expected claim amount is called the *pure premium* and it is the basis of the *gross premium* charged to the insured. This expected value is conditioned on information available about the insured and about the contract, which we call the *input profile*.

Automobile insurance ratemaking is a complex task for many reasons. First of all, many factors are relevant. Taking account of each of them individually, i.e., making independence assumptions, can be hurtful (Bailey and Simon (1960)). Taking account of all interactions is intractable and is sometimes referred to as the *curse of dimensionality* (Bellman (1957)). In practice, actuarial judgment is used to discover the most relevant of these interactions and feed them explicitly to the model. Neural networks, on the other hand, are well-known for their ability to represent high-order nonlinear interactions with a small number of parameters, i.e., they can automatically detect those most relevant interactions between variables (Rumelhart et al. (1986)). We explain how and why in section 4.

A second difficulty comes from the distribution of claims: asymmetric with fat tails with a large majority of zeros and a few unreliable and very large values, i.e., an asymmetric heavy tail extending out toward high positive values. Modeling data with such a distribution is essentially difficult because *outliers*, which are sampled from the tail of the distribution, have a strong influence on parameter estimation. When the distribution is symmetric around the mean, the problems caused by outliers can be reduced using *robust* estimation techniques (Huber (1982), Hampel et al. (1986), Rousseeuw and Leroy (1987)) which basically intend to ignore or down-weight outliers. Note that these techniques do not work for an asymmetric distribution: most outliers are on the same side of the mean, so down-weighting them introduces a strong bias on its estimation: the conditional expectation would be systematically underestimated. Recent developments for dealing with asymmetric heavy-tail distributions have been made (Takeuchi et al. (2002)).

The third difficulty is due to the non-stationary nature of the relationship between explanatory variables and the expected claim amount. This has an important effect on the methodology to use, in particular with respect to the task of *model selection*. We describe our methodology in section 3.

Fourth, from year to year, the general level of claims may fluctuate heavily, in particular in states and provinces where winter plays an important role in the frequency and severity of accidents. The growth of the economy and the price of gas can also affect these figures.

Fifth, one needs sufficient computational power to develop models: we had access to a large database of $\approx 8 \times 10^6$ records, and the training effort and numerical stability of some algorithms can be burdensome for such a large number of training examples. In particular, neural networks are computationally very demanding.

Sixth, the data may be of poor quality. In particular, there may be missing fields for many records. An actuary could systematically discard incomplete records but this leads to loss of information. Also, this strategy could induce a bias if the absence of a data is not random but rather correlated to some particular feature which affects the level of risk. Alternatively one could choose among known techniques for dealing with missing values (Dempster et al. (1977), Ghahramani and Jordan (1994), Bengio and Gingras (1996)).

Seventh, once the pure premiums have been established the actuary must properly allocate expenses and a reserve for profit among the different contracts in order to obtain the gross premium level that will be charged to the insureds. Finally, an actuary must account for competitive concerns: his company's strategic goals, other insurers' rate changes, projected renewal rates and market elasticity.

In this paper, we address the task of setting an appropriate pure premium level for each contract, i.e., difficulties one through four as described above. Our goal is to compare different models with respect to their performance in that regard, i.e., how well they are able to forecast the claim level associated to each contract. We chose several models within five broad categories: linear regressions, generalized linear models (McCullagh and Nelder (1989)), decision trees (Kass (1980)), neural networks and support vector machines (Vapnik (1998a)).

The rest of the paper is organized as follows: we start by describing the mathematical criteria underlying insurance premium estimation (section 2). Our methodology is described in section 3, followed by a review of the statistical learning algorithms that we consider in this study, including our best-performing mixture of positive-output neural networks (section 4). We then highlight our most important experimental results (section 5), and in view of them conclude with an examination of the prospects for applying statistical learning algorithms to insurance modeling (section 7).

2. Mathematical Objectives

The first goal of insurance premium modeling is to estimate the *expected claim amount* for a given insurance contract for a future period (usually one year). Here we consider that the amount is 0 when no claim is filed. Let $X \in \mathbf{R}^n$ denote the customer and contract *input profile*, a vector representing all the information known about the customer and the proposed insurance policy before the beginning of the contract. Let $A \in \mathbf{R}^+$ denote the amount that the customer claims during the contract period; we shall assume that A is non-negative. Our objective is to estimate this claim amount, which is the *pure premium* p_{pure} of a given contract x:[1]

$$p_{pure}(x) = E[A|X = x]. \tag{1}$$

where $E[\cdot]$ denotes expectation, i.e. the average over an infinite population, and $E[A|X = x]$ is a conditional expectation, i.e. the average over a subset of an infinite population, comprising only the cases satisfying the condition $X = x$.

2.1 The Precision Criterion

In practice, of course, we have no direct access to the quantity (1), which we must estimate. One possible criterion is to seek the *most precise* predictor, which minimizes the expected squared error (ESE) over the unknown distribution:

$$E[(p(X) - A)^2], \tag{2}$$

where $p(X)$ is a pure premium predictor and the expectation is taken over the random variables X (input profile) and A (total claim amount). Since the true joint distribution of A and X is unknown, we can **unbiasedly** estimate the ESE performance of an estimator $p(X)$ on a data set $D_{test} = \{\langle x_i, a_i \rangle\}_{i=1}^N$, as long as this data set is not used to choose p, using the **mean-squared error** on that data set:

$$\frac{1}{N} \sum_{\langle x_i, a_i \rangle \in D_{test}} (p(x_i; \theta) - a_i)^2, \tag{3}$$

1. The pure premium is distinguished from the premium actually charged to the customer, which must account for the underwriting costs (marketing, commissions, premium tax), administrative overhead, risk and profit loadings and other costs.

where θ is the vector of parameters of the model used to compute the premiums. The vector x_i represents the i^{th} input profile of dataset D_{test} and a_i is the claim amount associated to that input profile. Thus, D_{test} is a set of N insurance policies. For each policy, D_{test} holds the input profile and associated incurred amount. We will call the data set D_{test} a **test set**. It is used only to independently assess the performance of a predictor p. To **choose** p from a (usually infinite) set of possible predictors, one uses an *estimator* L, which obtains a predictor p from a given **training set** D. Such an estimator is really a **statistical learning algorithm** (Hastie et al. (2001)), yielding a predictor $p = L_D$ for a given data set D. What we call the **squared bias** of such an estimator is $(E[A|X] - E[L_D(X)])^2$, where $E[L_D(X)]$ is the average predictor obtained by considering all possible training sets D (sampled from $P(A, X)$). It represents how far the average estimated predictor deviates from the ideal pure premium, $E[A|X]$. What we call the **variance** of such an estimator is $E[(L_D(X) - E[L_D(X)])^2]$. It represents how the particular predictor obtained with some data set D deviates from the average of predictors over all data sets, i.e. it represents the sensitivity of the estimator to the variations in the training data and is related to the classical measure of **credibility**.

Is the mean-squared error (MSE) on a test set an appropriate criterion to evaluate the predictive power of a predictor p? First one should note that if p_1 and p_2 are two predictors of $E[A|X]$, then the MSE criterion is a good indication of how close they are to $E[A|X]$, since by the law of iterated expectations,

$$\begin{aligned} E[(p_1(X) - A)^2] - E[(p_2(X) - A)^2] &= E[(p_1(X) - E[A|X])^2] \\ &\quad - E[(p_2(X) - E[A|X])^2], \end{aligned}$$

and of course the expected MSE is minimized when $p(X) = E[A|X]$.

For the more mathematically-minded readers, we show that minimizing the expected squared error optimizes simultaneously both the precision (low bias) and the variance of the estimator. We denote E_D the expectation over the training set D. The expected squared error of an estimator L_D decomposes as follows:

$$\begin{aligned} E[(A - L_D(X))^2] &= E[((A - E[A|X]) + (E[A|X] - L_D(X)))^2] \\ &= \underbrace{E[(A - E[A|X])^2]}_{\text{noise}} + E[(E[A|X] - L_D(X))^2] \\ &\quad + \underbrace{2E[(A - E[A|X])(E[A|X] - L_D(X))]}_{\text{zero}} \\ &= \text{noise} + E[((E[A|X] - E_D[L_D(X)]) + (E_D[L_D(X)] - L_D(X)))^2] \\ &= \text{noise} + E[(E[A|X] - E_D[L_D(X)])^2] + E[(E_D[L_D(X)] - L_D(X))^2] \\ &\quad \underbrace{+2E[(E[A|X] - E_D[L_D(X)])(E_D[L_D(X)] - L_D(X))]}_{\text{zero}} \\ &= \text{noise} + \underbrace{E[(E[A|X] - E_D[L_D(X)])^2]}_{\text{bias}^2} + \underbrace{E[(E_D[L_D(X)] - L_D(X))^2]}_{\text{variance}}. \end{aligned}$$

Thus, algorithms that try to minimize the expected squared error simultaneously reduce both the bias and the variance of the estimators, striking a tradeoff that minimizes the sum of both (since the remainder is the noise, which cannot be reduced by the choice of predictor). On the other hand, with a rule such as minimum-bias used with table-based methods, cells are merged up to a point where each cell has sufficient credibility, i.e., where the variance is sufficiently low. Then, once the credibility (and variance) level is set fixed, the bias is minimized. On the contrary, by targeting minimization of the expected squared error one avoids this arbitrary setting of a credibility level.

In comparison to parametric approaches, this approach avoids distributional assumptions. Furthermore, it looks for an optimal trade-off between bias and variance, whereas parametric approaches typically focus on the unbiased estimators (within a class that is associated with a certain variance). Because of the above trade-off possibility, it is always possible (with a finite data set) to improve an unbiased estimator by trading a bit of bias increase for a lot of variance reduction (Hastie et al. (2001)).

2.2 The Fairness Criterion

In insurance policy pricing, the precision criterion is not the sole part of the picture; just as important is that the estimated premiums do not systematically discriminate against specific segments of the population. We call this objective the *fairness criterion*, sometimes referred to as *actuarial fairness*. We define the *bias of the premium* $b(P)$ to be the difference between the average pure premium and the average incurred amount, in a given sub-population P of dataset D:

$$b(P) = \frac{1}{|P|} \sum_{(x_i,a_i)\in P} p(x_i) - a_i, \tag{4}$$

where $|P|$ denotes the cardinality of the sub-population P, and $p(\cdot)$ is some premium estimation function. The vector x_i represents the i^{th} input profile of sub-population P and a_i is the claim amount associated to that input profile. A possible fairness criterion would be based on minimizing the sum, over a certain set of critical sub-populations $\{P_k\}$ of dataset D, of the square of the biases:

$$\sum_{k,P_k\in D} b^2(P_k) \tag{5}$$

In the particular case where one considers all sub-populations, then both the fairness and precision criterions lead to the same optimal solution, i.e., they are minimized when $p(x_i) = E[A_i|x_i]$, $\forall i$, i.e., for every insurance policy, the premium is equal to the conditional expectation of the claim amount. The proof is given in appendix A.

In order to measure the fairness criterion, we used the following methodology: after training a model to minimize the MSE criterion (3), we define a finite number of disjoint subsets (sub-populations) of test set D: $P_k \subset D, P_k \cap P_{j\neq k} = \emptyset$, and *verify* that the absolute bias is not significantly different from zero. The subsets P_k can be chosen at convenience; in our experiments, we considered 10 subsets of equal size delimited by the deciles of the test set premium distribution. In this way, we verify that, for example, for the group of contracts with a premium between the 5th and the 6th decile, the average premium matches, within statistical significance, the average claim amount.

2.3 Penalized Training Criterion and Bias-Variance Tradeoff

Although our objective is to minimize the expected out-of-sample squared error (ESE), it does not mean that we should minimize the in-sample (*training set*) mean-squared error (MSE):

$$\frac{1}{L} \sum_{(x_\ell, a_\ell) \in D_{train}} (p(x_i;\theta) - a_i)^2$$

in order to achieve that goal. The reason for that apparent discrepancy has to do with the bias-variance trade-off in generalization error (Geman et al. (1992)), and the fundamental principles of statistical learning theory (Vapnik (1998b)). To illustrate these ideas, let us consider the simple case of linear regression, which becomes ridge regression when the training criterion is penalized. Consider a class of linear predictive functions of the input x,

$$p(x) \;=\; \sum_{i=1}^{n} \beta_i x_i.$$

Instead of minimizing the training set mean-squared error (MSE), consider the following penalized criterion:

$$\frac{1}{L} \sum_{(x_\ell, a_\ell) \in D_{train}} (p(x_i;\theta) - a_i)^2 \;+\; \lambda \sum_i \beta_i^2$$

with $\lambda \geq 0$ and a minimum achieved at $\hat{\beta}_\lambda$. Thus $\hat{\beta}_0$ is the Ordinary Least Squares estimator. This minimum is always achieved with **shrinked** solutions, i.e. $||\hat{\beta}_\lambda|| < ||\hat{\beta}_0||$ for $\lambda > 0$. Note that this solution is generally **biased**, unlike $\hat{\beta}_0$, in the sense that if the data is generated from a multivariate normal distribution, the expected value of $\hat{\beta}_\lambda$ is smaller than the true value β from the underlying distribution. Note that the set of functions effectively allowed for a solution is smaller when λ is larger.

In the case where linear regression is the proper model (normally distributed data, with output variance σ^2), and the *amount of data l is finite*, it is easy to prove that the optimal fixed value of λ (in expectation over different training sets) is

$$\lambda^* = \frac{\sigma^2}{l\;||\beta||^2}.$$

Note therefore that **the optimal model is biased** (optimal in the sense of minimizing out-of-sample error).

This example illustrates the more general principle of bias-variance trade-off in generalization error, well discussed by Geman et al. (1992). Increasing λ corresponds to "smoothing more" in non-parametric statistics (choosing a simpler function) or to the choice of a smaller capacity ("smaller" class of functions) in Vapnik's VC-theory (Vapnik (1998b)). A too large value of λ corresponds to **underfitting** (too simple model, too much bias), whereas a too small value corresponds to **overfitting**) (too complex model, too much variance). Which value of λ should be chosen? It should be the one that strikes the optimal balance between bias and variance. This is the question that **model selection** algorithms address. Fortunately, the expected out-of-sample error has a unique minimum as a function of λ (or more generally of the capacity, or complexity of the class of functions). Concerning the above formula, note that unfortunately the data is generally not normal, and σ^2 and β are both unknown, so the above formula can't be used directly to choose λ. However, using a separate held-out data set (also called a *validation set*, here), and taking advantage of that unique minimum property (which is true for any data distribution), we can quickly select a good value of λ (essentially by searching), which approximately minimizes the estimated out-of-sample error on that validation set.

Note that we arrive at the conclusion that a biased model is preferable because we set as our goal to minimize out-of-sample error. If our goal was to **discover** the underlying "truth", and if we could make very strong assumptions about the true nature of the data distribution, then the more classical statistical approach based on minimum variance unbiased estimators would be more appropriate. However, in the context of practical insurance premium estimation, we don't really know the form of the true data distribution, and we really care about how the model is going to perform in the future (at least for ratemaking).

3. Methodology

A delicate problem to guard against when applying statistical learning algorithms is that of *overfitting*. It has precisely to do with striking the right trade-off between bias and variance (as introduced in the previous section), and is known in technical terms as *capacity control*. Figure 1 illustrates the problem: the two plots show empirical data points (black dots) that we are trying to approximate with a function (solid red curve). All points are sampled from the same underlying function (dashed blue curve), but are corrupted with noise; the dashed curve may be seen as the "true" function we are seeking to estimate.

The left plot shows the result of fitting a very flexible function, i.e. a high-order polynomial in this case, to the available data points. We see that the function fits the data points perfectly: there is zero error (distance) between the red curve and each of the black dots. However, the function oscillates wildly between those points; it has not captured any of the fundamental features of the underlying function. What is happening here is that the function has mostly *captured the noise* in the data: it overfits.

The right plot, on the other hand, shows the fitting of a less flexible function, i.e. a 2nd-order polynomial, which exhibits a small error with respect to each data point. However, by not fitting the noise (because it does not have the necessary degrees of freedom), the fitted function far better conveys the structural essence of the matter.

Capacity control lies at the heart of a sound methodology for data mining and statistical learning algorithms. The goal is simple: to choose a function class flexible enough (with enough capacity) to express a desired solution, but constrained enough that it does not fit the noise in the data points. In other words, we want to avoid overfitting **and** underfitting.

Figure 2 illustrates the basic steps that are commonly taken to resolve this issue: these are not the only means to prevent overfitting, but are the simplest to understand and use.

1. The full data set is randomly split into three *disjoint* subsets, respectively called the training, validation, and test sets.

2. The training set is used to fit a model with a chosen initial capacity.

3. The validation set is used to *evaluate the performance* of that fitted function, on **different data points** than used for the fitting. The key here is that a function

overfitting the training set will exhibit a low performance on the validation set, if it does not capture the underlying structure of the problem.

4. Depending on the validation set performance, the capacity of the model is adjusted (increased or reduced), and a new training phase (step 2) is attempted. This training–validation cycle is repeated multiple times and the capacity that provides the best validation performance is chosen.

5. Finally, the performance of the "ultimate" function (that coming out of the validation phase) is evaluated on data points never used previously—those in the test set—to give a completely unbiased measure of the performance that can be expected when the system is deployed in the field. This is called **generalization performance**.

4. Models

In this section, we describe various models that have been implemented and used for the purpose of ratemaking. We begin with the simplest model: charging a flat premium to every insured. Then, we gradually move on towards more complex models.

4.1 Constant Model

For benchmark evaluation purposes, we implemented the constant model. This consists of simply charging every single insurance policy a flat premium, regardless of the associated variable values. The premium is the mean of all incurred amounts as it is the constant value that minimizes the mean-squared error.

$$p(x) \quad = \quad \beta_0, \tag{6}$$

where β_0 is the mean and the premium $p(x)$ is independent of the input profile x. In figure 3, the constant model is viewed as a flat line when the premium value is plotted against one of the input variables.

4.2 Linear Model

We implemented a linear model which consists of a set of coefficients, one for each variable plus an intercept value, that minimize the mean-squared error,

$$p(x) \quad = \quad \beta_0 + \sum_{i=1}^{n} \beta_i x_i. \tag{7}$$

Figure 4 illustrates a linear model where the resulting premiums form a line, given one input variable value. With a two dimensional input variable space, a plane would be drawn. In higher dimension, the corresponding geometrical form is referred to as a hyper-plane.

There are two main ways to control the capacity of linear models when in presence of noisy data:

- using a subset of input variables; this directly reduces the number of coefficients (but choosing the best subset introduces another level of choice which is sometimes detrimental).

- penalizing the norm of the parameters (in general, intercept parameter β_0 is excluded from the penalty term); this is called *ridge regression* in statistics, and *weight decay* in the neural networks community. This was the main method used to control capacity of the linear model in our experiments (see above subsection 2.3).

It should be noted that the premium computed with the linear model can be negative (and negative values are indeed sometimes obtained with the trained linear models). This may happen even if there are no negative amounts in the data, simply because the model has no built-in positivity constraint (unlike the GLM and the softplus neural network described below).

4.3 Table-based methods

These more traditional ratemaking methods rely mainly on a classification system, base rates and relativities. The target function is approximated by constants over regular (finite) intervals. As shown on the figure, this gives rise to a typical staircase-like function, where each level of the staircase is given by the value in the corresponding cell in the table. A common refinement in one dimension is to perform a linear interpolation between neighboring cells, to smooth the resulting function somewhat. The table is not limited to two variables; however, when adding a new variable (dimension), the number of cells increases by a factor equal to the number of discretization steps in the new variable.

In order to use table-based methods to estimate a pure premium, find a certain number of variables deemed useful for the prediction, and discretize those variables if they are continuous. To fill out the table, compute over a number of years (using historical data) the total incurred claim amount for all customers whose profiles *fall within a given cell* of the table, and average the total within that cell. This gives the pure premium associated with each cell of the table.

Assuming that the i^{th} variable of profile x belongs to the j^{th} category, we obtain,

$$p(x) = \beta_0 \prod_{i=1}^{m} \beta_{i,j} + \sum_{i=m+1}^{n} \beta_{i,j}, \qquad (8)$$

where $\beta_{i,j}$ is the relativity for the j^{th} category of the i^{th} variable and β_0 is the standard premium. We consider the case where the first m factors are multiplicative and the last $n-m$ factors are additive.

The formula above assumes that all variables have been analyzed individually and independently. A great deal of effort is often put in trying to capture dependencies (or interactions) between some variables and to encode them into the premium model.

An extension of the above is to multiplicatively combine multiple tables associated to different subsets of variables. This is in effect a particular form of generalized linear model (see below), where each table represents the interdependence effects between some variables.

4.4 Greedy Multiplicative Model

Greedy learning algorithms "grow" a model by gradually adding one "piece" at a time to the model, but keeping the already chosen pieces fixed. At each step, the "piece" that is most helpful to minimize the training criterion is "added" to the model. This is how decision trees are typically built. Using the validation set performance we can decide when to stop adding pieces (when the estimated out-of-sample performance starts degrading).

The GLM described in the next section is a multiplicative model because the final premium function can be seen as a product of coefficients associated with each input variable. The basic idea of the Greedy Multiplicative Model is to add one of these multiplicative coefficients at a time. At each step, we have to choose one among the input variables. We choose the variable which would reduce most the training MSE. The coefficient for that component is easily obtained analytically by minimizing the MSE when all the previously obtained coefficients are kept fixed.

In the tables we use the short-hand name "CondMean" for this model because it estimates and combines many conditional means. Note that like the GLM, this model provides positive premiums.

4.5 Generalized Linear Model

Bailey and Simon (1960) introduced generalized linear models (GLM) to the actuarial community four decades ago. More recently, Brown (1988), Holler et al. (1999), Murphy et al. (2000) conducted experiments using such models. GLMs, at their roots, are simple linear models that are composed with a fixed nonlinearity (the so-called *link function*); a commonly-used link function is simply the exponential function e^x. GLMs (with the exponential link) are sometimes used in actuarial modeling since they naturally represent *multiplicative effects*, for example risk factors whose effects should combine multiplicatively rather than additively. They are attractive since they incorporate *problem-specific knowledge* directly into the model. These models can be used to obtain a pure premium, yielding such a formula,

$$p(x) = \exp\left(\beta_0 + \sum_{i=1}^{n} \beta_i x_i\right), \tag{9}$$

where, the exponentiation ensures that the resulting premiums are all positive. In figure 5, we can see that the model generates an exponential function in terms of the input variable.

In their favor, GLMs are quite easy to estimate[2], have interpretable parameters, can be associated to parametric noise models, and are not so affected when the number of explanatory variables increases, as long as the number of observations used in the estimation remains sufficient. Unfortunately, they are fairly restricted in the shape of the functions they can estimate.

The capacity of a GLM model can be controlled using the same techniques as those mentionned above (4.2) in the context of linear models. Again, note that the GLM always provides a positive premium.

4.6 CHAID decision trees

Decision trees split the variable space in smaller subspaces. Any input profile x fits into one and only one of those subspaces called *leaves*. To each leaf is associated a different premium

2. We have estimated the parameters to minimize the mean-squared error, but other training criteria have also been proposed in the GLM literature and this could be the subject of further studies.

level,

$$p(x) = \sum_{i=1}^{n_l} \mathbf{I}_{\{x \in l_i\}} \beta_i, \tag{10}$$

where $\mathbf{I}_{\{x \in l_i\}}$ is an indicator function equal to 1 if and only if x belongs to the i^{th} leaf. In that case, $\mathbf{I}_{\{x \in l_i\}} = 1$ and $p(x) = \beta_i$. Otherwise, $\mathbf{I}_{\{x \in l_i\}}$ is equal to zero, meaning x belongs to another leaf. The number of leaves is n_l. The premium level β_i is set equal to the average incurred amount of the policies for which the profile x belongs to the i^{th} leaf. In figure 6, the decision tree is viewed as generating a piecewise constant function. The task of the decision tree is to choose the "best" possible partition of the input variable space.

The basic way in which capacity is controlled is through several hyper-parameters: minimum population in each leaf, minimum population to consider splitting a node, maximum height of the decision tree and, in the case of CHAID decision trees (Kass (1980)), Chi-square statistic threshold value.

4.7 Combination of CHAID and Linear Model

This model is similar to the previous one except that, in each leaf, we have replaced the associated constant premium value with a linear regression. Each leaf has its own set of regression coefficients. There are thus n_l different linear regressions of $n + 1$ coefficients each.

$$p(x) = \sum_{i=1}^{n_l} \mathbf{I}_{\{x \in l_i\}} \left(\beta_{i,0} + \sum_{j=1}^{n} \beta_{i,j} x_j \right). \tag{11}$$

Each linear regression was fit to minimize the mean-squared error on the training cases that belong to its leaf. For reasons that are clear in the light of learning theory, a tree used in such a combination should have less leaves than an ordinary CHAID tree. In our experiments we have chosen the size of the tree based on the validation set MSE.

In these models, capacity is controlled with the same hyper-parameters as CHAID, and there is also the question of finding the right *weight decay* for the linear regression. Again, the validation set is used for this purpose.

4.8 Ordinary Neural Network

Ordinary neural networks consist of the clever combination and simultaneous training of a group of units or *neurons* that are individually quite simple. Figure 8 illustrates a typical *multi-layer feedforward architecture* such as the ones that were used for the current project.

We describe here the steps that lead to the computation of the final output of the neural network. First, we compute a series of linear combinations of the input variables:

$$v_i = \alpha_{i,0} + \sum_{j=1}^{n} \alpha_{i,j} x_j, \tag{12}$$

where x_j is the j^{th} out of n variables, $\alpha_{i,0}$ and $\alpha_{i,j}$ are the slope intercept and the weights of the i^{th} linear combination. The result of the linear combination, v_i, is often referred to as the *level of activation* in analogy to the neurons in the brain.

Then, a non-linear transform (called a *transfer function*) is applied to each of the linear combinations in order to obtain what are called the *hidden units*. We used the hyperbolic tangent function:

$$\begin{aligned} h_i &= \tanh(v_i) \\ &= \frac{e^{v_i} - e^{-v_i}}{e^{v_i} + e^{-v_i}}, \end{aligned} \tag{13}$$

where h_i is the i^{th} hidden unit. The use of such a transfer function with infinite expansion in its terms has an important role in helping the neural network capture nonlinear interactions and this is the subject of subsection 4.9.

Finally, the hidden units are linearly combined in order to compute the final output of the neural network:

$$p(x) = \beta_0 + \sum_{i=1}^{n_h} \beta_i h_i, \tag{14}$$

where $p(x)$ is the premium computed by the neural network, n_h is the number of hidden units and β_0 and β_i are the slope intercept and the weights of the final linear combination.

Put all together in a single equation, we obtain:

$$p(x) = \beta_0 + \sum_{i=1}^{n_h} \beta_i \tanh\left(\alpha_{i,0} + \sum_{j=1}^{n} \alpha_{i,j} x_j\right). \tag{15}$$

Figure 9 depicts a smooth non-linear function which could be generated by a neural network.

The number of hidden units (n_h above) plays a crucial role in our desire to control the capacity of the neural network. If we choose a too large value for n_h, then the number of parameters of the model increases and it becomes possible, during the parameter optimization phase, for the neural network to model noise or spurious relationships present in the data used for optimzation but that do not necessarily exist in other datasets. Conversely, if n_h is set to a low value, the number of parameters could be too small and the neural network would not capture all of the relevant interactions in order to properly compute the premiums. Choosing the optimal number of hidden units is an important part of modelling with neural networks. Another technique for controlling the capacity of a neural network is to use weight decay, i.e., a penalized training criterion as described in subsection 2.3 that limits the size of the parameters of the neural network.

Choosing the optimal values for the parameters is a complex task and out of the scope of this paper. Many different optimization algorithms and refinements have been suggested (Bishop (1995), Orr and Müller (1998)) but in practice, the simple stochastic gradient descent algorithm is still very popular and usually gives good performance.

Note that like the linear regression, this model can potentially yield negative premiums in some cases. We have observed much fewer such cases than with the linear regression.

4.9 How can Neural Networks Represent Nonlinear Interactions?

For the more mathematically-minded readers, we present a simple explanation of why neural networks are able to represent **nonlinear interactions between the input variables**. To simplify, suppose that we have only two input variables, x_1 and x_2. In classical linear regression, a common trick is to include fixed nonlinear combinations among the regressors, such as $x_1^2, x_2^2, x_1x_2, x_1^2x_2, \ldots$ However, it is obvious that this approach adds exponentially many terms to the regression, as one seeks higher powers of the input variables.

In contrast, consider a single hidden unit of a neural network, connected to two inputs. The adjustable network parameters are named, for simplicity, α_0, α_1 and α_2. A typical function computed by this unit is given by

$$\tanh(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2).$$

Here comes the central part of the argument: performing a Taylor series expansion of $\tanh(y + \alpha_0)$ in powers of y, and letting $\alpha_1 x_1 + \alpha_2 x_2$ stand for y, we obtain (where $\beta \equiv \tanh \alpha_0$),

$$\begin{aligned}
&\tanh(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2) = \\
&\quad \beta + (1-\beta^2)(\alpha_1 x_1 + \alpha_2 x_2) + (-\beta + \beta^3)(\alpha_1 x_1 + \alpha_2 x_2)^2 + \\
&\quad \left(-\tfrac{1}{3} + \tfrac{4\beta^2}{3} - \beta^4\right)(\alpha_1 x_1 + \alpha_2 x_2)^3 + \\
&\quad \left(\tfrac{2\beta}{3} - \tfrac{5\beta^3}{3} + \beta^5\right)(\alpha_1 x_1 + \alpha_2 x_2)^4 + O(\alpha_1 x_1 + \alpha_2 x_2)^5.
\end{aligned}$$

In fact the number of terms is infinite: the nonlinear function computed by this single hidden unit includes **all powers of the input variables**, but they cannot all be independently controlled. The terms that will ultimately stand out depend on the coefficients α_0, α_1, and α_2. Adding more hidden units increases the flexibility of the overall function computed by the network: each unit is connected to the input variables with its own set of coefficients, thereby allowing the network to capture as many (nonlinear) relationships between the variables as the number of units allows.

The coeffients linking the input variables to the hidden units can also be interpreted in terms of **projections** of the input variables. Each set of coefficients for one unit represents a direction of interest in input space. The values of the coefficients are found during the **network training** phase using iterative nonlinear optimization algorithms.

4.10 Softplus Neural Network

This new type of model was introduced precisely to make sure that positive premiums are obtained. The softplus function was recently introduced in Dugas et al. (2001) as a means to model a convex relationship between an output and one of its inputs. We modified the neural network architecture and included a softplus unit as a final transfer function. Figure 10 illustrates this new architecture we have introduced for the purpose of computing insurance premiums. The corresponding formula is as such:

$$p(x) = F\left(\beta_0 + \sum_{i=1}^{n_h} \beta_i \tanh\left(\alpha_{i,0} + \sum_{j=1}^{n} \alpha_{i,j} x_j\right)\right), \tag{16}$$

where $F(\cdot)$ is the softplus function which is actually simply the primitive (integral) function of the "sigmoid" function. Thus

$$F(y) = \log(1 + e^y). \tag{17}$$

The softplus function is convex and monotone increasing w.r.t. to its input and always strictly positive. Thus, as can be seen in Figure 11, this proposed architecture leads to strictly positive premiums.

In preliminary experiments we have also tried to use the exponential function (rather than the softplus function) as the final transfer function. However we obtained poor results due to difficulties in the optimization (probably due to the very large gradients obtained when the argument of the exponential is large).

The capacity of the softplus neural network is tuned just like that of an ordinary neural network. Note that this kind of neural network architecture is not available in commercial neural network packages.

4.11 Regression Support Vector Machine

Support Vector Machines (SVM) have recently been introduced as a very powerful set of non-parametric statistical learning algorithms (see Vapnik (1998a) and Schölkopf et al. (1998)). They have been very successful in classification tasks, but the framework has also been extended to perform regression. Like other *kernel methods* the class of functions has the following form:

$$p(x) = \sum_i \alpha_i K(x, x_i) \tag{18}$$

where x_i is the input profile associated with one of the training records, and α_i is a scalar coefficient that is learned by the algorithm and K is a *kernel function* that satisfies the Mercer condition (Cristianini and Shawe-Taylor (2000)):

$$\int_C K(x, y)\, g(x)\, g(y)\, dx\, dy \geq 0 \tag{19}$$

for any square integrable function $g(x)$ and compact subset C of $\mathbf{R}^n$. This Mercer condition ensures that the kernel function can be represented as a simple dot product:

$$K(x, y) = \phi(x) \cdot \phi(y) \tag{20}$$

where $\phi()$ is a function that projects the input profile vector into a (usually very) high-dimensional "feature" space, usually in a nonlinear fashion. This leads us, to a simple expression for the premium function:

$$\begin{aligned} p(x) &= \sum_i \alpha_i \phi(x) \cdot \phi(x_i) \\ &= \phi(x) \cdot \left(\sum_i \alpha_i \phi(x_i) \right) \\ &= w \cdot \phi(x). \end{aligned} \tag{21}$$

Thus, in order to compute the premium, one needs to project input profile x in its feature space and compute a dot product with vector w. This vector w depends only on a certain number of input profiles from the training dataset and their associated coefficients. These input profiles are referred to as the *support vectors* and have been selected, along with their associated coefficients by the optimization algorithm.

SVMs have several very attractive theoretical properties, including the fact that an exact solution to the optimization problem of minimizing the training criterion can be found, and the capacity of the model is automatically determined from the training data. In many applications, we also find that most of the α_i coefficients are zero.

However, in the case of insurance data, an important characteristic of regression SVMs is that they are NOT trained to minimize the training MSE. Instead they minimize the following criterion:

$$J = \frac{1}{2}||w||^2 + \lambda \sum_i |a_t - p(x_i)|_\epsilon \tag{22}$$

where $|e|_\epsilon = \max(0, |e| - \epsilon)$, λ and ϵ trade-off accuracy with complexity, a_i is the observed incurred claim amount for record i, x_i is the input profile for record i, and the vector w is defined in terms of the α_i coefficients above. It can therefore be seen that this algorithm minimizes something close to the ***absolute value of the error*** rather than the *squared error*. As a consequence, the SVM tends to find a solution that is close to the *conditional median* rather than the *conditional expectation*, the latter being what we want to evaluate in order to set the proper value for a premium. Furthermore, note that the insurance data display a highly asymmetric distribution, so the median and the mean are very different. In fact, the conditional median is often exactly zero. Capacity is controlled through the ϵ and λ coefficients.

4.12 Mixture Models

The *mixture of experts* has been proposed Jacobs et al. (1991) in the statistical learning litterature in order to decompose the learning problem, and it can be applied to regression as well as classification. The conditional expectation is expressed as a linear combination of the predictions of **expert models**, with weights determined by a **gater model**. The *experts* are specialized predictors that each estimate the pure premium for insureds that belong to a certain class. The *gater* attempts to predict to which class each insured belongs, with an estimator of the conditional probability of the class given the insured's input profile. For a mixture model, the premium can be expressed as

$$p(x) = \sum_c p(c|x) p_c(x) \tag{23}$$

where $p(c|x)$ is the probability that an insured with input profile x belongs to class c. This value is determined by the gater model. Also, $p_c(x)$ is the premium, as computed by the expert model of class c, associated to input profile x.

A trivial case occurs when the class c is deterministically found for any particular input profile x. In that case, we simply split the training database and train each expert model on a subset of the data. The gater then simply assigns a value of $p_c(x) = 1$ if c is the appropriate

Model	Train MSE	Valid MSE	Test MSE
Constant	56.1108	56.5744	67.1192
Linear	56.0780	56.5463	67.0909
GLM	56.0762	56.5498	67.0926
NN	56.0706	56.5468	67.0903
Softplus NN	56.0704	56.5480	67.0918
CHAID	56.0917	56.5657	67.1078
CondMean	56.0827	56.5508	67.0964
Mixture	56.0743	**56.5416**	**67.0851**

Table 1: Comparison between the main models, with MSE on the training set, validation set, and test sets. The MSE is with respect to claim amounts and premiums expressed in thousand of dollars.

class for input profile x and zero otherwise. This is in fact fundamentally equivalent to other techniques such as decision trees or table-based methods. A more general and powerful approach is to have the learning algorithm discover a relevant decomposition of the data into different regions of the input space which then become the classes and are encoded in the gater model. In that case, both the gater and the experts are trained together.

In this study both the experts and the gater are softplus neural networks, but any other model can be used. In Figure 12, we schematically illustrate a mixture model as the one that was used in the framework of this project.

5. Experimental Results

5.1 Mean-Squared Error Comparisons

Table 1 summarizes the main results concerning the comparison between different types of statistical machine learning algorithms. All the models have been trained using the same input profile variables. For each insurance policy, a total of 33 input variables were used and the total claims for an accident came from five main coverages: bodily injury, accident benefit, property damage, collision and comprehensive. Two other minor coverages were also included: death benefit and loss of use. In the table, *NN* stands for neural network, *GLM* for generalized linear model, and *CondMean* for the Greedy Multiplicative Model. The MSE on the training set, validation set and test set are shown for all models. The MSE is with respect to claim amounts and premiums expressed in **thousand of dollars**. The model with the lowest MSE is the "Mixture model", and it is the model that has been selected for the comparisons with the insurer's current rules for determining insurance premiums to which we shall refer as the *Rule-Based Model.*

One may wonder from the previous table why the MSE values are so similar across various models for each dataset and much different across the datasets. In particular, all models perform much worse on the testset (in terms of their MSE). There is a very simple explanation. The maximum incurred amount on the test set and on the validation set is around 3 million dollars. If there was one more such large claim in the test set than in

Model #1	Model #2	Mean	Standard Error	Z	p-value
Constant	Mixture	3.40709e-02	3.32724e-03	10.240000	**0**
Linear	Mixture	5.82350e-03	1.32211e-03	4.404700	**5.29653e-06**
GLM	Mixture	7.54013e-03	1.15020e-03	6.555500	**2.77278e-11**
NN	Mixture	5.23885e-03	1.41112e-03	3.712540	**1.02596e-04**
Softplus NN	Mixture	6.71066e-03	1.09351e-03	6.136810	**4.20977e-10**
CHAID	Mixture	2.35891e-02	2.57762e-03	9.151520	**0**

Table 2: Statistical Comparison Between Different Learning Models and the Mixture Model. The p-value is for the null hypothesis of no difference between Model #1 and the best mixture model. Note that ALL differences are statistically significant.

the validation set, one would expect the test MSE (calculated for premiums and amounts in thousand of dollars) to be larger by about 7 (these are in units of squared thousand dollars). Thus a difference of 11 can easily be explained by a couple of large claims. This is a reflection of the very thick right-hand tail of the incurred amount distribution (whose standard deviation is only of about 8 thousand dollars). Conversely, this also explains why all MSE are very similar across models for one particular dataset. The MSE values are all mainly driven by very large claims which no model could reliably forecast (no model could lead the insurer to charge one million dollars to a particular insured!) Consequently, truly significant differences between model performances are shadowed by the effect of very large claims on the MSE values. Although the differences between model performance are relatively small, we shall see next that careful statistical analysis allows us to discover that some of them are significant.

Figure 13 illustrates graphically the results of the table, with the models ordered according to the validation set MSE. One should note that within each class of models the capacity is tuned according to the performance on the validation set. On the test and validation sets, the Mixture model dominates all the others. Then come the ordinary neural network, linear model, and softplus neural network. Only slightly worse are the GLM and CondMean (the Greedy Multiplicative model). CHAID fared poorly on this dataset. Note that the CHAID + linear model described in section 4.7 performed worse than ordinary CHAID. Finally, the constant model is shown as a baseline (since it corresponds to assigning the same premium to every 1-year policy). It is also interesting to note from the figure that the model with the lowest training MSE is not necessarily the best out-of-sample (on the validation or test sets). The SVM performance was appalling and is not shown here; it did much worse than the constant model, because it is aiming for the conditional median rather the conditional expectation, which are very different for this kind of data.

Table 2 shows a statistical analysis to determine whether the differences in MSE between the Mixture model and each of the other models are significant. The *Mean* column shows the difference in MSE with the Mixture model. The next column shows the *Standard Error* of that mean. Dividing the mean by the standard error gives Z in the next column. The last column gives the *p-value* of the null hypothesis according to which the true expected squared errors for both models are the same. Conventionally, a value below 5% or 1% is interpreted as indicating a significant difference between the two models. The p-values

Model #1	Model #2	Mean	Standard Error	Z	p-value
Constant	CHAID	1.04818e-02	2.62416e-03	**3.994350**	**3.24368e-05**
CHAID	GLM	1.60490e-02	2.15109e-03	**7.460850**	**4.29823e-14**
GLM	Softplus NN	8.29468e-04	8.94764e-04	0.927025	1.76957e-01
Softplus NN	Linear	8.87159e-04	1.08802e-03	0.815392	2.07424e-01
Linear	NN	5.84651e-04	1.33283e-03	0.438653	3.30457e-01
NN	Mixture	5.23885e-03	1.41112e-03	**3.712540**	**1.02596e-04**

Table 3: Statistical Comparison Between Pairs of Learning Models. Models are ordered from worst to best. The test is for comparing the sum of MSEs. The p-value is for the null hypothesis of no difference between Model #1 and Model #2.

and Z corresponding to significant differences are highlighted. Therefore the differences in performance between the mixture and the other models are all statistically significant. As mentionned above, the MSE values are very much affected by large claims. Does such a sensitivity to very large claims make statistical comparisons between models incorrect? No. Fortunately all the comparisons are performed on **paired data** (the squared error for each individual policy), which cancel out the effect of these very large claims (since, for these special cases, the squared error will be huge for all models and of very close magnitude)

Table 3 has similar columns, but it provides a comparison of pairs of models, where the pairs are consecutive models in the order of validation set MSE. What can be seen is that the ordinary neural network (NN) is significantly better than the linear model, but the latter, the softplus neural network and GLM are not statistically distinguishable. Finally GLM is significantly better than CHAID, which is significantly better than the constant model. Note that although the softplus neural network alone is not doing very well here, it is doing very well within the Mixture model (it is the most successful one as a component of the mixture). The reason may be that within the mixture, the parameter estimation for model of the low incurred amounts is not polluted by the very large incurred amounts (which are learned in a separate model).

5.2 Evaluating Model Fairness

Although measuring the predictive accuracy—as done with the MSE in the previous section—is a useful first step in comparing models, it tells only part of the story. A given model could appear significantly better than its competitors *when averaging over all customers*, and yet perform miserably when restricting attention to a subset of customers.

We consider a model to be *fair* if different cross-sections of the population are not significantly biased against, compared with the overall population. Model fairness implies that the average premiums within each sub-group should be statistically close to the average incurred amount within that sub-group.

Obviously, it is nearly impossible to correct for any imaginable bias since there are *many* different criteria to choose from in order to divide the population into subgroups; for instance, we could split according to any single variable (e.g. premium charged, gender, rate group, territory) but also *combinations of variables* (e.g. all combinations of gender and

	Mixture Model		Rule-Based Model	
	Low	High	Low	High
Subgroup 1	50.81	166.24	139.27	245.0145
Subgroup 2	166.24	214.10	245.01	297.0435
Subgroup 3	214.10	259.74	297.04	336.7524
Subgroup 4	259.74	306.26	336.75	378.4123
Subgroup 5	306.27	357.18	378.41	417.5794
Subgroup 6	357.18	415.93	417.58	460.2658
Subgroup 7	415.93	490.34	460.26	507.0753
Subgroup 8	490.35	597.14	507.07	554.2909
Subgroup 9	597.14	783.90	554.29	617.1175
Subgroup 10	783.90	4296.78	617.14	3095.7861

Table 4: Subgroups used for evaluating model fairness, for the Mixture and Rule-Based Models. The lowest and highest premiums in the subgroups are given. Each subgroup contains the same number of observations, $\approx 28,000$.

territory, etc.). Ultimately, by combining enough variables, we end up identifying individual customers, and give up any hope of statistical reliability.

As a first step towards validating models and ensuring fairness, we choose the subgroups corresponding to the location of the deciles of the premium distribution. The i-th decile of a distribution is the point immediately above $10i\%$ of the individuals of the population. For example, the 9-th decile is the point such that 90% of the population come below it. In other words, the first subgroup contains the 10% of the customers who are given the lowest premiums by the model, the second subgroup contains the range 10%–20%, and so on.

The subgroups corresponding to the Mixture Model (the proposed model) differ slightly from those in the *Rule-Based Model* (the insurer's current rules for determining insurance premiums). Since the premium distribution for both models is not the same. The subgroups used for evaluating each model are given in Table 4. Since they correspond to the deciles of a distribution, all the subgroups contain approximately the same number of observations ($\approx 28,000$ on the 1998 test set).

The bias within each subgroup appears in Figure 14. It shows the average difference between the premiums and the incurred amounts, within each subgroup (recall that the subgroups are divided according to the premiums charged by each model, as per Table 4). A positive difference implies that the average premium within a subgroup is higher than the average incurred amount within the same subgroup. 95% confidence intervals on the mean difference are also given, to assess the statistical significance of the results.

Since subgroups for the two models do not exactly represent the same customers, we shall refrain from directly comparing the two models on a given subgroup. We note the following points:

- For most subgroups, the two models are being fair: the bias is usually not statistically significantly different from zero.

- More rarely, the bias is significantly positive (the models overcharge), but never significantly negative (models undercharge).
- The only subgroup for which both models undercharge is that of the highest-paying customers, the 10-th subgroup. This can be understood, as these customers represent the highest risk; a high degree of uncertainty is associated with them. This uncertainty is reflected in the huge confidence intervals on the mean difference, wide enough not to make the bias significantly different from zero in both cases. (The bias for the Rule-Based Model is nearly significant.)

From these results, we conclude that both models are usually fair to customers in all premium subgroups. A different type of analysis could also be pursued, asking a different question: "In which cases do the Mixture and the Rule-Based Models differ the most?" We address this issue in next section.

5.3 Comparison with Current Premiums

For this comparison, we used the best (on the validation set) *Mixture model* and compare it on the test data of 1998 against the insurer's Rule-Based Model. Note that for legislative reasons, the Rule-Based Model did not use the same variables as the proposed Mixture Model.

Histograms comparing the distribution of the premiums between the Rule-Based and the Mixture models appear in Figure 15. We observe that the premiums from the Mixture model is smoother and exhibits fatter tails (more probability mass in the right-hand side of the distribution, far from the mean). The Mixture model is better able to recognize risky customers and impose an appropriately-priced premium.

This observation is confirmed by looking at the distribution of the *premium difference* between the Rule-Based and Mixture models, as shown in Figure 16.

We note that this distribution is extremely skewed to the left. This means that for some customers, the Rule-Based model considerably under-charges with respect to the Mixture model. Yet, the median of the distribution is above zero, meaning that the *typical customer pays more under the Rule-Based model than under the Mixture model.* At the same time, the Mixture model achieves better prediction accuracy, as measured by the *Mean-Squared Error* (MSE) of the respective models, all the while remaining fair to customers in all categories.

Our overriding conclusion can be stated plainly: the Mixture model correctly charges less for typical customers, and correctly charges more for the "risky" ones. This may be due in part to the use of more variables, and in part to the use of a statistical learning algorithm which is better suited to capturing the dependencies between many variables.

6. Taking Advantage of Increased Discriminant Power

Neural networks have been known to perform well in tasks where discrimination is an important aspect of the task at hand and this has lead to many commercially successful application of these modelling tools (Keller (1997)). We have shown that, when applied properly while taking into account the particulars of insurance data, that ability to discriminate is also revealed with insurance data. When applied to automobile insurance ratemaking, they allow us to identify more precisely the true risk associated to each insured.

6.1 Application to Underwriting

Completely changing the rate structure of an insurer can be a costly enterprise, in particular when it involves significant changes in the computer systems handling transactions, or the relations with brokers. There are other applications of systems which improve the estimation of pure premium. In the States of Massachusetts, New Hampshire, North Carolina and the provinces of Québec and Ontario, improved discrimination can be used for the purpose of choosing the risks to be ceeded to the *risk-sharing pools* (actual terminology varies from one jurisdiction to another). According to these pool plans, an insurer can choose to ceed a portion of its book of business (5%-10%) to the pool by paying a portion of the gross premium that was charged to the insured. Then, in case an accident occurs, the pool assumes all claim payments. The losses in the pool are then shared between the insurers. Thus, for an insurer, the goal is to identify the risks that have been underpriced the most (i.e. those for which the difference between the true risk and the current premium is largest). There are a few reasons why such inadequately rated risks can be identified:

- legislation related to ratemaking could be more restrictive than the one that pertains to the risk-sharing pool,
- strategic marketing concerns may have forced the insurer to underprice a certain part of its book of business and,
- other concerns may not allow the insurer to use highly discriminative models for the purpose of ratemaking.

Better discrimination of risks can be used to identify, with higher confidence, the worst risks in a population and therefore improve the performance of an insurance company's underwriting team.

6.2 Application to Ratemaking and Marketing

The greatest benefit from an improved estimation of pure premium derives by considering its application to ratemaking. The main reason for these benefits is that a more discriminant predictor will identify a group of insureds that are significantly undercharged and a (much larger) group that is significantly overcharged. Identifying the *undercharged* will yield **increased profits**: increasing their premiums will either directly increase revenues (if they stay) or reduce underwriting losses (if they switch to another insurer). The advantage of identifying the insured profiles which correspond to *overcharged* premiums can be coupled with a marketing strategy in order attract new customers and **increase market share**, a very powerful engine for increased profitability of the insurer (because of the fixed costs being shared by a larger number of insureds).

To decide on the appropriate change in premium, one also needs to consider market effects. An elasticity model can be independently developped in order to characterize the relation between premium change and the probability of losing current customers or acquiring new customers. A pure premium model such as the one described in this paper can then be combined with the elasticity model, as well as pricing constraints (e.g. to prevent too much rate dislocation in premiums, or to satisfy some jurisdiction's regulations), in order to obtain a function that "optimally" chooses for each insured profile an appropriate change in gross premium, in order to maximize a financial criterion. We have successfully tested such an idea and the detailed analysis of these results will be the subject of a further paper.

7. Conclusion

In this paper, we have argued in favor of the use of statistical learning algorithms such as neural networks for automobile insurance ratemaking. We have described various candidate models and compared them qualitatively and numerically. We have found that the selected model has significantly outperformed all other models, including the current premium structure. We believe that their superior performance is mainly due to their ability to capture high-order dependencies between variables and to cope with the fat tail distribution of the claims. Other industries have adopted statistical learning algorithms in the last decade and we have shown them to be suited for the automobile insurance industry as well.

Appendix A. Proof of the equivalence of the fairness and precision criterions

In this section, we show that, when all subpopulations are considered to evaluate fairness, the precision criterion and the fairness criterion, as they were defined in section 2, both lead to the same premium function.

Theorem 1 *The premium function which maximizes precision (in the sense of equation 2) also maximizes fairness (in the sense of equation 5, when all subpopulations are considered), and it is the only one that does maximize it.*

Proof:

Let P be a subset of the domain of input profiles. Let q be a premium predictor function. The bias in P is defined by

$$b_q(P) = \frac{1}{|P|} \sum_{(x_i,a_i)\in P} (q(x_i) - a_i).$$

Let $F_q = -E[\sum_P b_q(P)^2]$ be the expected "fairness" criterion using premium function q, to be maximized (by choosing q appropriately).

Let $p(x) = E[a|x]$ be the optimal solution to the precision criterion, i.e. the minimizer of

$$E[(p(X) - A)^2].$$

Consider a particular population P. Let $q(P)$ denote the average premium for that population using the premium function $q(x)$,

$$q(P) = \frac{1}{|P|} \sum_{(x_i,a_i)\in P} q(x_i)$$

and similarly, define $a(P)$ the average claim amount for that population,

$$a(P) = \frac{1}{|P|} \sum_{(x_i,a_i)\in P} a_i$$

Then the expected squared bias for that population, using the premium function q, is

$$E[b_q(P)^2] = E[(q(P) - a(P))^2]$$

which is minimized for any q such that $q(P) = E[a(P)]$.

Note in particular that the optimal ESE solution, p, is such a minimizer of F_q, since

$$p(P) = \frac{1}{|P|} \sum_{(x_i,a_i) \in P} E[a_i|x_i] = E[\frac{1}{|P|} \sum_{(x_i,a_i) \in P} a_i] = E[a(P)]$$

We know therefore that $q = p$ is a minimizer of F_q, i.e. $\forall q, F_p \leq F_q$.

Are there other minimizers? Consider a function $q \neq p$, that is a minimizer for a particular population P_1. Since $q \neq p$, $\exists x$ s.t. $q(x) \neq p(x)$. Consider the particular singleton population $P_x = \{x\}$. On singleton populations, the expected squared bias is the same as the expected squared error. In fact, there is a component of F which contains only the squared biases for the singleton popuuations, and it is equal to the expected squared error. Therefore on that population (and any other singleton population for which $q \neq p$) there is only one minimizer of the expected squared bias, and it is the conditional expectation $p(x)$. So $E[(q(x)-A)^2|X = x] > E[(p(x)-A)^2|X = x]$ and therefore $E[b_q(P_x)] > E[b_p(P_x)]$. Since p is a maximiser of fairness for all populations, it is enough to prove that q is sub-optimal on one population to prove that the overall fairness of q is less than that of p, which is the main statement of our theorem:

$$\forall q \neq p, F_q > F_p.$$

References

R.A. Bailey and L. Simon. Two studies in automobile insurance ratemaking. *ASTIN Bulletin*, 1(4):192–217, 1960.

R.E. Bellman. *Dynamic Programming.* Princeton University Press, NJ, 1957.

Y. Bengio and F. Gingras. Recurrent neural networks for missing or asynchronous data. In M. Mozer, D.S. Touretzky, and M. Perrone, editors, *Advances in Neural Information Processing System*, volume 8, pages 395–401. MIT Press, Cambridge, MA, 1996.

C.M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, 1995.

R.L. Brown. Minimum bias with generalized linear models. In *Proceedings of the Casualty Actuarial Society*, 1988.

N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines.* Cambridge Press, 2000.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum-likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society B*, 39:1–38, 1977.

C. Dugas, Y. Bengio, F. Bélisle, C. Nadeau, and R. Garcia. A universal approximator of convex functions applied to option pricing. In *Advances in Neural Information Processing Systems*, volume 13, Denver, CO, 2001.

S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58, 1992.

Z. Ghahramani and M. I. Jordan. Supervised learning from incomplete data via an EM approach. In J.D. Cowan, G. Tesauro, and J. Alspector, editors, *Advances in Neural Information Processing Systems*, volume 6, San Mateo, CA, 1994. Morgan Kaufmann.

F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, and W.A. Stahel. *Robust Statistics, The Approach based on Influence Functions.* John Wiley & Sons, 1986.

T. Hastie, R. Tibshirani, and J. Friedman. *Data Mining, Inference and Prediction.* Springer, 2001.

K.D. Holler, D. Sommer, and G. Trahair. Something old, something new in classification ratemaking with a novel use of glms for credit insurance. *Casualty Actuarial Society Forum*, pages 31–84, 1999.

P.J. Huber. *Robust Statistics.* John Wiley & Sons Inc., 1982.

R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixture of local experts. *Neural Computation*, 3:79–87, 1991.

G.V. Kass. An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29(2):119–127, 1980.

Paul E. Keller. Neural networks: Commercial applications, 1997. http://www.emsl.pnl.gov:2080/proj/neuron/neural/products.

P. McCullagh and J. Nelder. *Generalized Linear Models.* Chapman and Hall, London, 1989.

K. Murphy, M.J. Brockman, and P.K.W. Lee. Using generalized linear models to build dynamic pricing systems. *Casualty Actuarial Society Forum*, pages 107–139, 2000.

G.B. Orr and K.-R. Müller. *Neural Networks: Tricks of the Trade.* Springer, 1998.

P.J. Rousseeuw and A.M. Leroy. *Robust Regression and Outlier Detection.* John Wiley & Sons Inc., 1987.

D.E. Rumelhart, G.E. Hinton, and R.J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986.

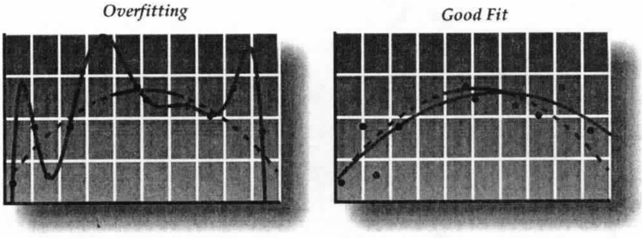
B. Schölkopf, C.J.C. Burges, and A.J. Smola. *Advances in kernel methods: support vector learning.* MIT Press, 1998.

I. Takeuchi, Y. Bengio, and T. Kanamori. Robust regression with asymmetric heavy-tail noise distributions. *Neural Computation*, 2002. to appear.

V. Vapnik. *Statistical Learning Theory.* John Wiley, Lecture Notes in Economics and Mathematical Systems, volume 454, 1998a.

V. Vapnik. *Statistical Learning Theory.* Wiley, Lecture Notes in Economics and Mathematical Systems, volume 454, 1998b.

Figure 1: Illustration of **overfitting**. The solid left curve fits the noise in the data points (black dots) and has not learned the underlying structure (dashed). The right curve, with less flexibility, does not overfit.



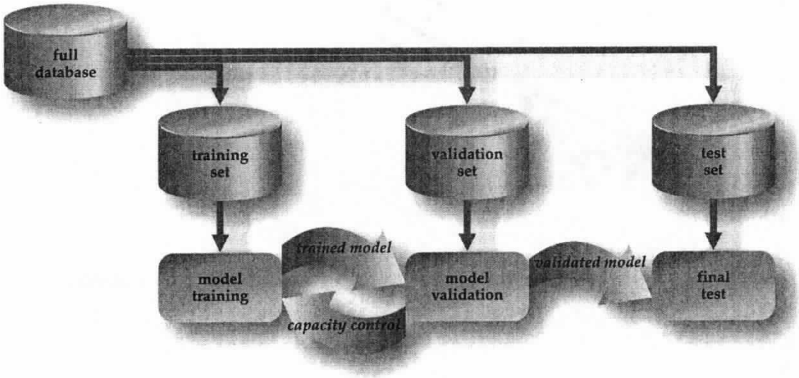


Figure 2: Methodology to prevent overfitting. Model capacity is controlled via a validation set, disjoint from the training set. The generalization performance estimator is obtained by final testing on the test set, disjoint from the first two.

Figure 3: The constant model fits the best horizontal line through the training data.

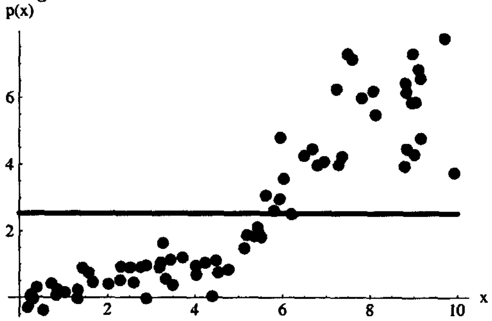


Figure 4: The linear model fits a straight line through the training data.

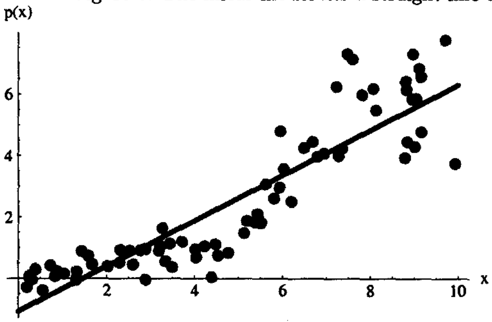


Figure 5: The generalized linear model fits an exponential of a linear transformation of the variables.

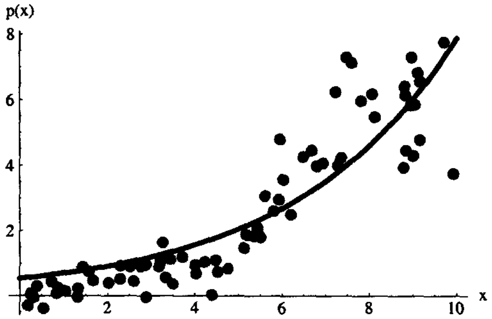


Figure 6: The CHAID model fits constants to partitions of the variables. The dashed lines in the figure delimit the partitions, and are found automatically by the CHAID algorithm.

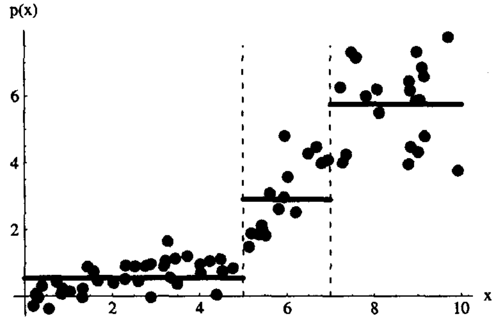


Figure 7: The CHAID+Linear model fits a straight line within each of the CHAID partitions of the variable space.

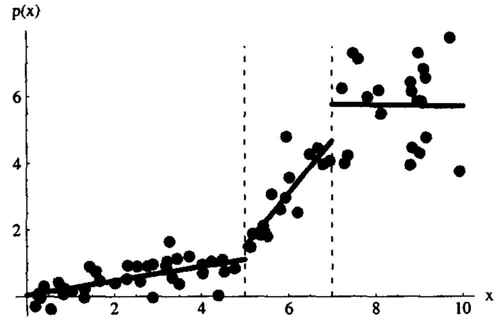


Figure 8: Topology of a one-hidden-layer neural network. In each unit of the hidden layer, the variables are linearly combined. The network then applies a non-linear transformation on those linear combinations. Finally, the resulting values of the hidden units are linearly combined in the output layer.

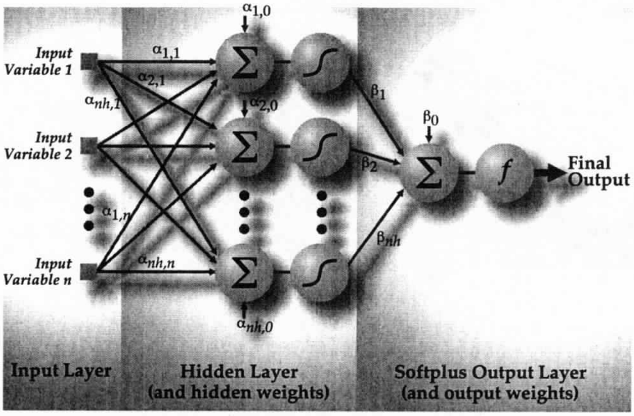


Figure 9: The neural network model learns a smooth non-linear function of the variables.

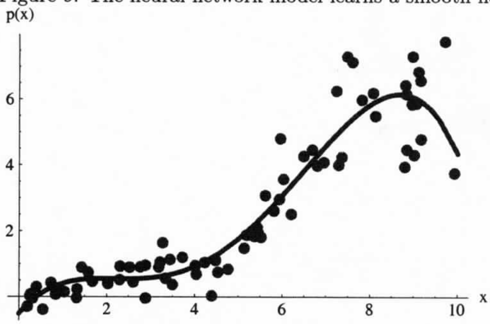


Figure 10: Topology of a one-hidden-layer softplus neural network. The hidden layer applies a non-linear transformation of the variables, whose results are linearly combined by the output layer. The softplus output function forces the function to be positive. To avoid cluttering, some weights linking the variables to the hidden layer are omitted on the figure.

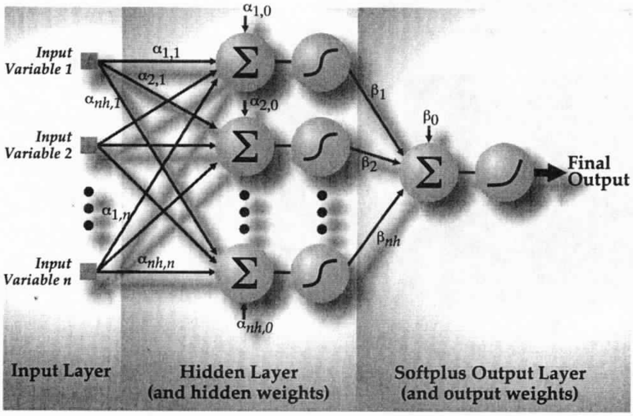


Figure 11: The softplus neural network model learns a smooth non-linear **positive** function of the variables. This positivity is desirable for estimating insurance premiums.

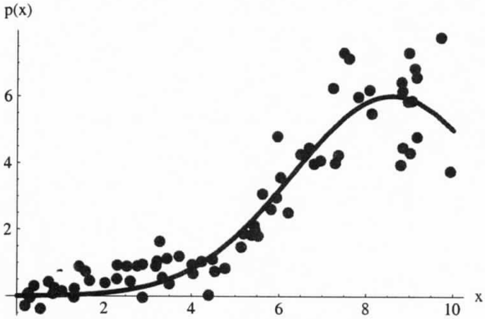


Figure 12: Schematic representation of the mixture model. The first-stage models each make an independent decision, which are linearly combined by a second-stage **gater**.

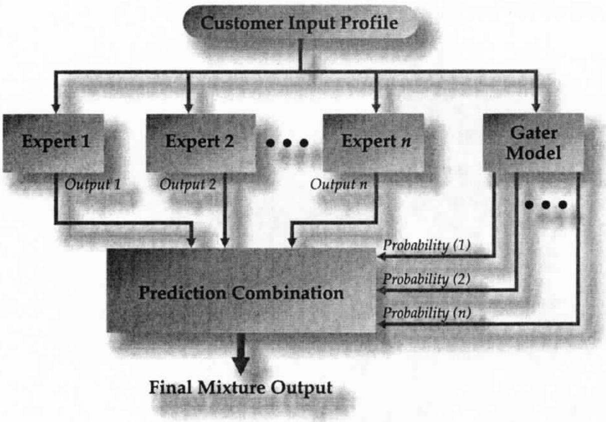
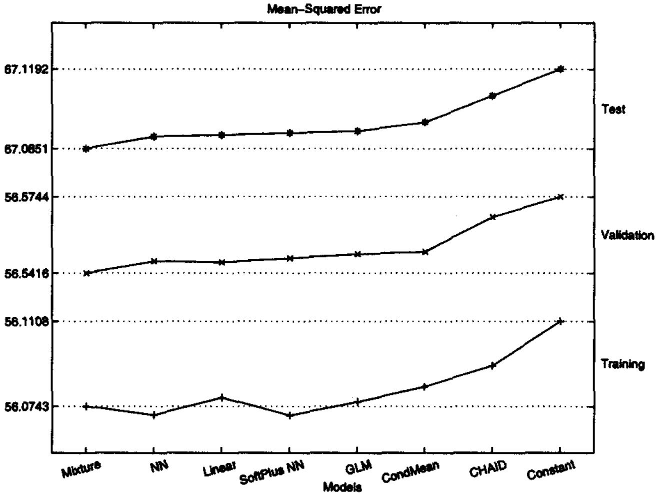


Figure 13: MSE results (from table 1) for eight models . Models have been sorted in ascending order of test results. The training, validation and test curves have been shifted closer together for visualization purposes. The out-of-sample test performance of the mixture model is significantly better than any of the other. Validation based model selection is confirmed on test results.



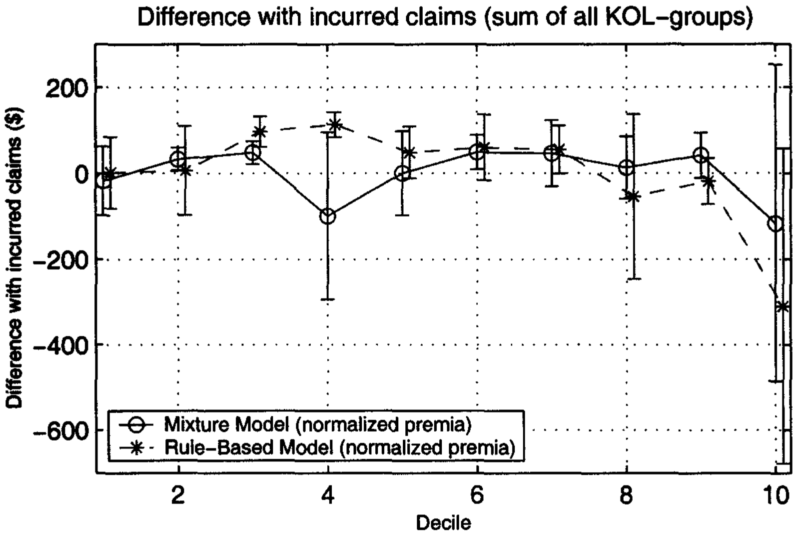


Figure 14: Average difference between premiums and incurred amounts (on the sum over all coverage groups), for the Mixture and Rule-Based models, for each decile of the models' respective premium distribution. We observe that both models are being fair to most customers, except those in the last decile, the highest-risk customers, where they appear to under-charge. The error bars represent 95% confidence intervals. (Each decile contains $\approx 28,000$ observations.)

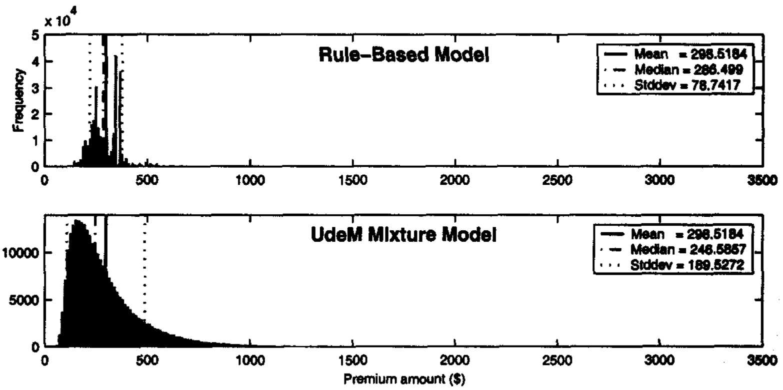
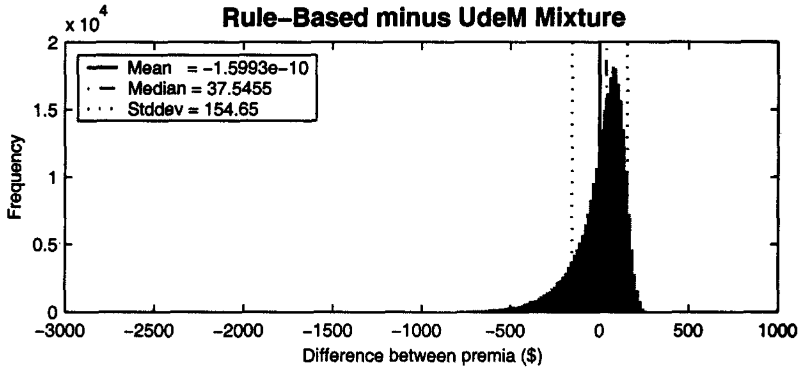


Figure 15: Comparison of the premium distribution for the current Rule-Based model and the Mixture model. The distributions are normalized to the same mean. The Mixture model distribution has fatter tails and is much smoother.

Figure 16: Distribution of the premium difference between the Rule-Based and Mixture models, for the sum of the first three coverage groups. The distribution is negatively skewed: the Rule-Based model severely under-charges for some customers.



Credibility Modeling via Spline Nonparametric Regression

Ashis Gangopadhyay and Wu-Chyuan Gau

Credibility Modeling via Spline Nonparametric Regression

Dr. Ashis Gangopadhyay

Associate Professor,

Mathematics and Statistics,

Boston University,

111 Cummington Street,

MA 02215, U.S.A.

Phone: 617-353-2560.

Email: ag@math.bu.edu.

Dr. Wu-Chyuan Gau

Assistant Professor,

Statistics and Actuarial Science,

University of Central Florida,

Orlando, FL 32816,

U.S.A.

Phone: 407-823-5530.

Email: wgau@mail.ucf.edu.

Credibility Modeling via Spline Nonparametric Regression

Abstract

Credibility modeling is a rate making process which allows actuaries to adjust future premiums according to the past experience of a risk or group of risks. Current methods in credibility theory often rely on parametric models. Bühlmann (1967) developed an approach based on the best linear approximation, which leads to an estimator that is a linear combination of current observations and past records. During the last decade, the existence of high speed computers and statistical software packages allowed the introduction of more sophisticated methodologies. Some of these techniques are based on Markov Chain Monte Carlo (MCMC) approach to Bayesian inference, which requires extensive computations. However, very few of these methods made use of the additional covariate information related to the risk, or group of risks; and at the same time account for the correlated structure in the data. In this paper, we consider a Bayesian nonparametric approach to the problem of risk modeling. The model incorporates past and present observations related to the risk, as well as relevant covariate information. The Bayesian modeling is carried out by sampling from a multivariate Gaussian prior, where the covariance structure is based on a thin-plate spline (Wahba, 1990). The model uses MCMC technique to compute the predictive distribution of the future claims based on the available data. Extensive data analysis is conducted to study the properties of the proposed estimator, and compare against the existing techniques.

Keywords: Credibility Modeling, Thin-plate Spline, MCMC, RKHS.

1 Introduction

The dictionary definition of a spline is "a thin strip of wood used in building construction." This in fact gives insight into the mathematical definition of splines. Historically, engineering draftsmen used long thin strips of wood called splines to draw a smooth curve between specified points. A mathematical spline is the solution to a constrained optimization problem.

In the credibility context, suppose we wish to determine how the current claim loss, Y_{ij}, depends on the past losses, say $Y_{i,j-1}$ and $Y_{i,j-2}$. Our approach is to consider the nonparametric regression model

$$y_{ij} = g(y_{i,j-1}, y_{i,j-2}) + \epsilon_i, \qquad i = 1, ..., n, \tag{1}$$

where g is a smooth function of its arguments. Our objective is to model the dependency between the current observations y_{ij} for all policyholders $i = 1, ..., n$, and those past losses $Y_{i,j-1}$ and $Y_{i,j-2}$ through a nonparametric regression at occasion j. The concept is similar to multiple linear regressions. Here, the dependent variable happens to be Y_{ij} while the past losses $Y_{i,j-1}$ and $Y_{i,j-2}$ are treated as covariates. For notational convenience, we let Y_i stands for the dependent variable (current observation) and use s_i or t_i for covariates (past losses). The key problem is to find a good approximation $\hat{g}$ of g. This is a tractable problem, and there are many different solutions to this problem. The purpose of this paper is to develop a methodology to estimate the function g given the data. We use a nonparametric Bayesian approach to estimate the multivariate regression model with Gaussian errors. In this approach, very little is assumed regarding the underlying model (signal); and we allow the data to "speak for itself".

Reproducing kernel Hilbert space (RKHS) models have been in use for at least ninety-five years. The systematic development of reproducing kernel Hilbert space theory is given by Aronszajn (1950). For further background the reader may refer to Weinert (1982) and Wahba (1990). A recent paper given by Evgeniou (2000) contains an introduction to RKHS, which we found to be useful for readers interested in further reading.

A reproducing kernel Hilbert space is a Hilbert function space characterized by the fact that it contains a kernel that reproduces (through a inner product) every function in the space, or, equivalently, by the fact that every point evaluation functional is bounded. RKHS models are useful in estimation problems because every covariance function is also a reproducing kernel for some RKHS. As a consequence, there is a close connection between a random process and the RKHS determined by its covariance function. These estimation problems can then be solved by evaluating a certain RKHS inner product. Thus it is necessary to be able to determine the form of inner product corresponding to a given reproducing kernel.

In optimal curve and surface fitting problems, in which one is reconstructing an unknown function based on the sample data, it is inevitable that the point evaluation functionals be bounded. Therefore, one is forced to express the problem in a RKHS whose inner product is determined by the quadratic cost functional that needs to be minimized. To solve these problems, one must find a basis for the range of a certain projection operator. One way to do this is to determine the reproducing kernel corresponding to the given inner product.

Consider an univariate model

$$y_i = f(t_i) + \epsilon_i, \; i = 1, 2, ...n \tag{2}$$

where $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)' \sim N(0, \sigma^2 I)$ and f is only known to be smooth. If f has $m-1$

continuous derivatives and is m^{th} derivative is square integrable, an estimate of f can be found by minimizing

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(t_i))^2 + \lambda \int_a^b (f^{(m)}(t))^2 dt, \tag{3}$$

for some $\lambda > 0$. The smoothing parameter λ controls the trade-off between smoothness and accuracy. A discrete version of problems such as (3) was considered in the actuarial literature by Whittaker (1923), who considered smoothing $y_1, ..., y_n$ discretely by finding $\mathbf{f} = (f_1, ..., f_n)$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f_i)^2 + \lambda\sum_{i=1}^{n-3}(f_{i+3} - 3f_{i+2} + 3f_{i+1} - f_i)^2. \tag{4}$$

If $m = 2$, (3) becomes the penalized residual sum of squares

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(t_i))^2 + \lambda \int_a^b (f''(t))^2 dx \tag{5}$$

where λ is a fixed constant, and $a \leqslant t_1 \leqslant \cdots \leqslant t_n \leqslant b$. If we consider all functions $f(t)$ with two continuous derivatives, it can be shown that (5) has an unique minimizer which is a nature cubic spline with knots at the unique values of t_i. As $\lambda \longrightarrow \infty$, it forces $f''(t) = 0$ everywhere, and the solution is the least-squares line. As $\lambda \longrightarrow 0$, the solution tends to an interpolating twice-differentiable function. The cubic spline can be generalized to two or higher dimensions. The thin-plate spline is one example. It derives from generalizing the second derivative penalty for smoothness to a two dimensional Laplacian penalty (Wendelberger, 1982).

This paper is organized as follows. Section 2 introduces the credibility problem. Section 3 reviews the thin-plate spline and the Bayesian model behind the smoothing spline. Section 4 introduces the basic idea of bivariate regression with Gaussian errors. Section 5 generalizes the results in section 4 to the trivariate case. Section 6 introduces the applications of the results developed in section 4 and section 5. All computations are carried out using Gibbs sampler. All functions, functionals, random variables, and function spaces in this paper will be real valued unless specifically noted otherwise.

2 The Credibility Problem

The classical data type in this area involves realizations from present and past experience of individual policyholders. Suppose we have n different risks (or policyholders) with a claims record over a certain number of years, say T;

$$Y_{11}, Y_{12}, ..., Y_{1T}$$
$$Y_{21}, Y_{22}, ..., Y_{2T}$$
$$..................$$
$$Y_{n1}, Y_{n2}, ..., Y_{nT}.$$

The data can be the amount of losses, the number of claims, or the loss ratio from insurance portfolios. Our goal is to estimate the amount or number of claims to be paid on a particular insurance policy in a future coverage period. The problem of interest is to model the relationship of $\mathbf{Y}_{T+1}$ to time and the past observed values of $\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_T$, i.e., to establish the relationship:

$$y_{ij} = f(t, y_{i1}, y_{i2}, ..., y_{i,j-1}) + e_{ij} \text{ for } i = 1, 2, ..., n;\ j = 1, 2, ..., T \quad (6)$$

where f is an unknown function and e_{ij} is a random error term.

We can also have a more general form of data configuration. Let $\{y_{ij} : j =$

$1, 2, \ldots, T_i; i = 1, 2, \ldots, n\}$ be a time series of length T_i over which the measurements of the i^{th} risk (or group of risks) were observed at time points $\{t_{ij} : j = 1, 2, ..., T_i\}$. In addition, let $\mathbf{X}_{ij}$ be the observed covariates, such as gender of a policyholder, or industry type of insureds etc., for the i^{th} risk (or group of risks) at time t_{ij}. In each risk (or group of risks), the data has the form

$$(y_{ij}, \mathbf{X}_{ij}, t_{ij}),\ j = 1, 2, ..., T_i;\ i = 1, 2, ..., n, \tag{7}$$

where $\mathbf{X}_{ij} = (X_{ij1}, X_{ij2}, ..., X_{ijd})$ are the d covariate variables measured at time t_{ij}. In this case, of interest is to study the association between the current response y_{ij} and the past responses $y^*_{i,j-1} = (y_{i,j-1}, y_{i,j-2}, ..., y_{i1})$ as well as the covariates and to examine how the association varies with time. Table 1 provides the data lay-out assuming $j = 1, 2, ..., T$.

	Occasion		
Subject	1	$\cdots$	T
1	$y_{11}, x_{11,1}, x_{11,2}, ..., x_{11,d}, t_{11}$	$\cdots$	$y_{1T}, x_{1T,1}, x_{1T,2}, ..., x_{1T,d}, t_{1T}$
$\vdots$	$\vdots$	$\cdots$	$\vdots$
i	$y_{i1}, x_{i1,1}, x_{i1,2}, ..., x_{i1,d}, t_{i1}$	$\cdots$	$y_{iT}, x_{i1,T}, x_{i1,T}, ..., x_{iT,d}, t_{iT}$
$\vdots$	$\vdots$	$\cdots$	$\vdots$
n	$y_{n1}, x_{n1,1}, x_{n1,2}, ..., x_{n1,d}, t_{n1}$	$\cdots$	$y_{nT}, x_{nT,1}, x_{nT,2}, ..., x_{nT,d}, t_{nT}$

Table 1: Data Configuration.

Therefore, we propose the following modified model, which is more general than (7),

$$y_{ij} = f(t_{ij}, \mathbf{X}_{ij}, y^*_{i,j-1}) + e_{ij} \text{ for } i = 1, 2, ..., n;\ j = 1, 2, ..., T_i. \tag{8}$$

In this paper, a new Bayesian approach is presented for nonparametric multivariate regression with Gaussian errors. A smoothness prior based on thin-plate splines is assumed for each component of the model. We use the reproducing kernel for a thin-plate spline for an unknown multivariate function as in Wahba (1990). All the computations are carried out using the Gibbs sampling schemes (Wood et. al., 2000). With a burn-in period, it is assumed that iterations have converged to draws from posterior distributions. A random sample from the convergence period are used to estimate characteristics of the posterior distribution. This model is used for estimation of function f and to predict for the future values. We analyze a real data from one Taiwan based insurance company. A comparison is being carried out between the proposed approach against other existing techniques.

3 The Thin-Plate Spline

RKHS methods have been successfully applied to a wide varieties of problems in the field of optimal approximation, which include interpolation and smoothing via spline function in one or more dimensions. The one dimensional case is generalized to the multidimensional case by Duchon (1977). Duchon's surface spline is called "thin plate" spline, because they approximate the equilibrium position of a thin plate deflected at scatter points. For an application of thin-plate splines to meteorological problems see Wahba and Wendelberger (1980).

3.1 The Thin-Plate Spline on E^d

The theoretical foundations for the thin-plate spline were from Duchon (1975, 1976, 1977) and Meinguet (1979), and some further results and applications to meteorological problems were given in Wahba and Wendelberger (1980) and Wood et. al. (2000).

Let us define the penalty functional

$$J_m(f) = \int_0^1 (f^{(m)}(u))^2 du.$$

It is assumed that data $\mathbf{y} = (y_1, \cdots, y_n)'$ follows the model

$$y_i = f(\mathbf{X}_i) + \epsilon_i, \quad i = 1, 2, \cdots, n,$$

where $\mathbf{X}_i \in \Omega$ and Ω is a general index set. The function f is assumed to be a smooth function in a reproducing kernel Hilbert space H of a real-valued functions on Ω. The $\{\epsilon_i\}$ are independent zero mean errors with common unknown variance. It is desired to find an estimate of f given $\mathbf{y} = (y_1, \cdots, y_n)'$. The estimate f_λ of f will be taken as the minimizer in H of

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(\mathbf{X}_i))^2 + \lambda J_m(f), \tag{9}$$

where $J_m(f)$ is a seminorm on H with M-dimensional null space spanned by $\phi_1, \cdots, \phi_M$, $M < n$. The seminaorm on the vector space H is a mapping $p : H \rightarrow \mathbf{R}$ satisfying $\|a\| \geqq 0$, $\|\alpha a\| = |\alpha|\|a\|$, and $\|a+b\| \leqq \|a\| + \|b\|$. Here a and b are arbitrary vectors in H and α is any scalar.

In the thin-plate spline case, we will assume $f \in \chi$, a space of functions whose partial derivatives of total order m are in $L_2(E^d)$. The data model is given by

$$y_i = f(x_1(i), \cdots, x_d(i)) + \epsilon_i, \quad i = 1, 2, \cdots, n, \tag{10}$$

where $f \in \chi$ and $\epsilon = (\epsilon_1, \cdots, \epsilon_n)' \sim N(0, \sigma^2 I)$. And $J(f) = J_m^d(f)$ is given by

$$J_m^d(f) = \sum_{\alpha_1+\cdots+\alpha_d=m} \frac{m!}{\alpha_1! \cdots \alpha_d!} \times \int \cdots \int (\frac{\partial^m f}{\partial x_1 \cdots \partial x_d})^2 dx_1 \cdots dx_d. \tag{11}$$

We want χ endowed with the seminorm $J_m^d(f)$ to be an RKHS (that is, for the evaluation functionals in χ to be bounded with respect to $J_m^d(f)$). Then, a thin-plate smoothing spline is the solution to the following variational problem. Find $f \in \chi$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_1(i), \cdots, x_d(i)))^2 + \lambda J_m^d(f). \tag{12}$$

Let us use the notation $t = (x_1, \cdots, x_d)'$ and $t_i = (x_1(i), \cdots, x_d(i))'$. The null space of the penalty functional $J_m^d(f)$ is the M-dimensional space spanned by the polynomials in d variables of total degree $\leqslant m-1$, where

$$M = \binom{d+m-1}{d}. \tag{13}$$

In the space $H = \{f : J_m^d(f) < \infty\}$ with $J_m^d(f)$ as a square semi norm, it is necessary that $2m - d > 0$ for the evaluation functional $L_t f = f(t)$ to be continuous; see Duchon (1977), Meinguet (1979), and Wahba and Wendelberger (1980). For $m = 2$, $d = 2$,

$$J_2^2(f) = \int\int [(\frac{\partial^2 f}{\partial x_1^2})^2 + 2(\frac{\partial^2 f}{\partial x_1 \partial x_2})^2 + (\frac{\partial^2 f}{\partial x_2^2})^2] dx_1 dx_2, \tag{14}$$

with $M = 3$, and the null space is spanned by ϕ_1, ϕ_2, ϕ_3 given by

$$\phi_1(x_1, x_2) = 1, \phi_2(x_1, x_2) = x_1, \phi_3(x_1, x_2) = x_2.$$

Before we go further, we need additional notations. Let $s, t \in E^d$, $s = (s_1, \cdots, s_d)'$ and $t = (t_1, \cdots, t_d)'$, then

$$| s - t | = (\sum_{i=1}^{d}(s_i - t_i)^2)^{1/2}.$$

We can define

$$E(r) = \theta_m^d r^{2m-d} \log r \quad \text{if } d \text{ even} \tag{15}$$
$$= \theta_m^d r^{2m-d} \quad \text{if } d \text{ odd,}$$

where

$$\theta_m^d = \frac{(-1)^{d/2+1}}{2^{2m-1}\pi^{d/2}(m-1)!(m-d/2)!} \quad \text{if } d \text{ even} \tag{16}$$
$$= \frac{(-1)^m \Gamma(d/2-m)}{2^{2m}\pi^{d/2}(m-1)!} \quad \text{if } d \text{ odd.}$$

We can also define

$$E_m(s,t) = E(| s - t |). \tag{17}$$

Duchon (1977) showed that, if $t_1, \cdots, t_n$ are such that least squares regression on $\phi_1, \cdots, \phi_M$ is unique, then (12) has an unique minimizer f_λ with representation

$$f_\lambda(t) = \sum_{\nu=1}^{M} d_\nu \phi_\nu + \sum_{i=1}^{n} c_i E_m(t, t_i). \tag{18}$$

Note that θ_m^d can be absorbed into c_i in (18). Let $u_1, \cdots, u_M$ be any fixed points in E^d such that least squares regression on the M-dimensional space of polynomials of total degree less than m at the points $u_1, \cdots, u_M$ is unique. Let $p_1, \cdots, p_M$

be the polynomials of total degree less than m satisfying $p_i(u_j) = 1, i = j$ and $p_i(u_j) = 0, i \neq j$. And let

$$K^1(s,t) = E_m(s,t) - \sum_{i=1}^{M} p_i(t) E_m(u_i, s) \tag{19}$$
$$- \sum_{j=1}^{M} p_j(s) E_m(t, u_j)$$
$$+ \sum_{i=1}^{M} \sum_{j=1}^{M} p_i(t) p_j(s) E_m(u_i, u_j).$$

It can be shown that K^1 is positive semidefinite and is a reproducing kernel for H_K and f_λ has a representation (Wahba, 1990)

$$f_\lambda = \sum_{\nu=1}^{M} d_\nu \phi_\nu + \sum_{i=1}^{n} c_i K^1_{t_i}(t), \tag{20}$$

where

$$K^1_t(\cdot) = K^1(t, \cdot).$$

The result from (20) can be shown to be the same as (18).

3.2 Bayes Model Behind The Thin-Plate Spline

Let us now take a look at the Bayes estimates behind the thin-plate spline. It is known that certain Bayes estimates are solutions to variational problems, and vice versa. Consider the random effect model

$$F(t) = \sum_{\nu=1}^{M} \theta_\nu \phi_\nu(t) + b^{1/2} X(t), \quad t \in [0,1], \tag{21}$$

$$Y_i = F(t_i) + \epsilon_i, \quad i = 1, \cdots, n.$$

Let $\{\phi_1, \cdots, \phi_M\}$ span H_0, the space of polynomials of total degree less than m, and H_1 be a RKHS with the reproducing kernel defined by

$$EX(s)X(t) = K^1(s,t),$$

where $K^1(s,t)$ given by (19). Then, the model in (21) will result in the thin-plate spline. To understand this result, let

$$\tilde{y}_i = y_i - \sum_{\nu=1}^{M} \theta_\nu \phi_\nu(t_i),$$

and set $f(t_i) = b^{1/2} X(t_i)$. Then (21) becomes

$$\tilde{y}_i = f(t_i) + \epsilon_i, \quad i = 1, \cdots, n,$$

with $\epsilon = (\epsilon_1, \cdots, \epsilon_n)' \sim N(0, \sigma^2 I)$ and

$$\begin{aligned} Ef(s)f(t) &= E[b^{1/2} X(s) b^{1/2} X(t)] \\ &= bE[X(s)X(t)] \\ &= bK^1(s,t). \end{aligned}$$

Then, from Wahba (2000),

$$E\left(\begin{bmatrix} f(t_1) \\ f(t_2) \\ \vdots \\ f(t_n) \end{bmatrix} \mid \tilde{\mathbf{y}}\right) = K^1(K^1 + \frac{\sigma^2}{b}\mathbf{I})^{-1}\tilde{\mathbf{y}} \tag{22}$$

$$\equiv A(\lambda)\tilde{\mathbf{y}}, \text{ with } \lambda = \frac{\sigma^2}{b}.$$

$A(\lambda)$ is known as the influence matrix and we will use the result later.

Now, consider the variational problem in H_1, we want to find f_λ to minimize

$$\frac{1}{n}\sum_{i=1}^{n}(\tilde{y}_i - f(t_i))^2 + \lambda \parallel f \parallel_{H_1}^2,$$

where $\parallel f \parallel_{H_1}^2$ is the squared norm in H_1. It can be shown that

$$E\left(\begin{bmatrix} f(t_1) \\ f(t_2) \\ \vdots \\ f(t_n) \end{bmatrix} \mid \tilde{\mathbf{y}}\right) = K^1(K^1 + \lambda\mathbf{I})^{-1}\tilde{\mathbf{y}}$$

$$\equiv A(\lambda)\tilde{\mathbf{y}}.$$

In summary, given the prior $\mathbf{f} \sim N(0, bK^1)$, a zero-mean Gaussian stochastic process with $\epsilon \doteq (\epsilon_1, \cdots, \epsilon_n)' \sim N(0, \sigma^2\mathbf{I})$, the posterior mean for $\mathbf{f}$ given $\tilde{\mathbf{y}}$ is the solution to a variational problem in an RKHS.

4 Bivariate Regression

Let us now return to credibility problems. Recall that, in the Bühlmann-Straub model, we wish to use the conditional distribution $f_{Y_{T+1}|\Theta}(y_{T+1}|\theta)$ or the hypothetical mean $E(Y_{T+1}|\Theta=\theta) \equiv \mu_{T+1}(\theta)$ for estimation of next year's claims. Since we have observed $\mathbf{y}$, one suggestion is to approximate $\mu_{T+1}(\theta)$ by a linear function of the past data. It turns out that the resulting credibility premium formula $Z\bar{Y} + (1-Z)\mu$ is of this form. The idea is to restrict estimators of the form $\alpha_0 + \sum_{t=1}^{T} \alpha_t Y_t$, where $\alpha_0, \alpha_1, ..., \alpha_T$ need to be chosen. We will choose the $\alpha's$ to minimize square error loss, that is,

$$Q = E\left\{[\mu_{T+1}(\theta) - \alpha_0 - \sum_{t=1}^{T} \alpha_t Y_t]^2\right\}.$$

We denote the result by $\widetilde{\alpha}_0, \widetilde{\alpha}_1, ..., \widetilde{\alpha}_T$ for the values of $\alpha_0, \alpha_1, ..., \alpha_T$ which minimize Q. Then the credibility premium can be written as:

$$\widetilde{\alpha}_0 + \sum_{t=1}^{T} \widetilde{\alpha}_t Y_t.$$

Meanwhile, the resulting $\widetilde{\alpha}_0, \widetilde{\alpha}_1, ..., \widetilde{\alpha}_T$ also minimize

$$Q_1 = E\left\{[E(Y_{T+1} \mid \mathbf{Y} = \mathbf{y}) - \alpha_0 - \sum_{t=1}^{T} \alpha_t Y_t]^2\right\}$$

and

$$Q_2 = E\left\{[Y_{T+1} - \alpha_0 - \sum_{t=1}^{T} \alpha_t Y_t]^2\right\}.$$

Hence, the credibility premium $\widetilde{\alpha}_0 + \sum_{t=1}^{T} \widetilde{\alpha}_t Y_t$ is the best linear estimator of each of the hypothetical mean $E(Y_{T+1}|\Theta=\theta)$, the Bayesian premium $E(Y_{T+1} \mid \mathbf{Y}=\mathbf{y})$, and Y_{T+1} in the sense of square error loss.

Now, we want to extend the standard credibility techniques into nonparametric regression models. In the credibility context, suppose we wish to determine how the current claim loss, Y_{ij}, depends on the past losses, say $Y_{i,j-1}$ and $Y_{i,j-2}$. Our approach is to establish a model as the nonparametric regression $y_{ij} = g(y_{i,j-1}, y_{i,j-2}) + \epsilon_i$, $i = 1, ..., n$, where g is a smooth function of its arguments. What we want to accomplish is to model the dependency between the current observations y_{ij} for all policyholders $i = 1, ..., n$, and those past losses $y_{i,j-1}$ and $y_{i,j-2}$ through a nonparametric regression at occasion j. Once the model is established, we can perform one-step ahead prediction on $y_{i,j+1}$ by using y_{ij} and $y_{i,j-1}$ as covariates. For notational convenience, we let y_i stands for the dependent variable (current observation) and use s_i or t_i for covariates (past losses). We develop a methodology to estimate the function g, given the data, from a nonparametric regression perspective. We will be using a Bayesian approach to fit the proposed model using a Gaussian prior on the unknown function g, which uses the reproducing kernel of a thin-plate spline as the covariance of the prior distribution (Wahba, 1990, p.30).

4.1 Model and Prior

Without loss of generality, we assume variables s_i, t_i lie in the interval $[0, 1]$. Consider the model from the bivariate regression model

$$y_i = g(s_i, t_i) + \epsilon_i, \quad i = 1, \cdots, n, \tag{23}$$

where g is a smooth regression of the variables s and t, and the errors ϵ_i are independent $N(0, \sigma^2)$. It is convenient to write (23) as

$$y_i = \alpha_0 + \alpha_1 s_i + \alpha_2 t_i + f(s_i, t_i) + \epsilon_i, \quad i = 1, \cdots, n, \tag{24}$$

with f having the zero initial conditions:

$$\begin{aligned} f(0,0) &= 0, \\ \frac{\partial f}{\partial s}(0,0) &= 0, \\ \frac{\partial f}{\partial t}(0,0) &= 0, \end{aligned} \tag{25}$$

which means that

$$\begin{aligned} \alpha_0 &= g(0,0), \\ \alpha_1 &= \frac{\partial g}{\partial s}(0,0), \\ \alpha_2 &= \frac{\partial g}{\partial t}(0,0). \end{aligned}$$

Model (24) has the same form as (21) with

$$\phi_1(s_i, t_i) = 1, \phi_2(s_i, t_i) = s_i, \phi_3(s_i, t_i) = t_i.$$

Now we can specify the prior on (24).

The prior for $f(s,t)$ is the reproducing kernel for the thin-plate spline in (19). This means that $f(s,t)$ is of zero-mean Gaussian random variables with the covariance function between $f(s_i,t_i)$ and $f(s_j,t_j)$ given by

$$cov\{f(s_i,t_i),f(s_j,t_j)\} = \tau^2\mathbf{K}\{(s_i,t_i),(s_j,t_j)\}. \tag{26}$$

Using results from (19), the kernel $\mathbf{K}$ is given by

$$\begin{aligned}\mathbf{K}\{(s_i,t_i),(s_j,t_j)\} &= E\{(s_i,t_i),(s_j,t_j)\} \\ &- \sum_{k=1}^{3} p_k(s_j,t_j)E\{u_k,(s_i,t_i)\} \\ &- \sum_{k=1}^{3} p_k(s_i,t_i)E\{(s_j,t_j),u_k\} \\ &+ \sum_{k=1}^{3}\sum_{l=1}^{3} p_k(s_i,t_i)p_l(s_j,t_j)E\{u_k,u_l\},\end{aligned} \tag{27}$$

where

$$E\{(s_i,t_i),(s_j,t_j)\} = r^2\log(r),\ r = \sqrt{(s_i-s_j)^2+(t_i-t_j)^2}. \tag{28}$$

This is because that, with $d=2$ and $m=2$, we have $J_2^2(f)$ given by (14). Obviously, $d/2+1$ is even in (16), so $E(r)$ is proportional to $r^2\log(r)$. Note that θ_m^d can be absorbed into c_i in (18). Furthermore, let

$$p_1(s_i,t_i) = -1+2s_i+2t_i,\ p_2(s_i,t_i) = 1-2s_i,\ p_3(s_i,t_i) = 1-2t_i. \tag{29}$$

By choosing

$$u_1 = (\frac{1}{2},\frac{1}{2}), u_2 = (0,\frac{1}{2}), u_3 = (\frac{1}{2},0),$$

we have p_1,p_2,p_3 be the polynomials of total degree less than 2 satisfying $p_i(u_j) =$

$1, i = j$ and $p_i(u_j) = 0, i \neq j$. Then we are now ready to apply the random effect model in (21).

To complete the prior specification for model (24), we take uninformative priors for all unknown parameters (Wood et. al., 2000). We take uniform independent prior on $[0, 10^{10}]$ for the smoothing parameter τ^2. The prior for $\alpha = (\alpha_0, \alpha_1, \alpha_2)'$ is

$$\alpha \sim \mathbf{N}(\mathbf{0}, c\mathbf{I}),$$

with $c \longrightarrow \infty$. The prior for σ^2 is

$$p(\sigma^2) \propto (\sigma^2)^{-1-10^{-8}} \exp(-10^{-10}/\sigma^2).$$

The resulting Bayes estimate will be the solution to the variational problem in (12) with $d = 2$ and $m = 2$.

4.2 Model Implementation

In this subsection, we will discuss the implementation of the model in (24). To make this model computationally feasible, we will consider a transformed model. As in Wood et. al. (2000), the sampling scheme requires factoring the covariance matrix $\mathbf{K}$ as $\mathbf{QDQ}'$, where $\mathbf{Q}$ is an orthonormal matrix and $\mathbf{D}$ is the diagonal matrix with diagonal elements, d_i, that are the eigenvalues of $\mathbf{K}$.

To ease the notation, we rewrite model in (24) as

$$y_i = \alpha_0 + \alpha_1 s_i + \alpha_2 t_i + f_i + \epsilon_i, \quad i = 1, \cdots, n, \tag{30}$$

where $f_i = f(s_i, t_i)$, and $\mathbf{f} = (f_1, \cdots, f_n)'$ is Gaussian with zero-mean and the covariance $\tau^2 K$. Let

$$\alpha = (\alpha_0, \alpha_1, \alpha_2)', \tag{31}$$

$$\mathbf{y} = (y_1, \cdots, y_n)',$$

$$\epsilon = (\epsilon_1, \cdots, \epsilon_n)',$$

and

$$\mathbf{Z} = \begin{bmatrix} 1 & s_1 & t_1 \\ 1 & s_2 & t_2 \\ \vdots & \cdots & \vdots \\ 1 & s_n & t_n \end{bmatrix}, \tag{32}$$

then we can write (30) in the matrix form

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & s_1 & t_1 \\ 1 & s_2 & t_2 \\ \vdots & \cdots & \vdots \\ 1 & s_n & t_n \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{bmatrix} + \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}, \tag{33}$$

that is,

$$\mathbf{y} = \mathbf{Z}\alpha + \mathbf{f} + \epsilon, \tag{34}$$

with the priors,

$$
\begin{aligned}
&\alpha \sim \mathbf{N}(\mathbf{0}, c\mathbf{I}), \\
&\mathbf{f} \sim \mathbf{N}(\mathbf{0}, \tau^2\mathbf{K}), \\
&\epsilon \sim \mathbf{N}(\mathbf{0}, \sigma^2\mathbf{I}), \\
&\tau^2 \sim unif[0, 10^{10}], \\
&\sigma^2 \sim IG(10^{-8}, 10^{-10}),
\end{aligned} \tag{35}
$$

where τ^2 is uniformly distributed in the interval $[0, 10^{10}]$ and σ^2 has a inverse Gamma distribution with parameters 10^{-8} and 10^{-10}. Let us return to the covariance matrix $\mathbf{K}$. Since $\mathbf{K}$ is positive definite, we can factor $\mathbf{K}$ as $\mathbf{QDQ}'$ such that

$$
\mathbf{QQ}' = \mathbf{I}. \tag{36}
$$

We can pre-multiply $\mathbf{Q}'$ to (34), so we have $\mathbf{y}^* = \mathbf{Q}'\mathbf{y}$. And the model becomes

$$
\mathbf{y}^* = \mathbf{Z}^*\alpha + \mathbf{f}^* + \epsilon^*, \tag{37}
$$

where

$$
\begin{aligned}
\mathbf{Z}^* &= \mathbf{Q}'\mathbf{Z}, \\
\mathbf{f}^* &= \mathbf{Q}'\mathbf{f}, \\
\epsilon^* &= \mathbf{Q}'\epsilon.
\end{aligned} \tag{38}
$$

The priors for α, τ^2, σ^2 will remain the same as in (35). Meanwhile, ϵ^* has the same distribution as $\epsilon \sim \mathbf{N}(\mathbf{0}, \sigma^2\mathbf{I})$ because of (36) in $\mathbf{N}(\mathbf{0}, \sigma^2\mathbf{Q}'\mathbf{I}\mathbf{Q})$. However, the prior for $\mathbf{f}^*$ becomes

$$\mathbf{f}^* \sim \mathbf{N}(\mathbf{0}, \tau^2\mathbf{D}), \tag{39}$$

because of

$$\begin{aligned} Var(\mathbf{Q}'\mathbf{f}) &= \mathbf{Q}'Var(\mathbf{f})\mathbf{Q}. \\ &= \mathbf{Q}'\tau^2\mathbf{K}\mathbf{Q} \\ &= \tau^2\mathbf{Q}'\mathbf{Q}\mathbf{D}\mathbf{Q}'\mathbf{Q} \\ &= \tau^2\mathbf{D}, \end{aligned}$$

where τ^2 and $\mathbf{D}$ as defined before.

4.3 Bivariate Regression for the Bühlmann-Straub Model

Consider data in Bühlmann-Straub Model, it allows different number of exposure units or different distribution of claim size across past policy years. This can be handled in model (30) by assuming

$$\epsilon_i \sim N(0, \frac{\sigma^2}{w_i}), \quad i = 1, \cdots, n, \tag{40}$$

where w_i is the corresponding weight for data value y_i. We can also have the same matrix form as (34),

$$\mathbf{y} = \mathbf{Z}\alpha + \mathbf{f} + \epsilon, \tag{41}$$

but with the priors

$$
\begin{aligned}
\alpha &\sim \mathbf{N}(\mathbf{0}, c\mathbf{I}), \\
\mathbf{f} &\sim \mathbf{N}(\mathbf{0}, \tau^2\mathbf{K}), \\
\epsilon &\sim \mathbf{N}(\mathbf{0}, \sigma^2\mathbf{W}^{-1}), \\
\tau^2 &\sim unif[0, 10^{10}], \\
\sigma^2 &\sim IG(10^{-8}, 10^{-10}).
\end{aligned} \tag{42}
$$

Here $\mathbf{W}$ is the diagonal matrix with diagonal elements, w_i, the corresponding weight for data value y_i.

This model can be easily transformed to a similar model as in (34) and (35). Then we can implement the modified model analogously as in Section 4.2. Now let

$$
\begin{aligned}
\mathbf{y}' &= \sqrt{\mathbf{W}}\mathbf{y} \\
\mathbf{Z}' &= \sqrt{\mathbf{W}}\mathbf{Z}, \\
\mathbf{f}' &= \sqrt{\mathbf{W}}\mathbf{f}, \\
\epsilon' &= \sqrt{\mathbf{W}}\epsilon.
\end{aligned} \tag{43}
$$

An interim model is given by

$$
\mathbf{y}' = \mathbf{Z}'\alpha + \mathbf{f}' + \epsilon', \tag{44}
$$

with the priors

$$\alpha \sim \mathbf{N}(\mathbf{0}, c\mathbf{I}), \tag{45}$$
$$\mathbf{f}' \sim \mathbf{N}(\mathbf{0}, \tau^2 \sqrt{\mathbf{W}}\mathbf{K}\sqrt{\mathbf{W}}),$$
$$\epsilon' \sim \mathbf{N}(\mathbf{0}, \sigma^2\mathbf{I}),$$
$$\tau^2 \sim unif[0, 10^{10}],$$
$$\sigma^2 \sim IG(10^{-8}, 10^{-10}).$$

We can then set $\mathbf{K}' = \sqrt{\mathbf{W}}\mathbf{K}\sqrt{\mathbf{W}}$. This means that $f'(s,t)$ is of zero-mean Gaussian random variables with the covariance function between $f'(s_i,t_i)$ and $f'(s_j,t_j)$ given by

$$cov\{f'(s_i,t_i), f'(s_j,t_j)\} = \mathbf{K}'\{(s_i,t_i),(s_j,t_j)\}.$$

This is just a different choice of the prior for the full bivariate surface $f'(s,t)$. We can then use results from (30) to (39) based on (44) to (45).

We use the Gibbs sampling scheme where the bivariate regression surface is modeled by the thin-plate spline prior as described earlier in the paper. A good introduction to the Gibbs sampler is given by Gelfand and Smith (1990). One of the advantages of Gibbs sampling is that it can take advantage of any additive structure in the model as explained in Wong and Kohn (1996). The sampling scheme is similar to the one used by Wood et. al. (2000) in a model selection context. In our case, the estimates of α, $\mathbf{f}'$, τ^2, and σ^2 are obtained by generating the iterations $\alpha^{[j]}$, $\mathbf{f}'^{[j]}$, $\tau^{2[j]}$, and $\sigma^{2[j]}$ from the sampling scheme described in (45). The constant c is chosen to be a large number ($c = 10^{20}$) so as to ensure that the prior for α is essentially a noninformative flat prior.

4.4 One-Step Ahead Prediction

Many problems in actuarial science involve the building of a mathematical model that can be used to predict insurance costs or to forecast losses in the future, particularly the short-term future. Our approach is to establish a nonparametric regression model $y_{ij} = g(y_{i,j-1}, y_{i,j-2}) + \epsilon_i$, $i = 1, ..., n$, where g is a smooth function of its arguments. This model allows us to describe the dependency between the current observations y_{ij} for all policyholders $i = 1, ..., n$, and those past losses $y_{i,j-1}$ and $y_{i,j-2}$ through a nonparametric regression function at occasion j.

Suppose that we are interested in one step ahead prediction of $Y_{i,j+1}$. We take the posterior mean $E(Y_{i,j+1} \mid \mathbf{y})$ as the best predictor of $Y_{i,j+1}$ and use the posterior $var(Y_{i,j+1} \mid \mathbf{y})$ to obtain the posterior pointwise prediction interval. For convenience, we estimate the posterior mean and variance of $Y_{i,j+1}$ using empirical estimates based on the values of $y_{i,j+1}$ generated during the sampling period by using the model in (30), that is,

$$y_{ij} = \alpha_0 + \alpha_1 y_{i,j-1} + \alpha_2 y_{i,j-2} + f_i + \epsilon_i, \quad i = 1, \cdots, n.$$

To generate $y_{i,j+1}$, we plug in y_{ij} and $y_{i,j-1}$ as covariates. For each iteration, with the generated values of α and $\mathbf{f}$ from the sampling scheme, we have

$$y_{i,j+1} = \alpha_0 + \alpha_1 y_{i,j} + \alpha_2 y_{i,j-1} + f_i + \epsilon_i, \quad i = 1, \cdots, n.$$

Therefore, the prediction is

$$\hat{y}_{i,j+1} = \hat{\alpha}_0 + \hat{\alpha}_1 y_{i,j} + \hat{\alpha}_2 y_{i,j-1} + \hat{f}_i,$$

where $\hat{f}_i$ is the expected noise on y_i given observed data from (22). After a burn-in period, it is assumed the iterations have converged to draws from the posterior distribution. We estimate the posterior mean and the posterior variance of $Y_{i,j+1}$ based on the values of $Y_{i,j+1}$ generated during the sampling period.

5 Nonparametric Regression with Higher Dimensions

Suppose the model is now extended to handle three variables. Similarly, we can treat s_i and t_i as the past losses and incorporate other relevant information as v_i. For example, v_i can be the number of years a policyholder remain in the same policy with the same insurer, or represents different driving age group in auto insurance. Then, the regression model is given by

$$y_i = g(s_i, t_i, v_i) + \epsilon_i, \quad i = 1, \cdots, n,$$

with ϵ_i independent $N(0, \sigma^2)$ and with $g(\cdot)$ of the form

$$y_i = \alpha_0 + \alpha_1 s_i + \alpha_2 t_i + \alpha_3 v_i + f(s_i, t_i, v_i) + \epsilon_i, \quad i = 1, \cdots, n.$$

The prior on f is specified similarly to (26) and (27), that is

$$cov\{f(s_i, t_i, v_i), f(s_j, t_j, v_j)\} = \tau^2 \mathbf{K}\{(s_i, t_i, v_i), (s_j, t_j, v_j)\}$$

where

$$\mathbf{K}\{(s_i,t_i,v_i),(s_j,t_j,v_j)\} = E\{(s_i,t_i,v_i),(s_j,t_j,v_j)\} - \sum_{k=1}^{4} p_k(s_j,t_j,v_j)E\{u_k,(s_i,t_i,v_i)\} - \sum_{k=1}^{4} p_k(s_i,t_i,v_i)E\{(s_j,t_j,v_j),u_k\} + \sum_{k=1}^{4}\sum_{l=1}^{4} p_k(s_i,t_i,v_i)p_l(s_j,t_j,v_j)E\{u_k,u_l\},$$

$$E\{(s_i,t_i,v_i),(s_j,t_j,v_j)\} = \tau^2\log(\tau), \quad \tau = \sqrt{(s_i-s_j)^2+(t_i-t_j)^2+(v_i-v_j)^2},$$

$$p_1(s_i,t_i,v_i) = -1 + 2s_i + 2t_i + 2v_i,$$
$$p_2(s_i,t_i,v_i) = 1 - 2s_i,$$
$$p_3(s_i,t_i,v_i) = 1 - 2t_i,$$
$$p_4(s_i,t_i,v_i) = 1 - 2v_i$$

and

$$u_1 = (\frac{1}{2},\frac{1}{2},\frac{1}{2}), u_2 = (0,\frac{1}{2},\frac{1}{2}), u_3 = (\frac{1}{2},0,\frac{1}{2}), u_4 = (\frac{1}{2},\frac{1}{2},0).$$

This is because of (13), where $m = 2$, $d = 3$, and

$$
\begin{aligned}
M &= \binom{d+m-1}{d} \\
&= \binom{4}{3} \\
&= 4.
\end{aligned}
$$

Without loss of generality, we assume that the variables s, t, and v all lie in the interval $[0, 1]$.

Let $\alpha = (\alpha_0, \alpha_1, \alpha_2, \alpha_3)'$ be the vector of linear regression parameters, and let

$$
\mathbf{Z} = \begin{bmatrix} 1 & s_1 & t_1 & v_1 \\ 1 & s_2 & t_2 & v_2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & s_n & t_n & v_n \end{bmatrix}
$$

The priors for α, the smoothing parameter τ^2, and σ^2 are the same as in (35) which gives us

$$
\begin{aligned}
&\alpha \sim \mathbf{N}(\mathbf{0}, c\mathbf{I}), \\
&\mathbf{f} \sim \mathbf{N}(0, \tau^2\mathbf{K}), \\
&\epsilon \sim \mathbf{N}(0, \sigma^2\mathbf{I}), \\
&\tau^2 \sim unif[0, 10^{10}], \\
&\sigma^2 \sim IG(10^{-8}, 10^{-10}).
\end{aligned}
$$

The model implementation and the sampling scheme will be exactly the same as in the bivariate model.

6 Application to medical insurance data

In this section, the results of Bayesian nonparametric regression model for the Bühlmann-Straub type data with unequal exposure units will be illustrated by an application to a collective medical health insurance data from an insurance company in Taiwan. We consider a portfolio consisting of thirty-five group policyholders that has been observed for a period of three years. The claim associated with group j $(= 1, \cdots, 35)$ in year of observation t $(= 1, 2, 3)$ is represented by the random variable Y_{jt}, which is an average taken over w_{jt} employee. We choose groups with moderate group size (23 to 80 individuals), and assume that the number of employee does not change over the periods. Therefore, we have $w_{jt} = w_j$ for all t, and the claim Y_{jt} with weights w_j fulfill the Bühlmann-Straub assumptions. Table 2 shows some observed realizations of the Y_{jt}, and the numbers of employee w_j. We want to determine the estimated premium to be charged to each group in year 4. The data is in new Taiwan dollar (NTD). The exchange rate is about 1 US dollar to 35.16 new Taiwan dollar in March, 2002.

We consider the semiparametric regression model discussed in section 4.3,

$$y_{j,t} = \alpha_0 + \alpha_1 y_{j,t-1} + \alpha_2 y_{j,t-2} + f(y_{j,t-1}, y_{j,t-2}) + \epsilon_i, \quad j = 1, \cdots, n, \tag{46}$$

where $\epsilon_i \sim N(0, \frac{\sigma^2}{w_i})$, $i = 1, \cdots, n$. The function $f(y_{i,t-1}, y_{i,t-2})$ has zero initial conditions and is estimated nonparametrically using the priors in (42). The scatter plot, fitted surface by (46), and fitted surface by the Bühlmann-Straub (BS) model are shown in Figure 1. The contour plots shown in Figure 2 provide a better look of different levels of surfaces.

To examine the performance of the regression function, the average squared error (ASE) was calculated for the estimates of the regression functions. The ASE is calculated as follows

$$ASE = \frac{1}{\sum_{j=1}^{n} w_j} \sum_{j=1}^{n} w_j (\hat{g}(s_j, t_j) - g(s_j, t_j))^2. \tag{47}$$

The ASE of the bivariate spline nonparametric regression model is about 214.26, while that of the Bühlmann-Straub model is 2208.07. Clearly, the bivariate spline nonparametric regression model outperforms the Bühlmann-Straub model.

Our goal is to determine the estimated premium to be charged to each group in year 4. We perform one-step ahead prediction discussed in section 4.5. We estimate the posterior mean and variance of $Y_{j,4}$ using empirical estimates based on the values of $y_{j,4}$ generated during the sampling period. Figure 3 shows some of the posterior distributions of $Y_{j,4}$. Some of the estimated premiums together with 95 percent posterior pointwise prediction intervals (in parenthesis) are shown in Table 3. For example, for group 3, the estimated premium is 5965.36 NTD for each employee in this group, and the total estimated premium is $49 \times (5965.36) = 292302.64$ NTD. Similar calculations can be done for other groups.

7 Conclusions

Many problems in actuarial science involve the building mathematical models that can be used to predict insurance cost in the future, particularly the short-term future. A Bayesian nonparametric approach is proposed to the problem of risk modeling. The model incorporates past and present observations related to the risk, as well as relevant covariate information, and uses MCMC technique to compute the predictive distribution of the future claims based on the available data, where the covariance structure is based on a thin-plate spline (Wahba, 1990).

We have illustrated applications of Gibbs sampling within the context of non-parametric regression and smoothing. Gibbs sampling provides feasible approach to the computation of posterior distributions. Combined with assumed thin-plate spline structure of the regression surface and the computational availability of the bivariate or trivariate surface estimation, this methodology opens up a new dimension to credibility literature. Although our discussion concentrates primarily on two and three-dimensional applications, the technique can be easily extended to higher dimensional problems. Our investigation shows that this method performs at a superior level compared to the existing techniques in the credibility literature.

In this paper, we have outlined a new approach to modeling actuarial and financial data. The model uses a Bayesian nonparametric procedure in a novel manner by incorporating a Gaussian prior on function space. We believe that this procedure provides a flexible approach to function estimation and can be used successfully in the statistical analyses of a wide range of important problems.

	Policyholder	Year 1	Year 2	Year 3	Year 4
Average claim	1	5419.09	1691.38	5984.65	?
No. in group		74	74	74	74
Average claim	2	5603.50	4150.12	5797.48	?
No. in group		52	52	52	52
⋮	⋮	⋮	...	⋮	⋮
Average claim	35	4554.38	4646.96	5059.80	?
No. in group		80	80	80	80

Table 2: Average claims in group policyholders during three years.

	Policyholder	Year 4
Average claim	3	5965.36 (3080.30, 8877.31)
No. in group		49
Average claim	5	5485.61 (1429.58, 9398.640)
No. in group		45
Average claim	9	5024.05 (−595.66, 10495.44)
No. in group		42
Average claim	14	5000.68 (2173.55, 7648.52)
No. in group		31
Average claim	20	6437.74 (2558.72, 7116.71)
No. in group		36
Average claim	24	5217.51 (−361.08, 10462.63)
No. in group		35
Average claim	26	4959.50 (308.02, 9345.16)
No in group		42
Average claim	35	5074.70 (4620.35, 5497.48)
No. in group		80

Table 3: Estimated average claim for year 4 with 95 percent posterior pointwise prediction interval in parenthesis.

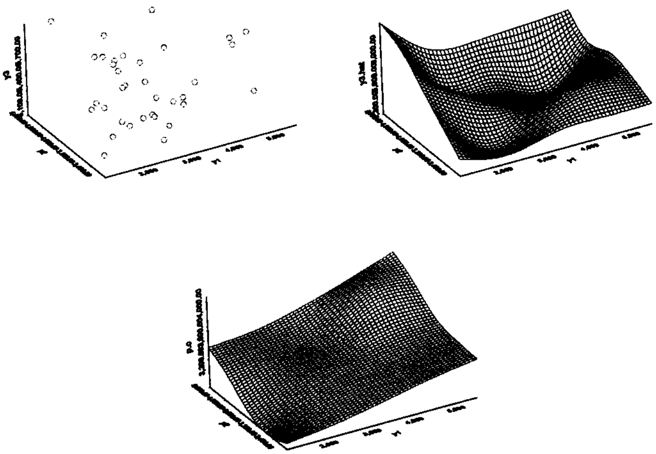


Figure 1: Surface Plots. (a) Scatter plot. (b) Plot of regression surface as a function of y_{t-1} and y_{t-2}. (c) Plot of BS model surface.

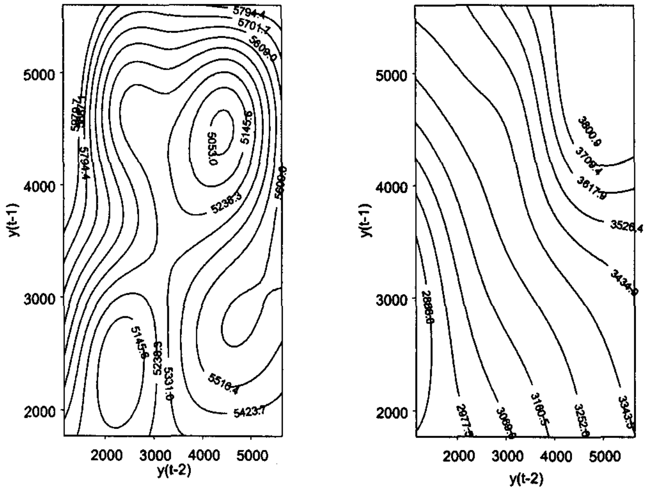


Figure 2: Contour Plots. (a) Plot of regression function as a function of y_{t-1} and y_{t-2}. (b) Plot of BS model.

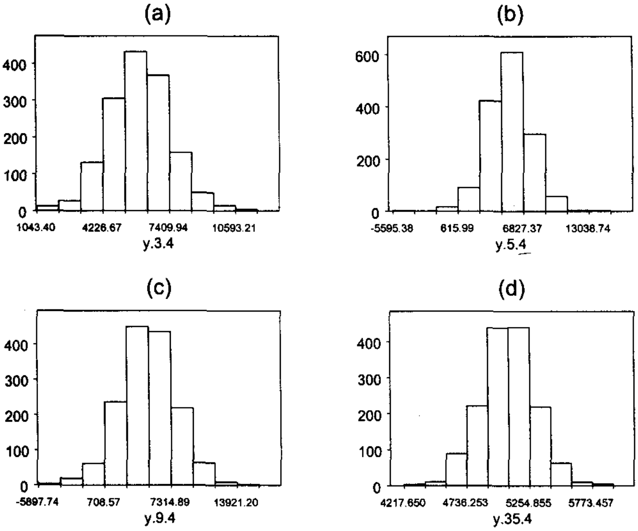


Figure 3: Posterior Plots. (a) $Y_{3,4}$. (b) $Y_{5,4}$. (c) $Y_{9,4}$. (d) $Y_{35,4}$.

References

1. Aronszajn, N. Theory of Reproducing Kernels. Transactions of the American Mathematical Society, 68:337-404, 1950.

2. Bühlmann, H. Experience Rating anf Credibility. ASTIN Bulletin, 4:199-207, 1967.

3. Duchon, J. Functions Splines et Vecteurs Aleatoires. Technical Report 213, Seminaire d'Analyse Numeriaue, 1975.

4. Duchon, J. Functions-spline et Esperances Conditionnelles de Champs Gaussians, Annal Science University Clermont Ferrand II Mathematics, 14:19-27. 1976.

5. Duchon, J. Spline Minimizing Rotation-Invariant Semi-Norms in Sobolev Spaces. in Constructive Theory of Functions of Several Variables, Springer-Verlag, Berlin, 1977.

6. Evgeniou, T., Pontil, M., and Poggio, T. Regularization Networks and Support Vector Machines. Advances in Computational Mathematics, 13:1-50, 2000.

7. Gelfand, A.E. and Smith, A.F.M. Sampling-based Approaches to Calculating Marginal Densities. Journal of the American Statistical Association, 85:398-409, 1990.

8. Meinguet, J. Multivariate Interpolation at Arbitrary Points Made Simple. Journal of Applied Mathematical Physics (ZAMP), 30:292-304, 1979.

9. Wahba, G. Spline Models for Observational Data. SIAM, Philadephia, 69, 1990.

10. Wahba, G. An Introduction to Model Building with Reproducing Kernel Hilbert Space. Technical Report NO. 1020, Department of Statistics, University of Wisconsin, Madison, WI, 2000.

11. Wahba, G. and Wendelberger, J. Some New Mathematical Methods for Variational Objective Analysis using Splines and Cross-Validation. Monthly Weather Review, 108:1122-1145, 1980.

12. Weinert, H.L (Editor). Reproducing Kernel Hilbert Spaces: applications in statistical signal processing. Hutchinson Ross Pub. Co., 1982.

13. Wendelberger, J. Smothing Noisy Data with Multidimensional Splines and Generalied Cross Validation. Ph.D. Thesis. Department of Statistics, University of Wisconsin, Madison, WI, 1982.

14. Whittaker, E.T. On A New Method of Graduation. Proceedings of Edinburgh Mathematical Society, 41:63-75, 1923.

15. Wang, C. and Kohn, R. A Bayesian Approach to Additive Semiparametric Regression. Journal of Econometrics, 7:209-223, 1996.

16. Wood, S., Kohn, R., Shively, T., and Jiang, W. Model Selection in Spline Nonparametric Regression. Preprint, 2000.

Classification Ratemaking Using Decision Trees

Nasser Hadidi, Ph.D., FCAS, MAAA

Classification Ratemaking Using Decision Trees

By

Nasser Hadidi, Ph.D., FCAS, MAAA

Introduction

Manual rating of specific risks begin with a base rate, which is then modified by appropriate relativity factors depending on characteristics of each risk. Classical methods of deriving indicated relativities, are described by McClenahan (1996) and Finger (1996). A number of different modeling procedures are described in Brown's (1988) "minimum bias" paper and Venter's (1990) review of Brown's paper. These methods generally rely on the "multiplicative" or "additive" assumptions, which may not be reasonable for all types of risk. In this paper an alternative method of calculating indicated relativities is described, and demonstrated using a commercial Business Owners' Product (BOP) data set. Accident year 1997 to 2000 data is used to describe the method. The results are then applied to claims with accident year 2001. The derived 2001 relativities are compared with observed relativities, thereby demonstrating the extent of suitability of this method. It should be stressed that the intent here is entirely demonstration of a procedure. For actual practical implementation, modification would be required.

First a few words about the terminology and the data set. Relativities are based on grouping of risks with similar risk characteristics. This is essentially a classification problem. The purpose of any classification procedure is partitioning of objects - in our case risks - into demonstrably more homogeneous groups. For the BOP data we seek groups of risks with significantly differing claim frequencies, severities, pure premiums or loss ratios. Typically partitioning is based on a number of risk factors, which for BOP might be Coverage, Risk State, ISO territory, ISO coverage code, Property versus Liability, etc.

Distinctions must be made between these rating factors, which are used to group risks together and variables such as frequency, severity, pure premium or loss ratio, which must be estimated partially based on these risk factors. The former are independent or predictor variables while the latter are dependent variables. Both of these variables may be assumed to be either categorical or vary continuously in a given interval. The statistics used as the basis of classification depends on whether the dependent and/or the independent variables are categorical or interval scale. For risk classification the independent variables are typically categorical. For example the BOP data includes losses for three different coverages, in 51 different risk states, and with 178 different ISO territory codes, 6 different ISO construction codes, etc. It is customary to say that the classification variable – risk state - has 51 different levels. Similarly, there are six levels of ISO construction code, etc.

The BOP data set under consideration includes 27,854 claims with accident years 1997 through 2000, and 9011 claims with accident year 2001. These are broken down by:

	Number of Levels
Risk State	51
Company	4
ISO protection code	1014
ISO territory code	178
ISO construction code	6
Property/Liability indicator	2
Building/Content/Other indicator	3
Coverage code	3
ISO subline	2
ISO coverage code	4,

as well as a few other variables.

The data includes paid and incurred loss and expense (combined), basic limit (300k) loss and expense, and excess as well as cat losses.

The customary and classical method of splitting data for the purposes of deriving relativities is to consider all possible combinations of levels of permissible classification factors. For auto coverage, for example, in some states a 2x2 table of gender by rural/urban may be considered. Or in this case for any risk state, company and territory we may have a typical 2x3 table of property versus liability by building versus content versus other. The biggest drawback of such splits is that a specific rating factor may have a large impact on one level of another rating factor, but no impact at all on other levels. Or if it does, the nature of the impact may be quite different. For example ISO construction code may influence losses in one way in one ISO territory and in no way or a completely different way in another territory. The methodology described here is intended to avoid this drawback without the necessity of introducing numerous "interaction" parameters.

Rather than splitting the data in all possible ways, the data can be sequentially and selectively split in mutually exclusive homogeneous groups. Splits would be made only if clearly indicated by data to be meaningful. Meaningful in the sense that, risks are indeed more homogeneous in whatever it is that we are measuring- typically frequency or severity. The term decision "tree" is derived from this sequential approach. That is, start with the whole data - appropriately split the data into branches by one "best" rating factor to begin with. Then concentrate on one branch by itself, disregarding the rest of the branches, and find the next "best" rating factor for it, thereby creating new branches for this branch. Then concentrate on the next branch and find the next "best" rating factor for it, which may be – and in fact often is - a different rating factor. New branches are then created for this second branch and so on. Details of how exactly this can be done in practice is given below, but first it should be stated that this method is based on principles underpinning a large number of classification procedures – referred to as classification or decision tree procedures – which are amply reported in the statistical data mining literature.

These procedures appear with a variety of acronyms including:

AID Automatic Interaction Detection

THAID	Theta AID
CHAID	Chi-Squared Automatic Interaction Detection
Exhaustive CHAID	Modified CHAID
C&RT / CART	Classification And Regression Trees
QUEST	Quick, Unbiased, Efficient Statistical Tree
FACT	Fast Algorithm for Classification Trees
FIRM	Formal Inference-based Recursive Modeling
C4.5	A set of computer programs that construct classification models
ID3	The predecessor of C4.5
GUIDE	Generalized Unbiased Interaction Detection and Estimation
CRUISE	Classification Rule with Unbiased Interaction Selection and Estimation

Etc.

AID, THAID, CHAID and Exhaustive CHAID

AID was first described by Morgan and Sonquist in1963 as a sequential procedure for analysis of survey data. It is intended to avoid the problem of interaction between variables used for classification. In the case of classifying risks, the problem of interaction translates into the possibility that type of coverage, for example, may have a different impact on rates for one territory as opposed to another territory. They propose *bisecting* the data sequentially, one factor at a time, based on maximizing the between levels sum of squared deviation. This is somewhat similar to the ordinary analysis of variance procedure, though in their 1963 paper they propose stopping the splits simply when the reduction in error sum of squares is less than a specified value. THAID, which was proposed by Messenger and Mandell in 1972, similarly *bisects* data, but based on a different statistic. This statistic, which they call THeta, is related to the proportional reduction in misclassification errors. CHAID was described by Kass in (1980). As the acronym indicates the predominant statistic used for splits in this procedure is the Chi-square statistic. Whereas for AID the dependent variable is interval scaled, here the dependent variable is nominal. As an example one would look at the overall proportion of policies that resulted in none, one, or 2 or more claims; that is three categories. And then split policies in groups so that the proportion of policies in each category would differ significantly between groups. Independent variables also being categorical, typical two way contingency tables are constructed and Chi-square statistics are calculated which form the basis of the splits. This procedure is demonstrated by Gallager Monroe, and Fish (2001) for private passenger automobile experience. Exhaustive CHAID (spss.com) is based on Biggs, de Ville, and Suen. (1991). It is a refinement and expansion of the method given by Kass. The significance level of the utilized Chi-square statistic is appropriately adjusted for the number of independent variables.

CART, C&RT, QUEST, FACT , GUIDE and CRUISE

CART, (cart.com) was introduced by Breiman, Friedman, Olshen, and Stone in (1984). It is similar to AID in that to achieve the final classification a series of binary splits are made. But it is far different from AID or CHAID in the splitting mechanism. Here at each step of the classification a series of queries is made regarding the value of each of the independent variables. For categorical independent variables such queries take the form of whether or not each case belongs to a given subset of levels of each independent variable. All possible subsets of all independent variables are considered. For interval scale independent variables the percentage of cases with values less than all observed values of this variable are considered. A misclassification cost is then calculated and the split is based on minimizing that misclassification cost. C&RT is another vendor's (spss.com) version of CART. QUEST is proposed by of Loh and Shih (1997). It is intended to reduce the bias in favor of splits that are based on independent variables for which more branching is possible. Categorical independent variables with more levels and interval scale independent variables with more distinct values are more likely to be selected first in classification tree procedure. This bias is a frequently recurring criticism of these classification procedures and several efforts in minimizing the bias are reported in the literature. FACT is also proposed by Loh and Vanichsetakul, (1988). It is described as an algorithm combining CART and Linear Discriminant Analysis (LDA). Discriminant analysis is the classical method of predicting group membership based on predictive characteristics. Depending on the number of groups one or more discriminant functions are estimated from the data. These functions are linear combinations of independent variables, and are in turn used to predict group membership. The values of these functions are, ideally, substantially different for each group. This would be the case if predictors have sufficient discriminating information. FACT differs from CART in that it uses a different misclassification cost based on these discriminant functions. GUIDE is proposed by Loh (2002). It is also intended to eliminate the variable selection bias. As mentioned for QUEST this bias refers to the fact that categorical independent variables with more levels as well as interval scale variables with more distinct observed values are more likely to be selected first in the tree structure. The bias is eliminated by an adjustment to the Chi-square p-value. CRUISE is the described by Kim and Loh (2001). It borrows ideas from FACT, QUEST, GUIDE, and CART, and is claimed to be faster and further reduce the variable selection bias.

FIRM , C4.5 and ID3

FIRM is a collection of codes presented by Hawkins (1990) for implementation of CHAID. Two versions, CATFIRM and CONFIRM, are given respectively for categorical and interval scale dependent variable. Details of the procedure are given in Hawkins and Kass (1982). Here essentially the interval scale variables are converted to categorical variables by clustering adjacent values in one category. C4.5 and its predecessor ID3 are presented by Quinlan (1993). They are a collection of computer programs that construct classification trees. The construction method is based on what they refer to as "divide and conquer algorithm" which uses the "gain" criterion. They refer to the data as "training set" and for any split of the training set the gain is defined in terms of the information or entropy obtained thereby.

Procedure Description

Almost all of the above procedures are packaged, some more elegantly than others, and are available commercially. But none can be used without modifications with actuarial data since they are not specifically designed as such. Many of the splits automatically tested in these procedures are meaningless for actuarial data. CART would routinely test if a BOP policy belongs to *all subsets* of the rating variable ISO construction code. Clearly only subsets including only one element are meaningful for actuarial data. None address the credibility issue. Furthermore routine 'black box' style use of these packages usually mask the statistics used as splitting criteria. It is therefore not clear whether the assumptions required, especially with regard to the distribution of such statistics are indeed valid for the data at hand. Therefore, with actuarial data the common underlying principles of these procedures should be grasped, modified appropriately and implemented directly.

These underlying principles are the sequential consideration of the rating factors, splitting the data based on an appropriate metric and at each split combining the levels of each rating factor as long as they are not significantly different. How this can be done in practice would now be demonstrated using the described BOP data. For this illustration the natural log of basic limit losses is considered the dependent variable, and the following factors are independent variables:

Coverage code, Risk State, Company, ISO protection code, ISO territory code, ISO construction code, Property/Liability indicator, Building/Content/Other indicator, ISO Subline, ISO coverage code

The selected metric is the F statistics (or equivalently its p-value) given by the ratio of between and within mean squares as described below. The choice of this statistic is justified by the fact that basic limit losses here very closely follow the lognormal distribution. It is essential to check this lognormal assumption which results in the F distribution for the mean square ratios when we use log of losses.

Splits will not be made if the p-value of the F statistic is more than 0.01 or the resulting splits will have less than 200 claims.

As stated earlier risk state, ISO protection code, and ISO territory code have, 51, 1014, and 178 levels respectively. Most of these levels have very few claims. Risk state '54' (Alaska) and '99' (miscellaneous) have 3 and 1 claims respectively. Therefore before any analysis, for each factor the number of levels is reduced by appropriate level combinations and/or introduction of an 'all other' category. The exact number of levels to reduce to is not crucial at this stage of the analysis since groups will be recombined objectively with the tree structure in the next steps. Simply inspecting the mean severity by level of each factor *along with the standard errors* of each mean (confidence interval) provides an adequate means of combining levels.

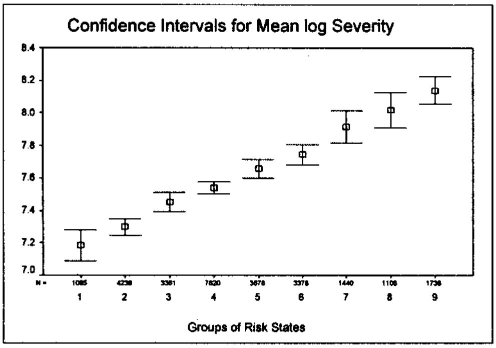
In this fashion the rating factor risk state was combined into nine distinct groups. Similarly the levels of ISO protection code, and ISO territory were regrouped into 9 and 8 levels respectively. Exhibit 3 is a description of these recodes along with the number of claims in each group.

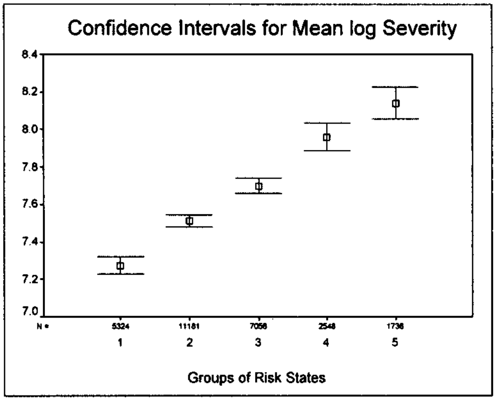
A standard multivariate split of this data would result in at most
(9 Risk State)x(4 Company)x (9 ISO protection code)x(8 ISO territory code)x
(6 ISO construction code)x(2 Property/Liability indicator)x(3 Building/Content/Other indicator)x(3 Coverage code)x(2 ISO subline)x(4 ISO coverage code)=2,239,488 cells,
The whole idea of this method is which of these 2,239,488 cells are indeed materially different from the rest, and must be evaluated individually. The tree structure is intended to isolate these significantly different cells from the total. As shown below for the BOP data only 21 tiers or nodes , need be considered separately. Obviously each risk would belong to one and only one of these 21 tiers. Here is how the procedure works.

For the entire data, the so called node 0, the mean and standard deviation of log severity based on 27,845 claims is 7.5944 and 1.7692 respectively. At this stage the "best" predictor of this log severity is risk state. The best predictor means the predictor producing the highest F ratio, which here is the selected metric of choice for splitting, or equivalently the lowest p-value for the F statistic. This factor would divide risks into five groups:

Risk States	log Severity	Number of claims
9	8.1405	1736
7,8	7.9598	2548
5,6	7.6994	7056
3,4	7.5127	11181
1,2	7.2742	5324
All	7.5944	27845

The following graphs demonstrate the reasoning behind these combinations.





From the first graph it is observed that there is no clear reason not to combine 1 and 2, 3 and 4, 5 and 6, 7,8 and 9. But once 7 and 8 are combined, 9 would be significantly different from that combination. From the second graph it is clear that the five new groups have significantly different means. Graphical descriptions aside, the F statistic is the ratio of between mean square

$[1736(8.1405-7.5944)^2 + 2548(7.9598-7.5944)^2 + 7056(7.6994-7.5944)^2 + 11181(7.5127-7.5944)^2 + 5324(7.2742-7.5944)^2]/4 = 389.0511$

and within (error) mean square

$(27844x1.7692^2 - 4x389.0511)/27840= 3.0746$

This ratio equals 126.5, which is of course highly significant leading to the necessity of the above split.

Let us now concentrate on states 3 and 4, disregarding other states for now. For this branch the next best predictor is property/liability indicator, which again based on the F ratio of 69.0 divides risk into 2 branches:

	log Severity	Number of claims
Property	7.2632	2533
Liability	7.5858	8648
All	7.5127	11181

Next, consider liability claims in states 3 and 4. These have to be broken down by ISO territory. The resulting F statistic is 28.5 based on 3 distinct groups of territories:

Territory	log Severity	Number of claims
2,6,7	7.6988	4357
1,3,4,5	7.5001	3961
8	7.1206	330
All	7.5858	8648

Next, consider liability claims in states 3 and 4 and ISO territories 1,3,4,5. These have to be broken down by ISO constuction code. The resulting F statistic is 16.35.

Construction	log Severity	Number of claims
3,4,5,6	7.6257	1632
1,2	7.4121	2329
All	7.5001	3961

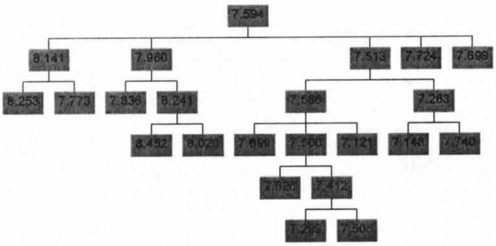
Next, consider liability claims in states 3 and 4, ISO territories 1,3,4,5 and ISO construction code 1,2. These have to be broken down by building/content indicator. The resulting F statistic is 9.61.

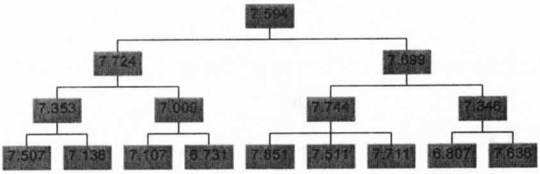
	log Severity	Number of claims
Building	7.2992	1050
Cotent	7.5049	1279
All	7.4121	2329

With the constraint of a p-value less than 0.01 and at least 200 claims, no further splits based on any other independent variable is implied by this procedure.

In this manner a number of distinct tiers, or the so called terminal nodes can be identified.

It is customary to depict these tiers with a tree structure as follows:





There are corresponding structures containing tier standard deviations and sample sizes needed for credibility adjustments, which are given in Exhibits 1 and 2.

The point is that while the overall mean is 7.594, 21 distinct tiers have in this manner been identified with means ranging from as low as 6.731 and as high as 8.452.

The profiles of risks in each tier are listed below:

Tier	Profile	log Severity	Number of Claims
1	Risk State 1 Liability	6.7313	317
2	Risk State 5,6 ISO Territories 4,6,8 ISO Subline 2	6.8069	279
3	Risk State 2 Liability	7.1072	898
4	Risk State 3,4 ISO Territories 8 Property	7.1206	330
5	Risk State 1,2 ISO Territories 2,5,6,8 Property	7.1355	1708
6	Risk States 3,4 Liability Building, Other	7.1477	2039

7	Risk State 3,4 ISO Territories 1,3,4,5 Property ISO Construction 1,2 Building, Other	7.2992	1050
8	Risk State 3,4 ISO Territories 1,3,4,5 Property ISO Construction Code 1,2 Content	7.5049	1279
9	Risk State 1,2 ISO Territories 1,3,4,7 Property	7.5071	2401
10	Risk State 5,6 ISO Territory Code 1,2,3,5,7 ISO Construction Code 1	7.5107	989
11	Risk State 3,4 ISO Territory Code 2,6,7 Property	7.6988	4357
12	Risk States 3,4 ISO Territories 1,3,4,5 Property ISO Construction Code 3,4,5,6	7.6257	1632
13	Risk State 5,6 ISO Territories 4,6,8 ISO Subline 1	7.6359	518

14	Risk State 5,6 ISO Territories 1,2,3,5,7 ISO Construction Code 2,3,5	7.7114	2359
15	Risk State 3,4 Liability Content	7.7398	494
16	Risk State 9 Building	7.7726	407
17	Risk State 5,6 ISO Territories 1,2,3,5,7 ISO Construction Code 4,6	7.8505	2911
18	Risk State 7,8 Coverage 1,3	7.8360	1769
19	Risk State 7 Coverage 2	8.0208	381
20	Risk State 9 Content, Other	8.2532	1329
21	Risk State 8 Coverage 2	8.4521	398
All		7.5944	27,845

Based on this procedure, certain independent variables do not impact claim costs at most levels of other independent variables. Coverage for example is only relevant for risk state groups 7 and 8.

Credibility Adjustments

Consider tier 20 with mean and standard deviations equal to 8.2532 and 1.8208 respectively. The estimated mean here is

$$e^{8.2532+0.5\times1.8208^{\wedge}2} = 20{,}148$$

which once compared with the overall mean of

$$e^{7.5944+0.5\times1.7692^{\wedge}2} = 9{,}504$$

results in relativity of 2.120. This figure is based on 1329 claims. So it is not fully credible.

The standard of full credibility utilized here is to be within one percent of the estimated mean with a probability of 0.99. As stated before limited losses being very closely distributed as a lognormal random variable, for this class the full credibility standard would be at least

$$(2.575\text{x}1.8208/0.01\text{x}8.2532)^2 = 3{,}227$$

claims. Using the square root rule, the partial credibility of 2.120 is thus

$$(1329/3227)^{0.5} = 0.642\,.$$

The complement of credibility is assigned here to the non-terminal node immediately preceding this tier, the so called parent node. If there are not sufficient claims in this parent node to attain full credibility one can first adjust this relativity with its own parent node before using it as a complement. In this case the parent of tier 20 has log severity mean and standard deviation of 8.1405 and 1.8078 respectively, giving the severity mean of

$$e^{8.1405+0.5\times1.8078^2} = 17{,}581$$

and relativity of 1.850. But the number of claims here is only 1736 not reaching its own full credibility standard of

$$(2.575\text{x}1.8078/0.01\text{x}8.1405)^2 = 3{,}270.$$

The 1.850 estimate therefore has partial credibility of 0.729. The node immediately preceding this node has a relativity of 1, which has full credibility. Therefore the complement of credibility for tier 20 is attached to

0.729x1.850 + 0.271x1 = 1.619

Hence the credibility adjusted estimate of relativity for tier 20 is

0.642x2.120 + 0.358x1.619 = 1.941.

The necessary calculations for all 21 tiers are given in Exhibits 4-9.

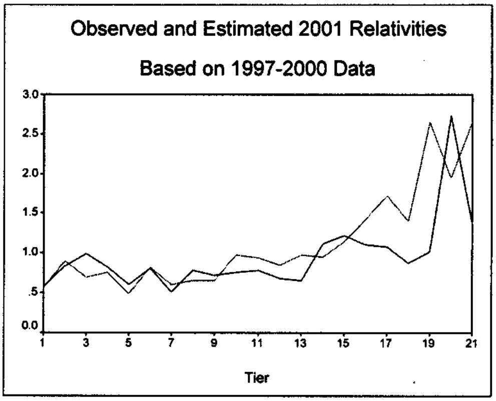
Cross Validation

As stated earlier the BOP data set includes 9011 claims with accident year 2001 which were not used for this classification scheme. These were deliberately left out for cross validation of the procedure. The accident year 2001 claims are grouped into 21 tiers based on the above scheme. For example coverage 2 claims in risk state 8 form tier 21, etc. Tier 1 includes 126 claims with the observed mean of $7,053. This value compared with the overall observed mean of $12,253 gives an observed relativity of **0.575.** How does this compare with the estimated relativity based on 1997-2000 data? The unadjusted relativity in that tier computed as explained above and listed in Exhibit 4, is 0.392 with a partial credibility of $(317/4368)^{0.5} = 0.269$. The complement of this credibility is assigned to the adjusted relativity of the parent node, which is

0.509x0.664 + 0.491x0.604 = 0.634. Thus the credibility adjusted estimated relativity of tier 1 is 0.269x0.392 + 0.731x0.634 = **0.569.**

This sort of calculation and comparison of course has to be done for all tiers. Details are given in Exhibits 4-9 in the appendix. The resulting credibility adjusted relativities and the actual observed 2001 relativities are listed below. The extent of association between these values can be observed from the chart.

Accident Year 2001 Observed Relativities	Credibility Adjusted Relativities
0.575	0.569
0.833	0.904
0.994	0.700
0.820	0.760
0.602	0.492
0.813	0.820
0.510	0.602
0.791	0.652
0.717	0.658
0.758	0.975
0.781	0.920
0.677	0.853
0.655	0.982
1.116	0.951
1.217	1.148
1.107	1.431
1.086	1.729
0.870	1.405
1.021	2.651
2.725	1.941
1.384	2.655



Summary and Conclusions

Classical methods of deriving rate relativities, are based on either a univariate or multivariate analysis of the data. The former requires the additive or multiplicative assumption and the latter may require estimation of numerous interaction parameters. An alternative method based on classification tree procedures is described in this paper. It is shown how risks with homogeneous loss severities can be grouped, based on appropriate combinations of levels of rating factors. Using accident year 1997-2000 data for a particular product, relativities are computed for 2001 accident year claims. With appropriate adjustment for credibility, these relativities are then compared with actual observed relativities demonstrating the suitability of this method. Because of the particular description of risk tiers, results obtained by these procedures might be somewhat difficult to implement. However, as an additional underwriting guideline, especially when deviating from manual rates, these procedures can be quite useful.

Acknowledgements

Valuable suggestions made by Christopher Monsour, and Charles H. Boucek of the CAS ratemaking committee and Jim Mohl is hereby gratefully acknowledged.

References

Biggs, D., B. de Ville, and E. Suen. (1991). A method of choosing multiway partitions for classification and decision trees, Journal of Applied Statistics, 18: 49-62.

Breiman L., Friedman J.H., Olshen, R.A. and Stone C.J. (1984), Classification and Regression Trees, Wadsworth, Inc.

Brown R.L. (1988) Minimum Bias with Generalized Linear Models, PCAS LXXV, pp 187-218.

Finger R.J. (1996) Risk Classification, Chapter 5, Foundation of Casualty Actuarial Science, Third Edition, Casualty Actuarial Society, 233-281.

Gallager C.A., H. M. Monroe, and J.L. Fish (2001) An Iterative Approach to Classification Analysis.

Hawkins D.M. and G.V. Kass (1982) 'Automatic Interaction Detection' in Topics in Applied Multivariate Analysis, ed. D. M. Hawkins, Cambridge University Press.

Hawkins D.M. (1990), FIRM Formal Inference-based Recursive Modeling, Technical Report #546, Department of Applied Statistics, University of Minnesota, St. Paul.

Kass G.V. (1980), An Exploratory Technique for Investigating Large Quantities of Categorical Data, 29, No. 2, 119-172.

Kim H. and Loh W.Y. (2001) Classification Trees with Unbiased Multiway Splits, Journal of American Statistical Association, 96, 589-604.

Loh W.Y. and Vanichsetakul N. (1988) Tree-Structured Classification Via Generalized Discriminant Analysis, Journal of American Statistical Association, 83, 715-728.

Loh W. Y. and Shih Y. S. (1997) Split Selection Methods for Classification trees Statistica Sinica 7, 815-840.

Loh W. Y. (2002) Regression Trees with Unbiased Variable Selection and Interaction Detection, Statistica Sinica, to appear.

McClenahan C.L. (1996) Ratemaking, Chapter 2, Foundation of Casualty Actuarial Science, Third Edition, Casualty Actuarial Society, 25-89

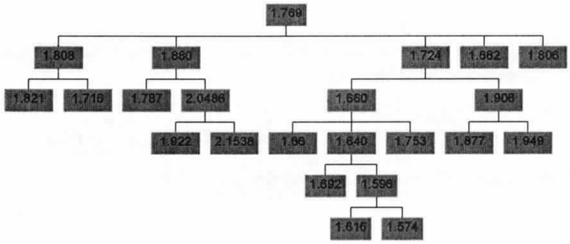
Messenger R.C. and Mandell L. M. (1972), A model search technique for predictive nominal scale multivariate analysis, Journal of American Statistical Association, 67, 768-772.

Morgan J. A. and Sonquist J. N. (1963), Problems in the analysis of survey data: and a proposal, Journal of American Statistical Association, 58, 415-434.

Quinlan J. R. (1993) C4.5 Programs for Machine Learning San Mateo, CA , Morgan Kautmann.

Venter G.G. (1990) Discussion Paper, PCAS LXXVII, pp 337-349.

Exhibit 1
Node log Severity Standard Deviations



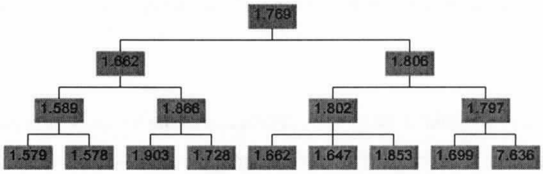
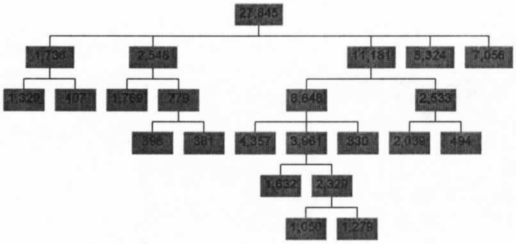


Exhibit 2
Node Number of Claims



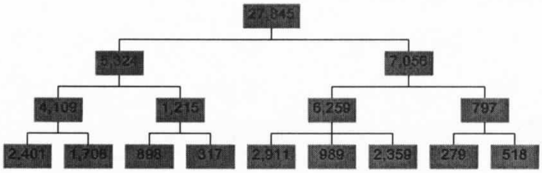


Exhibit 3

Recoded Levels of Risk Factors

Risk State

5,45,32,49,36,3,26,18	1
10,37,35,22,28	2
47,2,25,30,43,15	3
29,17	4
6,16,21,8,38,42	5
12,19,41,27,46,23,24,1,39,44,33,54	6
31	7
9,20,4,7	8
34,14,11,13,48,99,40	9

Company

BD	1
BE	2
BG	3
Others	4

ISO Protection Code

1,10	1
2	2
3	3
4	4
5	5
6	6
7,11	7
8,9	8
Missing	9

ISO Territory Code

1	1
2	2
3	3
4	4
5	5
6	6
Missing	8

Others	7

ISO Construction Code

1	1
2	2
3	3
4	4
Missing	6
Others	5

ISO Coverage Code

21	1
22	2
Missing	3
Others	4

Coverage

81	1
84	2
Others	3

ISO Subline

915	1
Missing	2

Exhibit 4

Unadjusted Observed Relativities by Tier

Tier	(a_1) Underlying Normal Mean	(b_1) Number of Claims	(c_1) Underlying Normal Std. Deviation	(d_1) Mean Severity	(e_1) Unadjusted Relativity
1	6.7313	317	1.7276	3,728	0.392
2	6.8069	279	1.8531	5,034	0.530
3	7.1072	898	1.9032	7,467	0.786
4	7.1206	330	1.7528	5,749	0.605
5	7.1355	1708	1.5776	4,359	0.459
6	7.1477	2039	1.8772	7,403	0.779
7	7.2992	1050	1.6163	5,461	0.575
8	7.5049	1279	1.5741	6,272	0.660
9	7.5071	2401	1.5785	6,329	0.666
10	7.5107	989	1.6621	7,273	0.765
11	7.6988	4357	1.6600	8,748	0.920
12	7.6257	1632	1.6927	8,590	0.904
13	7.6359	518	1.6985	8,764	0.922
14	7.7114	2359	1.6474	8,676	0.913
15	7.7398	494	1.9485	15,339	1.614
16	7.7726	407	1.7156	10,345	1.088
17	7.8505	2911	1.9540	17,319	1.822
18	7.8360	1769	1.7872	12,494	1.315
19	8.0208	381	2.1538	30,953	3.257
20	8.2532	1329	1.8208	20,148	2.120
21	8.4521	398	1.9216	29,684	3.123
Total	7.5944	27845	1.7692	9,504	1.000

$(d_1) = \text{Exp}[(a_1)+0.5\text{x}(c_1)^2]$

$(e_1) = (d_1)/[\text{Total Entry of } (d_1)]$

Exhibit 5
Required Number of Claims for Full Credibility by Tier

	(a2)	(b2)
Tier	Number of Claims for full Credibility	Partial Credibility
1	4,368	0.269
2	4,914	0.238
3	4,755	0.435
4	4,018	0.287
5	3,241	0.726
6	4,573	0.668
7	3,251	0.568
8	2,917	0.662
9	2,932	0.905
10	3,247	0.552
11	3,083	1.000
12	3,267	0.707
13	3,281	0.397
14	3,026	0.883
15	4,202	0.343
16	3,230	0.355
17	4,108	0.842
18	3,449	0.716
19	4,781	0.282
20	3,227	0.642
21	3,427	0.341

$(a_2)=[2.575x(c_1)/(.01)x(a_1)]^2$
$(b_2)=[(b_1)/(a_2)]^{0.5}$

Exhibit 6
Parent Node Partial Credibility

Tier	Parent Node Normal Mean	Std. Deviation	Number of Claims	Parent Node Mean Severity	(a₃) Unadjusted Relativity	(b3) Number of claims for Full Credibility	Partial Credibility
1	7.0092	1.8656	1,215	6,307	0.664	4,697	0.509
2	7.3457	1.7971	797	7,789	0.819	3,969	0.448
3	7.0092	1.8656	1,215	6,307	0.664	4,697	0.509
4	7.5858	1.6597	8,648	7,810	0.822	3,174	1.000
5	7.3526	1.5885	4,109	5,510	0.580	3,095	1.000
6	7.2632	1.9055	2,533	8,766	0.922	4,564	0.745
7	7.4121	1.5962	2,329	5,920	0.623	3,075	0.870
8	7.4121	1.5962	2,329	5,920	0.623	3,075	0.870
9	7.3526	1.5885	4,109	5,510	0.580	3,095	1.000
10	7.7444	1.8023	6,259	11,714	1.233	3,591	1.000
11	7.5858	1.6597	8,648	7,810	0.822	3,174	1.000
12	7.5001	1.6398	3,961	6,937	0.730	3,170	1.000
13	7.3457	1.7971	797	7,789	0.820	3,969	0.448
14	7.7477	1.8023	6,259	11,753	1.237	3,588	1.000
15	7.2632	1.9055	2,533	8,766	0.922	4,564	0.745
16	8.1405	1.8078	1,736	17,581	1.850	3,270	0.729
17	7.7444	1.8023	6,259	11,714	1.233	3,591	1.000
18	7.9598	1.8798	2,548	16,758	1.763	3,698	0.830
19	8.2409	2.0486	779	30,924	3.254	4,098	0.436
20	8.1405	1.8078	1,736	17,581	1.850	3,270	0.729
21	8.2409	2.0486	779	30,924	3.254	4,098	0.436

(a_3) and (b_3) are calculated as in Exhibit 4

Exhibit 7
Parent of Parent Node Relativity

(a4)

Tier	Parent of Parent Node Mean	Std. Deviation	Number of Claims	Mean Severity	Relativity*
1	7.2742	1.6619	5,324	5,740	0.604
2	7.6994	1.8060	7,056	11,274	1.186
3	7.2742	1.6619	5,324	5,740	0.604
4					
5					
6	7.5127	1.7237	11,181	8,089	0.851
7	7.5001	1.6398	3,961	6,937	0.730
8	7.5001	1.6398	3,961	6,937	0.730
9					
10					
11					
12					
13	7.6994	1.8060	7,056	11,274	1.186
14	7.6994	1.8060	7,056	11,274	1.186
15	7.5127	1.7237	11,181	8,089	0.851
16	7.5944	1.7692	27,845	9,504	1.000
17	7.5127	1.7237	11,181	8,089	0.851
18	7.5944	1.7692	27,845	9,504	1.000
19	7.9598	1.8798	2,548	16,758	1.763
20	7.5944	1.7692	27,845	9,504	1.000
21	7.9598	1.8798	2,548	16,758	1.763

*) Same as Exhibit 3 and 4

Exhibit 8
Observed Claims and Relativities for
Accident Year 2001

Tier	Observed Mean Severity	Number of Claims	Observed Relativity
1	7,053	126	0.575
2	10,213	69	0.833
3	12,179	339	0.994
4	10,052	75	0.820
5	7,380	344	0.602
6	9,968	344	0.813
7	6,250	216	0.510
8	9,689	229	0.791
9	8,788	775	0.717
10	9,287	300	0.758
11	9,578	1443	0.781
12	8,297	456	0.677
13	8,033	157	0.655
14	13,679	691	1.116
15	14,913	489	1.217
16	13,574	131	1.107
17	13,314	1072	1.086
18	10,661	810	0.870
19	12,518	234	1.021
20	33,395	579	2.725
21	16,964	132	1.384
All	12,253	9011	1.000

Exhibit 9
Comparison of 2001 Observed and 1997-2000 Credibility Adjusted Relativities

Accident Year 2001 Observed Relativities	(a_5) Credibility Adjusted Relativities
0.575	0.569
0.833	0.904
0.994	0.700
0.820	0.760
0.602	0.492
0.813	0.820
0.510	0.602
0.791	0.652
0.717	0.658
0.758	0.975
0.781	0.920
0.677	0.853
0.655	0.982
1.116	0.951
1.217	1.148
1.107	1.431
1.086	1.729
0.870	1.405
1.021	2.651
2.725	1.941
1.384	2.655

$(a_5) = (e_1)(b_2)+[1-(b_2)][(a_3)(b_3) +\{1-(b_3)\}(a_4)]$

Quantifying the Impact of Non-Modeled Catastrophes on Homeowners Experience

Israel Krakowski, FCAS, MAAA

Quantifying the Impact of Non-modeled Catastrophes on Homeowners Experience

Israel Krakowski

Abstract

Much has been done in recent years to quantify the impact of hurricanes and earthquakes on Homeowners loss experience, primarily through the construction of simulation models. Non-modeled catastrophes, primarily Wind, have retained the standard catastrophe ratemaking methodology. This paper examines various different ways of improving that methodology via the incorporation of other states' data.

ACKNOWLEDGEMENT

Many thanks to Joel Atkins, Erik Bouvin, Dave Chernick, Fred Cripe, Kevin Dickson, Sara Drexler, Christopher Monsour, and Fei Zeng, and readers of the paper for valuable input and suggestions.

INTRODUCTION

The 1990's saw considerable attention paid by the actuarial community to natural catastrophes, that is: hurricanes and earthquakes. The impetus for this focus was the gargantuan losses incurred by these perils, most dramatically by Andrew and Northridge. The most significant pricing related development to evolve from this attention has been the creation of catastrophe models by various entities (insurance companies, consulting firms, and companies whose primary product are these models). To a large extent these catastrophe models are black box simulations, not typical actuarial models. They have, however, by now gained a wide measure of acceptance by all segments of the industry. And their use has been addressed by Actuarial Standard of Practice #38.

This paper will not deal with the aforementioned models, about which a considerable amount has been written.[1] It will rather focus on other natural catastrophes, of which the most significant peril, from an insured loss perspective, is (non-hurricane) wind and hail (which are lumped together and referred to as "wind" in the balance of the paper); though fire, water, and explosion, can cause substantial damage as well. To distinguish the catastrophe losses generated by these perils from hurricane and earthquake losses, they shall be referred to as "non-modeled catastrophes". The criteria for what constitutes a catastrophe vary by company. Typically there will be a dollar threshold (not increased nearly often enough) and

[1] See for instance Burger, et. al. [1], Chernick [2], Walters and Morin [3]

some other criteria, such as more than one insured sustaining a loss. On an industrywide basis Property Claims Service (PCS) assigns catastrophe numbers to natural events based on its estimate of total damage.

In the last few years two phenomena have begun to focus attention on non-modeled catastrophes. The first is that by mere virtue of the fact that there exist models that quantify hurricane and earthquake catastrophe risk, the "remainder" has taken on an identity of its own, and become the subject of distinct analyses. From an operations perspective a similar process has occurred. Companies have mitigated their hurricane and earthquake exposures via reinsurance—private and governmental, higher mandatory deductibles, limiting writings in designated areas, etc.. Homeowners insurance, which has not in recent memory been significantly profitable, has had particularly poor results recently. For 1991 through 2000 the industry has run a 110.9 operating ratio (combined ratio after dividends and investment income); for only one of these years has the ratio been less than 100.[2] Non-modeled catastrophes, now separated from hurricanes and earthquakes, have drawn attention as a distinct and significant contributor to these poor losses.

The second phenomenon is that not only have the non-modeled losses begun to stand out in relief as a distinct peril (to use the term broadly) worthy of study, but the actual quantity of losses derived from these events have been rising, when measured over the long term (see for example Figure 1 and Exhibit 2). This is so whether one measures losses in absolute dollars, per dollar of premium, per house year or per amount of insurance year (these latter being the natural exposure bases).

2. CURRENT METHODOLOGY

Prior to the creation of the hurricane and earthquake models, the most widely used method of quantifying catastrophe risk was the ISO excess wind methodology. While many of its faults were clearly understood, it nevertheless remained the best that could be done. As simulation models have gained popularity for the hurricane peril, the ISO methodology, or one of its many variants, has been the primary methodology for quantifying what's left over. (Note that sometimes this procedure is applied to non-hurricane catastrophe wind losses only, and sometimes to all non-hurricane, non-earthquake catastrophes.) The basic concept is to take a long-term ratio of catastrophe losses to non-catastrophe losses. Thus the ISO excess wind methodology takes as excess losses all wind losses in excess of the long term historical median ratio of wind to non-wind losses, but only for years in which that wind/non-wind ratio is in excess of 1.5 times the historical median ratio. These excess losses are then spread to all years. Again, the basic concept, for this and the many variants, is to take a long-term ratio of catastrophe losses—however defined-- to non-catastrophe losses, and spread the losses across years (or equivalently load in the average).

There are many problems with these procedures. Some of them are.[3]

---

[2] And 2001 has likewise been a poor year.

[3] Some of the points are mentioned already in Hays and Farris[4], and McCarthy[5] and Chernick[2].

- The impact of distributional changes over time: in policy forms, geography, etc.. Changes over time in policy form, such as actual cash value vs. replacement cost, and coverages can affect the extent to which a natural event will yield covered losses. Even more significantly changes in exposure concentration over time will affect loss potential.

- The impact of changes in the definition and coding of catastrophes. PCS has gone from being in excess of 1 to 5 to 25 million as its definition of catastrophe. It is safe to conjecture that all major companies periodically change their definitions as well.

- Even what is considered long term (e.g. 30 years) for the calculation of the catastrophe factors, is not long enough, for a given state.

- Adjustments that are typically made to numbers in the rate analysis process, such as trend and loss development, should probably be done separately for the catastrophe and non-catastrophe components. This is not so much a problem as a suggested refinement. The impact of severity trends and development on "excess" losses (in for instance the excess wind procedure) call for individualized attention. Similarly, frequency of catastrophe events might possibly not track with the frequency of non-catastrophe events.

- Changes in premium adequacy over time, if the statistic one is using is loss ratio, should be adjusted for. A very poor loss ratio could be a function of very poor rates and not unusually large losses. Capping should not be a function of premium adequacy.

- The non-catastrophe losses that form the base for the excess ratio comprehend multiple perils. Trends in some of these perils, such as liability and crime, may have no correlation to catastrophes, and cause distortion in excess ratios. Thus if liability losses become a much greater proportion of all coverages, then the ratio of catastrophe to non-catastrophe losses will artificially appear to go down, all other things being equal.

- For those procedures that apply to excess wind only, there needs to be some adjustment for non-wind catastrophe such as fire, explosion, and water.

We shall stop at this summary description of the current methodology. Those wishing further details can see Chernick [2], Hays and Farris [4], and Homan [6]. Bradshaw and Homan [7] suggest a variation which incorporates the output of a simulation model. McCarthy [5] recommends a procedure, which develops the catastrophe load based on non-hurricane wind loss frequency. Dean, Hafling, Wegner and Wilson [8], suggest a variation wherein capping is done below as well as above. (This list is not necessarily comprehensive).

3. GENERAL CONSIDERATIONS

The primary focus of this paper will be on methods for analyzing a given state's non-modeled catastrophe experience by incorporating other states' data. The lack of such external data is a deficiency in most currently used methods. We will not be discussing how to take the indicated non-modeled catastrophe damage ratio, and incorporate it into an overall (loss ratio or pure premium) rate indication methodology. (Damage ratio is defined as losses over AIY—amount of insurance years. This is the primary statistic we will be dealing with.) While there are details to be worked out, the overall procedure should be fairly straightforward. Exhibit 1 gives one such way.

Before discussing specific methods some general comments are in order. First, all the methods to be presented had, in their creation, various externally imposed requirements. 1) That the separate indicated state damage ratios sum to a reasonable countrywide damage ratio. 2) If credibility is used, states with very stable damage ratios over time, even if small, should have relatively high credibilities. (In many of these states we would have good reason to believe that the state' process variance is lower than average, to put it in these terms, based on external—e.g. meteorological—considerations.) 3) States which had (what appears to be) an extreme (once in a hundred year or greater, say) event, should not be unduly penalized for said event.

Most of the methods examined were constrained by the nature of the data available. An analysis with better or more restricted data can adjust the methods accordingly. The primary data used consisted of a summary, by state and calendar year, of various Allstate companies catastrophe losses and amount of insurance years from 1971 to 2000.[4] These losses are for Homeowners, Renters and Condo. In addition, for the years 1988 to 2000, data was available in some further detail. For these latter years thought was given to segregating other than wind (non-modeled) catastrophes, and perhaps having a separate load for these in selected states.[5]

Some of the methods to be presented below (e.g. the trended method) will likely strike the reader as having problems which make them less optimal than the other methods, since they yield results that are significantly unintuitive for specific states. For the remaining methods though, it is not obvious which one is best Aside from meeting the above externally imposed requirements there are three primary criteria by which a method is judged. The first is accuracy. This is the most important but most difficult to apply, since by their very nature, the existence or absence of catastrophes in a state in years subsequent to the predicted indication, do not necessarily bear directly on the accuracy of the indication. The second criterion is stability. This is easier to measure, and some exhibits will be presented below. The final is "sellability". Especially given the lack of a good test for accuracy, characteristics deemed unacceptable to either regulators or parts of one's internal organization, will count

[4] Because these were calendar year one does get, in a few instances, odd results such as negative loss numbers.
[5] This option was rejected since it did not seem to improve the results. Water catastrophe losses correlate with wind catastrophe losses, and other catastrophe perils—with exceptions in a few states--are usually small.

heavily against a method. One instance of such a characteristic is having losses spread from one state to another.

Since it is unclear which of the methods to be presented is optimum, this paper should be viewed as providing ideas on how to improve the non-modeled catastrophe component of the rate indication process. For this reason, and to keep the number of permutations down, not every modification or refinement (e.g., of credibility) is presented for each method.

4. SIMULATION MODELS

Since our goal is to improve on the current methodology, we briefly note a potential methodology, not discussed in detail, which--once fully developed--may be the most accurate. That method is to construct simulation models for the non-hurricane wind peril analogous to those developed for hurricanes. Such models are in fact actively being worked on by the various modeling firms, and some of the first (Beta) versions are being released.

The most glaring problem of other methods, including those to be presented below, is the omission from the analyses of change over time in exposure concentration, in areas that are likely to have windstorms or other natural disasters. No doubt increases in non-modeled catastrophe losses are to a significant extent driven by increases in these concentrations. To quantify the impact of increases in concentration we need the likelihood of natural events for each geographic area, where what constitutes a geographic area varies by the type of natural event. We need to understand how losses caused by different types of events are differentially impacted by the interaction of changes in exposure concentration and topography. In short, we need a simulation model that, in its very broad outline, is similar to hurricane models.

Why not then use the soon to be available commercial models? First one should never use a Beta version of anything. Secondly, hurricane models required quite a few iterations until they reached their present state. The non-hurricane wind models will be, it appears, even harder to get right than hurricane models because of the different sorts of events and the high level of resolution needed[6]. The combinations of types of event and topography are numerous, and the amount of historical data needed for accurate simulations great indeed. Nevertheless one can be (cautiously) optimistic that eventually we will have a workable model. In the meantime the methods presented below may be of some use.

[6] Even restricting ourselves to windstorms there are hailstorms, tornadoes, and straight line windstorms; each of these have a wide range of intensities and interact differently to the geographical environment.

5. TRENDED METHOD

The first method to be discussed is the "trended method". It might be conjectured that data going back as far as 1971 would be sufficient to calculate each state's own indicated damage ratio. One major problem (not discussed in the literature) is the calculation of trend factors for catastrophe data at the state level. Catastrophe losses are dramatically more volatile than non-catastrophe losses, and fitting trends to an individual state's catastrophe data does not give reliable results. Nor would applying a countrywide trend to each state's be appropriate, since the true trends (which are indiscernible with the data we have) will clearly vary by state. Credibility weighting trend (of which there are a few methods) might have been pursued, but without a good understanding of the drivers underlying these trends, would not likely result in reliable estimates: Methods of credibility weighting trend lines usually assign credibility as some function of the variability of the trend (parameter) estimate. Catastrophe experience at the state level would tend to be very variable, and one would like to be able to distinguish the noise from true trends. Typically the level of exposure in a state would be a factor in estimation variability. If, however, because of concentration impacts additional exposure does not yield less variability (and more credibility) to damage ratios—as would be typically assumed, then one should know what these increases in concentration are, and what their impact is, before assigning a credibility weight to a trend indication.[7]

One solution is as follows: first calculate a countrywide (linear) trend in damage ratios, weighted by amount of insurance years (AIY). Exhibit 2 gives the calculation of the countrywide trend. Note that the numbers are unadjusted (e.g., for development, change in threshold, etc.). The trend is projected out to the average loss date under consideration.

The ratio of this countrywide trended damage ratio to the countrywide arithmetic mean damage ratio is calculated. This ratio, 1.701 from line 10 of Exhibit 2, is then applied to the arithmetic mean damage ratio of each state, to produce the indicated non-modeled catastrophe damage ratio. This method applies a "trend" factor to state data, while mitigating many of the problems with a straightforward trend calculation. Thus, since it is only one of 30 years, any outlier in a given state will not significantly distort the state indication, as the direct application of trend to state data frequently does.

There do remain significant problems with the method. States which appear to have no trend, or much higher trend than countrywide, are multiplied by a seemingly inappropriate factor.[8] Further, distributional shifts alone could and do distort the indications. Thus there is about a five point difference between the countrywide trended damage ratio, and the sum of the state damage ratios derived by using the "trended" methodology, when weighted by 2000 AIY. A large part of this discrepancy is due to a distributional shift, caused by much higher growth than average in the most recent years in a state having particularly poor catastrophe experience. In short, while initially promising, there still remain problems with this method.

---

[7] It must be admitted though, that further investigation along these lines—even with the data at hand—might prove fruitful.

[8] Though one typically tends to hear complaints only when the factor was too high.

6. REGIONWIDE METHODS

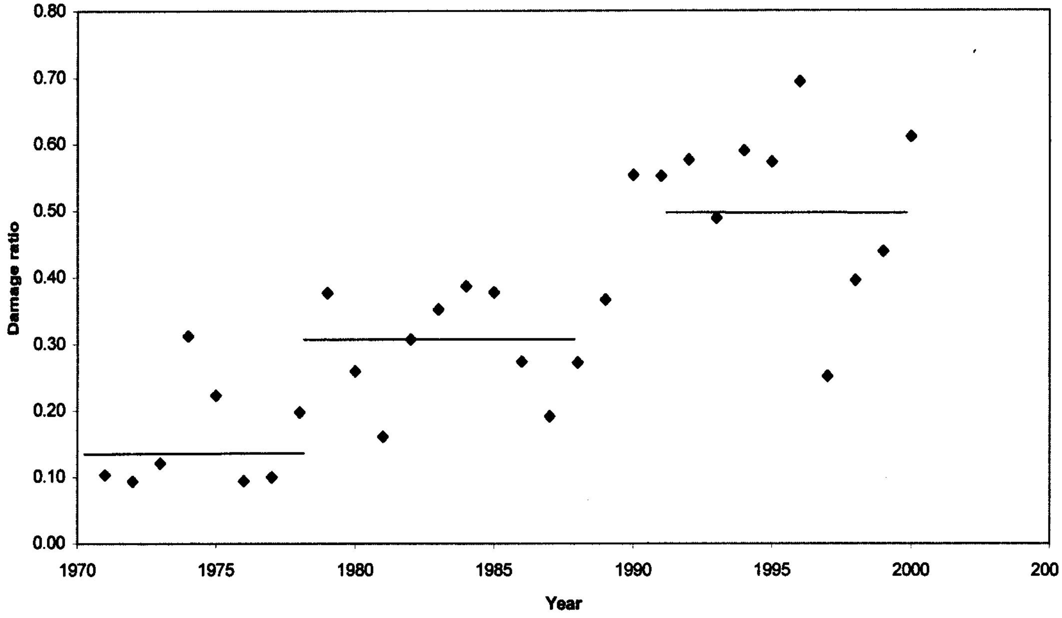
All of the subsequent methods—barring the last which skips the second step--begin with the same first two steps. Step one calculates a countrywide indicated damage ratio. Step two divides the country into "regions", i.e., collections of states, and then calculates an indicated damage ratio for each such region. These region-wide indications are rebalanced, based on the *most recent* year's AIY (which should take care of the problem of distributional shifts), to the overall countrywide indicated damage ratio. These rebalanced indicated regional damage ratios are the beginning points for all subsequent calculations. It should be noted that the actual method used here of calculating the indicated regional and countrywide damage ratios (as discussed immediately below), are not an essential component of the methodologies. One could, if so inclined, use more elaborate procedures. What is essential is that there be a countrywide indication and regional indications to be balanced back to it.

A. Countrywide Indication

The first step then is to calculate a countrywide indicated damage ratio, to be balanced back to. Figure 1 shows the raw countrywide damage ratios over time. The increases appear to come in steps and from 1990 on there appears to be no trend. As noted at the bottom of the graph the trend for each segment is very close to 0, and the means for each segment (the horizontal bars on the graph) are significantly different. Consequently the arithmetic mean of years 1990 on is used[9]. The mean is not weighted (by AIY), for the accuracy of each estimate (year) is, as far as can be told, independent of the size of that year. From the exhibit it is apparent that the 1990-2000 numbers already incorporate all the "trend" of prior years. Consequently the state indications, which ultimately balance back to the countrywide indication, are "trended" without the explicit application of trend on an individual state basis. Again, this is particularly significant given that data was not available on the previously mentioned drivers of these trends (on a countrywide or individual state basis) such as exposure concentration, or on the nature of data related distortions.

[9] In practice a (somewhat arbitrary) load of 4% was tacked on to the countrywide damage ratio. This was done to recognize the fact that deviation is likely to be on the upside: I.e., it is much less likely that countrywide damage ratios will systematically start going down.

FIGURE 1
CW Cat damage ratios
including 2000



Years	Linear trend
1971-1978	0.006
1979-1989	0.000
1990-1999	-0.019
1990-2000	-0.010

It will not have escaped the observant reader's attention that the endpoints on Figure 1 were selected so as to make the trends look flat and have discrete jumps, and could have been made to look dramatically otherwise had the endpoints been chosen differently. This is correct. The crucial point is that given countrywide data the best guess as to the following year's damage ratio is the average of the years since 1990, or so we would claim.[10]

There is one other minor point to be addressed. Given the above should one use 1990 forward for the countrywide indication, or take a rolling 10 years (1991-2000 currently) average. On the one hand we want to use all relevant points that have information. On the other hand, 10 years is a more standard choice (e.g. in a filing). And if there is in fact a trend in countrywide numbers going forward, a rolling average will pick it up better. There does not seem to be a clear-cut answer to this question.

B. State Groupings

The second step is to group the states into regions based on their catastrophe experience (historical damage ratios). Since we only have aggregated data by state and year, there is a limit to the analysis that can be done. Basically, contiguous states with similar damage ratios are grouped together. A few large states are standalone groupings. Where a state looks as if it might reasonably go into more than one region, historical correlations of damage ratios are used to decide the issue.[11]

Some more sophisticated method of grouping than above might conceivably produce better results, but with the data available it is doubtful. Grouping states will be problematic no matter how it is done, for natural catastrophes do not obey arbitrarily drawn political boundaries. Frequently, an appropriate meteorological territory will cross state lines, and there may be natural breaks within a state. (E.g., Colorado which is divided into mountain ranges and plains with very different weather patterns). Indeed one could argue for needing different regions for different catastrophe perils. Since wind is such a dominant component of the losses, in practice this is not an issue.

---

[10] It is not that we have totally discounted the possibility of damage ratios trending up further: whether continuously or via a "jump"(see the previous footnote). But we should need evidence. Certainly over time one would think that a constant threshold and monetary inflation would, cetirus paribus, cause damage ratios to rise. But thresholds can be changed, and monetary inflation affects the denominator (AIYs) as well as the numerator (catastrophe losses)--though the effects on losses have historically been greater than on AIYs. More importantly, there are presumably more significant forces affecting the overall catastrophe damage ratios: Frequency and severity of natural events, changes in concentration (have our writings in concentrated areas remained relatively constant, or even gone down, over the last 10, 11 years, where it had been increasing previously?), and so on. The impact of these will clearly swamp the impact of, e.g., pure monetary inflation (especially in the current monetary environment). So given the data we have, the most rational assumption would be to take a recent average, until the evidence argues otherwise.

[11] This procedure is due to Kevin Dickson (personal communication). I do not give additional details, since the actual grouping process is somewhat tangential to the main concerns.

Even when working within our constraint of using state and not topographic groupings what looks like an obvious grouping based on the data, may not be optimal.[12] In order to circumvent these problems and create more refined territories one would need to have both meteorological data and geographic exposure distributions: i.e., would be back to the simulation model data requirements.

Once the "regions" are constructed, indicated damage ratios are then calculated. Once again a straight arithmetic mean of the latest 10 or 11 years was chosen as the indicated damage ratio. These average damage ratios are rebalanced to the countrywide indication based on the most recent year's AIY distribution. Exhibit 3 provides state groupings and the rebalanced damage ratios for use in one of the methods below. Other equally reasonable groupings could have been chosen.

7. DUAL CAPPING METHODOLOGY

The first method considered is a modification of one proposed by Dean et al [8]. Their method is a variation of the excess wind methodology, but with loss ratios censored below as well as above. In a given state non-modeled catastrophe loss ratios--catastrophe losses divided by total earned premium--are calculated for each year (of seventeen). These are ranked from low to high. A low and high loss ratio is chosen (corresponding to percentiles previously decided on). Any loss ratios below or above these two designated loss ratios are "capped" at the low and high loss ratios respectively. The net of losses excluded from above minus losses excluded from below are "excess" losses. These excess losses are summed and divided by the total earned premium for all seventeen years, to yield a load factor. In the overall rate indication calculation, wind losses are again capped above and below for each year at the chosen loss ratios, and the previously calculated load is factored in; these adjusted wind loss ratios are then added back to the loss ratios for all other perils.

While two sided censoring is certainly an improvement over the traditional method, there still remain problems. First the use of earned premium could distort the procedure if there are substantial differences in premium adequacy over the years. Changing to damage ratios, as we do below, addresses this issue. The next problem, and it is a large one, is that the losses ratios are either trended or they are not. (Their paper does not say, a reasonable guess is that they are.) At the state level, as discussed above, trending catastrophe losses is a problematic exercise: changes in catastrophe thresholds and definitions; changes in storm frequency and severity; changes in concentration of exposure, all need to be taken into account. If the numbers are not trended, then there is an inconsistency with the non-catastrophe indications, wherein the losses are standardly trended. Further the capping procedure itself will be

---

[12] Two states may be close geographically, have approximately the same level of damage ratio, with a relatively consistent pattern over a short period of time, and still not in reality belong in the same group. One state may have, for example, an increase in its catastrophe damage ratio due primarily to a large increase in concentration of exposures, while the frequency of catastrophes remains constant; while a second state could have no or a negative increase in concentration, but have an increase (random or not) in the frequency of natural events causing catastrophes. Though their numbers make them look similar, they might be more appropriately slotted in different groupings. But without more detailed data there is no way to tell.

distorted: The most recent and oldest years will more likely be capped—from above and below, respectively—than the other years. Finally, as the authors recognize, seventeen years, while a considerable amount for most purposes, is not adequate for a catastrophe load in any one given state.

A synthesis of two-sided capping method with some of the components from above, neutralizes some of these problems. Rather than going back in time as far as possible, one gets more data points by using all the damage ratios from a region. For each state in such a region one takes the latest ten years damage ratios (so there will be 10 x the number of states, points). One then ranks these ratios irrespective of state. The two-sided capping procedure is then applied to the damage ratios within the region, so ranked. The method then proceeds as in the original paper with a load calculated and incorporated into the rate indication process. Exhibit 4 provides an example for one region of the calculation of an excess load factor which would be applied to the wind peril in a Homeowners indication calculation.

This modification has various benefits. First, since it uses damage ratios and not loss ratios, premium adequacy is not an issue. Secondly, it typically contains more points: in Exhibit 4 there are ninety versus seventeen in the original paper.[13] Finally, since we are assuming that trend is already incorporated into the most recent ten years, there is no trending problem.

There are some remaining problems. As discussed, there is no perfect grouping of states. Consequently, some states may seem out of place, having lower or higher damage ratios on average than the rest of the states in its region. (Percentiles should be chosen so that the capping procedure does not penalize or reward a particular state because of this phenomenon.) Further the procedure might have a difficult time gaining acceptance because it appears too much to just be spreading losses from one state to another.

8. CREDIBILITY WEIGHTING STATE INDICATIONS

The remaining four methods all credibility weight individual state indications: the first two use actual damage ratios, the latter two relativities. The first three weight against the previously referenced "regional" (rebalanced) indications, the last directly against a countrywide indication.

The use of credibility in these methods proceeded in an extremely pragmatic and somewhat ad-hoc fashion (some would say alarmingly so). Given the external constraints listed above the rational for a given formula is often to a large extent teleological. Further, it is clear (see below) that the standard Buhlman-Straub formulation is not appropriate in the present context. Mahler [9] in a recent comprehensive paper, expounds--among many other things—on how one might adjust for different behavior for different size risks; for parameter shifts over time, for parameter uncertainty and for use of external state data. All these adjustments are potentially applicable, with modifications, to our case. However because of the summary nature of the data, credibilities so adjusted often can not be derived; and even where

[13] And our working assumption is that the region consists of states with roughly similar catastrophe exposures; to the extent that this is true the points can be considered drawn from the same population.

quantities could be calculated, they would not be trustworthy since, once again, the primary drivers of the variance and covariance of the damage ratios, e.g. increases in concentration of exposure,[14] are not known.

For our purposes then our pragmatic approach, which is somewhat forced upon us, does not lose us much, especially since it has been shown[15] that within a wide range of values the use of any credibility weighting will be superior to none (even if it is not the best). Our adjustments were to the calculation of the process variance, and to the variance of the hypothetical means[16]. In the presentation of these methods various credibility adjustments have been made not because a particular adjustment is necessarily tied to the method in which it is presented, but simply as a way of presenting examples.

A. Credibility Weighting Damage Ratios with the Region As Complement

The first credibility method uses the latest 11 years of catastrophe damage ratios for each state, (circumventing trending problems). The unweighted mean damage ratios for each state, is credibility weighted against the average (unweighted) damage ratio for its region. The standard calculation of process variance, which weights by exposures (in this case AIY) would be inappropriate here, as it would also be in the next three methods. An assumption underlying the weighting is that the process variance (in this case of the damage ratio) is inversely proportional to the exposures; and that in turn assumes that exposures are (more or less) independent. With the catastrophe peril this is very often not the case: additional exposures, in an already concentrated area, are very much correlated. And one should not expect a proportional decrease in variance with additional exposures, in such a case.[17] Hence the credibility formula used here is $z=y/(y+k)$ where y is the number of years, and k is the ratio of the expected process variance to the variance of the hypothetical means, without consideration of exposure level. This is a case where, given the data we have, the decision not to use exposures as weights seems theoretically as well as pragmatically correct.[18]

Because some regions have what seems clearly to be different process variances by state, the first credibility method presented uses a weighting of each state's own calculated process variance with the average process variance (See exhibit 5); This might be thought of as a very crude attempt to capture some of the additional structure in the data.[19]

[14] Even if an effort were made to gather data on the change in exposure concentration, there is not currently a clear conception of what the appropriate level of detail is: is it relevant how concentrated one's become in a state, a county, a zip, or a census track? The answer no doubt varies with topography.
[15] See Loss Models [10] pp 451-454
[16] There does not seem much point to making other refinements, such as using the credibility weighted overall mean as the complement of credibility.
[17] There is some, admittedly weak, evidence for this. See table 2 below.
[18] There are various empirical tests one could attempt to estimate the relationship between size and variance (See Mahler[9]). Because of the aggregated nature of the data, and more importantly because, as mentioned, the process variance is among other things a function of size and concentration—which we do not have—one would have to be very suspicious of any quantitative inferences about how the process variance should vary with exposure. Hence assuming no relationship seemed safest.
[19] While in general using the average expected value of the process variance is mathematically less variable than the using each state's own estimated process variance, in the present case our adjustment will hopefully yield a

Three different estimates of the variance of the hypothetical means are calculated. 1) The variance of the state mean damage ratios; 2) The difference between the total variance and the average process variance; 3) The covariance of the sum of the first five years damage ratios and the second five years, across the states in the region. The covariance estimator has in other contexts proven to be quite good. In the present context it was not, yielding wide swings by region, including negative results. The other two calculations also yielded slightly negative numbers in some cases. The standard interpretation is that this entails that each state should get 0 credibility. This is a hard conclusion to accept, particularly given the somewhat arbitrary way some of the regions were put together. Consequently the bias adjustment from the estimate of the variance of the means (i.e. subtracting the process variance/number of years) was eliminated from the first estimate, and a floor of zero was put on the second; the average of the two was then taken, as can be seen in exhibit 5.[20]

Once the credibility weighted state damage ratios--with the complement being the region-wide (unweighted) average damage ratio—are calculated, they are adjusted to the chosen regional factor. First the damage ratio for each state is multiplied by its most recent year's AIY. These are summed and compared to the latest year's losses implied by the previously calculated regional factor (The regional latest year AIY x the indicated regional damage ratio). The difference between these two is spread back to each state on a flat percentage basis. Exhibit 5 provides details of this procedure for one region.

B. Including Non-Hurricane Wind Data

---

more accurate estimate.(About twenty years ago I asked Glenn Meyers why a particular ISO credibility procedure had settled on the average process variance for the expected process variance, rather than have each state (or class—I don't recall) use it's own calculated variance. His reply was—if memory serves me correctly—that ISO had indeed looked into that alternative, but the results were too variable.)

Our case might be thought similar to Mahler's cases of heterogeneity; his example is a large WC insured with several locations. These locations might share some risk characteristics, and be different on others. (His other example is commercial auto.) He derives formulas that give less credibility to heterogeneous risks by virtue of the variance of the hypothetical means increasing less slowly that as the square of the sizes of the risk (as it would in Buhlman credibility).

Let us take an auto example, where we have divided the populations into various classes based on some subjective criteria, and wherein each insured is assumed to have a Poisson distribution. Each class is certainly still heterogeneous to a certain extent, so if we could estimate the various parameters in Mahler's procedure we could apply that procedure. But it is difficult to estimate the parameters. Another possibility is to focus in on the classes themselves; think of them as indivisible entities, and assume, say, that they have Negative Binomial distributions—as is often done. In that case differences in heterogeneity will manifest themselves in different Negative binomial parameters for the classes and hence in different process variances. One way to accommodate these differences would be to take into account a class's own (sample) process variance as well as the average. This is what we have done above—where, by the way, it would be well nigh impossible to come up with an estimate of heterogeneity.

[20] This "adjustment" is indeed arbitrary. However, even though estimating the variance of the hypothetical means with the correction included, is an unbiased estimate, the consequent estimation of the credibility factor Z remains biased.(See Ventner[11] pp 440-446). Since it is reasonable that each state does have some credibility, my adjustments do not seem unreasonable.

The next regional method, is really not so much a change in method as a change in data. Rather than non-modeled catastrophes only, losses are taken for all non-hurricane wind plus all non-modeled catastrophes.[21] Combining the two might seem to run counter to the standard rational for a separate catastrophe analysis that catastrophe losses make indications too variable: that by analyzing the catastrophe losses separately we cap the underlying wind losses and hence provide more stable indications for that segment; while the catastrophe losses can then be grouped (across many years, many states, etc.). We loose refinement, but we gain stability.

While the above is true, there are various practical considerations arguing for combining non-hurricane wind with non-modeled catastrophe losses. First there are the standard coding problems that will misclassify catastrophe losses. Further, catastrophe thresholds and definitions typically vary over time and are not necessarily consistently applied across all states. Combining eliminates these potential, and frequently occurring, distortions.

Territorial indications, also become considerably more stable and reasonable when wind and catastrophe are combined. Finally because of the additional ballast provided by the wind numbers, the standard deviations and coefficients of variation for the indications are substantially reduced for the catastrophe portion of the indications(though not necessarily for the indication process in aggregate), as is presented in Table 1.

Table 1

	State SD	State CV	Year SD	Year CV
Cat	0.61	1.21	0.12	0.25
Cat+wind	0.78	0.82	0.12	0.16

For this second method, the process variance is calculated slightly differently. It reflects the consideration that even a state which has been very stable could have a huge catastrophe; that there is an element of randomness in one particular state within a region having had the once in a hundred year event rather than the others (which is why they were grouped into the same region).[22] Therefore the expected process variance was calculated as a weighting of the average process variance with the maximum process variance of any state in the region.

Exhibits 6 presents the results of using the procedure on the combined non-hurricane wind and non-modeled catastrophe losses.

[21] Again, we have an apples and oranges situation somewhat. The catastrophes contain perils other than wind. The justification for this is that the other peril catastrophes are too small to be analyzed on a standalone basis, and do not seem to distort the indications here.
[22] Many regions had at least one state that had a huge catastrophe (10 times the median for that state as an approximation)

C. Credibility Weighting Relativities with the Region As Complement

The final two methods use non-modeled catastrophe [23]damage ratio relativities rather than the damage ratios themselves. The motivation for using relativities is that though for a given state the damage ratios typically vary significantly over the long term, we might expect the relativities to be more stable: even if there is trend in the damage ratios we might hope for none in the relativities, thus allowing for a longer time period for each state's data; and indeed there is in general no significant trend as can be seen from the line labeled "linear trend," on exhibit 7,[24] which contains other descriptive statistics as well for examining the reasonableness of using relativities[25].

While data is available from 1971 on, the early years are too sparse and variable even when using relativities (some years have 0 losses), as can be seen in table 2[26]. Therefore only 1981 and subsequent is used.

Table 2

Average state variance of relativities

Years	Average Variance
1971-1980	11.07
1981-1990	1.87
1991-2000	4.24

The procedure proceeds, as can be seen on Exhibit 8, along the same lines as the first regional method, but uses relativities as the statistic. Once a credibility weighted relativity is calculated the estimated damage ratio is calculated by multiplying the estimated relativity factor by the indicated region-wide damage factor. They are then rebalanced as before.[27]

One additional detail which needs to be addressed when using relativities is the impact of distributional shifts in exposure between states. These can, and on occasion do, have significant impacts. This problem is addressed by adjusting all relativities to the 2000 AIY

[23] Because of the greater number of years used, non-hurricane wind could not be included.

[24] These relativities are to adjusted countrywide damage ratios; relativities to region-wide damage ratios should be even more stable.

[25] The R-squared that goes along with the trend is given. Standard deviation and coefficients of variations of the relativities with which to measure the variability of relativities by state are given. The correlation of each states damage ratio (not relativity) to the countrywide (adjusted) damage ratios, is also given. All rows are labeled.

[26] One would conjecture that the variance has gone up in the most recent years due to increases in concentration. But the data is not available to test this hypothesis.

[27] In the calculation of credibility the process variance was again calculated as a weighting of the maximum process variance for any state within the region with the average process variance of all states.

level. Thus let w_i be the 2000 AIY for state i; let d_{ij} be the damage ratio for state i in year j; then the adjusted region-wide damage ratio for year j is $A_j = \Sigma_i (w_i \times d_{ij})/\Sigma_i w_i$. Relativities are then taken to this adjusted region-wide damage ratio; that is the relativity for state i and year j is d_{ij} / A_j. A similar adjustment is made to the relativities in the next method. For a simple numerical example assume there are 3 states in the region (or countrywide for the next case).

Year	State1		State 2		State 3		Regionwide	Adjusted
	AIY	DR	AIY	DR	AIY	DR		
1981	10	.3	10	.6	10	.9	.6	.7
|								
2000	20	.05	40	.1	60	.20	.142	.142

Here DR represents the damage ratio relativity. The 1981 regionwide adjusted relativity would be (20*.3+40*.6+60*.9)/120 = .7. The relativity for State 1 in 1981 would be .3/.7 .

The three previously referenced estimates of the variance of the hypothetical means come out to be quite close (for this and the next method) and the first estimate of the variance was used. Exhibit 8 gives the results of these calculations for one region.

D. Credibility Weighting Relativities with Countrywide As Complement

The final method calculates relativities in a year as each state's damage ratio divided by the countrywide adjusted damage ratio.[28] For the countrywide damage ratio (by which the final calculated state relativities are multiplied to obtain the final state indicated damage ratios) it was also necessary to use the mean of the adjusted countrywide averages, rather than the mean of the raw countrywide averages. This had the effect of changing the countrywide damage ratio to approximately .56 from approximately .52. Texas had tripled its AIY between 1990 and 2000 while all other states had on average approximately only doubled. This state had huge catastrophes in two of the last 10 years, and had significant impact on the countrywide average. Without readjusting to the 2000 AIY distribution the relativities as well as the countrywide average would have been distorted.

Using countrywide relativities has one major drawback. The complement is biased. That is, it is clear that many states have an expected relativity, while not known precisely, much different than 1. Indeed, a significant argument for using regional relativities, is that it eliminates (imperfectly) this problem. Exacerbating this problem is the desiderata that a given method should not overly penalize a state for a one in a hundred (or greater) year event.[29] For these extreme states one desires lower credibility than would otherwise be obtained given the wide spread of countrywide relativities.

---

[28] With the same adjustment, but with the adjusted countrywide damage ratio replacing the regionwide damage ratio in the above explanation.

[29] These losses could not simply be eliminated, since another desiderata was that on a countrywide level, they be accommodated.

This problem was resolved by having unusually large relativities capped and the off balance spread back to each state in proportion to the state's relativity standard deviation (after capping), as measured in 2000 expected losses. This method of rebalancing gives states with high average relativities (excluding the impact of once in a hundred year events) more load and those with lower relativities less, counterbalancing the impact of the biased complement. Further the loading back of these losses (a catastrophe load for catastrophe losses, if you will), is a function of the state's own characteristics. States with lower variance get less load and states smaller in absolute size (AIY) get less load, and vice versa. This procedure is intuitively appealing: smaller states, all else being equal, should get a commensurately smaller load, and less variable states, all else being equal, should get smaller loads. This characteristic should also make the capping and spreading more palatable to outside parties.

The capping procedure calls for some comment. Within each state, the standard deviation of the relativities before capping is calculated. If any relativity for that state is greater than the arithmetic mean relativity plus three standard deviations, that relativity is capped (between 1 and 2% of the points were capped). The relativity exceeding the cap is changed not to the mean plus 3 standard deviations, but to the highest actual relativity lower than the cap- almost always the next lower actual relativity. To cap the losses more conventionally would not have accomplished much, since the standard deviation calculation included the extreme event, and the conventionally capped number would have been much higher than desired. Further, on an intuitive basis, this procedure replaces an extreme year with an estimate of a typical (once in 20) really bad year. While the proposed method does have the disadvantage that a state with a damage ratio slightly beneath the cap might easily have a higher indication than if it had come in slightly above the cap, this is not a major problem in practice.

The spreading back of losses is calculated as follow (references are to exhibit 9). The indicated damage ratio relativity for state i is d_i (line 7); expected 2000 losses for the state i (line 8) is the indicated damage ratio relativity times the countrywide chosen damage ratio times its 2000 AIY: d_i x CWD x AIY_i. One standard deviations worth of these losses are (line 9) line 8 x sd_i . The off balance, calculated in a manner similar to previous methods, is spread back in proportion to line 9. The detailed steps for one state are given in Exhibit 9.

10. CONCLUSION

Various methods have been presented which use additional states' data to calculate non-modeled catastrophe loads. Every one of these methods is an improvement over current methodology, and each has its own strengths and weaknesses. The trended and dual capping methods have problems (discussed previously) that the other methods do not. Of the remaining four, using relativities, for either the regional or countrywide method, would seem superior to using damage ratios since it allows for the inclusion of many more years without concern about adjusting the numbers for trend. On the other hand, in our case we can no longer include ground up wind experience in the data; if one does have wind data going back that far, then incorporating it into the relativity methods would be optimal.

This leaves the regional and countrywide relativity methods, with wind data if one has it. Exhibit 11 gives a comparison of the results by state for these two methods as well as the trended method and the "Agg/Agg" method, which is simply a weighted average of all years of a state's (untrended) damage ratios. How do these two remaining methods compare on the criteria delineated at the beginning of the paper?

The first, and most important criteria is accuracy; as mentioned we unfortunately know of no way to measure this, even on a relative basis. While, as can be seen on Exhibit 10, there are some significant differences in estimates, especially for the larger damage ratio states, even had we the results of the next few years because of the nature of catastrophes we could not assess how well each method has done. The second criterion is stability. Exhibit 11 provides a test of the stability of the countrywide relativity method: i.e. the change in state indications between 1999 and 2000. The results are much more stable than the trended and Agg/Agg method to which it is compared. Similar results obtain for the regionwide method. And indeed the regionwide method is superior in this regard. Because of the capping process used in the countrywide method a capped year might become uncapped the next calendar year and vice versa. While this is not necessarily a drawback--our assessment of what is extreme will change with new information--it does cause less stable results. The third criterion is sellability. Here the countrywide method comes out ahead. Most audiences understand and are willing to accept the concept of a relativity to countrywide. However, somewhat paradoxically, once the relativity is to a region, there is, based on informal observation, more of a (negative) flavor of spreading losses. So there remains in the end no clear cut winner.

REFERENCES

[1] Burger, George; Fitzgerald, Beth E; White, Jonathan; Woods, Patrick B., "Incorporating a Hurricane Model into Property Ratemaking," Casualty Actuarial Society Forum 1996 Winter, pages 129-190.

[2] Chernick, David R., "Pricing the Hurricane Peril - Change is Overdue," Casualty Actuarial Society Forum 1998 Winter , pages 23-54.

[3] Walters, Michael A.; and Morin, Francois, "Catastrophe Ratemaking Revisited (Use of Computer Models to Estimate Loss Costs)," Casualty Actuarial Society Forum 1996 Winter pages 347-382.

[4] Hays, David H., and W. Scott Farris, "Pricing the Catastrophe Exposure in Homeowner's Ratemaking," Casualty Actuarial Society *Discussion Paper Program*, 1990, pp. 559-604.

[5] McCarthy, Tim, "A Frequency Based Model for Excess Wind in Property Ratemaking," Casualty Actuarial Society *Forum*, Winter 1998, pp. 209-238.

[6] Homan, Mark, Homeowners Insurance Pricing—An Update", *Pricing Issues,* Casualty Actuarial Society *Discussion Paper Program*, 1990, pp. 719-780.

[7] Bradshaw, John, and Mark Homan, "Homeowners Excess Wind Loads: Augmenting the ISPO Wind Procedure," Casualty Actuarial Society *Forum*, Summer 1993, pp. 339-350.

[8] Dean, Curtis Gary, David N. Hafling, Mark S. Wegner, and William F. Wilson, "Smoothing Weather Losses: A Two Sided Percentile Method," *PCAS* LXXV 1998 pp. 89-100.

[9] Mahler, Howard C., "Credibility With Shifting Risk Parameters, Risk Heterogeneity, and Parameter Uncertainty," *PCAS* LXXXV 1998 pp. 455-653.

[10] Klugman, Stuart A., Harry H. Panjer, and Gordon E. Willmot, *Loss Models: From Data to Decisions,* New York: John Wiley and Sons, 1998.

[11] Venter, Gary G., "Credibility," Foundations of Casualty Actuarial Society Vol 3, pp381-490.

EXHIBIT 1
HOMEOWNERS

DEVELOPMENT OF EXPECTED CATASTROPHE INCURRED LOSS RATIO

(1) Average Earned AIY* for 12 month period ending 3/31/2001	135.88
(2) Factor to Adjust AIY @ 01/01/2003	1.046
(3) Average AIY Trended to 01/01/2003 (1) x (2)	142.13
(4) Total Dollar Catastrophe Provision Per AIY including all LAE (Refer to the Dev. of Total Catastrophe Provision Exhibit)	0.431
(5) Expected Catastrophe Losses (3) x (4)	$61.26
(6) Average Earned Premium @CRL	$443.87
(7) Factor to Adjust Premium for Premium Trend @ 01/01/2003	1.046
(8) Trended Average Earned Premium @CRL (6) x (7)	$464.29
(9) Expected Catastrophe Loss Ratio including all LAE (5) / (8)	13.20%

*1 AIY = $1000 Of Coverage in Force for One Year

EXHIBIT 2

HOMEOWNERS
COUNTRYWIDE
DEVELOPMENT OF CATASTROPHE TREND FACTOR

(1) CALENDAR YEAR	(2) AMOUNT OF INSURANCE YEARS	(3) CATASTROPHE INCURRED LOSS	(4) CATASTROPHE RATIO (3) / (2)	(5) FITTED CATASTROPHE RATIO
1971	50,744,591	5,574,000	0.11	0.155
1972	56,809,992	5,357,000	0.094	0.168
1973	63,630,027	8,119,000	0.128	0.182
1974	71,301,809	23,660,000	0.332	0.196
1975	79,935,311	18,550,000	0.232	0.209
1976	92,593,646	9,278,000	0.1	0.223
1977	109,629,993	11,545,000	0.105	0.237
1978	140,793,253	29,102,000	0.207	0.25
1979	172,171,716	67,836,000	0.394	0.264
1980	205,704,018	56,214,000	0.273	0.278
1981	229,742,921	37,883,000	0.165	0.291
1982	244,770,419	74,005,000	0.302	0.305
1983	259,520,483	91,019,000	0.351	0.319
1984	282,063,918	107,694,000	0.382	0.332
1985	309,884,767	116,237,000	0.375	0.346
1986	352,952,506	95,634,000	0.271	0.36
1987	400,596,851	75,712,000	0.189	0.373
1988	447,064,515	121,665,000	0.272	0.387
1989	503,736,622	184,044,000	0.365	0.401
1990	551,875,055	299,840,000	0.543	0.414
1991	604,545,778	328,134,000	0.543	0.428
1992	628,498,039	357,020,000	0.568	0.442
1993	643,057,601	312,072,000	0.485	0.456
1994	673,490,999	394,674,000	0.586	0.469
1995	709,520,629	405,451,000	0.571	0.483
1996	743,945,331	513,895,000	0.691	0.497
1997	783,663,555	195,818,000	0.25	0.51
1998	831,623,953	328,613,000	0.395	0.524
1999	878,902,781	385,679,000	0.439	0.538
2000	927,355,116	566,488,000	0.611	0.551

1)	Projected Catastrophe Ratio	0.585
2)	Average Catastrophe Ratio	0.344
3)	Catastrophe Trend Factor	1.701

EXHIBIT 3

NON-CATASTROPHE WIND PLUS NON-MODELED CATASTROPHES

	(1) 1990-2000 AIY	(2) 1990-2000 Damage ratio	(3) 2000 AIY	(4) (2) x (3)	(5) (4) x (6) Adj. Dam Ratio	(6) (3)/(5) Adjustment
1	1,459,116,130	0.66	167,735,984	111,360,643	0.62	
2	625,160,130	1.22	73,452,055	89,944,077	1.14	
3	299,750,783	2.03	36,862,828	74,758,372	1.88	
4	968,763,429	0.59	109,989,815	64,731,746	0.55	
5	1,768,915,559	0.37	191,130,811	71,461,447	0.35	
6	278,441,837	0.28	28,216,696	7,898,130	0.26	
7	322,105,688	0.44	43,943,005	19,175,022	0.41	
8	307,439,130	0.37	39,331,209	14,450,301	0.34	
9	1,443,547,377	0.60	158,041,453	94,466,352	0.56	
10	415,847,509	2.65	67,147,524	177,723,848	2.46	
11	43,338,576	0.28	5,551,336	1,561,510	0.26	
Countrywide	7,932,426,148	0.734	921,402,716	727,531,450	0.79	0.93

EXHIBIT 4

WIND+NON-MODELED CATASTROPHE DAMAGE RATIOS
CALCULATION OF EXCESS LOAD

Scrambled State	Year	AIY	Ratio	Normalized	Difference	Load	Ratio with Excess Load
7	1997	4932842	0.13	0.35	-0.22	-1105963	0.46
2	1992	1558165	0.13	0.35	-0.22	-338129	0.46
19	1990	4968539	0.18	0.35	-0.17	-854218	0.46
19	1995	6103028	0.18	0.35	-0.17	-1013489	0.46
1	1995	47046433	0.20	0.35	-0.15	-7132006	0.46
19	1993	5512493	0.20	0.35	-0.15	-819112	0.46
6	1995	6222733	0.21	0.35	-0.14	-884108	0.46
12	1991	21348083	0.22	0.35	-0.13	-2861658	0.46
1	1992	39866439	0.22	0.35	-0.13	-5273210	0.46
6	1990	17708641	0.22	0.35	-0.13	-2244416	0.46
19	1996	6446360	0.22	0.35	-0.13	-805847	0.46
2	1993	1555988	0.23	0.35	-0.12	-194307	0.46
19	1997	6768754	0.24	0.35	-0.11	-752187	0.46
12	1998	26544315	0.24	0.35	-0.11	-2893222	0.46
19	1999	7309175	0.24	0.35	-0.11	-770019	0.46
6	1993	21492188	0.25	0.35	-0.10	-2236552	0.46
6	1992	20173868	0.25	0.35	-0.10	-1922780	0.46
3	1992	5222934	0.26	0.35	-0.09	-471473	0.46
19	1994	5750330	0.27	0.35	-0.08	-487845	0.46
19	1992	5397392	0.29	0.35	-0.06	-340511	0.46
1	1991	38340227	0.31	0.35	-0.04	-1667870	0.46
7	1992	4171362	0.31	0.35	-0.04	-164660	0.46
6	1990	5365445	0.31	0.35	-0.04	-199001	0.46
6	1993	5859889	0.32	0.35	-0.03	-169550	0.46
6	1992	5719858	0.33	0.35	-0.02	-108512	0.46
6	1991	5519497	0.34	0.35	-0.01	-47540	0.46
6	1994	5901932	0.34	0.35	-0.01	-35195	0.46
1	1998	54028023	0.35	0.35	0.00	-221306	0.46
3	1991	5025826	0.36	0.36	0.00	0	0.47
3	1999	7300808	0.36	0.36	0.00	0	0.48
12	1995	23459255	0.37	0.37	0.00	0	0.48
1	1994	43867236	0.38	0.38	0.00	0	0.49
2	1995	10422946	0.40	0.40	0.00	0	0.51
2	1996	1748661	0.40	0.40	0.00	0	0.51
1	1993	41086429	0.40	0.40	0.00	0	0.52
19	1991	5207609	0.42	0.42	0.00	0	0.53
12	1997	25333505	0.43	0.43	0.00	0	0.54
2	1990	1369293	0.44	0.44	0.00	0	0.55
6	1994	23166192	0.46	0.46	0.00	0	0.57
12	1992	21804382	0.47	0.47	0.00	0	0.58
12	1990	19740580	0.47	0.47	0.00	0	0.58
3	1990	4442279	0.48	0.48	0.00	0	0.59
7	1999	6011216	0.49	0.49	0.00	0	0.60

State	Year	AIY	Ratio	Normalized	Difference	Load	Excess Load
7	1998	5459104	0.50	0.50	0.00	0	0.61
2	1994	9659257	0.52	0.52	0.00	0	0.64
6	1991	19287517	0.53	0.53	0.00	0	0.64
1	1999	55360869	0.53	0.53	0.00	0	0.64
6	1996	6543222	0.54	0.54	0.00	0	0.65
6	1995	25501750	0.54	0.54	0.00	0	0.65
1	1997	52321296	0.55	0.55	0.00	0	0.66
2	1995	1655078	0.57	0.57	0.00	0	0.68
12	1999	27636612	0.58	0.58	0.00	0	0.69
12	1993	22146895	0.59	0.59	0.00	0	0.70
2	1990	8596674	0.59	0.59	0.00	0	0.70
2	1998	12701378	0.60	0.60	0.00	0	0.71
6	1996	28761106	0.61	0.61	0.00	0	0.72
7	1990	4083888	0.62	0.62	0.00	0	0.73
1	1990	34421881	0.64	0.64	0.00	0	0.75
2	1999	2224422	0.66	0.66	0.00	0	0.77
2	1991	9124509	0.68	0.68	0.00	0	0.79
12	1994	22671807	0.70	0.70	0.00	0	0.82
1	1996	49718593	0.74	0.74	0.00	0	0.86
6	1997	30502724	0.76	0.76	0.00	0	0.87
2	1992	9038422	0.77	0.77	0.00	0	0.89
2	1993	9178447	0.79	0.79	0.00	0	0.91
7	1993	3993066	0.81	0.81	0.00	0	0.92
2	1997	12162994	0.81	0.81	0.00	0	0.92
7	1991	4326154	0.83	0.83	0.00	0	0.95
6	1998	31253404	0.87	0.87	0.00	0	0.98
3	1997	6474537	0.90	0.90	0.00	0	1.01
3	1995	5867127	0.92	0.92	0.00	0	1.03
12	1996	24394643	0.92	0.92	0.00	0	1.04
2	1997	1839770	0.95	0.95	0.00	0	1.06
6	1997	6853448	1.08	0.95	0.13	868074	1.06
7	1994	4099556	1.16	0.95	0.21	868622	1.06
2	1991	1521504	1.16	0.95	0.21	323906	1.06
3	1994	5506161	1.17	0.95	0.22	1199413	1.06
7	1995	4282047	1.21	0.95	0.26	1106249	1.06
7	1996	4565126	1.30	0.95	0.35	1582846	1.06
3	1993	5258085	1.31	0.95	0.36	1898605	1.06
6	1999	32900006	1.41	0.95	0.46	15220088	1.06
19	1998	7099487	1.43	0.95	0.48	3399160	1.06
6	1999	7355554	1.74	0.95	0.79	5802451	1.06
3	1998	6946682	2.15	0.95	1.20	8309594	1.06
2	1999	13029501	2.16	0.95	1.21	15829750	1.06
2	1998	1982210	2.21	0.95	1.26	2488907	1.06
2	1994	1581060	3.39	0.95	2.44	3854807	1.06
2	1996	11328705	3.49	0.95	2.54	28789905	1.06
3	1996	6226123	5.19	0.95	4.24	26405747	1.06
6	1998	7116595	9.86	0.95	8.91	63430190	1.06
Totals		1291380146			0.11	145363632	

EXHIBIT 5

Credibility weighting by State grouping
States1,2,3,4,5,6
Non-Modeled Catatrophes

	1	2	3	4	5	6		
1990	0.29	0.86	0.00	1.11	0.59	0.43		
1991	0.41	0.72	1.32	1.02	0.25	0.43		
1992	0.18	0.61	0.16	0.18	0.36	0.01		
1993	0.03	0.19	0.09	0.93	0.90	0.32		
1994	0.76	0.09	1.12	0.82	0.69	1.82		
1995	0.56	1.69	0.23	0.89	0.18	0.70		
1996	5.65	0.93	1.57	1.15	0.24	0.60		
1997	0.87	0.37	0.07	0.23	0.12	0.81		
1998	0.07	0.27	0.45	1.60	1.26	1.33		
1999	3.06	0.46	0.12	0.04	0.21	0.96	Regional	Variance of
2000	2.00	5.21	0.14	0.96	0.45	0.15	Mean	Means
Artihmetic m	1.26	1.04	0.48	0.81	0.48	0.69	0.79	0.098
							Average Process Variance	
Process variance	2.971	2.111	0.328	0.225	0.126	0.281	1.007	
Estimated Process variance	1.989	1.559	0.667	0.616	0.566	0.644		

Total variance	1.012
Estimated VHM	0.052

	1	2	3	4	5	6	SUM
(1) Cred estimate	0.207	0.249	0.437	0.457	0.478	0.446	
(2) Damage ratio estimate	0.89	0.85	0.66	0.80	0.64	0.75	
(3) 2000 AIY	3153771	9386777	2389888	7501237	15807745	7830024	46069442
(4) (2)*(3)	2805442	8007591	1566336	6003659	10142149	5834552	34359729
Balanced estimates	0.82	0.79	0.61	0.74	0.59	0.69	

0.746 Implied 2000 regional damage ratio
0.69 Chosen regional damage ratio
0.925 Balancing adjustment

EXHIBIT 6

Credibility weighting by State grouping
1,2,3,4,5,6
Non-modeled Catastrophes and Non-Hurricane Wind

		1	2	3	4	5	6			
1990		1.10	1.42	0.89	1.90	1.02	0.86			
1991		1.41	1.71	2.82	1.59	0.68	0.86			
1992		1.13	1.34	0.97	0.64	0.73	0.34			
1993		0.74	0.74	0.74	1.39	1.33	0.76			
1994		1.55	0.39	1.80	1.20	1.01	2.11			
1995		1.17	2.02	0.85	1.47	0.61	1.05			
1996		6.40	1.39	2.17	1.61	0.54	1.07			
1997		1.34	0.89	0.74	0.88	0.49	1.27			
1998		0.55	0.83	1.20	2.37	1.86	2.28			
1999		3.67	0.92	0.69	0.71	0.64	1.32			
2000		2.58	5.78	0.67	1.36	0.96	0.57			
Arth. Mean		1.97	1.58	1.23	1.37	0.90	1.14	Grand Mean:	1.37	
								Variance of means:	0.14	
Process variance		2.94	2.15	0.52	0.26	0.17	0.36	Average Process Var	1.07	
Estimated Process variance		2.00	2.00	2.00	2.00	2.00	2.00			
Total variance	1.1026									
Estimated VHM	0.0881									
(1) Cred estimate		0.3259	0.3259	0.3259	0.3259	0.3259	0.3259			
(2) Damage ratio estimate		1.56	1.44	1.32	1.37	1.21	1.29		**Totals**	
(3) 2000 AIY		3,153,771	9,386,777	2,389,888	7,501,237	15,807,745	7,830,024		46,069,442	
(4) (2)*(3)		4,924,742	13,485,689	3,157,599	10,263,966	19,177,172	10,106,926		61,116,094	
									1.3266	Implied 2000 regional damage ratio
Balanced estimates		1.34	1.23	1.14	1.18	1.04	1.11		1.1400	Chosen regional damage ratio
									0.859	Balancing adjustment

EXHIBIT 7

COUNTYWIDE NON-MODELED CATASTROPHE RELATIVITIES

	CALENDAR YEAR	State 1	State 2	State 3	State 4	State 5	State 6	State 7	State 8	State 9	State 10	State 11	State 12	State 13
	1981	0.17	0.27	0.01	0.23	0.63	0.52	1.04	4.53	0.14	0.41	0.22	0.67	0.10
	1982	0.60	0.78	7.99	0.00	0.79	1.69	0.41	5.41	1.29	0.56	0.03	0.77	0.74
	1983	0.31	1.10	1.75	0.02	0.32	0.26	0.58	1.31	0.40	2.05	-0.01	0.60	2.73
	1984	1.37	1.70	0.42	0.11	1.48	0.55	0.70	1.95	0.85	0.68	0.03	0.22	0.80
	1985	0.12	1.17	0.30	0.00	0.40	0.26	0.15	2.82	0.74	0.74	0.02	0.33	0.45
	1986	0.00	0.20	0.18	0.20	0.07	0.31	3.50	1.63	4.75	0.97	0.50	0.01	0.13
	1987	0.22	0.08	1.07	1.47	0.74	1.03	1.25	6.22	1.13	1.72	0.15	1.16	0.12
	1988	0.75	0.29	-0.45	0.04	0.61	1.47	2.51	0.30	0.22	1.44	0.02	0.42	0.06
	1989	0.43	0.85	0.01	0.17	0.38	2.19	0.29	0.36	0.31	3.00	0.02	0.54	0.00
	1990	0.00	0.92	0.00	0.18	0.76	0.61	0.47	1.94	0.34	1.36	0.04	0.11	0.04
	1991	0.01	0.43	0.01	0.29	0.26	0.87	1.68	8.27	0.20	1.24	0.03	0.01	0.55
	1992	0.16	0.50	0.02	0.20	0.13	0.84	0.05	17.13	0.03	0.84	-0.01	0.02	0.07
	1993	0.28	1.70	-0.03	0.03	0.47	1.16	0.12	6.49	1.95	0.36	0.07	0.35	0.00
	1994	0.78	1.02	0.00	0.00	0.23	0.45	4.82	2.96	1.22	0.14	0.15	0.71	0.12
	1995	0.00	0.26	0.00	0.12	0.08	0.27	0.71	2.88	0.99	2.50	0.34	0.12	0.28
	1996	0.25	0.33	0.00	0.18	0.81	4.52	0.38	1.73	6.91	1.31	0.03	0.95	0.36
	1997	0.05	0.47	0.00	0.13	1.56	2.23	3.21	0.72	2.55	1.43	-0.03	0.46	1.84
	1998	0.20	3.16	0.00	0.01	0.51	1.01	5.15	2.98	4.74	0.69	4.20	-0.05	1.63
	1999	-0.01	0.48	0.00	0.00	0.73	4.34	0.98	5.05	0.20	1.04	0.01	0.18	2.44
	2000	0.01	0.74	0.00	0.01	1.71	1.94	0.53	-0.15	1.97	8.53	0.12	0.42	0.51
Arithmetic mean		0.28	0.82	0.56	0.17	0.63	1.33	1.43	3.73	1.55	1.55	0.30	0.40	0.65
Weighted mean		0.26	0.90	0.19	0.12	0.75	1.68	1.59	4.01	1.91	2.00	0.35	0.40	0.69
	Max	1.37	3.16	7.99	1.47	1.71	4.52	5.15	17.13	6.91	8.53	4.20	1.16	2.73
	St Dev	0.35	0.72	1.81	0.32	0.47	1.24	1.57	3.90	1.87	1.79	0.93	0.34	0.83
	CV	1.24	0.87	3.21	1.90	0.75	0.93	1.10	1.05	1.21	1.16	3.13	0.84	1.28
	Linear Trend	-0.02	0.02	-0.15	-0.01	0.00	0.11	0.10	0.05	0.12	0.01	0.05	-0.02	0.03
	R-Squared	0.12	0.02	0.20	0.02	0.00	0.26	0.12	0.00	0.13	0.00	0.10	0.08	0.05
Correlation of Damage Ratios to Adjusted CW		0.14	0.32	-0.24	0.08	0.26	0.40	0.14	0.47	0.33	0.37	-0.04	0.28	-0.08
Process variance		0.12	0.52	3.27	0.10	0.23	1.53	2.45	15.25	3.49	3.22	0.86	0.11	0.70

EXHIBIT 8

Region 1 Non-modeled Catastrophe Relativities
(Adjusted to regional 2000 AIY distribution)

	1	2	3	4	5	6	7	8	9		
1981	3.044	2.110	3.271	0.275	0.519	1.059	1.272	0.547	0.213		
1982	0.161	0.426	4.293	1.347	0.881	1.764	0.829	0.238	0.771		
1983	1.301	0.649	1.238	0.453	0.399	0.293	0.354	0.230	3.057		
1984	0.633	0.671	3.923	0.811	0.214	0.529	1.423	0.784	0.770		
1985	0.033	0.234	3.149	1.148	2.294	0.397	0.620	0.623	0.705		
1986	0.220	8.671	1.208	11.792	0.420	0.758	0.170	0.020	0.332		
1987	1.664	1.980	1.503	1.795	0.933	1.642	1.170	0.260	0.188		
1988	0.016	5.905	0.286	0.528	0.279	3.460	1.446	0.148	0.153		
1989	0.425	0.778	0.571	0.845	0.156	5.884	1.020	1.002	0.009		
1990	0.464	1.050	1.282	0.756	0.905	1.359	1.678	0.218	0.091		
1991	0.607	4.260	2.401	0.522	0.126	2.227	0.669	1.262	1.409		
1992	0.472	0.214	1.061	0.126	1.636	3.825	0.603	1.302	0.316		
1993	0.257	0.229	1.992	3.697	1.370	2.202	0.898	0.177	0.007		
1994	0.200	10.870	3.221	2.761	1.373	1.008	0.525	0.137	0.265		
1995	0.006	2.776	5.405	3.900	0.823	1.050	0.301	0.002	1.109		
1996	0.351	0.296	1.162	5.388	0.672	3.528	0.632	0.114	0.284		
1997	1.388	2.092	-0.119	1.660	0.585	1.454	1.015	0.266	1.199		
1998	10.964	2.323	0.371	2.140	0.089	0.457	0.229	1.442	0.736		
1999	2.345	0.673	0.326	0.134	0.559	2.972	0.500	0.084	1.674		
2000	1.587	0.429	0.014	1.595	0.544	1.569	1.386	1.325	0.415		
Arth. Mean	1.308	2.333	1.828	2.084	0.739	1.872	0.837	0.509	0.685	Grand Mean:	1.355
										Variance of means:	0.320
Process variance	5.884	8.799	2.472	7.199	0.321	2.053	0.200	0.244	0.541	Average Process Var	3.079
Estimated Process variance	4.223	4.223	4.223	4.223	4.223	4.223	4.223	4.223	4.223		

Total variance	3.244
Estimated VHM	0.320

	1	2	3	4	5	6	7	8	9		
(1) Cred estimate	0.635	0.635	0.635	0.635	0.635	0.635	0.635	0.455	0.455		
(2)Relativity estimate	1.325	1.976	1.655	1.818	0.963	1.684	1.026	0.970	1.050	**Totals**	
(3) 2000 AIY	7639139	2693808	6656491	7677641	28965458	13675577	58042460	7548625	34836785	167,735,984	
(4) (2)*(3)*Chosen Regional	4397677	2312986	4787313	6064115	12123511	10002277	25867396	3181945	15896752	84,633,973	
Damage Ratio										0.505	Implied 2000 reg
Balanced estimates	0.496	0.739	0.619	0.680	0.360	0.630	0.384	0.363	0.393	0.434	Chosen regional
										0.861	Balancing adjustm

EXHIBIT 9

Sample of Proposed Non-modeled Cat loading calculation

Year	Damage Ratio	CW Damage Ratio (reweighted)	Relativity to CW	Capped Relativities
1981	|	|	|	|
|	|	|	|	|
1990	6.728	0.634	10.619	6.250
1991	3.639	0.582	6.250	6.250
1992	0.970	0.721	1.346	1.346
1993	0.634	0.531	1.195	1.195
1994	2.008	0.680	2.954	2.954
1995	0.117	0.675	0.173	0.173
1996	0.762	0.712	1.070	1.070
1997	0.824	0.257	3.206	3.206
1998	0.296	0.397	0.745	0.745
1999	0.297	0.443	0.670	0.670
2000	0.205	0.610	0.336	0.336
(1) arithmetic mean			2.650	2.394
(2) standard deviation			2.527	
(3) max relativity			10.619	
(4) relativity cap:(1)+3*(2)			10.232	
(5) process variance				3.68
(6) credibility				0.901
(7) credibility weighted relativity				2.290
(8) implied 2000 losses				19,794,656
(9) implied 2000 1 sd loss				37,976,702
(10) rebalanced, to 1.0, relativity				2.645
(11) balanced damage ratio				1.367

NOTES:

(4) Three standard deviations (calculated from the unadjusted numbers) above the arithmetic mean

(5) The process variance is the unweighted variance of the adjusted 20 year relativities.
The average across states of this number is the process variance used in the credibility formula

(6) Credibility is 20/(20+K); K is the ratio of the above process variance and the VHP(not shown)

(7) (1)*(6)+(1-(6))*1

(8) is (7)*selected CW damage ratio*State 2000AIY

(9) is sqrt((5))*(8)

(10) adjusts individual state indicated relativities for the difference between the weighted (by AIY) CW relativity and 1. This adjustment by state is done in proportion to (9)

11) The balanced damage ratio is (10)*the selected CW damage ratio

Exhibit 10

COMPARISON OF METHODS

STATE	Agg-to-agg	Trended	Relativity Regional	Relativity Countrywide
1	1.984	2.522	1.71	2.184
2	1.872	2.555	1.60	1.861
3	1.683	2.337	2.20	2.114
4	1.133	1.524	1.30	1.367
5	1.073	1.578	1.43	1.384
6	1.008	1.57	1.35	1.073
7	0.900	0.908	0.47	0.502
8	0.868	1.994	1.60	1.024
9	0.819	0.98	0.60	0.889
10	0.720	1.024	0.78	0.816
11	0.698	0.963	0.62	0.888
12	0.693	0.865	1.05	0.801
13	0.689	0.743	0.58	0.755
14	0.642	0.795	1.15	0.859
15	0.614	0.783	1.03	0.798
16	0.567	0.7	0.53	0.671
17	0.539	0.668	0.60	0.718
18	0.525	0.83	0.52	0.602
19	0.515	0.98	0.56	0.659
20	0.449	0.549	0.91	0.578
21	0.422	0.516	0.52	0.557
22	0.405	0.397	0.30	0.450
23	0.391	0.495	0.45	0.442
24	0.356	0.505	0.43	0.534
25	0.335	0.412	0.29	0.364
26	0.296	0.413	0.35	0.471
27	0.296	0.359	0.40	0.395
28	0.284	0.288	0.27	0.261
29	0.269	0.426	0.40	0.382
30	0.239	0.271	0.32	0.320
31	0.236	0.384	0.38	0.309
32	0.236	0.274	0.34	0.291
33	0.226	0.308	0.30	0.186
34	0.210	0.359	0.20	0.147
35	0.202	0.247	0.23	0.232
36	0.201	0.208	0.25	0.212
37	0.189	0.256	0.38	0.249
38	0.184	0.24	0.28	0.331
39	0.173	0.252	0.24	0.261
40	0.164	0.199	0.16	0.261
41	0.158	0.136	0.15	0.113
42	0.144	0.165	0.14	0.173
43	0.141	0.171	0.21	0.176
44	0.120	0.197	0.28	0.298
45	0.099	0.133	0.17	0.147
46	0.090	0.368	0.085	0.187
47	0.058	0.069	0.15	0.116
48	0.048	0.051	0.08	0.099
49	0.033	0.042	0.08	0.098

EXHIBIT 11

SENSITIVITY: CHANGE IN INDICATED DAMAGE RATIO BETWEEN 1999 AND 2000

State	Trended	AGG/AGG	CW Relativity Method
1	-2.4%	-9.0%	2.2%
2	-1.5%	-9.0%	-3.5%
3	-1.4%	-7.3%	-4.0%
4	-1.4%	-6.4%	-4.3%
5	-1.4%	-10.6%	-3.1%
6	-1.3%	-7.1%	-2.6%
7	-1.3%	-10.9%	-12.0%
8	-1.2%	-9.8%	-2.9%
9	-1.2%	-5.9%	-4.0%
10	-1.1%	-7.4%	-4.0%
11	-1.1%	-6.7%	-3.7%
12	-1.0%	-8.0%	-3.8%
13	-0.9%	-7.4%	-3.9%
14	-0.7%	-7.1%	-3.3%
15	-0.6%	-7.1%	-2.9%
16	-0.5%	-6.8%	-3.4%
17	-0.4%	-6.5%	-2.6%
18	-0.1%	-5.2%	-2.8%
19	0.0%	-6.4%	-1.6%
20	0.0%	-4.8%	-2.3%
21	0.4%	-4.3%	-2.3%
22	0.5%	-5.0%	2.8%
23	0.8%	-3.6%	-1.9%
24	1.2%	-3.3%	-1.8%
25	1.3%	-2.9%	-2.0%
26	1.4%	-4.0%	-1.3%
27	2.2%	-3.9%	1.0%
28	2.8%	-2.0%	-0.4%
29	3.0%	-0.5%	-0.4%
30	3.6%	0.4%	0.0%
31	3.8%	1.3%	0.6%
32	3.9%	2.4%	-0.2%
33	4.0%	3.5%	0.9%
34	4.0%	2.3%	0.8%
35	4.4%	4.9%	2.7%
36	4.5%	2.8%	1.6%
37	5.3%	7.5%	0.7%
38	5.6%	3.6%	2.2%
39	5.7%	4.5%	2.2%
40	5.7%	1.7%	1.4%
41	6.0%	1.9%	0.7%
42	6.1%	2.2%	7.5%
43	7.5%	5.3%	3.2%
44	9.6%	7.2%	4.6%
45	12.2%	21.2%	9.7%
46	16.4%	34.1%	16.3%
47	16.6%	20.2%	8.7%
48	20.3%	29.0%	32.0%
49	41.5%	61.4%	8.1%
CW	3.9%	1.9%	0.6%

Countrywide numbers weighted by 2000 AIY's

A Unifying Approach to Pricing Insurance and Financial Risk

Andreas Kull

A Unifying Approach to Pricing Insurance and Financial Risk

September 10, 2002

Andreas Kull*

Abstract

The actuarial and the financial approach to the pricing of risk remain different despite the increasingly direct interconnection of financial and insurance markets. The difference can be summarized as pricing based on classical risk theory (insurance) vs. non-arbitrage pricing (finance). However, comparable pricing principles are of importance when it comes to transferring insurance risk to financial markets and vice versa as it is done e.g. by alternative risk-transfer instruments or derivative products. Incompatibilities blur business opportunities or may open up the possibility to arbitrage.

For these situations, the paper aims to bridge the gap between insurance and finance by extending the non-arbitrage pricing principle to insurance. The main obstacle that has to be tackled is related to the incompleteness of the insurance market. It implies that equivalent martingale probabilities are not uniquely defined. By the information theoretical maximum entropy principle a sensible way to choose a particular equivalent martingale measure is found. This approach parallels the successful application of the maximum entropy principle in finance.

The paper pays special attention to the role that investment opportunities beyond risk-free investments play for insurance operations. Equivalent martingale probabilities for the combination of insurance operation risk-free investment and a risky investment are determined. They turn out to be connected to the Esscher measure. This recovers a generalized form of a well-known actuarial premium calculation principle.

The sketched approach is further investigated for typical reinsurance structures like stop-loss and excess-of-loss reinsurance. Arbitrage-free reinsurance premiums are calculated. A numerical example stresses the influence that characteristics of risky investment opportunities have on arbitrage-free premiums.

* Converium Ltd, General Guisan-Quai 26, CH-8022 Zurich, Switzerland (e-mail: Andreas.Kull@converium.com)

1 Introduction

The interconnection of financial and insurance markets has become more direct during the past two decades. This is due to the repositioning of insurance companies as integral financial service providers, increasing exposures that tend to exceed the capacity of the insurance market and finance-related insurance products (e.g. catastrophe-bonds). Presumably, the convergence of insurance and finance has attained its most advanced level in alternative risk transfer contracts, where insurance and financial risks are covered jointly.

Financial and insurance markets and the pricing of respective products differ in many respects. On the market side, the main difference can be summarized as warehousing of risk (insurance) versus intermediation of risk (finance). Insurance pricing on the other hand is in general based on classical risk theory while finance relies on non-arbitrage pricing. Establishing comparable pricing principles is, however, of importance when it comes to transferring insurance risk to financial markets and vice versa. Incompatibilities blur business opportunities or open up the possibility to arbitrage. This is a main motivation for this paper.

Non-arbitrage pricing relies on liquid and efficient markets. Clearly, most of the insurance market is neither liquid nor efficient. Nevertheless, there are situations where in addition to pricing based on classical risk theory the corresponding non-arbitrage value of an insurance contract is of interest, e.g. when part of the risk is transferred to the financial markets or insurance risk is traded. With a growing interconnection of financial and insurance markets this situation becomes more frequent. A pioneering area in this respect is the insurance of credit risk, where warehousing and intermediation of risk overlap in a very natural way.

Pricing of insurance contracts is commonly based on real probabilities P, i.e. probabilities reflecting the actual likelihood of loss events. In the simplest one period case, the premium for an insurance contract covering losses X according to the equivalence principle is

$$\text{premium} = \frac{1}{1+r_l} E^P[X] + S[X]$$

(1.1)

where $E^P[X]$ is the loss expectation calculated under the real probability measure P, discounted with a risk-discount rate r_l chosen according to actuarial judgement. $S[X]$ is the safety loading or risk premium. A particular choice of r_l and $S[X]$ should reflect the overall risk related to the contract, the risk-free interest rate, the cost of capital, the expected investment return, market conditions etc. For complex contracts this is usually not an easy task.

Following the seminal work of Black, Scholes and Merton in the early 70's, in finance, the non-arbitrage pricing principle lead to a shift away from the real probability measure P to equivalent martingale measures Q. Valuation of a future (stochastic) cash flow F under the equivalent martingale measure Q takes the form

$$\text{price} = \frac{1}{1+r_{rf}} E^Q[F]$$

(1.2)

where $E^Q[\cdot]$ denotes the expectation operator under the measure Q. Pricing in this context thus invokes again the equivalence principle, however, with respect to the equivalent martingale measure Q. The discounting is performed with the directly observable risk-free interest rate r_{rf} and there is no modification due to a loading. An appealing feature of (1.2) is that once the equivalent martingale measure Q is known, the valuation of the cash flow F is performed without recourse to subjective criteria. Economically, non-arbitrage pricing derives its justification from the existence of a hedge portfolio that creates an overall risk-less position. The existence of a hedge portfolio in turn relies on liquid and efficient markets. Arbitrage-free pricing was pioneered by Black and Scholes (1973). Cox and Ross (1976) and Harrison and Kreps (1979) established the sound theoretical basis in terms of risk-neutral valuation and equivalent martingale measures.

The main difference between classical risk theory and the non-arbitrage approach to pricing is that the non-arbitrage approach substitutes real probabilities and expert knowledge for 'preference-free' probabilities that comply with the non-arbitrage assumption. The fundamental task in both cases remains to find the corresponding probabilities.

Gerber (1973) has introduced martingale methods to risk theory. Since then, a number of papers have been investigating martingales in risk theory. Mainly, these papers deal with assessing ruin probabilities. For a review of insurance related use of martingals see e.g. Schmidli (1996). The martingale approach to premium calculation, which is considered here, has been pioneered by Delbaen and Haezendonck (1989). In this paper it was shown that common premium principles can be recovered by martingale methods. Another important paper in this context is by Sondermann (1991), who considers arbitrage-free pricing for reinsurance.

The emphasis of this paper lies on the practical application of martingales to the pricing of insurance contracts whose performances depend on financial markets. The main problem arising is that, due to the incompleteness of the insurance market, equivalent martingale probabilities are not uniquely defined. Another difficulty arises since often the return distribution of insurance contracts is fundamentally different from that of asset returns ('heavy tails'). Thus it is not straightforward to apply standard financial techniques as e.g. the Black-Scholes-Merton framework. We show how the information theoretical maximum entropy principle can be applied to choose in a sensible way a particular equivalent martingale measure in this situation.

The paper pays special attention to the role that investment opportunities beyond risk-free investments play for insurance operations. Equivalent martingale probabilities for the combination of insurance operation risk-free investment and a risky investment are determined. They turn out to be connected to the Esscher measure. This recovers a generalized form of a well-known actuarial premium calculation principle.

The maximum entropy principle has been successfully applied in other fields and to similar problems in finance (e.g. Stutzer, 1996). The methods used here thus are not new. Also, other approaches to unify the actuarial and financial pricing based on different methodology (e.g. deflators, Jarvis et al. 2001) exist.

The paper is organized as follows: Section 2 gives a very brief review of basic concepts used in this paper. In Section 3 a non-arbitrage condition for insurance contracts is formulated. Section 4 tackles the problem of determining equivalent martingale probabilities in an incomplete market. Equivalent martingale probabilities are determined by making use of the maximum entropy principle. Section 5 generalizes the theory to include investment opportunities. Section 6 discusses issues related to unique valuation and implied discounting rates. As an example, simple reinsurance structures are considered in Section 7. This section also contains results of numerical simulations that illustrate some of the main results. Section 8 extends the one-period case of Section 7 to a multi-period framework. The last section presents conclusions.

2 Brief Review of Basic Concepts used in this Paper

This paper aims at bridging the gap between financial and insurance pricing by introducing some of the concepts of modern finance and information theory into insurance. For our purpose, the concepts of non-arbitrage pricing and equivalent martingale probabilities are of importance. They are covered by standard textbooks like e.g. Hull (2000) or Copeland and Weston (1992). In short, non-arbitrage pricing relies on the insight[1] that, in the absence of arbitrage opportunities, the price (or premium) of some contingent claim should match the price for a position perfectly hedging its risk. As has first been shown by Cox and Ross (1976) and Harrison and Kreps (1979), this insight can be reformulated mathematically in terms of 'equivalent martingale probabilities', i.e. in terms of a probability measure Q satisfying for a given random process X

$$E^{Q}[X_T] = X_t \tag{2.1}$$

[1] It was this insight (and its mathematical formulation) that gained Merton and Scholes the Nobel price in 1997.

Equation (2.1) has a straightforward interpretation: The best forecast of future values X_T at $T > t$ is the observed value X_t.

When pricing contingent claims in the non-arbitrage framework, equivalent martingale probabilities substitute the 'real' probabilities P e.g. derived from historical data. Dealing with equivalent martingale probabilities one should keep in mind that they emerge as a consequence of the non-arbitrage assumption and that they're 'artificial' insofar as they do not need to correspond to any real probabilities or beliefs. Second, it is important to note that only for complete and efficient markets there exists a unique equivalent martingale measure. In incomplete markets, however, equivalent martingale measures are not uniquely defined. This reflects the fact that contingent claims can be hedged only partially.

In our context of incomplete insurance markets, basic concepts of information theory come into play when a particular equivalent martingale measure Q is chosen from the infinite possibilities. Information theoretical concepts prove useful by providing a measure of the information a particular equivalent martingale measure Q embodies. In the absence of information other than the non-arbitrage assumption, the rationale is to choose the distribution that embodies least additional information. In a discrete setting, this is achieved by maximizing the entropy[2]

$$S = -\sum_i q_i \cdot \ln[q_i] \tag{2.2}$$

where the q_i's are the probabilities associated with a particular equivalent martingale measure Q Defining probability distributions by maximizing (2.2) has a longstanding history in information theory and statistical physics. The method is known as the *Information Theoretical Maximum Entropy Principle.* A more comprehensive review of information theoretical concepts lies beyond the scope of this paper. However, there are accessible textbooks covering these topics (see e.g. Cover and Thomas (1991)).

3 Formulating the Non-Arbitrage Constraint

Consider the simplest form of insurance: An insurer accepts the liability to pay for the compound loss

$$X = \sum_{j=1}^{N} L_j \tag{3.1}$$

occurring over a period $[0,T]$ where L_j is the (random) jth claim amount during the time period and N is the random number of claims. Here we will assume that the claim amounts L_j are independently and identically distributed and $L_j \equiv L$ refers to the claim distribution without deductible or limit. In exchange for this liability, the insurer receives a premium b. For simplicity it is assumed that there are no costs or investment returns and payments are made at time T only (for a generalization see Section 4). Then, the premium b can be interpreted as a risky asset generating one-period returns

$$R = \frac{b - X}{b} \tag{3.2}$$

The definition of the insurance related return R directly depends on b. To stress this point, b will be referred to as the reference premium in the following.

The non-arbitrage theorem can be expressed in different ways. An intuitive formulation in a discrete setting refers to different states i of the world, each of which is characterized by a set of payoffs. These payoffs originate from assets, e.g. in our case from the premium b or a risk-free bond. To formulate the non-arbitrage condition, realizations $l_{j,i}$ and n_i of the single claim amount L and the claim number N are considered. They translate by (3.1) and (3.2) into realizations x_i and r_i of the compound loss X and return R respectively. In addition, a risk-free

[2] Behind this lies the equivalence of the expression for information content and the expression for entropy as

bond with return r_{rf} is considered. Following standard procedures (see e.g. Neftci (2000)) in finance, $(1+r_i)$ and $(1+r_{rf})$ are identified with the insurance and bond related payoffs in the ith state. Grouping the payoffs in a matrix $D \in \Re^{2,K}$ where K is the total number of states leads to

$$D = \begin{bmatrix} (1+r_{rf}) & (1+r_{rf}) & \Lambda & (1+r_{rf}) \\ (1+r_1) & (1+r_2) & \Lambda & (1+r_K) \end{bmatrix}$$

(3.3)

The non-arbitrage theorem now states that if there is no arbitrage, positive constants ψ_i exist such that

$$\begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} (1+r_{rf}) & (1+r_{rf}) & \Lambda & (1+r_{rf}) \\ (1+r_1) & (1+r_2) & \Lambda & (1+r_K) \end{bmatrix} \cdot \begin{bmatrix} \psi_1 \\ \psi_2 \\ \mathrm{M} \\ \psi_K \end{bmatrix}$$

(3.4)

holds. With $q_i = \psi_i \cdot (1+r_{rf})$ relation (3.4) becomes

$$\begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & \mathrm{L} & 1 \\ \frac{(1+r_1)}{(1+r_{rf})} & \frac{(1+r_2)}{(1+r_{rf})} & \mathrm{L} & \frac{(1+r_K)}{(1+r_{rf})} \end{bmatrix} \cdot \begin{bmatrix} q_1 \\ q_2 \\ \mathrm{M} \\ q_K \end{bmatrix}$$

(3.5)

from which it is evident that the quantities q_i can be interpreted as probabilities (as the inspection of the first component demonstrates the positive q_i's sum up to one). The q_i's are interpreted as 'risk-adjusted' or 'risk-neutral' probabilities. From (3.5) it follows that under the probability measure Q we have

---

discovered by Shannon (1948) who identified the generation of information with the reduction of entropy.

$$E^{Q}\left[\frac{1+R}{1+r_{rf}}\right]=\frac{1}{1+r_{rf}}E^{Q}\left[\frac{X-b}{b}+1\right]=1$$

(3.6)

i.e. under the measure Q, $(1+R)/(1+r_{rf})$ is a martingale and the q_i's are equivalent martingale probabilities.

Given the probabilities q_i, the arbitrage-free pricing of an arbitrary insurance contract referring to the losses L and the claim number N is straightforward. It takes the form

$$\text{non}-\text{arbitrage premium}=\frac{1}{1+r_{rf}}E^{Q}[f(L,N)]$$

(3.7)

where $f(L,N)$ stands for the loss amount to be paid according to the insurance contract. Relation (3.7) is formally identical to (1.2). According to it, the non-arbitrage premium is the expectation value calculated under the equivalent martingale measure Q, discounted by the risk-free interest rate r_{rf}. No risk premium is added since the risk adjustment is internalized by the change of the probability measure. Pricing any insurance contract whose loss can be written as $f(L,N)$ thus reduces to determine the equivalent martingale probabilities q_i.

In complete and efficient markets, the strict economic justification of (3.7) relies on the fact that the non-arbitrage premium coincides with the (unique) price of a hedging portfolio that yields an overall risk-less position. In incomplete markets (e.g. insurance markets) considered here, however, the situation is more complex. It can be shown that no unique and perfect hedging portfolio exists. This is reflected by the fact that in general the equivalent martingale probabilities q_i are not uniquely defined. The question that arises is how the martingale probabilities should be defined, or, in other words, what particular equivalent martingale measure Q should be chosen out of the infinitely many possibilities. We will address this question in the next section.

4 Equivalent Martingale Measure

The insurance market is neither liquid, effective, nor it is complete. An immediate consequence is that pricing contingent claims by their replication cost is not possible which implies that no unique equivalent martingale measure exists. Other practical difficulties arise since often the loss distribution of L is fundamentally different from that of asset returns. In particular, the loss distribution L may often show heavy tails, i.e. relatively high probabilities of large losses[3]. Thus it is not straightforward to apply standard financial techniques like the Black-Scholes-Merton framework, which implicitly relies on complete markets and log-normal distributed random variables. How can, nevertheless, as much as possible of the appealing properties of non-arbitrage pricing be recovered in this situation?

The equivalent martingale probabilities q_i have to fulfill according to the non-arbitrage assumption the relation (3.5) representing a linear system of two equations and K unknowns q_i. The system is of the form

$$S = \frac{1}{1+r_0} D * q \tag{4.1}$$

For given $S = [1,1]^T$ and D (specified by K realizations r_i), q is not determined uniquely. Indeed, there exist an infinite number of solutions for q_i which reflects market incompleteness.

The underdetermined nature of (4.1) is formally known as the 'Inverse Problem'. Several methods exist to deal with such situations. One of the most elegant and well-founded ways is provided by the maximum entropy principle of information theory, which we will follow here. In finance, the maximum entropy principle has been successfully applied in a similar context to problems related to option pricing. Rubinstein (1994) used the maximum entropy principle to

---

[3] A characteristic property of heavy tailed distributions is the non-existence of higher moments $E\left[X^k\right]$ for some $k > k_0$. As an example, consider the Pareto distribution whose second and higher moments do not exist if $\alpha < 2$.

infer martingale measures from observed option prices (Rubinstein 1994). Buchen and Kelly (1996), Stutzer (1996) and Gulko (1999) worked out generalizations and applications of this idea referring to option pricing. In the following, we use to a large extent the same well known information theoretic methods as Stutzer (1996).

4.1 Determining Equivalent Martingale Probabilities

The only information available about the equivalent martingale probabilities q_i is the non-arbitrage constraint as specified by (3.5) and the assumption that all realizations r_i are equally alike. Any other additional information (e.g. the volatility of the equivalent martingale probabilities q_i) would be imposed on the martingale probabilities q_i without justification since relation (3.5) is the only constraint defining properties of the equivalent martingale probabilities. A consistent criterion to make a choice along this line is the additional information a particular distribution is adding besides the constraint (3.5). One should try to minimize it. Minimizing additional information is equivalent to maximize the entropy

$$S = -\sum_i q_i \cdot \ln[q_i] \tag{4.2}$$

associated with a particular distribution $\{q_i\}$ (Jaynes, 1957a,b). This information theoretical *Maximum Entropy Principle* guarantees the chosen distribution to incorporate no information other than specified by the constraint (3.5), which is equivalent to choose the distribution that is 'most unbiased' or is 'embodying least structure'.

The problem of maximizing the entropy associated with the distribution q_i is a constraint maximizing problem:

$$\max S = -\sum_i q_i \cdot \ln[q_i] \tag{4.3}$$

under the constraints

I: $\sum_i q_i = 1$ (normalization)

(4.4)

II: $q_i \geq 0, \ \forall i$ (positivity)

(4.5)

III: $E^Q\left[\frac{1+R}{1+r_{rf}}\right] = \frac{1}{1+r_{rf}} \sum_i q_i (1+r_i) = 1$ (non-arbitrage condition)

(4.6)

It is solved by using Lagrangian multiplier techniques, i.e. maximizing the expression

$$L = -\sum_i q_i \cdot \ln[q_i] + \gamma\left(\frac{1}{1+r_{rf}} \sum_i q_i (1+r_i)\right) + \beta\left(\sum_i q_i\right)$$

(4.7)

where, γ and β represent the Lagrangian multipliers associated with the constraints (4.4) and (4.6). Maximizing (4.7) leads to

$$q_i = \frac{\mathrm{Exp}\left[\hat{\gamma}\frac{1+r_i}{1+r_{rf}}\right]}{\sum_j \mathrm{Exp}\left[\hat{\gamma}\frac{1+r_j}{1+r_{rf}}\right]}$$

(4.8)

where $\hat{\beta}^{-1} = \sum \mathrm{Exp}[\hat{\gamma}(1+r_i)/(1+r_{rf})]$ (constraint (4.4)). The parameter $\hat{\gamma}$ is to be determined numerically by the non-arbitrage constraint (4.6). Since the distribution (4.8) is an exponential, constraint (4.5) is automatically fulfilled.

To recapitulate: The distribution (4.8) is the maximum entropy probability distribution (or 'most unbiased' probability distribution). It has been deduced from the maximum entropy principle and the non-arbitrage constraint. By construction, the probabilities (4.8) are equivalent martingale probabilities. Their distribution is consistent with the non-arbitrage theorem and is otherwise assumption-free.

4.2 A Information Theoretical Justification of the Esscher Premium Principle

As a side remark note that (4.8) is equivalent to a special case of the Esscher Transform

$$F(x,\gamma) = \Pr[X < x,\gamma] = \frac{\int_{-\infty}^{x} \mathrm{Exp}[\gamma \cdot y] \cdot dF(y,t)}{\int_{-\infty}^{\infty} \mathrm{Exp}[\gamma \cdot y] \cdot dF(y,t)} \tag{4.9}$$

where the parameter γ can be chosen to be consistent with a non-arbitrage condition similar to (4.6). As discussed by Gerber and Shiu (1994), resulting equivalent martingale probabilities can be used for option pricing. For discrete probabilities q_i as discussed here, the Esscher Transform becomes

$$F(r,\gamma) = \Pr[R < r,\gamma] = \frac{\sum_{r_i < r} \mathrm{Exp}\left[\gamma \frac{1+r_i}{1+r_{rf}}\right]}{\sum_{r_i} \mathrm{Exp}\left[\gamma \frac{1+r_i}{1+r_{rf}}\right]} \tag{4.10}$$

With $\gamma = \hat{\gamma}$, (4.10) is consistent with the non-arbitrage constraint and (4.8) is recovered. This shows the special case of the Esscher transform to be equivalent to the maximum entropy probability distribution. In this sense, the Esscher transformation (4.10), which lies at the heart of the well-known Esscher premium principle[4], has an information theoretical justification.

5 Including Costs and Investment Return

Costs and investments play a crucial role for the profitability of insurance operations. While costs enter premium calculations in a straightforward way, this is less obvious for investments. Pricing of contracts and assessing reserves depends directly on the anticipated investment return. Usually, investment opportunities are taken into account by a risk-adjusted discounting factor in (1.1) whose determination is left for actuarial judgment. From the market perspective, the difficulty lies in finding a market conforming discount factor for the combined risk of insurance and investment operation or, in other terms, the market price of the overall risk. A way to address this problem is to consider arbitrage-free pricing of insurance operations that reflects and incorporates investment possibilities.

The maximum entropy approach provides an elegant way to account for investment returns and cost simply by modifying or adding constraints. Fixed costs c modify the return related to insurance (risky asset b) according to

$$R = \frac{b - c - X}{b} \tag{5.1}$$

Further, assume that the insurer is investing the amount b during the period $[0,T]$ in which no payments are made in a risky asset with (stochastic) return Y. The overall return of the insurance operation then becomes

[4] The Esscher premium principle defines (as e.g. the Variance premium principle) a particular way of calculating a safety loading or risk premium. Interestingly enough, the Esscher premium principle relies on 'distorted' probabilities in much the same way as non-arbitrage pricing relies on equivalent martingale probabilities. Indeed, in our case here, the martingale probabilities can be seen as originating as a special case from the Esscher premium principle. Besides this, there exist other economically relevant links (e.g. to the framework of Pareto-optimal risk exchange, see Bühlmann (1980)). I thank Peter Blum for stressing this point.
The connection of the Esscher premium principle to information theory, non-arbitrage pricing, and Pareto-optimal risk exchange certainly warrants a more detailed study.

$$R^{tot} = \frac{b-c-X}{b} + Y$$

(5.2)

Note that the overall return R^{tot} of the insurance operation will reflect diversification effects that are present due to the non (or only partially) correlated nature of the returns R and Y.

Again, equivalent martingale probabilities are calculated as outlined in Section 2 and 3. The only differences are that the constraint (4.6) is modified to reflect the cost c and that a new non-arbitrage constraint

III':
$$E^Q\left[\frac{1+Y}{1+r_{rf}}\right] = \frac{1}{1+r_{rf}}\sum_i q_i(1+y_i) \,\&1$$

(5.3)

for the investment return is added. Because of linearity, $(1+R^{tot})/(1+r_{rf})$ is automatically a martingale, i.e. no modification for R^{tot} is needed. Maximizing the entropy in analogy to (4.7) leads to

$$q_i = \frac{\mathrm{Exp}\left[(1+r_{rf})^{-1}(\hat{\gamma}_1(1+r_i)+\hat{\gamma}_2(1+y_i))\right]}{\sum_i \mathrm{Exp}\left[(1+r_{rf})^{-1}(\hat{\gamma}_1(1+r_i)+\hat{\gamma}_2(1+y_i))\right]}$$

(5.4)

where the parameters γ_1 and γ_2 are determined by the constraints (4.6) and (5.3), respectively. The non-arbitrage value of an insurance contract is defined according to (3.7) again. Note that the discounting is done with respect to the risk-free interest rate r_{rf}, i.e. the dependence on the investment return is internalized.

Evaluating (3.7) with the equivalent martingale measure (5.4) will account for investment opportunities. Specifically, it will yield the non-arbitrage value of the combination of insurance and investment operation and, by this, the market price of the overall risk. Thus, the non-arbitrage pricing principle (3.7) together with the equivalent martingale measure (5.4) provides a unified valuation of assets and liabilities.

The equivalent martingale measure (5.4) will also reflect diversification effects present due to the non (or only partially) correlated nature of the returns R and Y. To see this, consider that (5.4) depends on the correlation of R and Y. How the correlation affects the non-arbitrage premium, however, is not easy to guess since it depends on the parameters γ_1 and γ_2 that depend on the correlation themselves. What can be inferred from (5.4) is that for negative, fixed $\hat{\gamma}_1$ and $\hat{\gamma}_2$ a positive correlation between R and Y results in relatively more weight being put on low returns. Vice versa, no correlation or negative correlation will decrease the relative weight on low returns. This feature of (5.4) can be interpreted in terms of a portfolio effect. We will come back to these issues in Section 7.

To end this Section, note that relation (5.4) can be interpreted as a generalization of the Esscher transformation (4.10). The transformed measure Q in (5.3) depends not on a single measure P but on two measures P_1 and P_2 related to insurance and investment returns. Relation (5.4) can easily be extended to cover more complex situation with additional investment (or insurance) opportunities (provided that the number of states of the world is bigger than the number of incorporated risk factors plus two).

6 Unique Valuation and Implied Discounting

In comparison to real probabilities P, the equivalent martingale probabilities Q will put more weight on low returns. This connects in a natural way to the insurance related concept of the safety loading. In Delbaen and Haezendonck (1989) it was shown that common premium and loading principles can be recovered by suitably choosing martingale probabilities. The case of a compound Poisson process was considered. The situation here is different in two respects: First, in addition to the insurance return, the stochastic investment return is considered. Second, we will be interested in the discounting factor that is implied by the overall non-arbitrage value of the contract and a particular loading. The motivation for this stems from the fact that according to standard finance, the non-arbitrage value coincidences with the (unique) market price. While this interpretation relies on liquidity, effectiveness and completeness of the market that don't fully apply in the current case, the implied discounting rate may nevertheless shed some light on how the discounting factor in (1.1) has to be chosen.

The implied discounting factor is obtained by solving the equation

$$\frac{1}{1+r_i}\mathrm{E}^P[f(L)]+S[f(L)]=\frac{1}{1+r_{rf}}\mathrm{E}^Q[f(L)] \tag{6.1}$$

for $(1+r_i)$ where $f(L)$ stands for the loss occurred by an insurance structure. It reconnects the actuarial and financial view. By definition, it is consistent with the chosen loading principle and the market value of the overall risk of insurance and investment operation as implied by non-arbitrage pricing.

Relation (6.1) defines the implied discounting factor $(1+r_i)$ in a unique way. One has to keep in mind however that the right hand side of equation (6.1) depends on the martingale measure Q, which is, due to the incompleteness of the insurance market, not uniquely defined. As well, Q is linked by construction to the reference premium b. Thus, the implied discounting factor $(1+r_i)$ is relative to the reference premium b as defined 'ground-up' by (1.1)

$$B=\frac{1}{1+r_i}E^P[X]+S[X]$$

Other definitions for b can be considered. And indeed, in a real life situation likely a reference premium would be considered that is referring to $E^P[f'(L)]$ and $S(f'(L))$, or simply to a premium of an already placed contract.

7 An Example: Simple Reinsurance Structures

An illustrative example of non-arbitrage pricing with methods discussed here is reinsurance. Sonderman (1991) has pioneered pricing of reinsurance contracts with martingales. Premiums are calculated under the assumption of an arbitrage-free reinsurance market and in the context of the continuous Lundberg model. Considering discounted cash flows based on stochastic interest rates links the insurance and finance side. However, there is no explicit non-arbitrage condition for investments in financial assets. By these assumptions, the model of Sondermann (1991) differs in two respects from the approach followed here. First, we developed a unified view of insurance and investment operations. There is no discounting of cash flows that creates the somehow artificial bridge between insurance and finance. Investments in financial and insurance-related risky assets are considered on an equal footage. Second, a discrete setting is investigated which reflects the reality of basic reinsurance policies that are renewable once a year. The quarterly traded catastrophe insurance futures traded at the Chicago Board of Trade give a comparable real-life example.

Two simple reinsurance structures will be considered here, one of which is related to compound claim amounts X (stop-loss insurance) and the other one to single claim amounts L_i (excess of loss insurance). Both refer to the same underlying asset b with stochastic one-period returns defined by (3.2).

7.1 Stop-Loss Reinsurance

In the case of stop-loss reinsurance, an insurer cedes the risk of being exposed to compound losses X surpassing some threshold d to a reinsurer. For simplicity, it is again assumed that no payments are made in the time period $[0,T]$. What is the non-arbitrage value of reinsurance in this situation and its relation to the premium b? The incurred loss, which at time T is ceded to the reinsurer, is

$$f(L,N) = f(X) = \max(X - d, 0) \tag{7.1}$$

where d is the deductible. Note that (7.1) is basically the payoff of a call option. On the other hand, the insurer incurs the loss

$$\min(X, d) \tag{7.2}$$

Taking the expectation value of (7.1) with respect to the equivalent martingale measure Q and discounting with the risk-free rate leads to the non-arbitrage reinsurance premium

$$b_{reins} = \frac{\mathrm{E}^{Q}[\max(X-d,0)]}{1+r_{rf}} \tag{7.3}$$

Because of the relation

$$\max(X-d,0) = X - \min(X,d) = X - (d - \max(d-X,0)) \tag{7.4}$$

the reinsurance premium b_{reins} can also be expressed as

$$b_{reins} = \frac{(b-d)}{1+r_{rf}} + \frac{\mathrm{E}^{Q}[\max(d-X,0)]}{1+r_{rf}} \tag{7.5}$$

The first term in (7.5) is the discounted value of a cash amount $(b-d)$ while the second term corresponds to the value of a 'put option on insured losses' with strike d. Thus, the value of reinsurance for the insurer equals to holding (or lending) the cash amount $(b-d)$ at the risk-free rate r_{rf} and selling a put option with strike d. In this context it is important to keep in mind that by the equivalent martingale measure Q expression (7.3) (or (7.5)) for the reinsurance premium b_{reins} implicitly takes into account the investment return Y which means that the equivalent martingale measure Q depends on both, the insurance return related to X and the investment return Y.

Relation (7.5) shows how the insurance premium b and reinsurance premium b_{reins} are interconnected when no arbitrage is present. Other relations can be deduced. An illustrative reminiscence to the well known 'call–put parity' in finance following from (7.4) is

$$b + \mathrm{E}^{Q}[\max(d - X, 0)] = d + \mathrm{E}^{Q}[\max(X - d, 0)] \tag{7.6}$$

With respect to the premium b and the deductible d, equation (6.6) interconnects the value of put and call options with equal strike (i.e. deductible) on compound losses X. Similar relations for more complicated reinsurance structures can be deduced.

7.2 Excess-of-Loss Reinsurance

Up to now, only a compound claim amount X has been considered. In the most common cases of insurance and reinsurance, however, the amount to be paid does depend on single claim amounts. As an example consider excess-of-loss reinsurance where the amount for the ith claim to be paid by the reinsurer is

$$\max(L_i - d, 0) \tag{7.7}$$

The total amount paid in the period $[0, T]$ is obtained by summing over i

$$f(L, N) = \sum_{i=1}^{N} \max(L_i - d, 0) \tag{7.8}$$

Here, N is the (random) number of claims occurring in the period $[0, T]$. Interpreting (7.8) as a payoff function $f(L, N)$ of the underlying asset b, the non-arbitrage premium is calculated by taking the expectation value of (7.9) with respect to the equivalent martingale measure Q

$$b_{reins} = \frac{\mathrm{E}^{Q}\left[\sum_{i=1}^{N} \max(L_i - d, 0)\right]}{1 + r_{rf}} \tag{7.9}$$

The expectation value $E^Q[\cdot]$ in expression (7.9) can be written (see e.g. Daykin, Pentikäienen and Pensonen (1994)) as

$$b_{reins} = \frac{E^Q[N] \cdot E^Q[\max(L_i - d, 0)]}{1 + r_{rf}} \tag{7.10}$$

and, in analogy to Section 7.1, an equivalent expression

$$b_{reins} = \frac{(b - E^Q[N] \cdot d)}{1 + r_{rf}} + \frac{E^Q[N] \cdot E^Q[\max(d - L, 0)]}{1 + r_{rf}} \tag{7.11}$$

can be obtained. In the same manner, relation (7.6) now becomes

$$E^Q[\max(L - d, 0)] = \left(\frac{b}{E^Q[N]} - d \right) + E^Q[\max(d - L, 0)] \tag{7.12}$$

The interpretation is similar to the one in the case of compound losses. With respect to the premium b, the deductible d and the expected value of number of claims evaluated under the equivalent martingale measure Q, equation (7.12) interconnects the value of put and call options with equal strike (i.e. deductible) on individual losses L_i.

7.3 Numerical Examples

For illustrative purposes, this section presents numerical results for the reinsurance examples of Section 7.1 and 7.2. Arbitrage-free reinsurance premiums are calculated as a function of the deductible d. Particular attention is paid to the role of investment possibilities.

We assume that the investment return Y over the period T is a normally distributed with mean μ_I and standard deviation σ_I. The dependence of non-arbitrage premiums on investment opportunities is investigated by comparing non-arbitrage premiums for different parameter values μ_I and σ_I. The numerical examples are based on loss amounts generated from Poisson distributed claim numbers N and lognormal distributed claim amounts L. The corresponding parameter values are given in Table 1, together with the characterization of the investment return Y. Investment returns Y for different parameter values μ_I and σ_I have been obtained by scaling and shifting. The numerical results are based on 50 scenarios (i.e. states of the world) that correspond to about 200 single losses[5]. These numbers are sufficient to sample a large part of the compound loss distribution X.

The compound loss distribution is shown in Figure 1 together with an example of equivalent martingale probabilities. The equivalent martingale probabilities are shown for parameter values μ_I =6% and σ_I = 5%. Figure 2 and Figure 3 present non-arbitrage premiums for stop-loss reinsurance as a function of the deductible d and various combinations of μ_I and σ_I. Corresponding non-arbitrage premiums for excess-of-loss reinsurance are shown in Figure 4 and Figure 5, again as a function of the deductible d. This time however the deductible d refers to single claim amounts. The Figures show generic features: By construction of the equivalent martingale probabilities, all non-arbitrage premiums coincide for d =0. This reflects the fact that as discussed in Section 6 non-arbitrage premiums are defined relative to a given premium b. It should be noted that the particular choice d =0 is arbitrary. In principle, corresponding results for any reference premium b' can be derived. For $d \neq 0$ the non-arbitrage premiums reflect the investment characteristics: For a fixed volatility σ_I the higher the average investment return μ_I is, the higher is the non-arbitrage premium, i.e. the arbitrage-free value of the contract. For a fixed average return μ_I and varying volatility σ_I, a reversed situation is encountered: The higher the volatility σ_I, the lower is the non-arbitrage premium. These relations are the same

[5] The investment return time-series and the loss scenarios correspond to the typical data that is needed in real life examples. Historical data can be used in a straightforward way.

independently of the nature of the reinsurance contract (i.e. excess-of-loss or stop-loss) considered here.

Figure 2 to Figure 5 in addition show the expected loss $E^P[f(X)]$ as a function of the deductible d and sample insurance premiums calculated from the variance premium principle

$$\text{premium} = \frac{1}{1+r_{rf}} E^P[f(X)] + S[f(X)] = \frac{1}{1+r_{rf}} E^P[f(X)] + \alpha \cdot Var[f(X)]$$

with α =0.015. By equating the insurance premium with the non-arbitrage premium, the implied discounting factor $(1+r_i)$ can be deduced. Implied discount factors corresponding to Figure 5 are plotted in Figure 6. The arbitrage-free premiums (and the related implied discounting factor $(1+r_i)$) have to be understood as originating from the combined insurance and investment operations. This explains the somewhat counterintuitive fact the higher investment returns correspond to lower implied discounting rates. In the present one-period case, these implied discount factors have little meaning besides reconnecting the actuarial and financial premium calculation.

The example considered here is one in which the insurance premium is in general higher than the non-arbitrage premium. Obviously, depending on the loading $S[f(X)]$ and other parameters, the situation could be reversed. In such a situation, the question of how investment related value could be transferred back to the insured by lowering the premium becomes important. A discussion of these issues lies beyond the scope of this paper. However we note that a consistent way of valuing the passing of investment generated value to the primary insurer should be based on the equivalent martingale measure Q, i.e. on an arbitrage-free valuation.

Finally, Figure 7 demonstrates the presence of a portfolio effect. It shows excess-of-loss non-arbitrage premiums for different correlations between insurance and investment return. The different cases have been obtained by considering insurance and investment returns with rank correlation 1, approximately, 0 and –1, respectively. These cases thus represent the extremes. The graphs show a pronounced and expected dependence on the correlation between insurance and investment return.

8 Multi-Period Contracts

The one-period setting considered here is easily generalized to the multi-period case. Basically, the constraints (3.6) and (5.3) have to capture multi-period payoffs. As well, the discounting with respect to the risk-free interest rate r_{rf} should account for multiple periods. In the simplest case where no term structure is present, constraint (3.6) e.g. becomes

$$E^Q\left[\sum_t \frac{1+R_t}{(1+r_{rf})^t}\right] = 1$$

Considering multiple periods will allow for the valuation of long-term contracts, which often come with complex option-like structures (e.g. commutation features, guarantees). Recently, with the pressure from under-performing financial markets, issues related to the valuation of options became especially important in life insurance.

9 Conclusions

The relevance of arbitrage-free pricing relies on the liquidity and efficiency of the insurance market. Liquidity and efficiency are not features of the insurance market in general and thus the non-arbitrage approach to pricing of insurance contracts may be inadequate. The situation is different when insurance risk is traded or transferred to the financial markets and vice versa. In this case, the implicit assumptions underlying the non-arbitrage approach may be satisfied in that a (at least partial) hedge can be set up. Even for traditional insurance structures the financial risk due to investments can become comparable to the insurance risk itself (e.g. life insurance). In these cases the price for the overall risk should reflect in some way the price financial markets are putting on risk. In this paper we tried to tackle this problem by extending the non-arbitrage principle to the insurance side. Due to incompleteness of the insurance market, no unique equivalent martingale measure exists. The information theoretical maximum entropy principle is applied to make a sensible choice in this situation.

Equivalent martingale probabilities consistent with the combined non-arbitrage conditions for insurance and investment operations for a one-period horizon are calculated. They turn out to be linked to a generalized form of the Esscher measure and show correlation dependence. The latter illustrates the presence of a portfolio effect. These findings are illustrated by a numerical example referring to common reinsurance-like structures.

In summary, a practicable and well-motivated way to infer overall market prices for combined insurance and investment operations based on the non-arbitrage pricing principle has been outlined. While the relevance of this approach may be limited for traditional insurance structures with only limited exposure to financial risk, the market price is clearly of relevance for insurance structures whose performances depend heavily on the performance of financial markets.

Acknowledgements

I thank Peter Blum (Converium), Victor Choi and Sergei Esipov (Centre Solutions) for valuable comments and stimulating discussions. Part of this work has been performed at Centre Solutions. The views expressed here are those of the author and not Converiums' or Centre Solutions'. All remaining errors are the authors' responsibility.

References

Black, F. and M. Scholes (1973), The Pricing of Options and Corporate Liabilities, *Journal of Political Economy*, Vol. 81, pp. 637-659.

Buchen, P., and M. Kelly (1996), The maximum entropy distribution of an asset inferred from option prices, Journal of Financial and Quantitative Analysis, Vol. 31, pp. 143 – 159.

Bühlmann, H. (1980), An Economic Premium Principle, *ASTIN Bulletin*, Vol. 11, 1980, pp. 52-60.

Copeland, T. E. and Weston, J. F. (1992), *Financial Theory and Corporate Policy*, Addison Wesley.

Cover, T. M. and Thomas, J. A. (1991), *Elements of Information Theory*, John Wiley.

Daykin, C.D., Pentikainen, T. and Pesonen, M. (1994), *Practical Risk Theory for Actuaries.*, Chapman & Hall: London, New York.

Delbaen, F. and Haezendonck, J. M., (1989), A martingale approach to premium calculation principles in an arbitrage-free market, *Insurance: Mathematics and Economics,* Vol. 8, 269-277.

Gerber, H. U. and Shiu, S. W. (1994), Martingale Approach to Pricing Perpetual American Options, *ASTIN Bulletin*, Vol. 24, 1994, pp. 195-220.

Gerber, H. U. and Shiu, S. W. (1994), Option Pricing by Esscher Transforms, *TSA*, Vol. 16, pp. 99-140.

Golan, A., G. G. Judge, and D. Miller (1997), The Maximum Entropy Approach To Estimation and Inference: An Overview, *Advances in Econometrics*, Vol. 12, pp. 3-24.

Gulko, L. (1997), Dart Board and Asset Prices: Introducing the Entropy Pricing Theory, *Advances in Econometrics*, Vol. 12, pp. 237-276.

Harrison, J. M., and Kreps, D. M., 1979, Martingales and Arbitrage in Multiperiod Securities Markets, *Journal of Economic Theory,* Vol. 20, 381-408.

Hull, J.C. (2000), *Options, Futures and other Derivatives*, Prentice Hall.

Jaynes, E. T. (1957a), Information Theory and Statistical Mechanics, *Physics Review*, Vol. 106, pp. 620-630.

Jaynes, E. T. (1957b), Information Theory and Statistical Mechanics II, *Physics Review*, Vol. 108, pp. 171-190.

Jarvis S., Southall F. and Varnell E. (2001), Modern Valuation Techniques, available at www.sias.org.uk/papers/mvt.pdf

Neftci, S. (2000), *Mathematics of Financial Derivatives*, Academic Press.

Schmidli, H., 1996, Martingales and insurance risk. In: Obretenov A. (ed.) *Lecture notes of the 8-th international summer school on probability and mathematical statistics* (Varna). Science Culture Technology Publishing, Singapore, 155-188.

Shannon, C. E., 1948, A Mathematical Theory of Information, *The Bell System Technical Journal,* 27, 379-423 and 623-656.

Stutzer, M. (1996), A Simple Nonparametric Approach to Derivative Security Valuation, *Journal of Finance*, Vol. 51, pp. 1633-1652.

Sondermann D., 1991, Reinsurance in arbitrage-free markets. *Insurance: Mathematics and Economics*, Vol. 10, 191-202.

Rubinstein, M. (1994), Implied Binominal Trees, *Journal of Finance*, Vol. 49, pp. 771-818.

	Loss Number	Loss Severity	Invest. Return	Risk-free Rate
Type	Poisson	LogNormal	Normal	Constant
μ	4.0	1.0	4%, 5%, 6%	3%
σ	-	8.0	3%,4%,5%,7%	

Table 1 Distribution types and parameter values referring to numerical examples discussed in Section 7.

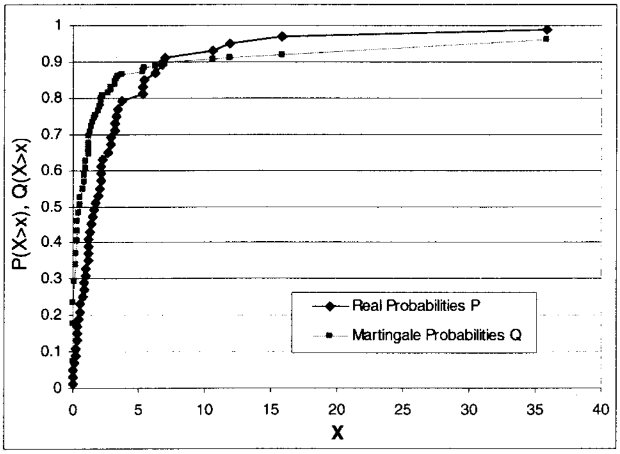


Figure 1 Real probabilities P and corresponding equivalent martingale probabilities Q referring to an investment return μ_I =6% with volatility σ_I =5% (see text).

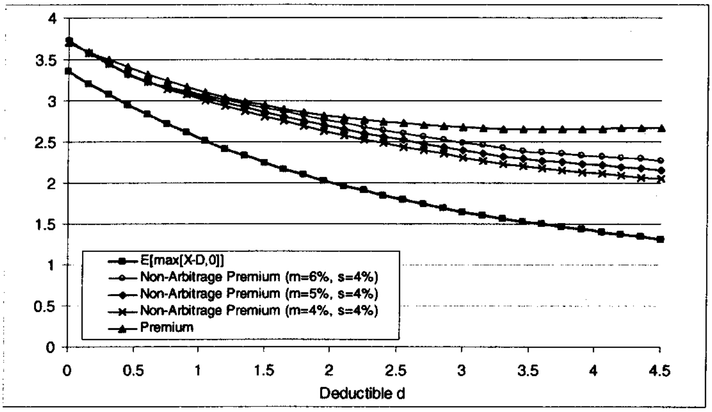


Figure 2 Stop-loss structure. Expected loss, premium and non-arbitrage premiums for different investment opportunities as a function of the deductible *d*.

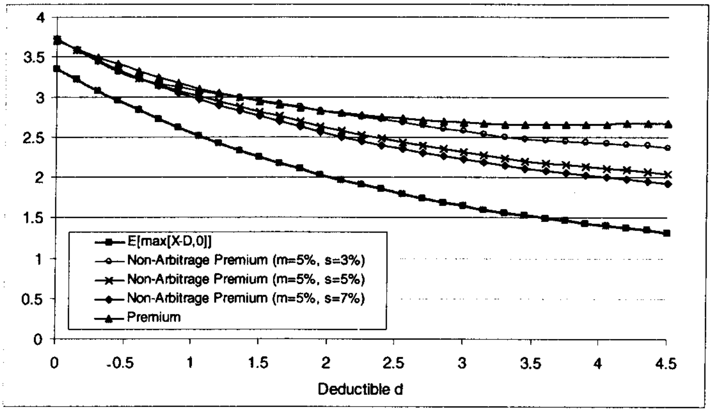


Figure 3 Stop-loss structure. Expected loss, premium and non-arbitrage premiums for different investment opportunities as a function of the deductible *d*.

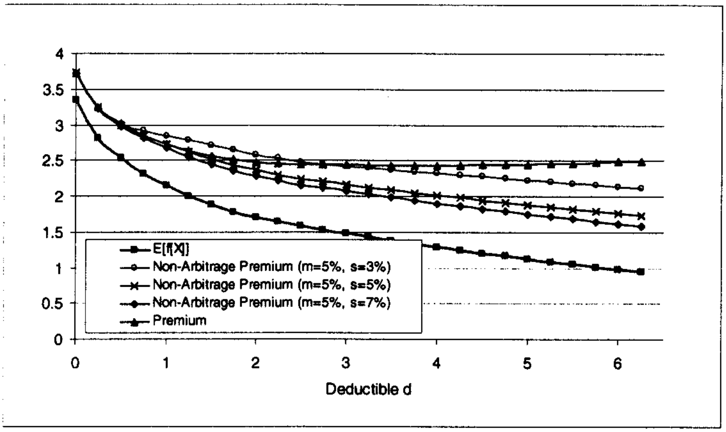


Figure 4 Excess-of-loss structure. Expected loss, premium and non-arbitrage premiums for different investment opportunities as a function of the deductible *d*.

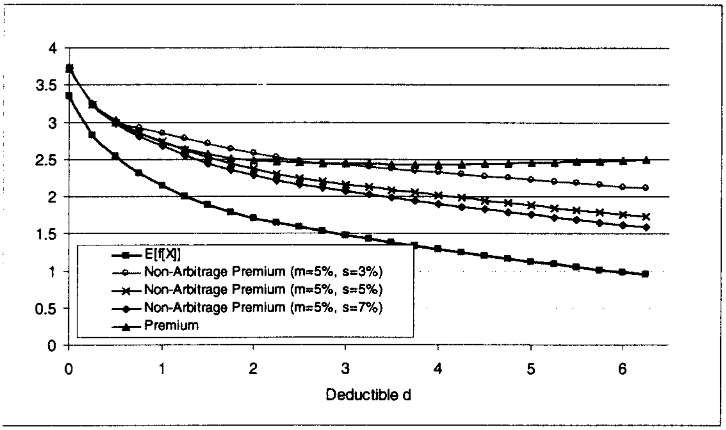


Figure 5 Excess-of-loss structure. Expected loss, premium and non-arbitrage premiums for different investment opportunities as a function of the deductible *d*.

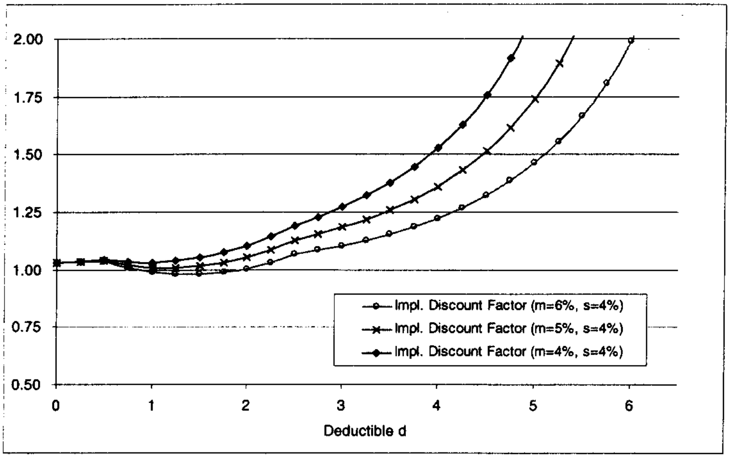


Figure 6 Implied discount factors (corresponding to Figure 5) as a function of the deductible *d*.

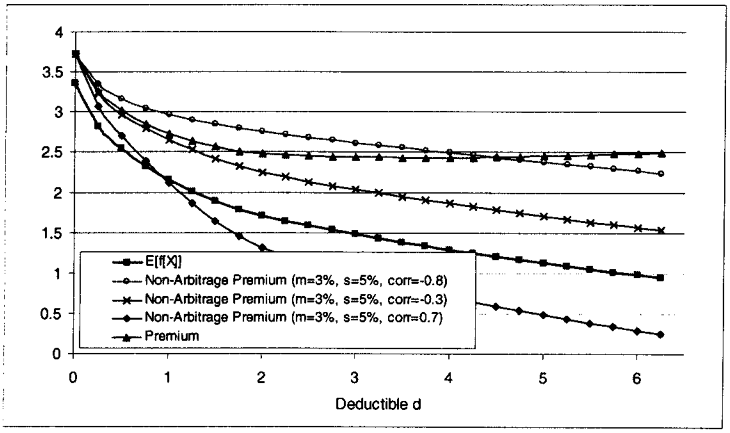


Figure 7 Effect of correlation between insurance and investment returns on non-arbitrage premiums for the excess-of loss structure.

Capital Consumption: An Alternative Methodology for Pricing Reinsurance

Donald F. Mango, FCAS, MAAA

Capital Consumption: An Alternative Methodology for Pricing Reinsurance

Donald Mango

Acknowledgements
I would like to thank Daniel Heyer, Eric Lemieux, Gene Chairmonte, Dave Sommer, Frank Schnapp, Leigh Halliwell, Dave Clark, Jim Sandor, Glenn Meyers, Lee Van Slyke, and the members of the CAS Committee on Theory of Risk for their assistance in reviewing earlier drafts and challenging the ideas. Such interaction is critical to research, as it forced me to refine and clarify my arguments, all of which improved the end product.

1. Introduction

This paper introduces a *capital consumption* methodology for the price evaluation of reinsurance in a stochastic environment. It differs from the common practice of risk-based capital allocation and release by: (i) evaluating the actual contract cash flows at the scenario level; (ii) eliminating the need for contract-level supporting capital allocation and release; (iii) evaluating each scenario's operating deficits as contingent capital calls on the company capital pool; and (iv) reflecting the expected cost of contingent capital calls as an expense load.

This method eliminates the need for capital allocation and release; creates scenarios that more closely model actual contract capital usage; allows more flexibility in stochastic modeling; and makes risk-return preferences an explicit part of the pricing decision.

Section 2 begins with an overview of the capital consumption approach, framing the major differences from capital allocation. Section 3 then presents the details of the approach. Section 4 delves further into the concept of contingent capital consumption and its costs. Section 5 shows examples of price evaluation using this approach. Section 6 concludes with linkages to other current research efforts. Appendix A addresses the question: Does insurance capital allocation make sense? Appendix B demonstrates one approach for calibrating to the portfolio level.

2. Capital Consumption Overview

This paper challenges many fundamental conceptual underpinnings of reinsurance pricing. Any attempt at an overview will be difficult. As a start, we will outline the major differences in treatment of capital under an allocation versus consumption framework by considering four questions:

1. What happens to the total capital?
2. How are the segments evaluated?
3. What does being in a portfolio mean?
4. How is relative risk contribution reflected?

Question 1: What happens to the total capital?	
Allocation	**Consumption**
• Divided up among the segments. • Either by explicit allocation, or assignment of the marginal change in the total capital requirement from adding the segment to the remaining portfolio	• Left intact • Each segment has the right to "call" upon the total capital to pay its operating deficits or shortfalls

Allocation splits up the total capital and doles it out to segments. Two critical assumptions underlie this approach: that the capital itself is divisible; and that, similar to manufactured products, insurance products require up-front capital investment to produce. Consumption instead recognizes the right (widely acknowledged among capital allocation proponents) of any contract to consume potentially all the company's capital.

Question 2: How are the segments evaluated?	
Allocation	**Consumption**
• Give the allocations to each segment • Evaluate each segment's return on their allocated capital • Must clear their hurdle rate	• Give each segment "access rights" to the entire capital • Evaluate each segment's potential calls (both likelihood and magnitude) on the total capital • Must pay for the likelihood and magnitude of their potential calls

Allocation proponents ask, without capital allocation, how can you make either performance evaluations or investment decisions? How can you decide where to grow or shrink your book? How can you divide a bonus pool? They also advocate this as a translation vehicle to results from other industries.

Consumption is a valid alternative for either performance evaluation or portfolio composition decision-making. Both approaches are based upon the premise that riskier segments must pay for their risk. Both approaches are also dependent upon a sound portfolio risk model, the true foundation of stochastic reinsurance pricing.

Question 3: What does being in a portfolio mean?	
Allocation	**Consumption**
• Being standalone with less capital • But still having access to all the capital if necessary, although it is unclear how this is reflected	• Being standalone with potential access to all the capital • But all other segments have similar access rights

This is **the** critical difference. Allocation treats segments as if standalone, with less capital. This means being in a portfolio is like being on your own, but you have to support less capital. Consumption on the other hand treats being in a portfolio like being standalone, with access to potentially all the capital, but with the added wrinkle that all the other segments have similar access rights.

Question 4: How is relative risk contribution reflected?	
Allocation	**Consumption**
• Use a single risk measure to determine required capital • Select a dependence structure for the aggregation of segment distributions into a portfolio aggregate distribution • The marginal impact of adding a segment to the remaining portfolio is that segment's risk contribution	• Use scenario-level detail generated by stochastic modeling • Use explicit risk-return evaluation via utility function • Segment's risk contribution is determined at the scenario level, then aggregated over all scenarios

This is a deep question, one that will be covered in extensive detail in the remainder of the paper. The essential point: once you move to the modeled scenario level, capital allocation becomes increasingly difficult to meaningfully interpret. Allocated capital is determined based on a risk measure of the distribution in total. For any given scenario, though, this overall amount is never the actual required amount — namely, the modeled operating deficit. The allocated capital is excessive for favorable scenarios, and grossly inadequate in severe loss scenarios (unless the capital equals the policy limit)[1].

3. Details of the Capital Consumption Approach

We will demonstrate the capital consumption approach within a stochastic contract analysis framework. We will cover the three major differences from typical risk-based capital allocation approaches: (i) analyzing the contract outcome at the scenario level, (ii) discounting at a default-free rate, and (iii) calculating the contract's capital consumption within each scenario.

Scenario Analysis

The first modification requires maintaining the scenario detail, and analyzing the contract's outcomes at the scenario level. Stochastic modeling is basically scenario analysis extended to a high level of granularity. Modeling thousands of points of a contract outcome distribution means generating thousands of scenarios. This extension to scenario detail may appear trivial. If the functions are linear, or the distributions symmetric, no benefit will be gained by expanding the detail. Expected values are sufficient for decision-making. Jensen's inequality[2] becomes Jensen's "equality" in these conditions:

$$g(E[x]) = E[g(x)]$$

However, reinsurance contracts have non-linear contract features such as aggregate deductibles, caps, corridors, and co-participations. They also have extremely skewed distributions. In such conditions, we must evaluate $\boldsymbol{E[g(x)]}$ to get an accurate result. Evaluating each point of the distribution requires maintenance and use of the scenario detail.

Default-Free Discounting

The second modification involves discounting cash flows at a default-free rate. Scenario analysis (indeed, simulation modeling) is built upon the premise that possible, realizable, plausible outcomes can be generated and analyzed. For the entire process to work, each generated scenario is "**conditionally certain**": given the scenario occurs, its outcome is certain. Where it is not, the entire practice of

---

[1] This problem is particularly striking in the evaluation of catastrophe reinsurance contracts, with small probabilities of a full contract limit loss. A "risk-based" capital amount might be some small fraction of the limit—say 5%. What sense does this capital amount make in the limit loss scenario? We held 5% of the limit as capital? And then how exactly did we fund the remaining loss amount? It came from company capital in total. The alternative—holding the full limit as capital—puts an unrealistic return burden on the contract. Current market price levels would likely make the contract look unattractive.

[2] Jensen's inequality states that for a cumulative probability distribution F(x) and a convex function g(x), E[g(x)] >= g(E[x]). There are countless references on this—e.g., Heyer [7], p. 98.

simulation modeling would be undermined by "meta-uncertainty." The scenarios themselves must withstand the scrutiny of a reality check.

Uncertainty for the contract in total is represented in the distribution across all modeled scenarios, and the probability weights assigned to those scenarios. In other words, *uncertainty is reflected between scenarios, not within them.* Given conditional certainty, scenario cash flows can be discounted at a default-free rate (or a simulated path of default-free forward rates).

Scenario Capital Consumption

The final, and perhaps most controversial, change involves the treatment of capital. Most methodologies focus on up-front allocation of supporting risk-based capital, and its release over time. The capital actually consumed (if any) by each modeled scenario of the contract is the focus here.

Capital is still required at the company level and still needs to be invested in an insurance company. However, it plays a fundamentally different role in an insurer than in a manufacturer. Capital investment in manufacturing is typically up-front, in equipment and raw materials. In contrast, insurance "products" are promises to pay contingent on valid claims. Thus the costs of insurance products are claim-related payments. They are not specific investments, and occur (if at all) in the future.

Insurers receive revenue in the form of premium that includes an estimated provision for their expected costs, plus some volatility loading. Insurance capital acts more like a "claims paying reservoir," an overall buffer for unpredictability and volatility of aggregated product results. This reservoir is subject to unpredictable future inflows and outflows. What has been termed an "allocation" of capital for underwriting new contracts is more like the granting of additional rights to draw upon future capital. The critical issue is, therefore, both the likelihood and magnitude of exposure of capital to possible consumption by contracts.

The cost of maintaining the capital reservoir is an overall cost of business — an overhead expense. This approach essentially assesses contracts for this overhead expense in a "risk-based" manner. The bases for the assessment are likelihood and magnitude of capital reservoir drawdowns.

Maintaining the scenario detail, and recognizing that scenarios are conditionally certain, we can evaluate the capital amounts *actually consumed* by each scenario. A contract can "pay its own way" if its total revenues exceed its total costs. Total revenues include premium and investment income on its own flows. Total costs means expenses and losses. Company capital is needed when the contract runs an **operating deficit** — when its costs exceed its revenues. Philbrick and Painter make this point well:

"When an insurance company writes a policy, a premium is received. A portion of this policy can be viewed as the loss component. When a particular policy incurs a loss, the company can look to three places to pay the loss. The first place is the loss component (together with the investment income earned) of the policy itself. In many cases, this will not be sufficient to pay the loss. The second source is unused loss components of other policies. In most cases, these two sources will be sufficient to pay the losses. In some years, it will not, and the company will have to look to a third source, the surplus, to pay the losses." [16, p. 124]

To evaluate scenario-level capital consumption and operating deficit, we look at the contract's **experience fund**. An experience fund is a concept from finite risk reinsurance. It is an account containing available revenue (premium net of expenses) plus investment income earned on the fund balance (at an assumed investment rate). All subject losses are paid from the fund. An experience fund allows us to calculate the contract's "terminal value" or cumulative operating result. Consider Example 1, the experience fund of a realistic long-tailed contract.

Example 1
Experience Fund for Long-tailed Contract
120% Loss Ratio Scenario

Probability 10.0%

Investment Rate 8.0% Loss Ratio 120.0% Ultimate Loss 120,000

1	2	3	4	5	6	7	8	9
Time	Beginning Fund Balance	Premiums	Expenses	Payment Pattern	Paid Losses	Investment Income	Ending Fund Balance	Capital Call
0	$ -	$ 100,000	$ 15,000	0.0%	$ -	$ -	$ 85,000	$ -
1	$ 85,000	$ -	$ -	50.0%	$ 60,000	$ 2,000	$ 27,000	$ -
2	$ 27,000	$ -	$ -	25.0%	$ 30,000	$ -	$ (3,000)	$ 3,000
3	$ -	$ -	$ -	12.0%	$ 14,400	$ -	$ (14,400)	$ 14,400
4	$ -	$ -	$ -	6.0%	$ 7,200	$ -	$ (7,200)	$ 7,200
5	$ -	$ -	$ -	4.0%	$ 4,800	$ -	$ (4,800)	$ 4,800
6	$ -	$ -	$ -	2.0%	$ 2,400	$ -	$ (2,400)	$ 2,400
7	$ -	$ -	$ -	1.0%	$ 1,200	$ -	$ (1,200)	$ 1,200
8	$ -	$ -	$ -	0.0%	$ -	$ -	$ -	$ -
9	$ -	$ -	$ -	0.0%	$ -	$ -	$ -	$ -
10	$ -	$ -	$ -	0.0%	$ -	$ -	$ -	$ -
TOTAL		$ 100,000	$ 15,000	100.0%	$ 120,000		$ -	$ 33,000
NPV		$ 100,000	$ 15,000	86.2%	$ 103,479			$ 24,775

Each column is explained in detail:

- Column 1 is time from inception of the contract in years.
- Column 2 is the fund balance at the beginning of each year.
- Column 3 is the premium flow into the fund.
- Column 4 is the expense flow out of the fund.
- Column 5 is the expected payment pattern as a percentage of ultimate loss. Ultimate loss is expressed as a ratio to premium. In this case, it is 120%.
- Column 6 is the product of ultimate losses $120,000 and the pattern in Column 5.

- Column 7 is the investment income earned on the end-of-year fund balance less all payments during the year, assuming those payments are made at the end of the year. This is assumed to go back into the fund for the next year. This could be adjusted to the midpoint of the year if desired.
- Column 8 is the end-of-year fund balance. It equals (2) + (3) – (4) – (6) + (7).
- Column 9 shows the **capital calls**.

Once Column 8 falls below zero, the contract is in an operating deficit position: the fund is empty, yet loss payments must be made. In order to make the payments, a capital call is made for the amount needed to make the required loss payment. Once the fund hits zero, it never rises above it again. Capital is only provided as needed to make the loss payments. Thus the contract makes what amount to a **series of capital calls** stretching into the future.

Time Profile of Capital Consumption

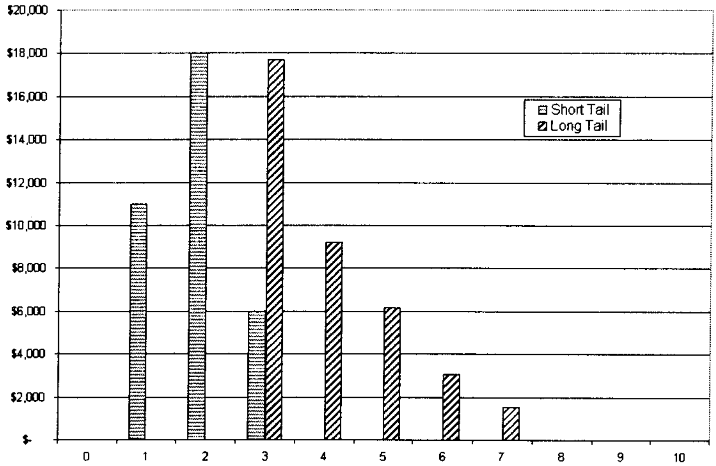
Compare Example 1A, which shows the capital calls for the contract in Example 1 with everything identical except a quicker payment pattern — a shorter tail.

Example 1A
Experience Fund for Short-tailed Contract
120% Loss Ratio Scenario

	Investment Rate		8.0%	Loss Ratio	120.0%	Ultimate Loss	120,000	
1	2	3	4	5	6	7	8	9
Time	Beginning Fund Balance	Premiums	Expenses	Payment Pattern	Paid Losses	Investment Income	Ending Fund Balance	Capital Call
0	$ -	$ 100,000	$ 15,000	0.0%	$ -	$ -	$ 85,000	$ -
1	$ 85,000	$ -	$ -	80.0%	$ 96,000	$ -	$ (11,000)	$ 11,000
2	$ -	$ -	$ -	15.0%	$ 18,000	$ -	$ (18,000)	$ 18,000
3	$ -	$ -	$ -	5.0%	$ 6,000	$ -	$ (6,000)	$ 6,000
4	$ -	$ -	$ -	0.0%	$ -	$ -	$ -	$ -
5	$ -	$ -	$ -	0.0%	$ -	$ -	$ -	$ -
6	$ -	$ -	$ -	0.0%	$ -	$ -	$ -	$ -
7	$ -	$ -	$ -	0.0%	$ -	$ -	$ -	$ -
8	$ -	$ -	$ -	0.0%	$ -	$ -	$ -	$ -
9	$ -	$ -	$ -	0.0%	$ -	$ -	$ -	$ -
10	$ -	$ -	$ -	0.0%	$ -	$ -	$ -	$ -
TOTAL		$ 100,000	$ 15,000	100.0%	$ 120,000			$ 35,000
NPV		$ 100,000	$ 15,000	90.9%	$ 109,084			$ 30,380

Because of the reduced investment income and shorter tail, the capital calls are larger ($35,000 vs $33,000) and sooner. Chart 1 shows the time profile comparison:

Chart 1: Capital Consumption Profile Over Time
Short versus Long Tail with 120% Loss Ratio



This chart visually depicts a major difference between short and long tail contracts that has yet to be fully understood and integrated into evaluation frameworks. Clearly the capital concept needs a significant extension over *time*. One can envision measures of concentration expanding from scalars to *vectors*, indexed into the future. For example, the impact on the company's future cash position could be reflected in the pricing and underwriting decision for a contract. A contract may have an attractive upside, but may have undesirable structural relationships to other portions of the portfolio (e.g., with respect to a large return premium or reserve adjustment). This concept will be elaborated on in future papers.

Reduced Operating Deficit

Under the 120% Loss Ratio scenario, the Long-tailed contract calls for a total of $33,000 in capital over time. If the loss ratio under another scenario were lower — say 100% — the contract would make smaller capital calls, since its operating deficit would be smaller. Consider Example 2:

Example 2
Experience Fund for Long-tailed Contract
100% Loss Ratio Scenario

		Probability	30.0%
Investment Rate	8.0%	Loss Ratio	100.0%
		Ultimate Loss	100,000

1 Time	2 Beginning Fund Balance	3 Premiums	4 Expenses	5 Payment Pattern	6 Paid Losses	7 Investment Income	8 Ending Fund Balance	9 Capital Call
0	$ -	$ 100,000	$ 15,000	0.0%	$ -	$ -	$ 85,000	$ -
1	$ 85,000	$ -	$ -	50.0%	$ 50,000	$ 2,800	$ 37,800	$ -
2	$ 37,800	$ -	$ -	25.0%	$ 25,000	$ 1,024	$ 13,824	$ -
3	$ 13,824	$ -	$ -	12.0%	$ 12,000	$ 146	$ 1,970	$ -
4	$ 1,970	$ -	$ -	6.0%	$ 6,000	$ -	$ (4,030)	$ 4,030
5	$ -	$ -	$ -	4.0%	$ 4,000	$ -	$ (4,000)	$ 4,000
6	$ -	$ -	$ -	2.0%	$ 2,000	$ -	$ (2,000)	$ 2,000
7	$ -	$ -	$ -	1.0%	$ 1,000	$ -	$ (1,000)	$ 1,000
8	$ -	$ -	$ -	0.0%	$ -	$ -	$ -	$ -
9	$ -	$ -	$ -	0.0%	$ -	$ -	$ -	$ -
10	$ -	$ -	$ -	0.0%	$ -	$ -	$ -	$ -
TOTAL		$ 100,000	$ 15,000	100.0%	$ 100,000		$ -	$ 11,030
NPV		$ 100,000	$ 15,000	86.2%	$ 86,232			$ 7,528

Now it only asks for $11,030. If the loss ratio were low enough, the contract would make no capital calls, as in Example 3:

Example 3
Experience Fund for Long-tailed Contract
80% Loss Ratio Scenario

		Probability	60.0%
Investment Rate	8.0%	Loss Ratio	80.0%
		Ultimate Loss	80,000

1 Time	2 Beginning Fund Balance	3 Premiums	4 Expenses	5 Payment Pattern	6 Paid Losses	7 Investment Income	8 Ending Fund Balance	9 Capital Call
0	$ -	$ 100,000	$ 15,000	0.0%	$ -	$ -	$ 85,000	$ -
1	$ 85,000	$ -	$ -	50.0%	$ 40,000	$ 3,600	$ 48,600	$ -
2	$ 48,600	$ -	$ -	25.0%	$ 20,000	$ 2,288	$ 30,888	$ -
3	$ 30,888	$ -	$ -	12.0%	$ 9,600	$ 1,703	$ 22,991	$ -
4	$ 22,991	$ -	$ -	6.0%	$ 4,800	$ 1,455	$ 19,646	$ -
5	$ 19,646	$ -	$ -	4.0%	$ 3,200	$ 1,316	$ 17,762	$ -
6	$ 17,762	$ -	$ -	2.0%	$ 1,600	$ 1,293	$ 17,455	$ -
7	$ 17,455	$ -	$ -	1.0%	$ 800	$ 1,332	$ 17,987	$ -
8	$ 17,987	$ -	$ -	0.0%	$ -	$ 1,439	$ 19,426	$ -
9	$ 19,426	$ -	$ -	0.0%	$ -	$ 1,554	$ 20,980	$ -
10	$ 20,980	$ -	$ -	0.0%	$ -	$ 1,678	$ 22,659	$ -
TOTAL		$ 100,000	$ 15,000	100.0%	$ 80,000		$ 22,659	$ -
NPV		$ 100,000	$ 15,000	86.2%	$ 68,986			$ -

The experience fund and capital calls give us the analytic framework. Now we consider *explicit valuation* of the calls, on our way to price determination.

4. Valuation of Contingent Capital Calls

The *concept* of contingent capital calls was presented in Philbrick and Painter (**emphasis mine**):

> "The entire surplus is available to every policy to pay losses in excess of the aggregate loss component. Some policies are more likely to create this need than others are, even if the expected loss portions are equal. Roughly speaking, for policies with similar expected losses, we would expect the policies with a large variability of possible results to require more contributions from surplus to pay the losses. We can envision an insurance company instituting a **charge for the access to the surplus. This charge should depend, not just on the likelihood that surplus might be needed, but on the amount of such a surplus call.**" [17, p. 124]

They continue (my inline comments are [***bold and italicized***]):

> "We can think of a capital allocation method as determining a charge to each line of business that is dependant on the need to access the surplus account [***contingent capital***]. Conceptually, we might want to allocate a specific cost to each line for the right to access the surplus account [***call***]. In practice though, we tend to express it by allocating a portion of surplus to the line, and then requiring that the line earn (on average) an adequate return on surplus. Lines with more of a need for surplus will have a larger portion allocated to them, and hence will have to charge more to the customers to earn an adequate rate of return on the surplus. Effectively, this will create a charge to each line for its fair share of the overall cost of capital." [17, p. 124]

Thus, Philbrick and Painter would like to charge a line of business for the right to access the surplus account — i.e., to make capital calls. However, they are still thinking in terms of a supporting capital allocation framework, within which no such concept exists. In contrast, the capital consumption approach is built around such a charge.

What exactly would such a charge mean? It could have several meanings simultaneously:

1. A risk-based overhead expense loading, though it may not be a payment to an outside entity.
2. A decision variable that influences the attractiveness of certain product types.
3. An explicit expression of the company's risk-return preferences by application of concepts from utility theory.

Each is considered in detail:

1. Risk-based Overhead Expense Loading
As requested by Philbrick and Painter, the charge is based on the magnitude and likelihood of calls upon the common capital pool. Since capital is at the company level only, and available to all contracts, the cost of that capital should be an overhead expense — like rent. However, unlike rent, the cost may not correspond to actual payments being made by the company. However, it is a cost of doing business, in that without capital in total to support the portfolio and guarantee a certain level of perceived claims-paying ability, the contracts could not be sold. All contracts partake of the benefits of the capital pool and, therefore, must be assessed some share of its maintenance cost.

2. Pricing Decision Variable
In order to make informed pricing decisions, product costs must be accurately and objectively assessed for all product types. The "science" of overhead expense allocation is far from exact, yet the stakes are high. Product viability decisions are driven to a large extent by expense figures. What company does not have product managers who feel the overhead cost allocations to their products are unfair or inaccurate? Yet without some kind of objective decision framework, the company may not be reflecting all the costs of a product when determining price adequacy. It is critical that the magnitude and likelihood of capital calls be assessed in a fair and reasonable manner, so that the cost of risk enters the pricing decision.

3. Application of Concepts from Utility Theory
Assessing a cost by scenario is akin to introducing a **utility function**[3]. The Faculty and Institute of Actuaries Subject 109 Financial Economics reading introduces utility as follows:

> "In the application of utility theory to finance it is assumed that a numerical value called the utility can be assigned to each possible value of an investor's wealth by what is known as a preference function or utility function....Decisions are made on the basis of maximising the expected value of utility under the investor's particular beliefs about the probability of different outcomes. Therefore the investor's risk-return preference will be described by the form of his utility function." [4, Unit 1, p. 1]

Introducing a utility function into reinsurance pricing analysis means the company is expressing its risk-return preferences in mathematical form. Borch stated the same thing forty years ago:

---

[3] Many papers have been written on the application of utility theory to insurance and reinsurance analysis. European actuaries include Karl Borch [1], Hans Gerber and Gerard Pafumi [6], and Hans Buhlmann [2]. In North America, Leigh Halliwell [8], Oakley Van Slyke [19], Alistair Longley-Cook [12], Daniel Heyer [7], and Frank Schnapp [18] have all published articles on utility theory and insurance.

> "To introduce a utility function which the company seeks to maximize, means only that such consistency requirements (in the various subjective judgements made by an insurance company) are put into mathematical form." [1, p. 23]

This appears to be a profound change in reinsurance pricing practice. However, the change really entails **making the implicit explicit**. Any reinsurance pricing practice includes a utility assumption buried within it. Consider the marginal standard deviation pricing formula from Kreps [10], a de facto reinsurance industry standard pricing method paraphrased here:

> *Our company values the risk of a contract using the marginal impact to the portfolio standard deviation. That is, we take the square root of the expected value of the square of deviations from the mean of the portfolio outcome distribution both with and without the new contract. This difference is used to determine the marginal capital requirement, to which we assess a cost of capital figure.*

Implicit in this method are the following utility assumptions — mathematical expressions of preferences:

> *The marginal impact on the portfolio standard deviation is our chosen functional form for transforming a given distribution of outcomes to a single risk measure.*
>
> *Risk is completely reflected, properly measured and valued by this transform.*
>
> *Upward deviations are treated the same as downward deviations.*

In fact, any risk-based pricing methodology has an implicit underlying utility function[4]. Utility is the mathematical valuation of uncertainty, the essence of reinsurance pricing. One wonders why an industry that exists to purchase risk has not made a point of being explicit about its risk-return preferences.

<u>*Cost Functions*</u>

How do we assess this capital call cost at the scenario level? We need a **cost function**. The simplest cost function would be a flat percentage of the capital call amount. Table 1 shows the costs for Examples 1 - 3, assuming the scenario probabilities shown in column 3, and a flat capital call cost of **150%** of the capital call amount:

[4] See Section 6.1 of Mango [13] for more on this.

Table 1
Sample Capital Call Costs

1	2	3	4	5
Example	**Loss Ratio**	**Probability**	**Total Capital Call Amount**	**Capital Call Cost = 150% of (4)**
1	120%	10%	$33,000	$49,500
2	100%	30%	$11,030	$16,545
3	80%	60%	$0	$0

Using the assumed probabilities, the expected capital call cost over the three scenarios would be:

(10% x $49,500) + (30% x $16,545) + (60% x $0)
= $9,914

This is ***E[f(x)]***, where ***f(x)*** is our cost function, which is a scenario-dependent, skewed cost function. We need the scenario detail, because the skewness and scenario-dependence imply that ***E[f(x)] >= f(E[x])***.

Kreps [11] proposes a similar approach, one much more deeply grounded in theory. The cost function example here would represent a simplistic special case of what Kreps proposes. His represents one of the few approaches to date in the actuarial literature that recommends modification of the outcomes to reflect risk. Specifically, he suggests:

> "[The] risk load is a probability-weighted average of riskiness over outcomes of the total net loss:
>
> $$R(X)=\int dx\, f(x)r(x) \quad \text{where } r(x)=(x-\mu)g(x) \qquad (1.17)$$
>
> The function g(x) can be thought of as the "riskiness leverage ratio" that multiplies the actual dollar excess that an outcome would entail to get the riskiness. It reflects that not all dollars are equal, especially dollars that trigger analyst or regulatory tests." [11, p. 4]

Risk Neutrality
The simplistic cost assessment is a flat charge: all capital calls cost **150%** of call amount. This is equivalent to a **risk-neutral** utility function. Implementing such a utility function suggests our attitude towards risk is linear with respect to capital call magnitude — e.g., a **$2M** capital call costs a scenario twice as much as a **$1M** call. Such linear scaling implies for example that a **$100M** deficit is "100 times worse" than a **$1M** deficit.

There are some benefits to a simple capital cost charge like this. It is easy to explain, and as Schnapp points out, makes prices additive – a desirable property

[18]. However, the linear charge also implies a constant cost of **marginal** capital utilization. This has troubling implications; for example,

> Two scenarios consuming an additional **$1M** in capital would be charged the same for that additional capital call magnitude, despite the fact that one is increasing its call from **$1M** to **$2M**, while the other is increasing from **$99M** to **$100M**.

Risk Aversion

We may in practice believe capital consumption costs are not linear with respect to magnitude. There are two arguments for considering a non-linear cost function.

First, at the company level, there are definitely non-linear effects of loss magnitude. Certain catastrophe loss scenarios are intolerable, because they will impair our ability to continue as a going concern. Losses of a certain size may also trigger rating agency downgrades, rendering us uncompetitive. We might even segment decreases to a company's capital into qualitative "bands" or "tiers" whose properties change non-linearly[5]:

1. *Acceptable* (0-10%) – acceptable variation, cost of doing business.
2. *Troubling* (10%-20%) – enough deviation to cause material concern, disclosure to shareholders and rating agencies.
3. *Impairment/rating downgrade* (20%-30%) – hinders functioning of firm as a going concern.
4. *Regulatory control* (30%-50%) – substantial intervention and rehabilitation.
5. *Insolvency* (>50%)

Second, reinsurance pricing actuaries know any pricing methodology implies preferences and cost allocations, creates incentives, and ultimately steers the composition of the portfolio. It is in essence a ranking and scoring scheme. Linear marginal consumption costs may steer us toward product lines with higher risk or greater downside potential than we are comfortable with. Whether or not there are non-linear effects observable at the contract level, from a consistency and portfolio management viewpoint, we may want to have non-linear cost assessment.

A non-linear, increasing marginal cost of capital is equivalent to a **risk-averse** utility function. Rather than having a constant implicit marginal cost of capital, a risk-averse utility function will increase the capital call cost rate (non-linearly) as a function of capital call magnitude.

A risk-averse utility function need not be expressed in a closed-form. A perfectly valid risk-averse capital call cost function can be a lookup table like Table 2:

[5] See Mango [13] for additional detail on this concept.

Table 2
Sample Risk-Averse Capital Call Function

Capital Call Magnitude	Capital Call Charge
$0 - $5,000	125%
$5,001 - $10,000	150%
$10,001 - $20,000	200%
Over $20,000	400%

Table 3 shows the capital call costs using the risk-averse capital call function.

Table 3
Sample Capital Call Costs
Using Risk-Averse Capital Call Function

1	2	3	4	5	5
Example	**Loss Ratio**	**Probability**	**Total Capital Call Amount**	**Capital Call Charge**	**Capital Call Cost**
1	120%	10%	$33,000	400%	$132,000
2	100%	30%	$11,030	200%	$22,060
3	80%	60%	$0	125%	$0

Now the expected cost is

(10% x $132,000) + (30% x $22,060) + (60% x $0)
= $19,818

Compared to the risk-neutral cost of **$9,914**, the risk-averse function resulted in a higher capital call cost. This is the response we would expect, since we are mathematically stating that we have an increasing aversion to risk.

Portfolio Calibration

Ultimately, the cost function is a critical portfolio management decision, since the implicit risk-return preferences embedded in it will heavily influence the eventual portfolio composition. It represents the mathematical expression of a firm's risk appetite. This represents perhaps the most dramatic recommendation in this paper. Critics may understandably argue it is too theoretical, that mathematical expression of preferences is practically impossible. Capital allocation techniques have the apparent advantage of "observability." A "cost of capital" or "risk-adjusted discount rate" can be derived using CAPM (see Section IV of Feldblum [5]), which *appears* to ground the result in the capital markets, giving many a sense of comfort.

Unfortunately, the comfort is illusory at best. Hanging pricing decisions on a CAPM-derived cost of capital merely pushes the parameterization problem onto the capital markets. Reality checks are of course important, as a firm that wishes to have returns far in excess of any of its competitors will be in for a rude awakening. Cost function calibration **will** be difficult; but, it is groundbreaking work, so it **should** be difficult. This is true research, and involves the elucidation of intuitive risk return preferences that guide a firm's decision process. This will require framing of the decision process in a progressively more refined, analytical manner. It will mean constant feedback loops, testing of assumptions, portrayal of tradeoffs via graphical depictions, and reframing of preferences to provide different perspectives. It will be an ongoing process involving a cross-functional team of senior personnel throughout the company.

Difficult calibration is also not unique to the capital consumption approach. In fact, the comparable calibration of allocated capital to total capital is at least as difficult in its own right. Here is a sampling of issues related to capital allocation which have yet to be adequately resolved.

- *Static or Dynamic?*
  Is capital allocated annually at plan time, or "real-time" as actual premium volumes by line come in? If it is annual, what happens when an underwriting unit "hits its goal"? Are the remaining contracts free?
- *Top-Down or Bottom-Up?*
  Perform a true allocation, or build up from contract or segment level values? This quickly becomes a calibration nightmare[6].
- *Capital Good or Bad?*
  Does allocated capital represent underwriting capacity or an expense burden? In other words, do underwriting units want more or less capital?
- *Ongoing or Runoff?*
  Should capital be allocated to reserves, assets, latent, or runoff lines?
- *Zero sum game?*
  That is, is the total capital fixed? If so, if one segment's capital requirement decreases, does that mean all the other segments' capital increases?
- *Additivity?*
  Do you allocate on a marginal basis? Do you re-balance so it adds up to the total? What about order dependency?

Calibration of a utility function should be no harder than calibration of a capital allocation exercise. The end result could arguably be of more value, being a tested, explicit, mathematical representation of a company's risk-return preferences. Appendix B explains one method that can be used to calibrate by-segment pricing targets to a portfolio measure.

---

[6] Mango and Sandor [12] explains in detail an experimental study of "bottom-up" capital allocation and calibration to a portfolio measure.

5. Examples: Reinsurance Price Evaluation

We have proposed the following principles of stochastic reinsurance contract evaluation:

- Contingent capital calls have a cost associated with them, which is a function of the magnitude of the call.
- This cost is assessed at the scenario level.
- The expected value of the cost over all scenarios is treated as an overhead expense loading in the contract pricing evaluation.
- We determine the risk-adjusted net present value of the contract as the expected net present value of contract cash flows minus the expected value of the capital call cost.

We will now demonstrate these principles on two example reinsurance contracts.

Long-Tail
We will look at the pricing of the Long-tailed contract from Examples 1 - 3, using Table 2, the Risk-Averse capital call cost function. For a $100,000 premium, we can pull the results from Table 3 with an additional column for NPV:

Table 3
Sample Capital Call Costs
Using Risk-Averse Capital Call Function

1	*2*	*3*	*4*	*5*
Example	**Loss Ratio**	**Probability**	**Capital Call Cost**	**NPV of Capital Call Cost**
1	120%	10%	$132,000	$99,099
2	100%	30%	$22,060	$15,057
3	80%	60%	$0	$0

The total costs on a discounted basis would be:

1	*2*	*3*	*4*	*5*	*6*	*7*
Example	**NPV Premium**	**NPV Expenses**	**NPV Losses**	**Underwriting NPV = (2) – (3) – (4)**	**NPV of Capital Call Cost**	**Overall NPV = (5) – (6)**
1	$100,000	$15,000	$103,479	($18,479)	$99,099	($117,578)
2	$100,000	$15,000	$86,232	($1,232)	$15,057	($16,289)
3	$100,000	$15,000	$68,986	$16,014	$0	16,014

The expected value of the Underwriting NPV is

10% * ($18,479) + 30% * ($1,232) + 60% * 16,014
= $7,391

The expected value of the Overall NPV *including capital costs* is

10% * ($117,578) + 30% * ($16,289) + 60% * 16,014
= ($7,036)

Thus, reflecting all costs, this deal is below break-even. Assuming constant expenses and the same ultimate loss dollars, we find the risk-adjusted "break-even" premium to be **$103,305.** The term "break-even" should not imply that overall the company is earning no return. The cost function can be calibrated to any desired level of portfolio return measure. A more appropriate term would probably be "target premium." Here are the figures at target:

1	*2*	*3*	*4*	*5*	*6*	*7*
Example	**NPV Premium**	**NPV Expenses**	**NPV Losses**	**Underwriting NPV = (2) – (3) – (4)**	**NPV of Capital Call Cost**	**Overall NPV = (5) – (6)**
1	$103,305	$15,000	$103,479	($15,173)	$86,858	($102,031)
2	$103,305	$15,000	$86,232	$2,073	$6,702	($4,629)
3	$103,305	$15,000	$68,986	$19,320	$0	19,320

The expected value of the Overall NPV *including capital costs* would be

10% * ($102,031) + 30% * ($4,629) + 60% * 19,320
= $0

Property Catastrophe
Consider a high-layer contract, with a 2% chance of being hit (1 in 50 years). However, when it is hit, it suffers a full limit loss. Example 4 shows the details:

Example 4
Property Catastrophe Contract

Premium	$	1,000,000
Limit	$	10,000,000

	No Loss Scenario		Loss Scenario	
Probability		98.0%		2.0%
Premiums	$	1,000,000	$	1,000,000
Expenses	$	-	$	-
Losses	$	-	$	10,000,000
Capital Call Amount	$	-	$	9,000,000
Capital Call Factor		0.0%		400.0%
Capital Call Charge	$	-	$	36,000,000
Expected NPV	$	800,000		
Expected Capital Call Cost	$	720,000		
Expected Risk-adjusted NPV	$	80,000		

In the full limit loss scenario, the capital call is for $9M. Hitting this with a 400% capital call charge factor, the expected risk-adjusted NPV is $80,000 – close to target.

6. Conclusions

This paper has ties to much current work in both actuarial and financial literature. In particular, it is linked to:

- o Pricing via probability measure change – from voluminous capital markets literature;
- o Utility theory in pricing – from Halliwell, Heyer and Schnapp;
- o The Wang Transform – from Wang;
- o The market cost of risk – from Van Slyke; and
- o Additive Co-Measures – from Kreps.

Of particular note are comments made by Kreps [11]:

> "[It] seems plausible that for managing the company the risk load for an outcome:
>
> (1) should be a down side measure (the accountant's point of view);
> (2) should be proportional to that excess over the mean for excess small compared to surplus (risk of not making plan, but also not a disaster);
> (3) should become much larger for excess significantly impacting surplus; and
> (4) should flatten out for excess significantly exceeding surplus – once you are buried it doesn't matter how much dirt is on top. " [11, p. 9]

The proposed approach focuses on downside, and can support discontinuous, non-linear risk measures as functions of surplus.

This approach also represents an attempt at conscious, intentional and explicit introduction of "information content" into reinsurance market prices. Perhaps one of the more dubious assumptions of competitive market theory is that market prices reflect all the information available. Even setting aside the enormous informational asymmetries in the reinsurance arrangement, one cannot ignore the large role of interpretation. In order for reinsurance market prices to contain all this information — i.e., to really "mean something" — the submission information must be converted into prices. The process is one of interpretation by reinsurance underwriters and actuaries, including subjective and objective considerations, market intelligence, internal strategy, tips and hints from the broker or client, relationship, bank.... With all this confluence of strategies and signals, the discipline of an explicit, objective utility approach seems desperately needed and sorely overdue.

References

[1] Borch, Karl, "The Utility Concept Applied to the Theory of Insurance," *ASTIN Bulletin*, Volume I, Part V, 1961, p. 245-255.

[2] Bühlmann, Hans, "An Economic Premium Principle," ASTIN Bulletin, 11(1), 1980, p. 52-60.

[3] CAS Task Force on Fair Value Liabilities, White Paper on Fair Valuing Property/Casualty Insurance Liabilities, http://www.casact.org/research/tffvl/index.htm.

[4] Faculty and Institute of Actuaries, "Subject 109 Financial Economics Core Reading 2000."

[5] Feldblum, Sholom, "Pricing Insurance Policies: The Internal Rate of Return Model," CAS Exam 9 Study Note, 1992.

[6] Gerber, Hans U., and Pafumi, Gérard, "Utility Functions: From Risk Theory To Finance," *North American Actuarial Journal*, Volume 2, Number 3, July 1998, p. 74-100.

[7] Heyer, Daniel D., "Stochastic Dominance: A Tool for Evaluating Reinsurance Alternatives," *CAS 2001 Summer Forum*, Reinsurance Call Paper Program, p. 95-118.

[8] Halliwell, Leigh J., "ROE, Utility, and the Pricing of Risk," *CAS 1999 Spring Forum*, Reinsurance Call Paper Program, p. 71-135.

[9] Halliwell, Leigh J., "A Critique of Risk-Adjusted Discounting," *2001 ASTIN Colloquim*, available at http://www.casact.org/coneduc/reinsure/astin/2000/halliwell1.doc.

[10] Kreps, Rodney, "Reinsurance Risk Loads from Marginal Surplus Requirements," *PCAS LXXX*, 1990, p. 196-203.

[11] Kreps, Rodney, "A Risk Class with Additive Co-measures," unpublished manuscript.

[12] Longley-Cook, Alastair G., "Risk Adjusted Economic Value Analysis," *North American Actuarial Journal*, Volume 2, Number1, January 1998, p. 87-100.

[13] Mango, Donald, and Sandor, James, "Dependence Models and the Portfolio Effect," *CAS 2002 Winter Forum*, Ratemaking Discussion Paper Program, p. 57-72.

[14] Mango, Donald, "Risk Load and the Default Rate of Surplus," *CAS 1999 Spring Forum*, Discussion Paper Program on Securitization of Risk, p. 175-222.

[15] Meyers, Glenn, "The Cost of Financing Insurance," *CAS 2001 Spring Forum*, DFA Call Paper Program, p. 221-264.

[16] Panjer, Harry H., et al (eds.), *Financial Economics*, 1998, The Actuarial Foundation.

[17] Philbrick, Stephen W., and Painter, Robert W., "DFA Insurance Company Case Study, Part II: Capital Adequacy and Capital Allocation," *CAS 2001 Spring Forum*, DFA Call Paper Program, p. 99-152.

[18] Schnapp, Frank, "The Diversification Property," presented to the 11th AFIR Colloquium, September 2001, available at http://www.actuaries.ca/meetings/afir/handouts/Schnapp - The Diversification Property.pdf.

[19] Van Slyke, Oakley E., "The Cost of Capital: An Axiomatic Approach," in Actuarial Considerations Regarding Risk and Return in Property-Casualty Insurance Pricing, CAS, 1999, p. 135-164.

[20] Wang, Shaun S., "Equilibrium Pricing Transforms: New Results Using Buhlmann's 1980 Model," to be published in *Journal of Financial Economics*.

Appendix A
Does Capital Allocation Make Sense for Insurance?

This appendix addresses whether the practice of capital allocation even makes sense for insurance. Capital allocation was originally applied to manufacturing firms. However, the nature of their usage of capital is fundamentally different from insurers.

Manufacturing

Consider a representative example from Halliwell:

> "A company is considering entering the widget business, which entails the purchase of a machine to produce widgets. The company estimates that the machine will last five years, and the profits from the sale of its widgets over those five years will be $100,000, $125,000, $125,000, $100,000, and $75,000." [8, p. 73]

This example typifies the manufacturing capital analysis framework. Capital is invested up front, and profits (hopefully) come in the future. This approach was designed for analysis of investment opportunities in industries where production comes before revenue collection. We might term such industries "*spend-then-receive*." These industries must invest capital into production and distribution costs before they can hope to collect revenue. There are no products to sell without capital. Manufacturers have the following time dynamic with respect to capital usage:

- Capital investment costs are mostly up-front and well known, while
- Revenues are in the future and unknown.

Manufacturing capital also must cover operating deficits. In order for the firm to continue operations when revenues are less than costs, additional capital must be invested[7].

It is important to note that capital investment represents a cost, an expenditure of a known amount. Large manufacturing organizations that "allocate capital" between business units actually spend the capital they are allocating. Capital is "consumed by production." There is nothing theoretical about either the total amount available, or the amounts allocated to various product lines. Since real spending of real money is involved, capital allocation decisions receive a tremendous amount of attention, scrutiny and peer review. They lie at the heart of strategic planning for manufacturing firms. Capital allocation is the lifeblood of a manufacturing business unit, the means to continue activities.

---

[7] Venture capitalists often refer to the "burn rate" of a start-up company: the rate at which the operation consumes capital during its start up period, when there are typically no revenues, only costs.

The manufacturer hopes this capital investment will be followed by profits in the future, but is uncertain how much revenue will come, or when. Typically, this uncertainty influences the decision in the form of a risk-adjusted discount rate applied to future revenue projections[8].

Insurance
In contrast to manufacturing, which is "*spend-then-receive*," insurance is a "*receive-then-spend*" industry. What we term production is really revenue collection. Our revenues are fairly predictable, even by product line. Demand is somewhat inelastic, given the legal and regulatory requirements. Insurers can plan their premium volume with a good degree of accuracy. They struggle to assess the loss cost of their products that come in the future. Comparing the time dynamic of insurance and manufacturing is illuminating:

Item	***Manufacturing***	***Insurance***
Revenue	In the future, unknown	Up front, well known
Costs	Up front, well known	In the future, unknown

Insurers are actually something of a "temporal mirror image" of manufacturers: they collect revenue up front, and hope the future costs aren't too high or too soon. There is no question that insurers need capital in total to secure claims paying ability. However, it is critical to recognize this distinction: insurers can collect revenue on products **without having had to invest any capital in product production**. Insurance "production costs" are actually loss payments.

It is surprising that such a striking difference in capital usage has not resulted in any materially different capital treatment in the insurance IRR framework. In fact, actuaries have kept the manufacturing capital usage profile, treating insurance products as if they require supporting capital to be invested up-front, then released. This insurance IRR framework is actually *pseudo-manufacturing*: the capital amount is "risk-based," derived from stochastic analysis; yet it is invested in an essentially deterministic framework that ignores the reality that insurance products do not require capital investment to produce.

Several major problems with this hybrid approach are immediately apparent.

First, when evaluated in a stochastic environment, **the allocated supporting capital makes no sense at the modeled scenario level**. The risk-based supporting capital is determined based on a risk measure of the distribution in

[8] The end goal of "dis-counting" – literally, reducing the value of – uncertain future revenues is appropriate. The method of risk-adjusted discounting – effecting that reduction in value by using compounded discounting at a higher rate – is unnecessary and (as Halliwell [8] has shown) fraught with inconsistencies. It represents an example of "overloading an operator," piling additional functional burden onto what should be a single purpose operator.

total. For any given scenario, though, this overall amount is never the actual required amount — namely, the modeled operating deficit. The allocated capital is excessive for favorable scenarios, and grossly inadequate in severe loss scenarios (unless capital equals the policy limit). This problem is particularly striking in the evaluation of catastrophe reinsurance contracts, with small probabilities of a full contract limit loss. A "risk-based" capital amount might be some small fraction of the limit — say 5%. What sense does this capital amount make in the limit loss scenario? We held 5% of the limit as capital? And then how exactly did we fund the remaining loss amount? It came from company capital in total. The alternative — holding the full limit as capital — puts an unrealistic return burden on the contract. Current market price levels would likely make the contract look unattractive.

Second, **insurance contracts use capital in the future**. Insurance "production" costs are in fact distribution and revenue collection costs. Manufacturing capital funds true production costs, as well as operating deficits. Since the vast majority of insurance costs come in the future in the form of loss payments, insurance capital usage belongs in the future as well. However, capital would only be needed if the contract began running an operating deficit or loss — a negative cash position reflecting all sources of revenue, including investment income.

Third, **allocated supporting capital is completely theoretical**. In contrast to manufacturing, where allocation means actual spending of actual known amounts, allocated supporting capital simply does not exist at the contract level. Nothing is actually spent or invested. The strong ties to reality inherent in the manufacturing framework have been lost, and with them go much of the discipline and meaning of capital allocation.

Finally, on a more philosophical level, **supporting capital is a portfolio concept, and may not be meaningfully divisible**. There is no question that supporting capital in total is *essential to the insurance operation*. The product we sell is current and future claims paying ability; however, this ability applies to the insurer in total as a going concern. Future claims paying ability is heavily dependent on total supporting capital. There is also no question that allocation of supporting capital is *possible*. The issue is with the meaning of that allocated capital. There are many holistic phenomena that have no divisible component pieces. An historical example is found in the ancient scientists' search in vain for the "seat of the soul" in the brain. They sought a physically grounded, identifiable location for what is now believed to be a "field" phenomenon. One might well consider trying to allocate life to the component organs of the body, or allocating the success of the Lakers to individual players: 38% to Shaq, 33% to Kobe.... Since actuaries like to communicate mathematically, in equation form:

$$\sum(\textit{Promises To Pay}) \neq \textit{Promise To Pay}\left(\sum\ \right)$$

Appendix B
One Approach to Portfolio Calibration

This appendix will outline one approach for calibration of a contract capital call cost function with a company total cost of capital. This approach can be used to develop risk-adjusted target combined ratios by LOB without allocating capital. No matter which cost function is used, a testing period is recommended where several possible functions and/or parameter sets can be evaluated across a significant sample of the portfolio. Once the results are aggregated, the total assessed cost of risk can be estimated and expressed as a percentage of a base such as expected losses or premium.

Here are the steps of the suggested process:

1. Generate modeled scenarios of company operating income and individual line of business underwriting result over a projected three calendar year period. Include reserves as well as prospective business in the definition of a line of business. Also include asset risk and linkages with generated economic scenarios.

2. Apply a risk-averse utility function to the company's operating income distribution to assess a "capital depletion" cost at the scenario level to those scenarios with negative operation income. Calibrate the expected value of this cost over all scenarios to a desired target cost of capital measure.

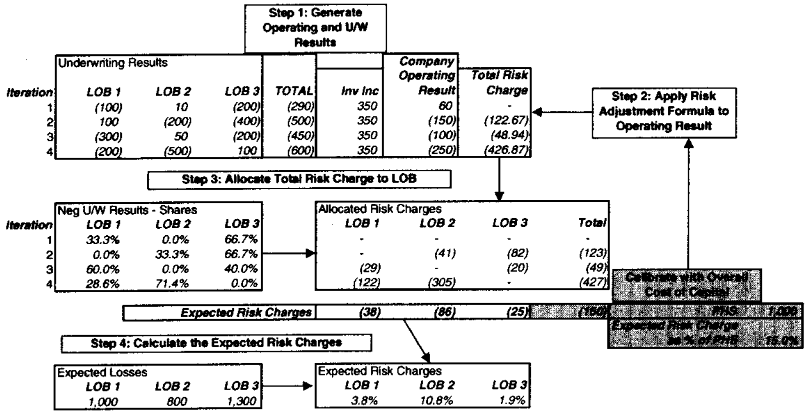
3. Allocate the scenario capital depletion costs back to line of business at the scenario level in proportion among all lines having an underwriting loss in that scenario.

4. Calculate the expected value of allocated depletion cost by line of business over all scenarios. Express this charge as a percentage of expected loss.

5. Generate the other components of a break-even risk-adjusted combined ratio, namely discounted loss ratio and expense ratio.

Example 5 shows a simplified flowchart of steps 1-4.

Example 5
Company Risk Adjustment Methodology Flowchart



We will cover each step in more detail.

Step 1: Model Company Operating Income and LOB Underwriting Income
We use company calendar year operating income as the risk measure at the scenario level. Negative operating income depletes capital, so the cost of capital depletion is assessed here, based on the magnitude of the depletion. We also model the line of business (LOB) calendar year underwriting income at the scenario level. The calendar year variation includes the random effects of reserve runoff for carried reserves as of the start of the simulation period.

The main strength of this approach lies in its inclusion of so many modeled dependencies and interactions: between reserve runoff LOB's, between reserves and prospective business; between liabilities and assets via the economic scenarios, etc.

Step 2: Apply a Utility Function to Assess the Cost of Capital Depletion
As can be seen in the demo flowchart, a risk-averse exponential utility function assesses a capital depletion charge to scenarios with negative operating income. The calibration of the expected assessed charge with a portfolio capital cost is straightforward.

Step 3: Allocate Capital Depletion Cost back to LOB
At the scenario level, the calculated depletion cost is allocated back to those LOB with underwriting losses, in proportion to their underwriting loss as a percent of the total of underwriting losses for LOB with an underwriting loss. This is one

allocation rule, and obviously not the only or even best. The point is that an allocation rule can be applied at the scenario level.

Step 4: Calculate Expected Capital Depletion Cost by LOB

Each LOB has an expected value of the allocated depletion costs over all scenarios. This figure is expressed as a percent of expected loss to facilitate inclusion in the break-even risk-adjusted combined ratio calculation.

Step 5: Calculate the Break-even Risk-adjusted Combined Ratio

The additional required elements are a discounted loss payment pattern and expense load. The goal is to calculate economically break-even combined ratios — i.e., 100% discounted combined ratio — including a load for the cost of capital.

Advantages

This approach shows one way to implement insurance portfolio management using a dynamic portfolio model. The approach links corporate cost of capital needs with LOB pricing targets in a simple, coherent framework, without allocating capital. A dynamic portfolio model also has other advantages, including the development of a more complete picture of current capital adequacy; the ability to introduce a time dimension into risk modeling; and a framework for introducing systematic risk from the insurance and capital markets.

Estimating the Cost of Commercial Airlines Catastrophes—A Stochastic Simulation Approach

Romel Salam, FCAS, MAAA

Estimating the Cost of Commercial Airlines Catastrophes - A Stochastic Simulation Approach

Romel Salam, FCAS, MAAA

Abstract

Actuaries are increasingly finding more applications for stochastic simulation in pricing, reserving, DFA, and other insurance and financial engineering problems. For instance, stochastic simulation has gained acceptance as a pricing tool for property catastrophe coverage in the insurance, reinsurance, broker, and investment communities. This has required primary companies to compile and provide information at a more detailed level than they did only a few years ago. Various commercial simulation products have emerged to help companies assess and price their property catastrophe exposures. Although there are many parallels between the catastrophe exposures of property and commercial aviation risks, the use of simulation is not widespread in the assessment of commercial aviation catastrophic exposures. In this paper, we present the framework for a simulation model for commercial aviation catastrophes and we discuss various aspects of designing such a model including the level and type of information needed.

Acknowledgement

We would like to thank Yipei Shen and Pavel Zhardetskiy for their many contributions to this project. We are also indebted to many of our other colleagues at Transatlantic Reinsurance Company who, directly or indirectly, knowingly or unsuspectingly, willingly or under physical threat, contributed to this paper. Of course, we remain solely liable for mistakes, misrepresentations, and any libelous material that appear in the following pages.

Introduction - The need for a stochastic model

The claims covered by a comprehensive commercial airline policy can be broken into two groups. The first group consists of trivial claims such as lost luggage, "slip and fall" accidents, or minor damage to the hull of an aircraft while the second group comprises catastrophic claims arising out of airplane crashes resulting in serious injuries, fatalities, property damage, and major or total loss of an aircraft. Most of the pricing tools that are used to price airline's hull and liability exposures tend to rely on experience rating techniques. Under a basic experience rating method, the projected losses are based on an average of past losses adjusted for trend and development. An experience rating approach may work relatively well when only the non-catastrophic exposure of airlines is considered. However, traditional experience rating methods would tend to overstate the expected loss when one or more catastrophes are included in the experience period, and, conversely would tend to understate the expected loss when there are no catastrophes in the experience period. Under a more sophisticated experience rating approach, losses are separated into their catastrophic and non-catastrophic components. The catastrophe losses are then compiled and averaged over a very long period of time in order to come up with an "expected catastrophe loss amount" similar to what is used in property ratemaking. Even under the latter approach, the question needs to be asked as to whether past catastrophe experience is representative of future experience. First, the frequency of catastrophic accidents may have changed over time due to such factors as improved aviation technology, better or worse safety regulation, or increased air traffic. Secondly, the costs of hull and liability coverage are indeed impacted by not only general inflationary trends which can be reflected within a traditional experience rating model, but also by changes in an airline's fleet, passenger load factors[1], destination and passenger profiles[2] which are harder to reflect in an experience rating exercise. Finally, as we look to the future, the introduction of new aircraft models such as the Airbus 380 model - which could

[1] Passenger load factor represents the average percentage of an airline's seating capacity that is filled.

[2] Destination profile for an airline refers to the countries to which the airline is flying. The liability damage award of accident victims may vary by country. For instance, a priori, an airline operating domestic flights solely in India would have a lower liability potential than one operating solely in the United States. Passenger profile refers to the age, occupation, income of passengers as these are all factors that can determine the level of compensatory damage of accident victims. A priori, an airline whose core clientele was made up of college freshmen going on vacation would have a lower liability potential than one whose core clientele consisted of well paid corporate managers going to business meetings.

accommodate up to 840 passengers in a single class configuration - to an airline's fleet, the addition of new set of routes and destinations, the continued evolution of contracts and laws establishing the compensation of victims of airline accidents will all make it less likely that traditional experience rating will remain an adequate forecasting tool.

The stochastic model that we present avoids most, if not all, of the pitfalls of traditional experience rating methods. While judiciously making use of historical data such as past accident rates, the model will rely on the most current information relating to an airline's fleet, passenger and destination profiles, number of departures (or miles flown), and passenger load factors. The model will also be flexible enough to allow the user to incorporate his/her views on the impact of legal changes on the cost of liability, hull or other related costs of accidents. This model will be especially well suited for analyzing contracts which carry a lot of bells and whistles. A simulation model that breaks down the loss process into its many components also forces the user to think about the different factors that impact on the costs of airline catastrophes. Perhaps, one drawback of such a model is that it requires a more detailed level of information than generally needed in performing an experience rating exercise. However, such information is generally available with a little bit of research.

This paper is organized into eight sections. In section 1, we present a schematic of a stochastic simulation model for evaluating the cost of passenger liability coverage. In section 2, we define commercial airlines and airline catastrophes. In section 3, we delve into the area of frequency, including the choice of an appropriate measure of exposure. In that section, we also explore the issue of classification by relying on work presented in "Reinventing Risk Classification – A Set Theory Approach" [6]. We revisit a statistic introduced in that paper for the purpose of making inferences about Poisson distributed events. We use this statistic to comment on a Wall Street Journal article, which sought to discuss the relative safety of several aircraft models. We then tackle the issue of whether the rate of airline catastrophes has changed at all over time in the same section. In section 4, we look at the cost of catastrophes for different coverages, including more easily determined costs such as those for hull coverage to more challenging ones such as passenger liability, third party liability, and products liability. We also briefly touch on the issue of classification relating to passenger liability costs in that section. In section 5, we discuss how various results coming from the model can be validated against actual historical data. In section 6, we offer some thoughts on how to incorporate the risks of terrorism or sabotage in the model. In section 7, we give an example of the simulation model for a cover for a hypothetical group of airlines. In section 8, we provide final thoughts.

1) Simulation Scheme

Figure 1.1 below shows how we would generate passenger liability losses using a simulation model. This scheme would vary depending on the level of information available and the coverage of interest.

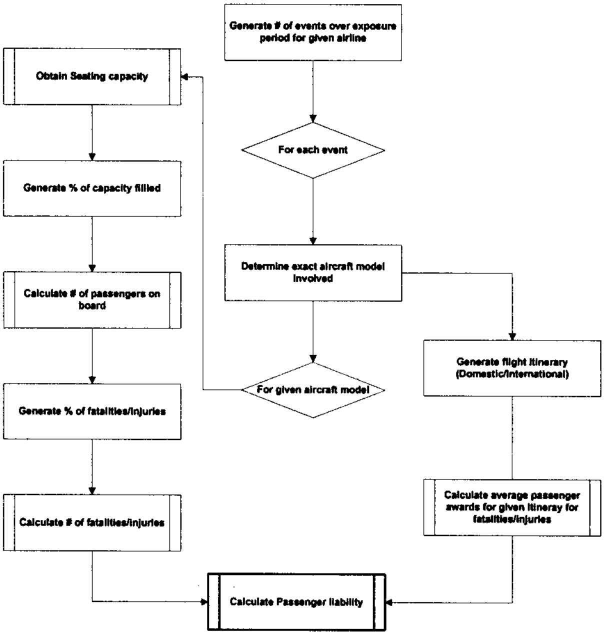
2) Definitions

Before we go too far into this discussion, let's try to agree on the topic of discussion itself by attempting to put some parameters around two of the terms that are central to this paper:

2.1) Commercial Airlines

Insurance underwriters generally differentiate between commercial and general aviation. General aviation typically encompasses operation of smaller airplanes used for leisure, industrial and agricultural purposes, or simply in the private transportation of individuals or employees. Helicopter and balloon operations are generally lumped into the general aviation category. Commercial aviation involves the transportation, for compensation or hire, of persons or cargo by aircraft. In the US, a commercial operator is one that has been certificated by the Federal Aviation Administration (FAA) under Code of Federal Regulation (CFR) part 121 (airlines) or CFR Part 135 (commuters) to provide air transport of passengers or cargo. So-called air taxis and commuters operate smaller aircrafts and carry few passengers per flight whereas airlines typically operate jet aircrafts that can carry large loads of passengers per flight. More recently, the line between commuters and airlines has been blurred by acquisitions as well as the amendment of some of the FAA codes. This paper is concerned mostly

Figure 1.1 – A simulation framework



with commercial airlines, as most of the statistics we will discuss will relate to airlines represented by US operators certificated under CFR Part 121 and operators in other countries with similar certification.

2.2) Airline Catastrophes

We have referred several times already to airline catastrophes as if the term was self-explanatory. In fact, one might take several views as to what constitutes a catastrophe. One perspective of catastrophes could be that of an excess of loss reinsurer who would typically be impacted only by events above a certain threshold. For instance, a reinsurer could *define a catastrophe as an accident, occurring between takeoff and landing, involving one or several aircrafts, and which results in major damage to or destruction of an aircraft's hull.* Under this definition, for instance, damage to an aircraft from a hailstorm or an earthquake while garaged would not be counted as a catastrophe, neither would fatalities or injuries occurring as a result of air turbulence, food poisoning, or falling luggage. A midair collision between two or more aircrafts would be counted as one catastrophe. Throughout this paper, we will use slightly different definitions of catastrophe and different data sources to illustrate different aspects of the simulation model. The exact definition used is of no particular importance since we are not trying to promote any one definition but rather trying to present a method by which the cost of such catastrophes, however defined, can be evaluated. It is, however, important that the data collected for the purpose of constructing a model be consistent with the definitions used in the contracts and products that are being evaluated.

3) Modeling the Frequency of Airline Catastrophes

How do we best model the number of airline catastrophes? Within the casualty and property actuarial practice, there are two distributions that are commonly used to represent the frequency distribution of accidents, namely the Poisson and the Negative Binomial distributions. We, a priori, will work with the Poisson distribution because of its simplicity and its intuitive appeal[3]. A modeler is free to choose other distributions that might work better or as well.

[3] There are three postulates implied by a Poisson process:

The Poisson model is defined as follows: $f(x) = \frac{(\lambda d)^x e^{-(\lambda d)}}{x!}$,

where x is the number of catastrophic accidents, λ is the expected rate of accident per exposure unit, and d is the number of exposure units for the period under review.

3.1) Picking an exposure base

The exposure base as will be used here is the unit upon which frequency will be measured. Airlines usually report the number of departures, miles (or kilometers) flown, hours flown for annual, quarterly, and even monthly periods. Any of these measures could serve as an exposure base since they are almost perfectly correlated. The modeler's decision as to which of these potential units of exposure to use may be based on which is found to be more accessible, more accurate, measured and defined more consistently overtime. The modeler needs to be well aware of potential distortions in the exposure data especially when using different sources to gather the data. We will use the number of departures as our unit of exposure because of some evidence showing that the risk of catastrophic occurrences is concentrated around takeoffs and landings. We, however, have not found any significant differences in our results when we used hours or miles flown as measures of exposure.

3.2) Classification

For the twenty-two year period from 1980 to 2001, commercial airlines catastrophes[4] occurred in the world at a frequency of 1.22 per million departures. Should we look at this frequency rate as being applicable to all commercial airlines and use it as the basis for the λ in the Poisson model for all airlines? This approach could potentially result in the underestimation of the accident risk for some groups of airlines while resulting in the overestimation of that risk for some other groups.

---

1) The numbers of occurrences in non-overlapping time intervals are independent.
2) The number of occurrences in a time interval has the same probability for all intervals.
3) The probability of two or more events in a small time interval is zero.

We believe that the occurrence of catastrophic airline accidents, excluding those caused by willful acts, satisfy all three postulates.

[4] Catastrophe here is defined as accidents resulting in total destruction of an aircraft. Data comes from a proprietary source and is based on Western-built aircrafts only.

Conversely, should we group airlines into cells[5] along certain risk characteristics[6] and proceed to calculate frequency rates for each cell based on the data within? Are we to then presume that airlines across these cells have fundamentally different rates of catastrophic accidents by virtue of our having devised the grouping scheme? Under this approach, we risk assigning different, perhaps even significantly different, frequency rates to cells of airlines that have essentially the same propensity for catastrophic accidents. Let's assume for a moment that airlines are grouped based on the subcontinent on which they are domiciled. In the twenty-two year period from 1980 to 2001, the average frequency of catastrophic accidents for North-American airlines has been around 0.5 per million of departures while that for Western-European airlines has been closer to 0.6 per million of departures. Did the difference in the observed accident rates arise out of the random nature of catastrophic accidents or did it arise out of a fundamental difference in the propensity of accident for the two groups? The search for answers to these questions spun a classification methodology introduced in the paper titled "Reinventing Risk Classification – A Set Theory Approach" [6]. We will use the procedures from this paper to give an example of a classification scheme for airlines. Before we do, however, we want to reacquaint the reader with a statistic, $\hat{R}_0$, introduced in the aforementioned paper and which was used to make inferences about Poisson distributed populations.

3.2.1) A review of the $\hat{R}_0$ statistic.

Let λ_A and λ_B represent the expected frequency rates for two Poisson populations A and B, with d_A and d_B units of exposure, respectively. Also, let $\hat{\lambda}_A$ and $\hat{\lambda}_B$, represent the maximum likelihood estimates of λ_A and λ_B, respectively. In [6, p. 105-114], we show that if $\lambda_A = \lambda_B$,

$$\hat{R}_0 = \frac{\hat{\lambda}_A - \hat{\lambda}_B}{\sqrt{\dfrac{\hat{\lambda}_A}{d_A} + \dfrac{\hat{\lambda}_B}{d_B}}} \rightarrow N(0,1)$$ for large[7] d_A and d_B values.

[5] A cell is a set of airlines with the same risk characteristics [6, p. 89].
[6] Risk characteristic is an attribute that identifies a risk or group of risks [6, p. 88].
[7] i.e. as d_A *and* $d_B \rightarrow \infty$ [6, p. 105-114].

We then use $\hat{R}_0$ to make inferences about the equality of the frequency rates underlying pairs of Poisson distributed populations. If we define the null hypothesis as $\lambda_A = \lambda_B$, then we will reject that hypothesis at the 10% significance level (90% confidence interval) if $\hat{R}_0$ falls outside the range of (-1.65, 1.65). We explain in [6, p 94] that $\hat{R}_0$ can be thought of as the observed distance between the two populations' samples. If that distance is small, we tend to accept the hypothesis that the populations have the same expected frequency. If it is large, we tend to reject the equality hypothesis. Observe that $\hat{R}_0$ depends not only on the MLE estimates of the population frequencies but also on the number of exposure units of the respective populations. For instance, the absolute value of $\hat{R}_0$ increases as the number of exposure units increases (everything else being equal).

A Wall Street Journal article [5] in which the author sought to demonstrate the poor safety record of the MD-11 aircraft relative to other similar models provides a good example of how $\hat{R}_0$ can be used to make inferences about Poisson populations. The article shows a graphic with accident rates by airplane types, which we summarize in table 3.1 below.

If we assume that the number of accidents for each of the aircraft models is Poisson distributed, we can use $\hat{R}_0$ to make inferences about the relative safety of these models. The exposure units are the number of millions of departures, while the maximum likelihood estimates of the expected frequencies are represented by the frequency per million departures. For instance, to test the hypothesis that the underlying accident rates for the MD-11 and the A300-Early are the same, we calculate $\hat{R}_0$, using an algebraic equivalent[8] of the formula introduced in section 3.2.1, as

follows: $\hat{R}_0 = \frac{6.54 - 1.29}{\sqrt{\frac{6.54^2}{5} + \frac{1.29^2}{7}}} = 1.771$.

Table 3.1- Accident Rates by Airplane Type[9]

Aircraft Model	(1)=(2)/(3) Million of Departures	(2) Hull Losses[10]	(3) Frequency per million Departures
B-707/720	17.8	115	6.46
DC-8	12.2	71	5.84
B-727	72.2	70	0.97
B-737-1 & 2	50.4	62	1.23
DC-9	58.1	75	1.29
BAC 1-11	8.3	22	2.64
F-28	8.1	32	3.94
B747-Early	11.1	21	1.90
DC-10	7.8	20	2.57
A300-Early	5.4	7	1.29
L-1011	5.2	4	0.77
MD-80/90	23.3	10	0.43
B-767	7.3	3	0.41
B-757	8.7	4	0.46
Bae146	5.1	3	0.59
A-310	2.9	4	1.40
A-300/600	2.2	3	1.34
B-737-3, 4 & 5	30.8	12	0.39
A-320/319/321	7.3	7	0.96
F-100	3.8	3	0.80
B747-400	2.0	1	0.49
MD-11	0.8	5	6.54

[8] $\hat{R}_0 = \frac{\hat{\lambda}_A - \hat{\lambda}_B}{\sqrt{\frac{\hat{\lambda}_A}{d_A} + \frac{\hat{\lambda}_B}{d_B}}} = \frac{\hat{\lambda}_A - \hat{\lambda}_B}{\sqrt{\frac{\hat{\lambda}_A^2}{n_A} + \frac{\hat{\lambda}_B^2}{n_B}}}$ where n_A and n_B represent the number of accidents for populations A and B, respectively, and $\hat{\lambda}_A = \frac{n_A}{d_A}$ and $\hat{\lambda}_B = \frac{n_B}{d_B}$.

[9] The article lists Boeing as the source of the data. The exposure units (million of departures) were not provided but calculated as the ratio of the number of accidents to the accident rate.

[10] The article defines hull losses as damage so severe the plane isn't repaired.

At a 10% significance level, we would reject the hypothesis that the two aircraft models have the same propensity for accident since $\hat{R}_0$ falls outside the interval (-1.65, 1.65). However, we would not be able to reject the hypothesis at a 5% significance level since $\hat{R}_0$ falls in the interval (-1.96, 1.96). In table 3.2, we calculate the $\hat{R}_0$ values between the MD-11 model and other aircraft models in table 3.1.

Table 3.2 - $\hat{R}_0$ values between $MD-11$ and other models

Aircraft Model	$\hat{R}_0/MD-11$
B-707/720	0.027
DC-8	0.233
B-727	1.903
B-737-1 & 2	1.813
DC-9	1.793
BAC 1-11	1.309
F-28	0.865
B747-Early	1.571
DC-10	1.332
A300-Early	1.771
L-1011	1.956
MD-80/90	2.087
B-767	2.089
B-757	2.072
BAe146	2.021
A-310	1.709
A-300/600	1.719
B-737-3, 4 & 5	2.101
A-320/319/321	1.893
F-100	1.939
B747-400	2.040

Before drawing any conclusions from the above table however, one would need to look into other factors that may impact on the accident rates. For instance, if the MD-11 losses were coming disproportionately from a particular operator or group of operators, the issue might be more specific to the operator or group of operators rather than to the aircraft model itself.

3.2.2) A classification scheme for the frequency of airline catastrophic accidents

A priori, one might expect airlines operating under similar jurisdictions and having similar types of operations, fleet, staff training, and safety procedures to display the same expected rate of accidents. The jurisdictions – which could be tantamount to countries – help explain the degree of oversight to which airlines are subject, the adequacy and competence of air traffic control, the level of competition in the market, the resources available to regulators to enforce safety rules, the degree of accountability of regulators and airlines to the public, and the public's attitude toward safety. Factors that may help delineate among jurisdictions include political system, economic standing, and judicial/tort system. Factors that may explain differences between airlines within the same jurisdictional group include size and years of operation. For instance, to the extent that there are economies of scale present in aircraft maintenance or staff training, larger airlines may exhibit a better safety record than smaller ones. For illustration purposes only, let's look at a two-dimensional classification scheme where jurisdiction and size of operations are the two classification variables. Then, we will define five jurisdictional groups and three sizes, which will result in fifteen cells for which the exposures and MLE estimates are shown in tables 3.3 and 3.4 below:

Table 3.3 – Departures in millions

	Jurisdiction 1	Jurisdiction 2	Jurisdiction 3	Jurisdiction 4	Jurisdiction 5
Large	56.9	120.8	-	4.8	15.8
Medium	33.5	8.9	7.4	12.5	15.5
Small	3.5	2.2	2.8	2.6	2.0

Table 3.4 – Initial MLE Estimates (Accidents / Million Departures)

	Jurisdiction 1	Jurisdiction 2	Jurisdiction 3	Jurisdiction 4	Jurisdiction 5
Large	0.527	0.356	N/A	3.120	2.019
Medium	0.507	1.679	4.305	2.800	2.710
Small	1.443	2.736	17.852	9.237	4.085

A revised set of estimates is obtained in table 3.7 below based on a procedure introduced in "Risk Classification – A Set Theory Approach" [6] and at the presentation of the paper at the winter 2002 meeting. Basically, each cell[5] defines a class[11] made up of the cell itself and possibly of other cells that

[11] Please refer to [6, p. 88 -90] for a definition of these terms.

are compatible[11] with it. Once the classes are defined, each cell within a class is given a credibility[11] weight and the revised estimate for each cell is the credibility weighted average of all the cells in its class. The classes defined by each cell and the credibility weights assigned to all the cells in each class are shown in tables 3.5 and 3.6 below, and again in exhibits 5 and 6 of Appendix A. The steps leading to table 3.5, 3.6, and 3.7 are detailed in Appendix A.

Table 3.5 - Classes defined by each cell

Cells	Classes
J1/L	{J1/L, J1/M, J1/S}
J2/L	{J2/L}
J4/L	{J4/L, J4/M}
J5/L	{J5/L}
J1/M	{J1/M, J1/L, J1/S}
J2/M	{J2/M, J2S}
J3/M	{J3/M}
J4/M	{J4/M, J4/L, J5M}
J5/M	{J5/M, J4/M, J5/S}
J1/S	{J1/S, J1/L, J1/M}
J2/S	{J2/S, J2/M, J5/S}
J3/S	{J3/S}
J4/S	{J4/S}
J5/S	{J5/S, J2/S, J5/M}

Table 3.6 - Credibility weights

Cells	Classes
J1/L	{J1/L .606, J1/M .357, J1/S .037}
J2/L	{J2/L, 1.00}
J4/L	{J4/L .278, J4/M .722}
J5/L	{J5/L, 1.000}
J1/M	{J1/M .357, J1/L .606, J1/S .037}
J2/M	{J2/M .803, J2S .197}
J3/M	{J3/M, 1.000}
J4/M	{J4/M .381, J4/L .147, J5M .472}
J5/M	{J5/M .517, J4/M .417, J5/S .065}
J1/S	{J1/S .037, J1/L .606, J1/M .357}
J2/S	{J2/S .168, J2/M .683, J5/S .150}
J3/S	{J3/S, 1.000}
J4/S	{J4/S, 1.000}
J5/S	{J5/S .112, J2/S .112, J5/M .789}

Table 3.5 – Revised MLE Estimates (Accidents / Million Departures)

	Jurisdiction 1	Jurisdiction 2	Jurisdiction 3	Jurisdiction 4	Jurisdiction 5
Large	0.554	0.356	N/A	2.889	2.019
Medium	0.554	1.887	4.305	2.804	2.837
Small	0.554	2.216	17.852	9.237	2.850

Unlike schemes that rely on arithmetic functions, this scheme does not force certain relationships to hold across jurisdictions or across size categories. For instance, while the frequency of small airlines is nearly four times that of medium size airlines in Jurisdiction 3, the difference is not nearly as pronounced in other jurisdictions. In fact, in jurisdiction 1, small, medium, and large airlines all have the same accident frequency.

The current classification scheme is one of many that could have been devised using our classification procedure. It should be compared to others to decide which is the most efficient. The notion of efficiency is addressed in [6, p 98].

3.3) Trend

Has the rate of accident changed overtime? If so, how has it changed? The answer to these questions has implications on how the accident rate is projected into the future. Other questions come up as well. Should we look at the change in the rate of accident over the entire cell universe or should we only be concerned with changes within individual cells or within individual classes? Can we even examine the issue of trend independently of that of classification? Should we look at time as one more variable in the classification scheme or do we examine changes over time after the scheme itself has been established? Isn't it possible for some cells to show improvement in frequency overtime while others show deterioration or no change in their accident frequency? Isn't it also possible for the accident rates of two cells or two classes to converge or diverge over time? We do not pretend to have the answers to all these questions. For now, we will look at the question of trend as a one-dimensional problem. In order to measure a trend pattern over time, we first need to specify a model as to how the frequency rate is changing over time. Just as importantly, we have to be able to estimate the parameters of the model and specify the distribution of these parameters.

We now use data from the National Transportation Safety Board shown in table 3.7 below to illustrate how a trend can be estimated and how we can make inferences about the significance of such trend. We will look at a linear and an exponential decay models described, respectively, by equations 1 and 2 below:

$$\hat{\lambda}_i = f_1(y_i) = \alpha + \beta y_i,\ i = 1,2,\cdots,n \qquad \text{Equation 1}$$

$$\hat{\lambda}_i = f_2(y_i) = \alpha' + \beta' e^{\delta'(y_i - y_1)},\ i = 1,2,\cdots,n \text{ where } \delta' \le 0, \alpha' \ge 0 \qquad \text{Equation 2}$$

Other than familiarity perhaps, there is not much rationale for choosing a linear model[12]. We show it here only because its use is so pervasive in casualty actuarial practice. In fact, the NTSB data below will show the potential fallacy of using a linear trend model. The exponential decay model is used commonly in biology and the rationale for its use in biology is applicable in the context of accident frequencies. We can think of the frequency as being made up of two components. The first one, represented by α', is the fixed, intrinsic, or ultimate portion of the accident rate due to a type of error that cannot be eliminated over time whereas the second part $\beta' e^{\delta y_i}$ is the variable portion of the accident rate due to the type of error that changes (decays) over time perhaps as a result of technological advances.

We know that the $\hat{\lambda}_i$'s do not have a constant variance[13]. Therefore, it would be inappropriate to estimate the parameters of equations 1 and 2 by using the ordinary least square function. Since the variance of each $\hat{\lambda}_i$ is inversely proportional[13] to the number of exposures d_i, we instead minimize the following weighted sum of square function: $WSS = \sum_{i=1}^{n} d_i(\hat{\lambda}_i - f_j(y_i))^2$, and we obtain the following weighted least square estimates for models 1 and 2:

[12] Even if a linear model is used, the parameters and their associated errors should not be estimated using simple linear regression. The assumptions of normality and of uniform variance that underlie the simple linear regression model do not hold for the $\hat{\lambda}_i$'s.

[13] $Var(\hat{\lambda}_i) = Var(\frac{x_i}{d_i}) = \frac{1}{d_i^2} Var(x_i) = \frac{\lambda_i d_i}{d_i^2} = \frac{\lambda_i}{d_i}$

Model 1: $\hat{\alpha} = 42.66$ and $\hat{\beta} = -.0212$.

Model 2: $\hat{\alpha}' = 0$, $\hat{\beta}' = .749$, $\hat{\delta}' = -.035$

Figure 3.5 shows the graph for the actual and least square estimates of models 1 and 2. Both models give similar results in the 1982 through 2000 period. However, the models diverge significantly beyond that period. Extrapolation from either model has to be done carefully and should not go out more than a couple of years. However, some situations may call on the modeler to extrapolate over a longer time horizon. For these situations, the exponential decay model may be more appropriate than the linear model. The indications from the linear model are counterintuitive in the long run as they indicate a negative frequency by the year 2018. For the NTSB data, the estimate of the ultimate frequency α' is zero. The indications from the exponential decay model taper off much more slowly and never quite reach zero. Taken at face value, this would be encouraging news for the probability of major accidents in the future.

In appendix B, we show the closed form formula for the weighted least square estimates $\hat{\alpha}$ and $\hat{\beta}$ for model 1 and we also show they are unbiased estimates of α and β. The weighted least square estimates of model 2 are obtained through numerical methods and no closed form formulas are available. What is the distribution of these parameter estimates? Are there statistics that can help us make inferences about the significance of these estimates? What is the error associated with the forecast based on these estimates? We will have to research further for answers to these questions.

Table 3.7 - Major Accident by NTSB Classification for US Air Carriers Operating under CFR 121[14]

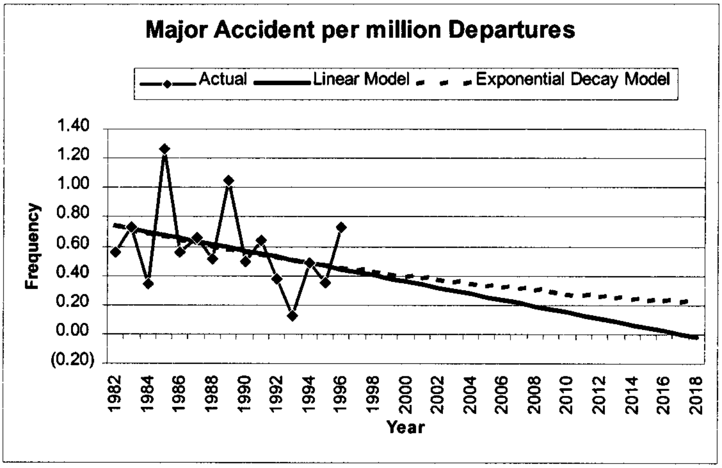
Year	Million of Departures	Major Accidents	Frequency per Million Departures
1982	5.35	3	0.56
1983	5.44	4	0.73
1984	5.90	2	0.34
1985	6.31	8	1.27
1986	7.20	4	0.56
1987	7.60	5	0.66
1988	7.72	4	0.52
1989	7.65	8	1.05
1990	8.09	4	0.49
1991	7.81	5	0.64
1992	7.88	3	0.38
1993	8.07	1	0.12
1994	8.24	4	0.49
1995	8.46	3	0.35
1996	8.23	6	0.73

[14]Source: Departures www.ntsb.gov/aviation/Table5.htm; Major Accidents www.ntsb.gov/aviation/Table2.htm.

The NTSB defines a major accident as one that meets any of the following three conditions:
a Part 121 aircraft was destroyed, or
there were multiple fatalities, or
there was one fatality and a Part 121 aircraft was substantially damaged.

The NTSB provides data through the 2001 year. However, starting in 1997, aircrafts with 10 or more seats used in scheduled passenger service began operating under 14 CFR 21. We did not want to analyze the data beyond 1996 as we were not sure whether the inclusion of this new category of aircrafts would distort the indicated trend.

Figure 3.4 – Accident per million departures 1982 – 1996



3.4) Modeling the number of aircrafts involved in accidents

Airline catastrophes may be the result of collisions involving several aircrafts. Most accidents involve the failure of a single aircraft. Accidents involving collision of multiple airplanes are relatively rare. However, these types of accidents have occurred and need to be reflected in the simulation model. It would be a mistake not to provide in the model for the possibility of two, three, and perhaps more airplanes being involved in a single collision. We are not talking about collisions triggered by willful acts of sabotage, war, or terrorism. Accidents caused by willful acts will be discussed in section 6. Multiple aircraft collisions can put serious financial strains on the insurers and reinsurers who are responsible for indemnifying the airlines. In addition to the probability distribution of the number of accidents, the modeler needs to specify a conditional distribution for the number of aircrafts involved once there has been an accident. The parameters of this distribution might need to be derived from a fair amount of judgment. A conditional probability distribution table for the number of aircraft involved in an accident is shown in table 3.8 below:

Table 3.8 - Conditional probability for the number of aircrafts involved in an accident

Number of Aircrafts	Probability
1	.97
2	.02
3	.01
4 or more	.00

If an accident is the result of a collision of multiple aircrafts, one also needs to determine the airlines and aircrafts involved. The modeler essentially needs to build yet another conditional probability table laying out the probability of collision between different airlines and aircrafts. Airlines that use a lot of the same airports are more likely to be in a collision than those that use few common airports. Hence, intuitively, these probabilities should be proportional to the proximity of the operation of the airlines and to their relative exposures (say, number of departures). This may seem like a daunting task given the low probability of such events and especially given that the exact identity of the other airlines involved may not be of interest in many applications. However, such table can be greatly simplified by making some broad assumptions. For instance, looking at the probability of two-aircraft collisions for a given airline, one may simply endeavor to compute the conditional probability of collisions involving only aircrafts from that airline. The complement of that probability would be the probability of collisions involving an aircraft from the given airline with one from any other airline.

4) Modeling the cost of catastrophes

The financial costs of airline accidents can be staggering and wide-ranging, affecting individuals, small business entities, corporations, and indeed entire financial markets. The insured portion of these costs is in principle bounded by the parameters of the insurance contract. It is this portion only that we hope to forecast. Here we discuss how to estimate the costs associated with coverage for hull, passenger liability, third party liability, and products liability. However, this model could be used to forecast the costs of other types of coverage such as accident and health, workers compensation, and cargo, for instance, and virtually any financial product where payment is dependent on catastrophic airline accidents.

4.1) Hull Costs

Once an accident has occurred, the hull costs are determined relatively quickly. The insured value of an airline's fleet is pre-determined by the insurance contract. The latter provides for a schedule of insured values for each aircraft in a fleet. The first four columns of table 4.1 below show a typical schedule of aircraft and insured values. This information often does not trickle down to reinsurers, retrocessionaires, or even to some of the smaller primary markets perhaps because it is not used in the rating process. Total fleet value, which is the aggregate of the insured values of individual aircrafts, is usually available but this information is mostly useless in the context of this type of simulation model. In order to accurately forecast the hull cost, the modeler needs to have some idea of the fleet and utilization profile[15] of the airline involved as well as the pre-agreed insured values. It behooves the modeler to make sure this information is obtained. The fleet and utilization profile of an airline is typically public information that can be obtained from the airline's website or from airline industry publications or regulatory agencies. However, insured values need to be obtained through insurance channels. As a substitute for actual insured values, one could estimate the hull cost using the price of a new similar aircraft (ballpark numbers are available from the manufacturers) and factor in a discount based on the aircraft age and configuration. This approach adds a layer of uncertainty in an area where there should be none.

Once the fleet distribution and utilization profile for a given airline is known, the conditional probability of a particular aircraft being involved in an accident can be determined. For instance, the conditional probabilities can simply be calculated as the ratio of each aircraft's projected number of departures to the total number of departures. One may want to factor in the age and model of aircraft in the determination of the conditional probabilities if one believes that these impact the probability of accidents. However, we will work from the basic premise that, for a given airline, the conditional probability of accident for a given aircraft is proportional to its utilization. The fleet and distribution profile, the hull values for a hypothetical airline are shown in table 4.1 below. The conditional probabilities are calculated as indicated above.

[15] The utilization profile refers to the number of hours flown or the number of departures within a period of time for each aircraft within a fleet.

4.1- Fleet, Utilization Profile, and Seating Capacity for a Hypothetical Airline

Aircraft Make and Model	Registration Number	Number of Seats excluding Crew Members	Insured Value	Projected Utilization (# Departures)	Conditional Probability
A-300-600	XXXXXX	298	118	739	.74%
B-717	XXXXXX	106	40	1,219	1.22%
B-727	XXXXXX	149	60	2,147	2.15%
Total				100,000	100%

4.2) Passenger Liability

There are two important variables in determining the total cost of passenger liability in the event of an accident. The first is the number of passengers involved in an accident while the second is the award per passenger.

4.2.1) Forecasting the number of passengers, survivors, and fatalities involved in an accident

The number of passengers involved in an accident depends on the seating capacity of the aircraft model involved, and the percentage of capacity filled. The seating capacity of each aircraft in a fleet is available in the schedule of aircraft and insured values. If the seating capacity of a given aircraft is not available, one can use the average seating capacity for that specific aircraft model, which can be obtained from many different sources. The actual seating capacity for a given aircraft model may vary based on the specific configuration for that aircraft. The larger the business class and first class sections, the smaller the overall seating capacity. The other important factor in determining the number of passengers involved in an accident is the passenger load factor, which is available through various airline industry publications. The modeler – having determined the aircraft model and therefore the passenger capacity involved in an accident – may use either a fixed or a random passenger load factor to determine the number of passengers on board the aircraft. The modeler may create a simple distribution based on the published load factor and an upper bound of 100%. For instance, if the published load factor is 85%, one might use a uniform distribution with lower and upper bound of 70% and 100%, respectively.

One may also be interested in projecting the number of survivors/fatalities in a given accident. For that, one could look at the survivability statistics for the type of accidents one is investigating and derive a probability distribution for the percentage of survivors/fatalities. That distribution will have a domain bounded by 0% and 100% and will likely be heavily weighted towards these two points. Survivability data can be obtained by reviewing fatality/survival ratios of individual accidents. In table 7.3 below, we show survivability statistics based on data published by the National Transportation Safety Board.

4.2.2) Cost per passenger

The determination of liability cost is more complicated and more involved than that of the number of passengers or injuries involved in an accident. If one tries to simulate the liability cost for each passenger, one has to know the jurisdiction in which compensation will be sought and information regarding the passenger including place of residence, age, marital status, current and projected net worth. Indeed, in the United States for instance, liability awards stemming from a given accident may vary significantly from one passenger to the next. This level of passenger profile detail is not only difficult to obtain but might be unnecessary in most applications. This information will only be relevant if the coverage depends on individual passenger payout such as a layer offering per passenger excess of loss protection. If we focus rather on forecasting the average liability cost per passenger rather than the actual award per passenger, the overriding consideration is the jurisdiction and the applicable laws under which compensation is sought. Such laws are complex, numerous, and constantly evolving. Accidents involving international flights are especially challenging, as even the jurisdiction in which compensation is sought is hard to determine. For instance, under what jurisdiction will compensation be sought in an accident occurring over Canadian land on a flight from New Delhi to New York with a stop in London? Passengers may have different recourses depending on their nationality, the place they purchased their ticket, their final destination on the trip. The modeler has to make some simplifying assumption as to which jurisdiction will be involved in the event of an accident on an international flight. For instance, the modeler may decide that the country on the itinerary with the higher award potential is the country in which suit will be brought. This is especially important for airlines where the liability awards are much smaller in their country of domicile than in some of their other destinations. For these airlines, the liability potential on domestic flights will be significantly lower than the potential on international flights. For accidents involving US airlines wherever occurring, one may simply assume that all suits will be brought in an American court. The modeler needs to have an idea of a carrier's percentage of domestic versus international flights. For international flights, the modeler needs to know the destination profile by country or group of countries. In turn, this information will be used to simulate the itinerary of a flight involved in an accident.

Does the modeler then need to know the distribution of the average liability award in every possible jurisdiction or country? This would be a daunting task even for the most industrious modeler. The modeler may instead look to group countries where the tort and compensation systems are similar. For instance, a group might be comprised of countries in the European Community, another of Mercosur countries[16]. Also, instead of looking at the distribution of actual average award, the modeler may choose instead to look at the average award as a ratio to the median income or the income per capita in a country. Assuming that the modeler can come up with such groups, there remains the challenge of using the historical data to come up with the average award distribution. Since liability claims can take an inordinate amount of time to settle, an average award distribution will need to be built largely on case estimates, the accuracy of which won't be known for a long time. For those claims that have already been settled, inflation, changes in law, voluntary agreements, and statutes may render them less relevant for the purpose of projecting the cost of future claims. The upshot of all this is that an average liability award distribution will involve significant judgment on the part of both the modeler and others. Once groups of countries have been defined, the modeler may try to find a distribution that fits the actual data adjusted for inflation and for past and expected future changes. Similarly to the frequency portion of the model, should the modeler devise some statistical tests to help him decide whether the distribution of liability awards (as a percentage of, say, median income) for the various groups are indeed dissimilar? For instance, upon close analysis, it may turn out that the average award potential in the European Community is not dissimilar to that in the Mercosur countries.

4.3) Third Party Liability

Third party liability costs are even more difficult to forecast. The range of scenarios for third party liability is obviously much wider than that for passenger liability. Also, relatively few commercial airline accidents result in injury or property damage to third parties. This is perhaps because the location of airports and the air routes have tended to steer airplanes away from populated areas. However, these events have occurred and need to be considered in the simulation model. One approach might be to look at the passenger and third party liability together. So instead of looking at the distribution of the average passenger liability per passenger as suggested in the preceding section, one would look at the average total liability per passenger. The tail of that distribution would be a lot more skewed than that of the average passenger liability award. Another approach is to estimate the number or percentage of accidents that will result in third party damage and to estimate the cost of such liabilities separately. This approach is better at allowing the modeler to factor in extreme scenarios. For instance, the modeler might include a scenario where, as a result of a midair collision, two jumbo airplanes plow onto a crowded area destroying life and property. There, considerable judgment might be used to determine the likelihood of different scenarios.

4.4) Products Liability

Defendants in lawsuits stemming from aircraft accidents include not only the airlines but also aircraft and parts manufacturers as well as other parties involved in the operation of the airline. Also, the airlines themselves can try to recoup losses by suing other parties not necessarily named in a suit. For this reason, aircraft and parts manufacturers require products liability to shield them from such suits. To understand the products liability exposure in the context of this simulation model, information about manufacturers and suppliers of engine, navigation equipment, electrical system and other components has to be collected for each insured aircraft. The identity of the aircraft manufacturer itself should be obvious. In the event of an accident, we would then have a list of potential defendants. We have

---

[16] Mercosur countries, as of the time of this writing, are made up of Agrentina, Brazil, Uruguay, and Paraguay.

already shown how the total liability from a given accident can be estimated. We then need to allocate that liability between airline operators and the product manufacturers. The actual allocation of liability will depend very much on what is determined to be the cause of accident, and, furthermore, may vary from one jurisdiction to the next. This is a very difficult area in need of much research. Considerable judgment or simplification might be used to come up with an allocation. In order to figure out the product liability exposure of a given manufacturer, we would then accumulate its exposure over the entire universe of airline operators.

5) Validation

Before using any model to forecast, the modeler needs to make sure that the model's assumptions are reasonable. The particular model we have presented makes many assumptions, one nested inside another. To develop any sense of how well the model will forecast the future, one can look at how well the model would have predicted past experience. Let's say that data from 1980 through 2000 is available. The modeler may endeavor to see how well the simulation model would have projected 1991 based on data through 1990, 1992 based on data through 1991, and so on, thus obtaining a comparison with actual data for ten years. The validation should be done in stages, starting with a look at the number of accidents, the number of passengers involved in accidents, the number of fatalities and injuries, and the insurance costs in that order. Doing the validation in stages allows one to identify where in the simulation process a bias may be occurring and to make the necessary adjustments. The results of the simulation will be a probability distribution for the projected variables similar to the one shown in table 7.5 of section 7 below. Focusing on the number of fatalities for instance, one would compare the actual number of fatalities in 1991, 1992 and subsequent years with the distribution predicted for each of these years by the simulation model. If, on one hand, the actual number of fatalities for the ten-year sample (1991 through 2000) looks like a random draw from each of the predicted distribution, this would tend to validate the simulation model. If, on the other hand, the actual number of fatalities for the ten-year sample tends to fall systematically either to the right of, say, the 95th percentile or to the left of, say, the 5th percentile of the predicted distributions, this would be a strong indication that the model's projections are biased.

6) Terrorism

The audacity and severity of the events of September 11th caught most insurance professionals off guard. Reportedly, aviation underwriters had been "giving away" for free the coverage for terrorist acts. In light of the devastating potential of terrorist acts, actuaries and underwriters have since been scrambling to put a price on terrorism coverage. While some pundits will offer nothing more than their eloquent prose enlightening us with revelations such as "...the risk [of terrorism] is real", underwriters and actuaries are left with the unenviable task of putting a price on the risk of terrorism. Perhaps, in no other area will the actuary need to use all available sources of information and rely on the expertise of others in order to try to quantify the risk of terrorism. Many of the assumptions we have made in relation to accidental airline catastrophes certainly don't apply to crashes occurring as a result of willful acts. We know that terrorist acts are neither random nor are they uncorrelated. This contradicts the assumptions implicit in our use of the Poisson distribution. We also know that history is perhaps a poor guide for figuring out future acts. The risk of terrorism is highly fluid as our geopolitical landscape changes constantly, and as airlines and law enforcement authorities learn how to better protect the public from such acts. In our preceding discussion, we rely extensively on historical data to derive expected frequencies. We could not do the same with terrorism although a look at the history can be instructive. Table 6.2 shows the number of hijackings perpetrated against US and foreign airlines, respectively, from 1970 through 2000. The sharp drop in the number of hijackings against US airlines, with none recorded between 1992 and 2000, testifies perhaps to the success US airlines and authorities have had in deterring and preventing such acts. The events of September 11th, 2001 serve as a staunch reminder that the probability of such acts is never quite zero. Although, we have focused here on hijackings, they are by no means the only terrorist threat facing airlines. We should also keep in mind that not all hijackings have resulted in death, injury, or destruction of property as people have sought to hijack airplanes for a variety of reasons. Hijackings need not be the result of some political conflicts. For instance, mentally deranged individuals with no apparent political motives have hijacked airplanes.

Although individual airlines may have their own security procedures, the modeler can work from the assumption that the risk of terrorist acts against an airline depends primarily on the level of security in the airports and countries where the airline operates. Actuaries could work with security experts in developing a grouping system to rate airports' and countries' potential for terrorist acts. The level of conflict in a country and its surrounding regions as well as a country's will and ability to effectively fight terrorism need to be factored in such a grouping. Actuaries could then try to formulate a probability of an airline being hit by a terrorist for each category in a grouping. For instance, table 6.1 below shows a hypothetical two-dimensional box which groups airports based on their level of security and the existence of a terrorist threat around them. The probability of an airline's being hit by a terrorist act would be calculated as the weighted average of the probabilities of each airport where it has exposure.

Table 6.1 – Probability (odds in 1 million) of a Terrorist Act in a 12 Month Period

Security / Terror Threat	Impenetrable	Strict	Adequate	Lax	Non-existent
Constant	10	100	250	1,000	10,000
Potential	4	50	200	800	9,500
Some	2	20	150	700	8,000
None	1	5	10	500	6,000

As we mentioned before, these probabilities are dynamic and should change as conflicts evolve or new ones emerge, as new information comes to light, and as new acts of terrorism are committed or attempted. This implies that the price for such coverage will, at least in theory, be dynamic, changing as the risk of terrorist acts is continuously reassessed.

Table 6.2 - Hijackings

Year	U.S.	Foreign	Year	U.S.	Foreign	Year	U.S.	Foreign
1970	25	49	1981	7	24	1992	0	12
1971	25	30	1982	9	23	1993	0	31
1972	26	30	1983	17	15	1994	0	23
1973	2	20	1984	5	21	1995	0	9
1974	3	17	1985	4	22	1996	0	14
1975	6	13	1986	2	5	1997	0	10
1976	2	14	1987	3	5	1998	0	9
1977	5	26	1988	1	10	1999	0	11
1978	7	16	1989	1	14	2000	0	20
1979	11	13	1990	1	39			
1980	21	18	1991	1	23			

Sources: 1970-1998 U.S. Department of Transportation, Federal Aviation Administration, Criminal Acts Against Civil Aviation - 1998, Charts and Graph; 1999-2000 http://cas.faa.gov/crimacts/doc/crim2000.doc, Appendices A&B

7) A simplified application of a simulation model

A simulation model has many applications for assessing the costs of insurance coverages and other financial instruments affected by airline catastrophes. Here, we present an example where we look at the costs of a cover that pays for the full insured value of a destroyed or damaged aircraft as well as $50,000 per fatality and $100,000 per injured passenger for a hypothetical group of airlines for the 2003 year. This coverage excludes acts of war and terrorism. The information and assumptions are set in the tables 7.1 through 7.4. The results of the simulation are presented in table 7.5.

Table 7.1 - Fleet, Utilization Profile, and Seating Capacity

Aircraft Type	Count	Seats	Insured Value (MM)	# of Departures	Prob
Airbus Industrie A300-600	79	298	118	58,390	0.69%
Airbus Industrie A300B2/B4	19	298	118	6,165	0.07%
Airbus Industrie A310	44	249	92	22,253	0.26%
Airbus Industrie A319	137	125	52	148,736	1.77%
Airbus Industrie A320	227	172	55	287,301	3.42%
Airbus Industrie A380	6	600	250	4,729	0.06%
Avro RJ Avroliner	36	70	26	71,171	0.85%
BAE SYSTEMS (HS) 146	18	94	40	47,420	0.56%
Boeing (McDonnell-Douglas) DC-10	239	264	110	123,748	1.47%
Boeing (McDonnell-Douglas) DC-8	194	146	60	92,469	1.10%
Boeing (McDonnell-Douglas) DC-9	430	115	50	684,705	8.15%
Boeing (McDonnell-Douglas) MD-11	66	325	150	41,923	0.50%
Boeing (McDonnell-Douglas) MD-80	670	155	60	1,056,052	12.57%
Boeing (McDonnell-Douglas) MD-90	21	163	60	34,382	0.41%
Boeing 717	31	106	40	37,779	0.45%
Boeing 727	729	167	60	715,326	8.51%
Boeing 737 (CFMI)	779	149	60	1,672,505	19.90%
Boeing 737 (JT8D)	248	149	60	518,652	6.17%
Boeing 737 (NG)	303	149	60	372,020	4.43%
Boeing 747 Classic	138	472	215	66,097	0.79%
Boeing 747-400	73	544	211	37,837	0.45%
Boeing 757	589	214	85	728,728	8.67%
Boeing 767	328	251	110	274,901	3.27%
Boeing 777	98	373	190	60,625	0.72%
Bombardier (Canadair) CRJ Regional Jet	273	50	25	543,579	6.47%
Embraer ERJ-135	53	36	14	64,287	0.76%
Embraer ERJ-145	169	50	20	274,355	3.26%
Fairchild/Dornier 328JET	22	30	13	20,837	0.25%
Fokker 100	123	113	50	234,759	2.79%
Fokker F.28	22	85	40	53,688	0.64%
Lockheed L-1011 TriStar	81	280	110	48,412	0.58%
Total	6,245	1,108,002	450,332	8,403,831	

Table 7.2 - Projected Exposures, Frequencies, and Passenger Loads

Year	2003
Projected Departures	8,918,213
Projected Average Passenger Load	0.65
Expected Frequency per million departures	0.45

- Distribution of survival ratio based on Beta distribution fitted to data in table 7.3 below:

Table 7.3 - Empirical Survival percentages from a sample of 27 accidents

Passengers	Survivors	Fatalities	% of Survivors
31	0	31	0.0%
25	0	25	0.0%
132	0	132	0.0%
68	0	68	0.0%
110	0	110	0.0%
230	0	230	0.0%
155	1	154	0.6%
71	1	70	1.4%
163	29	134	17.8%
57	20	37	35.1%
51	24	27	47.1%
296	185	111	62.5%
82	54	28	65.9%
89	67	22	75.3%
44	36	8	81.8%
108	94	14	87.0%
145	134	11	92.4%
33	32	1	97.0%
149	147	2	98.7%
142	142	0	100.0%
39	39	0	100.0%
23	23	0	100.0%
40	40	0	100.0%
102	102	0	100.0%
292	292	0	100.0%
62	62	0	100.0%
2,739	***1,524***	***1,215***	***55.6%***

Source: National Transportation Safety Board, "Survivability of Accidents Involving Part 121 U.S. Air Carrier Operations, 1983 Through 2000" Safety Report NTSB/SR-01/01, table 4, p. 14. http://www.ntsb.gov/Publictn/A_Stu.htm .

Table 7.4 - Conditional Probability for the Number of Aircrafts Involved in an Accident

Number of Aircrafts	Probability
1	.970
2	.029
3	.001
4 or more	.000

Additional Assumptions

- Destroyed aircraft is immediately replaced with similar aircraft.
- Projected departures, passenger loads and frequencies are fixed. More realistically, these should be allowed to vary.
- Collision occurs between aircrafts within group.
- Selections of parameters only loosely based on real data

Table 7.5 - Simulation Results from 5000 iterations

Variables	Accident Count	Aircraft Count	Passger Count	Injured Count	Death Count	Hull Cost (MM)	Passger Cost (MM)	Total Cost (MM)
Best Case	0	0	0	0	0	0	0	0
Wrst Case	13	14	1,790	1,015	1,012	1,131	1,376	2,507
Expected	3.59	3.69	375	197	179	233	286	519
Std Dev	1.91	1.99	230	154	144	139	182	315
5%	1	1	58	0	0	50	42	95
10%	1	1	109	18	8	60	76	149
15%	2	2	144	43	30	95	105	204
20%	2	2	176	65	51	115	128	248
25%	2	2	204	81	70	120	150	283
30%	2	3	231	97	86	145	172	322
35%	3	3	259	114	103	170	194	363
40%	3	3	287	131	121	180	214	398
45%	3	3	317	148	135	195	236	434
50%	3	4	347	167	150	216	258	474
55%	4	4	377	184	168	230	282	518
60%	4	4	407	209	190	252	305	559
65%	4	4	440	230	212	270	333	604
70%	4	5	474	254	233	290	364	652
75%	5	5	509	281	258	315	396	714
80%	5	5	554	314	287	345	430	772
85%	6	6	614	355	325	375	473	845
90%	6	6	681	407	371	415	531	932
95%	7	7	792	488	447	485	627	1,089

The distribution presented in table 7.5 above shows the variability in the loss process but does not incorporate parameter error. We have assumed for instance that the claim process follows a Poisson distribution with an expected frequency per million departures of .45. In reality, we will never know the true expected frequency of such distribution or the exact form of the distribution for that matter. At best, we will have an estimate of the frequency with an error margin. The conditional probability distribution for the number of aircrafts involved in an accident, the distribution of the passenger load factors and of the survival ratios are all subject to parameter error. One way to incorporate the parameter uncertainty into the results would be to allow the parameters to vary according to some distribution. A different but simpler approach might be to look at different combinations of parameter estimates and to run the simulation for each combination.

Applications for this type of simulation model extend much beyond the type of examples we have just presented. We have only scratched the surface of the possibilities that a simulation approach offers. The CAS's literature is replete with articles on how simulation models could be used to structure and price reinsurance products. Please see [1], [2], and [4] for a sample of such articles. Obviously the more detailed information available to the modeler, the more accurate the projections will be and the more specific the applications will be for this type of model. Modelers have to weigh the trouble of gathering the additional data against the additional accuracy and flexibility that would be gained.

8) Final Thoughts

Compared to a traditional experience rating approach, a simulation approach promises much more in terms of accuracy and range of applications. For instance, the pricing of reinsurance contracts that feature loss triggers, aggregate limits and deductibles, contingent profit will be more readily handled through simulation than through experience rating. The body of available actual experience would often not suffice to test the multitude of scenarios that can present themselves under such contracts. For the cover we introduced in the preceding section, assume we were interested in pricing an aggregate layer providing 1.0B in limit in excess of a 1.0B retention. According to the aggregate loss distribution in table 7.5 above, there is about a 5% chance that actual aggregate losses would exceed the 1.0B retention in any one year. Looking at the actual experience, over a five to ten year period, may not reveal any losses in the layer. Even when there would be losses in the layer, they may not have much predictive value.

Many of the benefits of using a stochastic simulation approach in the evaluation of property catastrophe extend to catastrophic airline exposures as well. Rade T. Musulin, in an article titled "Issues in the Regulatory Acceptance of Computer Modeling for Property Insurance Ratemaking", [3, p 354] lists the comprehensibility of prices, rational behavior, fair pricing, reduced information risk, and stable pricing as benefits of the improved estimates provided by a simulation model in evaluating property catastrophe exposures. It should be readily apparent how the use of a simulation approach in evaluating airline catastrophe would enhance the comprehensibility of prices, reduce information risk, and promote stable pricing.

Ultimately, whether simulation models gain acceptance in the commercial aviation realm will depend on whether the perceived benefits outweigh the additional effort required in implementing such models. In a sense, the widespread use and acceptance of property catastrophe modeling should have already paved the way for the use of simulation models not only in commercial aviation but also in other lines such as surety, credit, and workers compensation. These days, insurers and reinsurers have the means necessary to keep large amount of information about their property exposures at a zip code, and sometimes, finer level. More importantly, this level of detailed information is accepted as a normal course of doing business in the property catastrophe lines. The task of gathering information on individual aircrafts is relatively small due to the limited number of commercial airlines servicing the world and also the limited number of aircrafts they operate. For instance, the current fleet of US-domiciled airlines consists of less than 6,500 aircrafts of roughly 30 different models. Furthermore, two companies, Boeing and Airbus, manufacture 85% of these aircrafts. Once one goes through the trouble of tallying that information, subsequent updates should be relatively simple as airlines do not change their fleet drastically overnight.

We have left open a number of issues including inferences about the parameters of the exponential decay trend model. We also think that a substantial amount of work has to be done to make realistic projections for third party liability exposures and in the allocation of liability between airline operators and manufacturers. Finally, we have barely broached the issue of terrorism. However, we are confident that there will be a wealth of papers addressing these issues in a much more comprehensive fashion.

REFERENCES

[1] Bernes, Regina M, "Reinsurance Contracts with a Multi-Year Aggregate Limit", CAS Forum, Spring 1997.

[2] Daino, Robert A; Thayer, Charles A, "Comparing Reinsurance Programs – A Practical Actuary's System", CAS Forum, Spring 1997.

[3] Musulin, Rade, T, "Issues in the Regulatory Acceptance of Computer Modeling for Property Insurance Ratemaking", Journal of Insurance Regulation, Spring 1997.

[4] Papush, Dmitry E, "A Simulation Approach in Excess Reinsurance Pricing", CAS Forum, Spring 1997.

[5] Patzor, Andy, "Bumpy Ride", The Wall Street Journal (September 19, 2000), p A1.

[6] Salam, Romel G, "Reinventing Risk Classification – A Set Theory Approach", CAS Forum, Winter 2002.

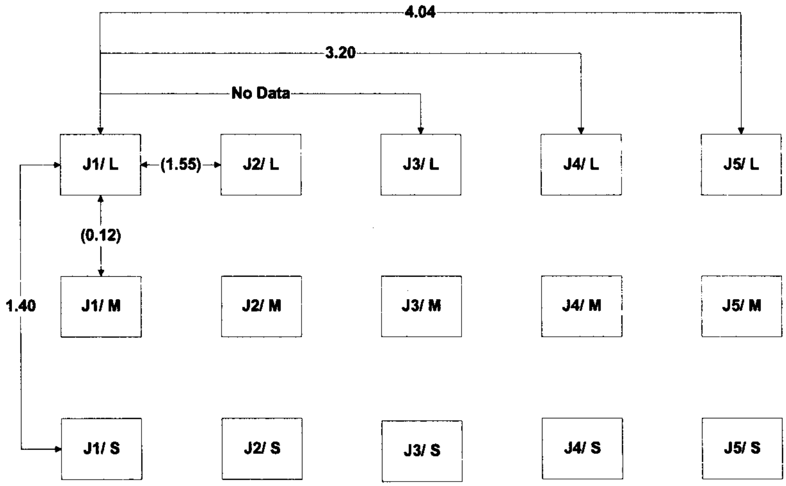
Appendix A

Risk Classification Procedure

Following the procedure introduced in "Risk Classification – A Set Theory Approach" and at the presentation of the paper at the winter 2002 meeting, we take the following steps:

i. We calculate the $\hat{R}_0$ values for all pairs of adjacent cells[1] as shown in exhibit 1. In figure A.1 below, we show the $\hat{R}_0$ values between the cell Jurisdiction 1/Large (J1/L) and the cells that are adjacent to it.

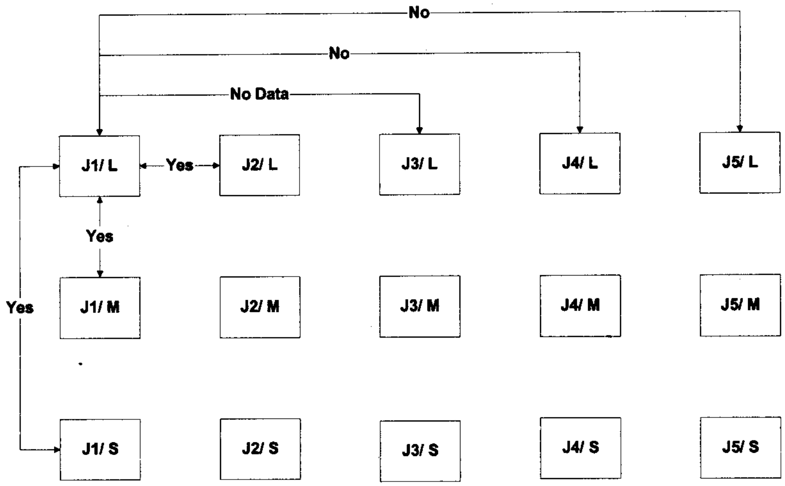
Figure A.1 - $\hat{R}_0$ between large airlines in jurisdiction 1 (J1/L) and those in adjacent cells



Those adjacent cells for which the $\hat{R}_0$ values fall within the interval (-1.65,1.65) are said to be compatible. All other pairs of cells are said to be incompatible. Exhibit 2 shows the compatibility relationship for all pair of cells. Below, in figure A.2, we answer the question of compatibility for large airlines in jurisdiction 1.

[1] Two cells are said to be adjacent if they have at least one common risk characteristic [6, p.89].

Figure A.2 – Are cells compatible with J1/L (before validation)?



ii. The values of $\hat{R}_0$ shown in exhibit 2 and in figure A.2 may have been the result of a chance occurrence or some oddity in the data and may not reflect the true relationship between cells. Rather than relying on just one drawing of $\hat{R}_0$, we repeat the calculation of the $\hat{R}_0$ values for 1,000 randomly selected samples from each cell in order to validate the compatibility relationships. A sample consists of a random draw of between 50% and 85% of the exposures (or airlines) within a cell. If the number of times $\hat{R}_0$ falls in the interval (-1.65, 1.65) for a given pair of cells is large (greater or equal to 875), then, compatibility is validated for that pair. Exhibits 3 and 4 show the number of times $\hat{R}_0$ falls in the interval for all pairs of adjacent cells and whether these cells are deemed compatible, respectively. This information is shown for the cell Jurisdiction 1/Large in figures A.3 and A.4 below.

Figure A.3 - Number of times out of 1,000 trials $\hat{R}_0$ falls in (-1.65, 1.65) interval

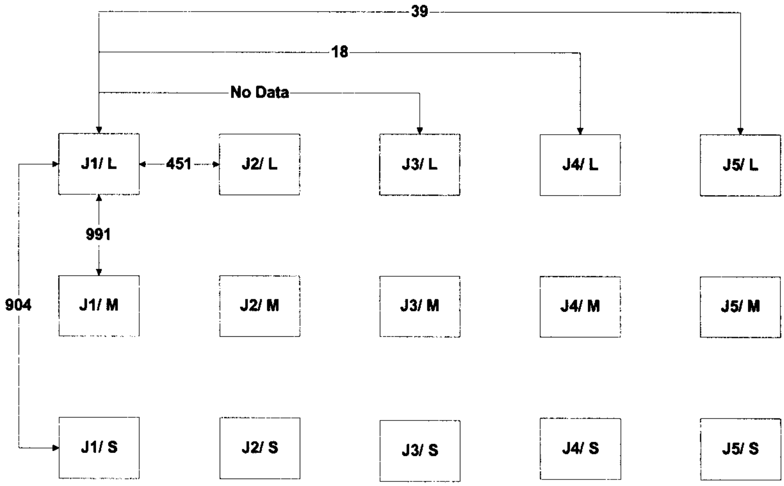
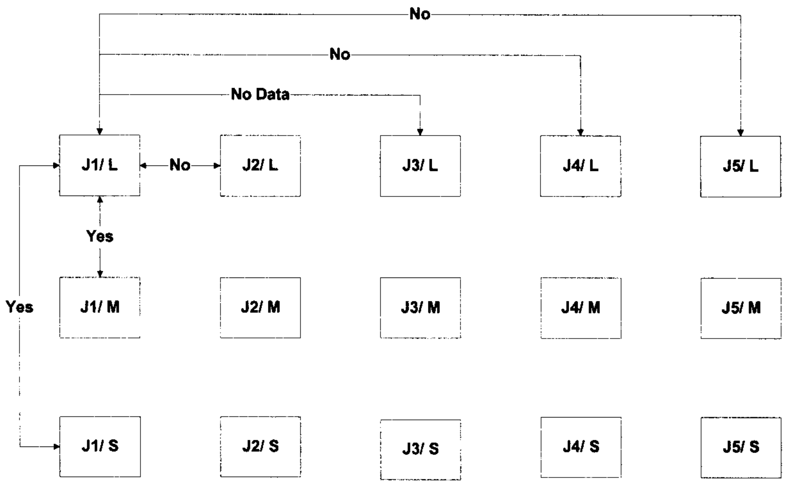


Figure A.4 – Are cells compatible with J1/L (after validation)?



iii. Each cell defines a class made up of these cells that are compatible with it, including the cell itself. In tables A.1 and A.2, respectively, we show the classes defined by each cell and the credibility weights assigned to each cell in a class. The class defined by large airlines in jurisdiction 1 consists of the following three cells: {Jurisdiction 1/Large, Jurisdiction 1/Medium, and Jurisdiction 1/Small}

iv. The revised MLE estimates for each cell is the weighted average of the MLE estimates of the cells in its class where the credibility weights are the exposures in each cell relative to the total exposures for the class.

Table A.1 – Calculation of Revised MLE Estimates for Jurisdiction 1/Large

	Jurisdiction 1 Large	Jurisdiction 1 Medium	Jurisdiction 1 Small	**Total/Weighted Average**
Initial MLE	0.527	.507	1.443	**.554**
Departures	56.93	33.50	3.46	**214.74**
Weights	0.61	0.36	0.04	**1.00**

Table A.2 – Revised MLE Estimates

	Jurisdiction 1	Jurisdiction 2	Jurisdiction 3	Jurisdiction 4	Jurisdiction 5
Large	0.554	0.356	N/A	2.889	2.019
Medium	0.554	1.887	4.305	2.804	2.837
Small	0.554	2.216	17.852	9.237	2.850

Appendix B

Derivation of parameter estimates for a linear trend Model

Let X_i be a Poisson distributed random variable for year y_i, $i = 1,2,\cdots,n$, with mean $\lambda_i d_i$, where d_i represents the number of exposures associated with year y_i.

We posit the following linear relationship between the λ_i's :

$$\lambda_i = \alpha + \beta y_i,\ i = 1,2,\cdots,n$$

Let $\Lambda_i = \dfrac{X_i}{d_i}$ be the random variable representing the maximum likelihood estimate of λ_i.

$$E(\Lambda_i) = E(\frac{X_i}{d_i}) = \lambda_i = \alpha + \beta y_i$$

Also, $Var(\Lambda_i) = \dfrac{\lambda_i}{d_i}$

We denote by x_i and $\hat{\lambda}_i = \dfrac{x_i}{d_i}$ the realizations of the random variables X_i and Λ_i, respectively.

We define the weighted least square error function:

$$WLSE = \sum_{i=1}^{n} d_i(\hat{\lambda}_i - \lambda_i)^2 = \sum_{i=1}^{n} d_i(\hat{\lambda}_i - \alpha - \beta y_i)^2$$

Let $\hat{\alpha}$ and $\hat{\beta}$ represent the values of α and β that minimize the weighted least square function. $\hat{\alpha}$ and $\hat{\beta}$ are obtained as follows:

$$\frac{dWLSE}{d\alpha} = -2\sum_{i=1}^{n} d_i(\hat{\lambda}_i - \alpha - \beta y_i) = 0 \qquad (1)$$

$$\frac{dWLSE}{d\beta} = -2\sum_{i=1}^{n} d_i y_i(\hat{\lambda}_i - \alpha - \beta y_i) = 0 \qquad (2)$$

Solving these two equations simultaneously yields:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} d_i({}_w y - y_i)\hat{\lambda}_i}{\sum_{i=1}^{n} d_i[{}_w y^2 - {}_w(y^2)]} \text{ and } \hat{\alpha} = {}_w\hat{\lambda} - \beta\,{}_w y \text{, where}$$

$${}_w y = \frac{\sum_{i=1}^{n} d_i y_i}{\sum_{i=1}^{n} d_i},\ {}_w y^2 = ({}_w y)^2,\ {}_w(y^2) = \frac{\sum_{i=1}^{n} d_i y_i^2}{\sum_{i=1}^{n} d_i},\ \text{and}\ {}_W\hat{\lambda} = \frac{\sum_{i=1}^{n} d_i \hat{\lambda}_i}{\sum_{i=1}^{n} d_i}$$

Let $\omega_i = \dfrac{d_i({}_W y - y_i)}{\sum_{i=1}^{n} d_i[{}_W y^2 - {}_W(y^2)]}$, we rewrite $\hat{\beta} = \sum_{i=1}^{n} \omega_i \hat{\lambda}_i$.

$\hat{\beta}$ is the realization of the random variable $\mathrm{B} = \sum_{i=1}^{n} \omega_i \Lambda_i$

$$E(\mathrm{B}) = \sum_{i=1}^{n} \omega_i E(\Lambda_i) = \sum_{i=1}^{n} \omega_i \lambda_i = \sum_{i=1}^{n} \omega_i E(\alpha + \beta y_i) = E(\alpha)\sum_{i=1}^{n} \omega_i + E(\beta)\sum_{i=1}^{n} \omega_i y_i = E(\mathrm{B}) = \beta$$

Therefore $\hat{\beta}$ is an unbiased estimate of β.

Also, $Var(\mathrm{B}) = \sum_{i=1}^{n} \omega_i^2 Var(\Lambda_i) = \sum_{i=1}^{n} \omega_i^2 \dfrac{\lambda_i}{d_i}$

Exhibit 1

Calculation of $\hat{R}_0$ between adjacent cells

	J1/L	J2/L	J4/L	J5/L	J1/M	J2/M	J3/M	J4/M	J5/M	J1/S	J2/S	J3/S	J4/S	J5/S
J1/L	-	1.55	(3.20)	(4.04)	0.12	//////	//////	//////	//////	(1.40)	//////	//////	//////	//////
J2/L	(1.55)	-	(3.42)	(4.61)	//////	(3.03)	//////	//////	//////	//////	(2.13)	//////	//////	//////
J4/L	3.20	3.42	-	1.25	//////	//////	//////	0.34	//////	//////	//////	//////	(2.98)	//////
J5/L	4.04	4.61	(1.25)	-	//////	//////	//////	//////	(1.26)	//////	//////	//////	//////	(1.39)
J1/M	(0.12)	//////	//////	//////	-	(2.60)	(4.93)	(4.69)	(5.05)	(1.42)	//////	//////	//////	//////
J2/M	//////	3.03	//////	//////	2.60	-	(3.00)	(1.75)	(1.71)	//////	(0.88)	//////	//////	//////
J3/M	//////	//////	//////	//////	4.93	3.00	-	1.68	1.84	//////	//////	(5.14)	//////	//////
J4/M	//////	//////	(0.34)	//////	4.69	1.75	(1.68)	-	0.14	//////	//////	//////	(3.31)	//////
J5/M	//////	//////	//////	1.26	5.05	1.71	(1.84)	(0.14)	-	//////	//////	//////	//////	(0.91)
J1/S	1.40	//////	//////	//////	1.42	//////	//////	//////	//////	-	(1.00)	(6.30)	(3.91)	(1.67)
J2/S	//////	2.13	//////	//////	//////	0.88	//////	//////	//////	1.00	-	(5.48)	(2.97)	(0.74)
J3/S	//////	//////	//////	//////	//////	//////	5.14	//////	//////	6.30	5.48	-	2.73	4.73
J4/S	//////	//////	2.98	//////	//////	//////	//////	3.31	//////	3.91	2.97	(2.73)	-	2.17
J5/S	//////	//////	//////	1.39	//////	//////	//////	//////	0.91	1.67	0.74	(4.73)	(2.17)	-

Exhibit 2

Are cells compatible (before validation)?

	J1/L	J2/L	J4/L	J5/L	J1/M	J2/M	J3/M	J4/M	J5/M	J1/S	J2/S	J3/S	J4/S	J5/S
J1/L	Yes	Yes	No	No	Yes	//////	//////	//////	//////	Yes	//////	//////	//////	//////
J2/L	Yes	Yes	No	No	//////	No	//////	//////	//////	//////	No	//////	//////	//////
J4/L	No	No	Yes	Yes	//////	//////	//////	Yes	//////	//////	//////	//////	No	//////
J5/L	No	No	Yes	Yes	//////	//////	//////	//////	Yes	//////	//////	//////	//////	Yes
J1/M	Yes	//////	//////	//////	Yes	No	No	No	No	Yes	//////	//////	//////	//////
J2/M	//////	No	//////	//////	No	Yes	No	No	No	//////	Yes	//////	//////	//////
J3/M	//////	//////	//////	//////	No	No	Yes	No	No	//////	//////	No	//////	//////
J4/M	//////	//////	Yes	//////	No	No	No	Yes	Yes	//////	//////	//////	No	//////
J5/M	//////	//////	//////	Yes	No	No	No	Yes	Yes	//////	//////	//////	//////	Yes
J1/S	Yes	//////	//////	//////	Yes	//////	//////	//////	//////	Yes	Yes	No	No	No
J2/S	//////	No	//////	//////	//////	Yes	//////	//////	//////	Yes	Yes	No	No	Yes
J3/S	//////	//////	//////	//////	//////	//////	No	//////	//////	No	No	Yes	No	No
J4/S	//////	//////	No	//////	//////	//////	//////	No	//////	No	No	No	Yes	No
J5/S	//////	//////	//////	Yes	//////	//////	//////	//////	Yes	No	Yes	No	No	Yes

Exhibit 3

Number of times $\hat{R}_0$ falls in (-1.65,1.65) interval

	J1/L	J2/L	J4/L	J5/L	J1/M	J2/M	J3/M	J4/M	J5/M	J1/S	J2/S	J3/S	J4/S	J5/S
J1/L	-	451	18	39	991	//////	//////	//////	//////	904	//////	//////	//////	//////
J2/L	451	-	0	18	//////	38	//////	//////	//////	//////	393	//////	//////	//////
J4/L	18	0	-	824	//////	//////	//////	966	//////	//////	//////	//////	129	//////
J5/L	39	18	824	-	//////	//////	//////	//////	791	//////	//////	//////	//////	861
J1/M	991	//////	//////	//////	-	189	1	4	10	905	//////	//////	//////	//////
J2/M	//////	38	//////	//////	189	-	170	642	760	//////	999	//////	//////	//////
J3/M	//////	//////	//////	//////	1	170	-	583	449	//////	//////	10	//////	//////
J4/M	//////	//////	966	//////	4	642	583	-	904	//////	//////	//////	75	//////
J5/M	//////	//////	//////	791	10	760	449	904	-	//////	//////	//////	//////	932
J1/S	904	//////	//////	//////	905	//////	//////	//////	//////	-	809	0	0	413
J2/S	//////	393	//////	//////	//////	999	//////	//////	//////	809	-	1	127	976
J3/S	//////	//////	//////	//////	//////	//////	10	//////	//////	0	1	-	315	20
J4/S	//////	//////	129	//////	//////	//////	//////	75	//////	0	127	315	-	455
J5/S	//////	//////	//////	861	//////	//////	//////	//////	932	413	976	20	455	-

Exhibit 4

Are cells compatible (after validation)?
Cut off point 875

	J1/L	J2/L	J4/L	J5/L	J1/M	J2/M	J3/M	J4/M	J5/M	J1/S	J2/S	J3/S	J4/S	J5/S
J1/L	Yes	No	No	No	Yes	//////	//////	//////	//////	Yes	//////	//////	//////	//////
J2/L	No	Yes	No	No	//////	No	//////	//////	//////	//////	No	//////	//////	//////
J4/L	No	No	Yes	No	//////	//////	//////	Yes	//////	//////	//////	//////	No	//////
J5/L	No	No	No	Yes	//////	//////	//////	//////	No	//////	//////	//////	//////	No
J1/M	Yes	//////	//////	//////	Yes	No	No	No	No	Yes	//////	//////	//////	//////
J2/M	//////	No	//////	//////	No	Yes	No	No	No	//////	Yes	//////	//////	//////
J3/M	//////	//////	//////	//////	No	No	Yes	No	No	//////	//////	No	//////	//////
J4/M	//////	//////	Yes	//////	No	No	No	Yes	Yes	//////	//////	//////	No	//////
J5/M	//////	//////	//////	No	No	No	No	Yes	Yes	//////	//////	//////	//////	Yes
J1/S	Yes	//////	//////	//////	Yes	//////	//////	//////	//////	Yes	No	No	No	No
J2/S	//////	No	//////	//////	//////	Yes	//////	//////	//////	No	Yes	No	No	Yes
J3/S	//////	//////	//////	//////	//////	//////	No	//////	//////	No	No	Yes	No	No
J4/S	//////	//////	No	//////	//////	//////	//////	No	//////	No	No	No	Yes	No
J5/S	//////	//////	//////	No	//////	//////	//////	//////	Yes	No	Yes	No	No	Yes

Exhibit 5

Classes defined by each cell

Cells	Classes
J1/L	{J1/L, J1/M, J1/S}
J2/L	{J2/L}
J4/L	{J4/L, J4/M}
J5/L	{J5/L}
J1/M	{J1/M, J1/L, J1/S}
J2/M	{J2/M, J2S}
J3/M	{J3/M}
J4/M	{J4/M, J4/L, J5M}
J5/M	{J5/M, J4/M, J5/S}
J1/S	{J1/S, J1/L, J1/M}
J2/S	{J2/S, J2/M, J5/S}
J3/S	{J3/S}
J4/S	{J4/S}
J5/S	{J5/S, J2/S, J5/M}

Exhibit 6

Credibility weights

Cells	Classes
J1/L	{J1/L .606, J1/M .357, J1/S .037}
J2/L	{J2/L, 1.00}
J4/L	{J4/L .278, J4/M .722}
J5/L	{J5/L, 1.000}
J1/M	{J1/M .357, J1/L .606, J1/S .037}
J2/M	{J2/M .803, J2S .197}
J3/M	{J3/M, 1.000}
J4/M	{J4/M .381, J4/L .147, J5M .472}
J5/M	{J5/M .517, J4/M .417, J5/S .065}
J1/S	{J1/S .037, J1/L .606, J1/M .357}
J2/S	{J2/S .168, J2/M .683, J5/S .150}
J3/S	{J3/S, 1.000}
J4/S	{J4/S, 1.000}
J5/S	{J5/S .112, J2/S .112, J5/M .789}

Paid Loss Development of Fixed Size Claims

Daniel R. Corro

Paid Loss Development of Fixed Size Claims[1]

Dan Corro
National Council on Compensation Insurance, Inc.
December, 2002

Abstract: *This paper considers a simple context in which we can quantify the impact of the payment schedule on paid loss development. To isolate the effect of the payment schedule, we restrict to the special case when all claims have the same incurred loss. We consider three simple periodic payment schedules: (1) a uniform payment schedule (2) an escalated (discounted) payment schedule and (3) a schedule that allows a single, fixed proportional adjustment to the payment amount. The paper defines a mathematical model for paid loss development and presents numeric examples to illustrate the sensitivity of paid loss development to the different schedules.*

It is apparent that the payment schedule influences paid loss development. In general a faster (slower) schedule will make losses develop faster (slower). While the direct nature of that relationship is apparent, it is not so apparent how to quantify it. This paper quantifies it in some very particular cases.

Let $S(t)$ denote a survival function on the time interval $(0,b)$.[2] We regard $S(t)$ as a distribution of closure times and let $F(t)=1-S(t)$ be the corresponding cumulative distribution function [CDF]. In effect, all claims are assumed to close on or before time b.

We are interested in a related CDF, which we denote by $\tilde{F}(t)$ to emphasize its relation with $F(t)$, which models the paid loss development as a function of time. More precisely, $\tilde{F}(t)$ is the proportion of total loss paid by time t, i.e. the proportion paid out during $(0,t)$ (without any discount adjustment). $\tilde{F}(t)$ is the reciprocal of the paid to ultimate loss development factor and we will refer to $\tilde{F}(t)$ as the paid loss development divisor [PLDD].[3]

In this note we make two basic assumptions on the size and the payment pattern of each claim:

- The same (undiscounted) amount is paid out on all claims.
- Payments are made continuously from a common the time of loss, $t=0$ to claim closure.

---

[1] The author expresses his thanks to Greg Engl, also of NCCI, who reviewed many versions of this paper, pointed out some serious errors, and made numerous suggestions for improvements.

[2] We are most interested in the case when $b<\infty$ is finite, although most of what we say applies to the case $b=\infty$. We are, however, admittedly rather cavalier about making whatever assumptions are needed to assure that all improper integrals exist and are finite.

[3] Gillam and Couret [4] consider the reciprocal of the loss development factor and call it the loss development divisor.

We consider first the "flat" case when all payments are of the same amount. We then weaken that assumption in a couple ways: first we allow the payments to vary at a constant rate of inflation—this is called a case of "COLA". Second we allow a single fixed proportional change in the payment amount, applicable during the unit of time just prior to claim closure--called a case of "Step". Some simple numeric examples are followed through the three cases. We begin the discussion with a general model for paid loss development.

Notation and Setup

With S, F, $\widetilde{F}$ and b as above, we also let $f(t)=\frac{dF}{dt}$ be the probability density function [PDF], $h(t)=\frac{f(t)}{S(t)}$ the hazard rate function, $CV=\frac{\sigma}{\mu}$ the coefficient of variation, and T the random variable that gives the "time" of closure t. We use those same letter symbols and "transparent" notation to specify the relationship between these functions. For example $\widetilde{h}_\alpha(t)$ denotes the hazard rate function of the PLDD $\widetilde{F}_\alpha(t)$ that corresponds to the claim survival function $S_\alpha(t)$ and $\widetilde{T}_\alpha$ the random variable with CDF $\widetilde{F}_\alpha(t)$.

In each of the cases we consider, the complete payment pattern of a claim is completely determined by the claim duration. So we make the assumption that for any time t, 0<t<b, all claims with duration t have the same pre-determined and differentiable payment pattern. We can capture this mathematically by defining the function

$G(x,t)$=amount paid through time t on a claim, conditional upon claim duration=x.

Then define

$g(x,t)$=partial derivative of $G(x,t)$ with respect to t.

We may interpret $g(x,t)$ as the rate of payment at time t on any claim of duration x. Both $G(x,t)$ and $g(x,t)$ are defined for x,t in $(0,b)$. Note that for $t>x$ we have $g(x,t)=0$ and $G(x,x)=G(x,t)=G(x,b)$= the ultimate incurred on any claim of duration x. In this paper we only consider the case when all claims have the same ultimate incurred cost. So without any real loss of generality we further make the assumption throughout the rest of this paper that $G(x,b)=1$ for all x (see [1] for a consideration of the more general case).

As noted, we refer to the case when the rate of payment $g(x,t)$ does not vary with time t as the "flat case". The "COLA case" means the rate corresponds to a fixed rate of inflation or discount and the "step case" provides for a one-time change in the rate $g(x,t)$—the precise meaning of those assumptions provided in their respective sections of the paper. We consider the cumulative payment for such a claim distribution in which all claims occur at time t=0 and conforming to these assumptions (sort of an accident instant, as opposed to an accident year). The only "stochastic" ingredient in this model is claim duration, for which the distribution $F(t)$ is specified. Under these assumptions, $F(t)$ determines not just closures but all payments. There is a well-defined expected cumulative paid loss $P(t)$ at any time t, from t=0 to ultimate paid at $t=b$. Indeed, we have:

$$P(t) = \int_0^t \int_0^b g(x,y) f(x) dx dy = \int_0^b f(x) \int_0^t g(x,y) dy dx = \int_0^b f(x) G(x,t) dx$$

$$= \int_0^t f(x) G(x,t) dx + \int_t^b f(x) G(x,t) dx$$

$$= \int_0^t f(x) dx + \int_t^b f(x) G(x,t) dx$$

$$= F(t) + \int_t^b f(x) G(x,t) dx$$

since $G(x,t) = 1$ for $t>x$. In particular, the expected ultimate loss per claim is normalized by our assumptions:

$$P(b) = \int_0^b f(x) dx = F(b) = 1$$

The (expected) ultimate paid loss development factor from time t is:

$$\lambda(t) = \frac{P(b)}{P(t)} = \frac{1}{P(t)}$$

and the inverse provides the PLDD on $(0,b)$ that is the focus of this study:

(*) $$\tilde{F}(t) = P(t) = F(t) + \int_t^b f(x) G(x,t) dx$$

For the PDF of the PLDD, we have, by the fundamental theorem of calculus:

(**) $$\tilde{f}(t) = \frac{d}{dt}\left(\int_0^t \int_0^b g(x,y) f(x) dx dy \right) = \int_0^b g(x,t) f(x) dx = \int_t^b g(x,t) f(x) dx$$

since $g(x,t) = 0$ for $t>x$.

Findings-Flat Case

In this section we assume a constant payment pattern. With the above notation, the following proposition documents some basic relationships between the duration density and the PLDD density:

Proposition 1: *Assume the "flat case" holds, then for* $t \in (0,b)$

i) $\tilde{F}(t) = F(t) + t\int_t^b \frac{f(x)}{x}dx = t\left(\frac{1}{b} + \int_t^b \frac{F(x)}{x^2}dx\right)$

ii) $\tilde{f}(t) = \int_t^b \frac{f(x)}{x}dx = \frac{1}{b} + \int_t^b \frac{F(x)}{x^2}dx - \frac{F(t)}{t} = \frac{S(t)-\tilde{S}(t)}{t} = \frac{\tilde{F}(t)-F(t)}{t}$

iii) $\tilde{h}(t) = \frac{\frac{S(t)}{\tilde{S}(t)} - 1}{t}$

iv) $E(\tilde{T}^k) = \frac{E(T^k)}{k+1} \quad k = 1,2,...$

v) $\tilde{S}(t) = t\int_t^b \frac{S(x)}{x^2}dx$

Proof: By our assumptions on the payment pattern, and using the above notation, we have:

$$G(x,t) = \begin{cases} \frac{t}{x} & 0 \le t \le x \\ 1 & x \le t \end{cases}$$

From that we confirm that:

$$g(x,t) = \begin{cases} \frac{1}{x} & 0 \le t \le x \\ 0 & x \le t \end{cases}$$

does not vary with t. The above equations (*) (**) show that in this flat case:

$$\tilde{F}(t) = F(t) + \int_t^b f(x)G(x,t)dx = F(t) + t\int_t^b \frac{f(x)}{x}dx$$

$$\tilde{f}(t) = \int_t^b g(x,t)f(x)dx = \int_t^b \frac{f(x)}{x}dx$$

Integration by parts gives:

$$\int_t^b \frac{f(x)}{x}dx = \int_t^b u dv \qquad u = x^{-1} \quad dv = f(x)dx \quad du = -x^{-2}dx \quad v = F(x)$$

$$= uv\Big]_t^b - \int_t^b v du = \frac{F(x)}{x}\Big]_t^b + \int_t^b \frac{F(x)}{x^2}dx = \frac{1}{b} - \frac{F(t)}{t} + \int_t^b \frac{F(x)}{x^2}dx$$

and we find that:

$$\tilde{F}(t) = F(t) + t\int_t^b \frac{f(x)}{x}dx$$

$$= F(t) + t\left(\frac{1}{b} - \frac{F(t)}{t} + \int_t^b \frac{F(x)}{x^2}dx\right) = t\left(\frac{1}{b} + \int_t^b \frac{F(x)}{x^2}dx\right)$$

proving (i) and the first two equations in (ii). For the rest of (ii) and (iii) we observe that:

$$\tilde{S}(t) = 1 - \tilde{F}(t) = 1 - F(t) - t\int_t^b \frac{f(x)}{x}dx = S(t) - t\tilde{f}(t).$$

Integration by parts also gives (c.f [2]):

$$E(T^k) = k\int_0^b t^{k-1}S(t)dt$$

And applying this to T and $\tilde{T}$:

$$E(\tilde{T}^k) = k\int_0^b t^{k-1}\tilde{S}(t)dt = k\int_0^b t^{k-1}S(t) - t^k\tilde{f}(t)dt = E(T^k) - kE(\tilde{T}^k)$$

and (iv) holds. For (v), substitute $1 - S$ for F in (ii):

$$S(t) - \tilde{S}(t) = t\left(\frac{1}{b} + \int_t^b \frac{F(x)}{x^2}dx - \frac{F(t)}{t}\right) = t\left(\frac{1}{b} + \int_t^b \frac{1-S(x)}{x^2}dx - \frac{1-S(t)}{t}\right)$$

$$= t\left(\frac{1}{b} + \left[\frac{-1}{x}\right]_t^b - \int_t^b \frac{S(x)}{x^2}dx - \frac{1-S(t)}{t}\right) = t\left(-\int_t^b \frac{S(x)}{x^2}dx + \frac{S(t)}{t}\right)$$

$$\Rightarrow \tilde{S}(t) = t\int_t^b \frac{S(x)}{x^2}dx$$

This completes the proof of Proposition 1.

Now we clearly have that the PDF $\tilde{f}(t)$ is decreasing, indeed $\frac{d\tilde{f}}{dt} = -\frac{f(t)}{t} \leq 0$ and so the mode of the PLDD $\tilde{F}(t)$ is 0. From the following Corollary, we see that the shift from $F(t)$ to $\tilde{F}(t)$ shrinks the mean and increases the coefficient of variation, but the effect on the variance depends on value of CV ($\tilde{\sigma} > \sigma \Leftrightarrow CV < \frac{1}{2}\sqrt{\frac{1}{2}}$).

Corollary 1.1:

i) $\tilde{\mu} = \frac{\mu}{2}$

ii) $\tilde{\sigma}^2 = \frac{\sigma^2}{3} + \frac{\mu^2}{12} = \frac{\sigma^2}{3}\left(1 + \frac{1}{4CV^2}\right)$

iii) $\tilde{CV} = \sqrt{\frac{4}{3}CV^2 + \frac{1}{3}}$

Proof: The proof is clear from the general observation that $\sigma^2 + \mu^2 = E(T^2)$ and Proposition 1 (iv).

In the WC work that motivated this, pension cases emerge as those that take longer to close and it is natural to try and use that as a way to isolate them. This leads us to consider what happens when there is a delay period to closure that applies to all pension claims, i.e. when $f(t) = 0$ for $t \in (0,a)$ where $0 \le a < b$. This is readily accommodated, as indicated in:

Corollary 1.2: *Suppose* $f(t) = 0$ *for* $t \in (0,a)$ *where* $0 \le a < b$ *then*

$$\tilde{F}(t) = \frac{t}{a}\tilde{F}(a) \quad \textit{for } t \in (0,a).$$

Proof: Under these assumptions, Proposition 1 (i) implies that.

$$\tilde{F}(a) = F(a) + a\int_a^b \frac{f(x)}{x}dx = a\int_a^b \frac{f(x)}{x}dx$$

but then for $t \in (0,a)$:

$$\tilde{F}(t) = t\int_a^b \frac{f(x)}{x}dx = \left(\frac{t}{a}\right)\left(a\int_a^b \frac{f(x)}{x}dx\right) = \left(\frac{t}{a}\right)\tilde{F}(a).$$

Probably the most useful family of distributions defined on a finite interval is the class of Beta densities on (0,1). Recall that the Beta distribution is a two-parameter, α, β, distribution that is usually defined in terms of its PDF:

$$f(\alpha,\beta;x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1} \quad x \in (0,1), \alpha > 0, \beta > 0$$

where B and Γ denote the usual Beta and Gamma functions (c.f. [1], [3]). The CDF of the Beta density is:

$$\mathrm{B}(\alpha,\beta;t) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^t x^{\alpha-1}(1-x)^{\beta-1}dx \quad t \in (0,1), \alpha > 0, \beta > 0$$

and we have:

Corollary 1.3: *For* $\alpha > 1, \beta > 0$ *let* $f(t) = f(\alpha, \beta; t), F(t) = \mathrm{B}(\alpha, \beta; t)$ *be the PDF and CDF of a beta density on* $(0,1)$*, as just defined, then:*

$$\tilde{f}(t) = \frac{\alpha+\beta-1}{\alpha-1}(1-\mathrm{B}(\alpha-1,\beta;t))$$
$$\tilde{F}(t) = \tilde{\mathrm{B}}(\alpha,\beta;t) = \mathrm{B}(\alpha,\beta;t) + \frac{(\alpha+\beta-1)t}{\alpha-1}(1-\mathrm{B}(\alpha-1,\beta;t)) \quad 0<t<1.$$

Proof: The proof is a straightforward application of Proposition 1. For the PDF:

$$\begin{aligned}\tilde{f}(t) &= \int_t^1 \frac{f(x)}{x}dx = \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)\int_t^1 \frac{x^{\alpha-1}(1-x)^{\beta-1}}{x}dx \\ &= \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)\int_t^1 x^{(\alpha-1)-1}(1-x)^{\beta-1}dx \\ &= \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)\left(\int_0^1 x^{(\alpha-1)-1}(1-x)^{\beta-1}dx - \int_0^t x^{(\alpha-1)-1}(1-x)^{\beta-1}dx\right) \\ &= \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)\left(\frac{\Gamma(\alpha-1)\Gamma(\beta)}{\Gamma(\alpha-1+\beta)} - \int_0^t x^{(\alpha-1)-1}(1-x)^{\beta-1}dx\right) \\ &= \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)\left(\frac{\Gamma(\alpha-1)\Gamma(\beta)}{\Gamma(\alpha-1+\beta)}\right)(1-\mathrm{B}(\alpha-1,\beta;t)) \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta-1)}\frac{\Gamma(\alpha-1)}{\Gamma(\alpha)}(1-\mathrm{B}(\alpha-1,\beta;t)) = \frac{\alpha+\beta-1}{\alpha-1}(1-\mathrm{B}(\alpha-1,\beta;t))\end{aligned}$$

And for the CDF, Proposition 1 gives:

$$\tilde{F}(t) = F(t) + t\tilde{f}(t) = \mathrm{B}(\alpha,\beta;t) + \frac{(\alpha+\beta-1)t}{\alpha-1}(1-\mathrm{B}(\alpha-1,\beta;t))$$

as claimed.

We next consider some more specific examples:

Example 1: Consider the case when $S(t) = \frac{b-t}{b}, b < \infty,$ is a DeMoivre survival curve. In this case (because we will be returning to these examples by number, we specify them with a subscript):

$$F_1(b;t) = F(t) = \frac{t}{b} \quad f(t) = \frac{1}{b} \quad \tilde{F}(t) = \frac{t}{b}\left(1-\ln\left(\frac{t}{b}\right)\right) \quad 0 \le t \le b$$

Example 2: It is easy to generalize the first example in a couple of ways. Let $0 < a \le b < \infty$ where we let a represent a potentially earlier time at which all claims close and pick $\varphi > 0$. Then consider the case when claim closure has the CDF:

$$F_2(\varphi,a,b;t) = F(t) = \begin{cases} \left(\frac{t}{a}\right)^{\varphi} & t \le a \\ 1 & a \le t \le b \end{cases}$$

When $\varphi = 1$, we readily see from Example 1 that

$$\tilde{F}(t) = \begin{cases} \frac{t}{a}\left(1 - \ln\left(\frac{t}{a}\right)\right) & t \le a \\ 1 & a \le t \le b \end{cases}$$

When $\varphi \ne 1$, we let the reader verify that:

$$\tilde{F}(t) = \begin{cases} \frac{1}{\varphi - 1}\left(\frac{\varphi t}{a} - \left(\frac{t}{a}\right)^{\varphi}\right) & t \le a \\ 1 & a \le t \le b \end{cases}$$

Example 3: Consider the case when fewer claims close over time according to a linear pattern:

$$f_3(t) = \begin{cases} \frac{2(a-t)}{a^2} & t \le a \\ 0 & a \le t \le b < \infty \end{cases}$$

We leave to the reader the straightforward verification that $f_3(t)$ is indeed a PDF on $[0,b]$ and that:

$$F_3(a,b;t) = \begin{cases} \frac{t}{a}\left(2 - \frac{t}{a}\right) & t \le a \\ 1 & a \le t \le b \end{cases}$$

Then Proposition 1 implies:

$$\tilde{F}_3(t) = \begin{cases} \frac{t}{a}\left(\frac{t}{a} - 2\ln\left(\frac{t}{a}\right)\right) & t \le a \\ 1 & a \le t \le b \end{cases}$$

While we are primarily interested in the case of finite support, it may be useful to consider a couple of examples when $b = \infty$.

Example 4: Consider the case of a single parameter Pareto (c.f. [6], p 584):

$$F_4(\alpha,\theta;t) = F(t) = 1 - \left(\frac{\theta}{t}\right)^{\alpha} \qquad f_4(\alpha,\theta;t) = f(t) = \frac{\alpha\theta^{\alpha}}{t^{\alpha+1}} \quad for\ t > \theta.$$

It is natural to extend the definition of the PDF to assign $f(t) = 0$ for $t \in (0,\theta]$. For $t \ge \theta$, Proposition 1 gives:

$$\tilde{F}(t) = F(t) + t\int_t^{\infty} \frac{f(x)}{x}dx = 1 - \left(\frac{\theta}{t}\right)^{\alpha} + t\int_t^{\infty} \frac{\alpha\theta^{\alpha}}{x^{\alpha+2}}dx = 1 - \left(\frac{\theta}{t}\right)^{\alpha} + \alpha\theta^{\alpha} t \frac{1}{(\alpha+1)t^{\alpha+1}}$$

$$= 1 - \frac{\left(\frac{\theta}{t}\right)^{\alpha}}{\alpha+1}.$$

And by Corollary 1.2 we have:

$$\tilde{F}_4(\alpha,\theta;t) = \begin{cases} \dfrac{\alpha t}{(\alpha+1)\theta} & 0 < t \le \theta \\ 1 - \dfrac{\left(\frac{\theta}{t}\right)^{\alpha}}{\alpha+1} = F_4(\alpha,\vartheta;t) & \theta \le t, \vartheta = \theta(\alpha+1)^{-\frac{1}{\alpha}} \end{cases}$$

For the final example, we recall the following integration formula (c.f. [6] page 570):

$$E_1(t) = \int_t^{\infty} \frac{e^{-u}}{u}du = -\gamma - \ln(t) - \sum_{n=1}^{\infty} \frac{(-1)^n t^n}{n \cdot n!}$$

Where $\gamma = 0.577215...$ is Euler's constant.

Example 5: Consider the case when claim closures follow an exponential density, so here again $b = \infty$. In this case, we have:

$$F_5(\theta;t) = F(t) = 1 - e^{-\frac{t}{\theta}}$$

$$f_5(\theta;t) = f(t) = \frac{e^{-\frac{t}{\theta}}}{\theta}$$

Then from Proposition 1 we have:

$$\tilde{f}_5(\theta;t) = \tilde{f}(t) = \int_t^{\infty} \frac{f(x)}{x}dx = \int_t^{\infty} \frac{e^{-\frac{x}{\theta}}}{\theta x}dx = \frac{1}{\theta}\int_t^{\infty} \frac{e^{-\frac{x}{\theta}}}{\frac{x}{\theta}}\frac{dx}{\theta} = \frac{1}{\theta}\int_{\frac{t}{\theta}}^{\infty} \frac{e^{-u}}{u}du = \frac{E_1\left(\frac{t}{\theta}\right)}{\theta}$$

$$\tilde{F}_5(\theta;t) = \tilde{F}(t) = F(t) + t\tilde{f}(t) = 1 - e^{-\frac{t}{\theta}} + \left(\frac{t}{\theta}\right)E_1\left(\frac{t}{\theta}\right)$$

Findings-COLA Case

In this section we replace the assumption of a flat payment rate with the assumption that payments are subject to a constant proportional adjustment equal to $1+\delta$ per unit of time. Payments are still assumed to be made continuously over the interval from the time of loss, $t=0$, to claim closure. Since we have considered the case $\delta = 0$ in the previous section, we will assume throughout this section that $\delta \neq 0$. It is more convenient to express findings in terms of the force of interest $\gamma = \ln(1+\delta) \neq 0$ (c.f. [5]).

Under these "COLA" case assumptions, it is again straightforward—just a bit messier--to determine the PLDD, again denoted $\widetilde{F}(t)$, on the time interval $(0,b)$. The basic result is:

Proposition 2: *Assume the "COLA case" holds, then for* $t \in (0,b)$

i) $\widetilde{F}(t) = F(t) + \left(e^{\gamma t} - 1\right)\int_t^b \frac{f(x)}{e^{\gamma x} - 1}dx$

ii) $\widetilde{f}(t) = \gamma e^{\gamma t}\int_t^b \frac{f(x)}{e^{\gamma x} - 1}dx = \frac{\gamma(S(t) - \widetilde{S}(t))}{1 - e^{-\gamma t}}$

iii) $\widetilde{h}(t) = \dfrac{\gamma\left(\frac{S(t)}{\widetilde{S}(t)} - 1\right)}{1 - e^{-\gamma t}}$

Proof: Observe that for any claim with closure at x, the amount paid to time $t \le x$ is in constant proportion to the continuous annuity (c.f. [5]):

$$\int_0^t (1+\delta)^w \, dw = \int_0^t e^{\gamma w} dw = \frac{e^{\gamma t} - 1}{\gamma}$$

By our assumptions on the payment pattern, then, for any claim with closure at x, we have:

$$G(x,t) = \begin{cases} \dfrac{e^{\gamma t} - 1}{e^{\gamma x} - 1} & 0 \le t \le x \\ 1 & x \le t \end{cases}$$

We again employ the earlier formula (*) for $\widetilde{F}(t)$:

$$\widetilde{F}(t) = P(t) = F(t) + \int_t^b f(x)G(x,t)dx = F(t) + \left(e^{\gamma} - 1\right)\int_t^b \frac{f(x)}{e^{\gamma x} - 1}dx$$

And for the PDF's we have:

$$g(x,t) = \begin{cases} \dfrac{\gamma e^{\gamma t}}{e^{\gamma x} - 1} & 0 \le t \le x \\ 0 & x \le t \end{cases}$$

and by (**):

$$\tilde{f}(t) = \int_t^b g(x,t) f(x) dx = \gamma e^{\gamma t} \int_t^b \frac{f(x)}{e^{\gamma x} - 1} dx$$

The second equation in (ii) now follows from:

$$\tilde{S}(t) = 1 - \tilde{F}(t) = 1 - F(t) - \left(e^{\gamma t} - 1\right) \int_t^b \frac{f(x)}{e^{\gamma x} - 1} dx = S(t) - \left(\frac{e^{\gamma t} - 1}{\gamma e^{\gamma t}} \right) \tilde{f}(t).$$

Finally, we have:

$$\tilde{h}(t) = \frac{\tilde{f}(t)}{\tilde{S}(t)} = \frac{\left(S(t) - \tilde{S}(t)\right)\left(\dfrac{\gamma}{1 - e^{-\gamma t}}\right)}{\tilde{S}(t)} = \frac{\gamma\left(\dfrac{S(t)}{\tilde{S}(t)} - 1\right)}{1 - e^{-\gamma t}}.$$

This completes the proof of Proposition 2.

Corollary 2.1: $\tilde{\mu} - \mu = \dfrac{M_{\tilde{T}}(-\gamma) - 1}{\gamma}$ *where* $M_{\tilde{T}}(z)$ *is the moment generating function of the PLDD*

Proof: We have:

$$\tilde{S}(t) - S(t) = -\left(\frac{e^{\gamma t} - 1}{\gamma e^{\gamma t}} \right) \tilde{f}(t) = \frac{-1 + e^{-\gamma t}}{\gamma} \tilde{f}(t)$$

and so:

$$\tilde{\mu} - \mu = \int_0^b \tilde{S}(t) - S(t) dt = \int_0^b \left(\frac{e^{-\gamma t} - 1}{\gamma} \right) \tilde{f}(t) dt = \frac{\int_0^b e^{-\gamma t} \tilde{f}(t) dt - 1}{\gamma} = \frac{E\left(e^{-\gamma \tilde{T}}\right) - 1}{\gamma}$$

and the result follows.

In the COLA case, we may regard the PLDD $\tilde{F}(t) = \tilde{F}(\gamma, t)$ as a function of γ, and we have:

$$\frac{\partial \tilde{F}}{\partial \gamma}=\frac{\partial}{\partial \gamma}\left(F(t)+\left(e^{\gamma}-1\right)\int_t^b \frac{f(x)}{e^{\gamma x}-1}dx\right)=\left(e^{\gamma}-1\right)\frac{\partial}{\partial \gamma}\left(\int_t^b \frac{f(x)}{e^{\gamma x}-1}dx\right)+\left(\int_t^b \frac{f(x)}{e^{\gamma x}-1}dx\right)\frac{\partial}{\partial \gamma}\left(e^{\gamma}-1\right)$$

$$=\left(e^{\gamma}-1\right)\left(\int_t^b \frac{\partial}{\partial \gamma}\left(\frac{f(x)}{e^{\gamma x}-1}\right)dx\right)+te^{\gamma t}\left(\int_t^b \frac{f(x)}{e^{\gamma x}-1}dx\right)=\left(1-e^{\gamma}\right)\int_t^b \frac{xe^{\gamma x}}{\left(e^{\gamma x}-1\right)^2}f(x)dx+te^{\gamma t}\left(\int_t^b \frac{f(x)}{e^{\gamma x}-1}dx\right)$$

$$=\int_t^b \frac{xe^{\gamma x}-te^{\gamma t}+(t-x)e^{\gamma(t+x)}}{\left(e^{\gamma x}-1\right)^2}f(x)dx.$$

This can be used to formally prove what is intuitively rather evident, namely that $\tilde{F}(t)=\tilde{F}(\gamma,t)$ is a decreasing function of γ. Indeed, for any $\alpha>0$:

$$e^{\alpha}=1+\alpha+\frac{\alpha^2}{2}+\ldots+\frac{\alpha n}{n!}+\ldots>1+\alpha$$

Then for all $x\geq t>0$:

$$e^{\gamma t}>1+\gamma t\Rightarrow \gamma\left(x-t+\frac{t}{1-e^{\gamma t}}\right)+1>0$$

$$1-e^{\gamma t}<0\Rightarrow \gamma\left(x-t+\frac{t}{1-e^{\gamma t}}\right)\left(1-e^{\gamma t}\right)+\left(1-e^{\gamma t}\right)<0$$

$$\Rightarrow \gamma x+1+\gamma te^{\gamma t}-\gamma xe^{\gamma t}-e^{\gamma t}=\gamma\left(x-t+\frac{t}{1-e^{\gamma t}}\right)\left(1-e^{\gamma t}\right)+\left(1-e^{\gamma t}\right)<0$$

Fix t and consider the function:

$$h(x)=xe^{\gamma x}-te^{\gamma t}+(t-x)e^{\gamma(t+x)}\qquad x\geq t.$$

Observe that $h(x)$ has the same sign as the integrand in the expression for $\frac{\partial \tilde{F}}{\partial \gamma}$. But we have:

$$\frac{dh}{dx}=\gamma x+1+\gamma te^{\gamma t}-\gamma xe^{\gamma t}-e^{\gamma t}<0$$

$$h(t)=0\Rightarrow h(x)<0\quad \text{for}\ x\geq t$$

and it follows that $\frac{\partial \tilde{F}}{\partial \gamma}\leq 0$.

Observe that:

$$\tilde{F}(t)\geq F(t)\Rightarrow \tilde{S}(t)\leq S(t)\Rightarrow \tilde{\mu}=\int_0^b \tilde{S}(t)dt\leq \int_0^b S(t)dt=\mu$$

with equality only when $\mu=0$. Combining these observations with the flat case, we have:

Corollary 2.2: *Assume* $\mu > 0$

$$\frac{\mu}{2} < \tilde{\mu} < \mu \Leftrightarrow \gamma > 0$$

$$\frac{\mu}{2} = \tilde{\mu} \qquad \Leftrightarrow \gamma = 0$$

$$0 < \tilde{\mu} < \frac{\mu}{2} \Leftrightarrow \gamma < 0$$

Proof: Indeed, consider the function $g(\gamma) = \tilde{\mu} = \int_0^b \tilde{S}(t)dt$. The above shows that $\tilde{F}(t) = \tilde{F}(\gamma, t)$ is a strictly decreasing function of γ when $\mu > 0$. But then $\tilde{S} = 1 - \tilde{F}$ is a strictly increasing function of γ. It follows that $g(\gamma)$ is monotonic increasing when $\mu > 0$. But we know from the flat case that $g(0) = \frac{\mu}{2}$ and Corollary 2.2 follows.

Note that as $\gamma \to \infty$ the PLDD distribution reflects payments more concentrated at time of closure, making that distribution approximate the distribution of closures, and we would expect $\lim_{\gamma \to \infty} \tilde{\mu} = \mu$. On the other hand, as $\gamma \to -\infty$ the PLDD $\tilde{F}(t)$ reflects payments becoming concentrated at time 0, suggesting that $\lim_{\gamma \to -\infty} \tilde{\mu} = 0$. More formally, we have:

Corollary 2.3: $\lim_{\gamma \to \infty} \tilde{\mu} = \mu$ *and* $\lim_{\gamma \to -\infty} \tilde{\mu} = 0$.

Proof: First assume $b < \infty$. Notice that for $0<t<x$:

$$\lim_{\gamma \to \infty} G(x,t) = \lim_{\gamma \to \infty} \frac{e^{\gamma t} - 1}{e^{\gamma x} - 1} = \lim_{\gamma \to \infty} \frac{e^{\gamma t}}{e^{\gamma x}} = \lim_{\gamma \to \infty} e^{\gamma (t-x)} = 0$$

Now $\tilde{S}(t)$ is continuous on $(0,b)$ and with $\tilde{S}(0) = 1, \tilde{S}(b) = 0$ it follows that $\tilde{S}(t)$ is uniformly continuous on the finite interval $[0,b]$. But then by Proposition 2 (i):

$$t > 0 \Rightarrow \lim_{\gamma \to \infty} \tilde{F}(t) = F(t) \Rightarrow \lim_{\gamma \to \infty} \tilde{S}(t) = S(t)$$

$$\Rightarrow \lim_{\gamma \to \infty} \tilde{\mu} = \lim_{\gamma \to \infty} \int_0^b \tilde{S}(t)dt = \int_0^b \left(\lim_{\gamma \to \infty} \tilde{S}(t) \right) dt = \int_0^b S(t)dt = \mu$$

Proposition 2 (i) also implies that

$$t > 0 \Rightarrow \lim_{\gamma \to -\infty} \tilde{F}(t) = F(t) + \lim_{\gamma \to -\infty} \left(e^{\gamma t} - 1\right) \int_t^b \frac{f(x)}{e^{\gamma x} - 1} dx$$

$$= F(t) + \lim_{\gamma \to -\infty} (-1) \int_t^b \frac{f(x)}{(-1)} dx = F(t) + \int_t^b f(x)dx = 1.$$

Whence:

$$\lim_{\gamma \to -\infty} \tilde{\mu} = \lim_{\gamma \to -\infty} \int_0^b \tilde{S}(t)dt = \int_0^b \left(\lim_{\gamma \to -\infty} \tilde{S}(t) \right) dt = \int_0^b 0\, dt = 0$$

For the case $b = \infty$, consider the "restriction" of the density $f(t)$ to $f_a(t) = \dfrac{f(t)}{F(a)}$ on the finite interval $[0,a]$. Note that Proposition 2 (i) implies that $\tilde{f}_a(t) = \dfrac{\tilde{f}(t)}{F(a)}$ which in turn implies that $\lim_{a\to\infty} \tilde{f}_a(t) = \tilde{f}(t)$. From the finite case we have for any $a>0$:

$$\lim_{\gamma\to\infty} \tilde{\mu}_a = \mu_a$$

Notice that:

$$\lim_{a\to\infty} \mu_a = \lim_{a\to\infty} \int_0^a xfa(x)dx = \lim_{a\to\infty} \int_0^a x\frac{f(x)}{F(a)}dx = \lim_{a\to\infty} \frac{1}{F(a)}\int_0^a xf(x)dx$$

$$= \lim_{a\to\infty} \frac{1}{F(a)} \lim_{a\to\infty} \int_0^a xf(x)dx = 1\int_0^\infty xf(x)dx = \mu$$

and by the same argument:

$$\lim_{a\to\infty} \tilde{\mu}_a = \tilde{\mu}.$$

Consider the mean of the restriction $g_2(a) = \mu_a$ as defining a function of a. It is intuitively clear that $g_2(a)$ is non-decreasing (adding larger observations cannot lower the mean); to verify this formally, note that

$$g_2(a) = \mu_a = \int_0^a \frac{xf(x)}{F(a)}dx = \frac{1}{F(a)}\int_0^a xf(x)dx$$

$$\Rightarrow \frac{dg_2}{da} = \frac{1}{F(a)}af(a) + \int_0^a xf(x)dx(-1)F(a)^{-2}f(a) = \frac{f(a)}{F(a)}\left(a - \frac{\int_0^a xf(x)dx}{F(a)}\right)$$

$$= \frac{f(a)}{F(a)}\left(a - \frac{\alpha\int_0^a f(x)dx}{F(a)}\right) \quad \text{for some } \alpha \in [0,a]$$

$$= \frac{f(a)}{F(a)}(a-\alpha) \geq 0.$$

Define the function $g(\gamma, a) = \tilde{\mu}_a$ and $g_1(\gamma) = \lim_{a\to\infty} g(\gamma, a) = \tilde{\mu}$. As in the proof of Corollary 2.2, $g(\gamma, a)$ is a non-decreasing function of γ and from we have just noted $g(\gamma, a)$ is also a non-decreasing function of a:

$$\frac{\partial g}{\partial \gamma} \geq 0 \quad \frac{\partial g}{\partial a} \geq 0.$$

Now we clearly have $g(\gamma, a) \leq \mu$ for all γ and a. So, by way of contradiction, suppose $\mu > \lim_{\gamma \to \infty} \tilde{\mu} = \lim_{\gamma \to \infty} g_1(\gamma) = \lim_{\gamma \to \infty} \lim_{a \to \infty} g(\gamma, a)$. That would imply that there is $\varepsilon > 0$ such that $g(\gamma, a) \leq \mu - \frac{\varepsilon}{2}$ for all γ and a. But then

$$\lim_{a \to \infty} \mu_a = \mu \Rightarrow \exists a_0 \text{ such that } |\mu - \mu_{a_0}| < \frac{\varepsilon}{4}$$

and by the finite case:

$$\lim_{\gamma \to \infty} g(\gamma, a_0) = \lim_{\gamma \to \infty} \tilde{\mu}_{a_0} = \mu_{a_0} \quad \Rightarrow \exists \gamma_0 \text{ such that } |g(\gamma_0, a_0) - \mu_{a_0}| < \frac{\varepsilon}{4}.$$

But then we have:

$$|g(\gamma_0, a_0) - \mu| = |g(\gamma_0, a_0) - \mu_{a_0} + \mu_{a_0} - \mu| \leq |g(\gamma_0, a_0) - \mu_{a_0}| + |\mu_{a_0} - \mu| < \frac{\varepsilon}{4} + \frac{\varepsilon}{4} = \frac{\varepsilon}{2}$$

$$\Rightarrow g(\gamma_0, a_0) > \mu - \frac{\varepsilon}{2}.$$

This contradiction shows that $\lim_{\gamma \to \infty} g(\gamma) = \mu$. Finally, suppose that $\lim_{\gamma \to -\infty} g(\gamma) = \alpha > 0$. This implies that:

$$\exists \gamma_0 \text{ such that } g(\gamma) > \frac{\alpha}{2} \quad \text{for all } \gamma \geq \gamma_0$$

$$\text{and } \mu = \int_0^\infty S(t)dt \Rightarrow \exists a_0 \text{ such that } \int_{a_0}^\infty S(t)dt < \frac{\alpha}{6}$$

Also

$$\lim_{\gamma \to -\infty} \tilde{S}\left(\frac{\alpha}{6}\right) = 0 \Rightarrow \exists \gamma_1 \text{ such that } \tilde{S}\left(\frac{\alpha}{6}\right) < \frac{\alpha}{6a_0} \quad \text{for all } \gamma \geq \gamma_0$$

$$\Rightarrow \tilde{S}(t) < \frac{\alpha}{6a_0} \quad \text{for } t \in \left[\frac{\alpha}{6}, a_0\right] \quad \text{and } \gamma \geq \gamma_0$$

Selecting $\gamma > Max(\gamma_0, \gamma_1)$, we find that:

$$\frac{\alpha}{2} < g(\gamma) = \int_0^\infty \tilde{S}(t)dt = \int_0^{\frac{\alpha}{6}} \tilde{S}(t)dt + \int_{\frac{\alpha}{6}}^{a_0} \tilde{S}(t)dt + \int_{a_0}^\infty \tilde{S}(t)dt$$

$$\leq \frac{\alpha}{6} + \frac{\alpha}{6a_0}\left(a_0 - \frac{\alpha}{6}\right) + \int_{a_0}^\infty S(t)dt \leq \frac{\alpha}{6} + \frac{\alpha}{6} + \frac{\alpha}{6} = \frac{\alpha}{2}$$

This contradiction shows that $\lim_{\gamma \to -\infty} g(\gamma) = 0$. This completes the proof of Corollary 2.3.

Example 1 (COLA case): Recall that here $S(t)$ is the DeMoivre survival curve $F_1(t)=\frac{t}{b}, f_1(t)=\frac{1}{b}, b<\infty$ and so Proposition 2 gives:

$$\widetilde{F}_1(t) = F_1(t) + \left(e^{\gamma t}-1\right)\int_t^b \frac{f_1(x)}{e^{\gamma x}-1}dx = \frac{t}{b} + \frac{e^{\gamma t}-1}{b}\int_t^b \frac{dx}{e^{\gamma x}-1}.$$

Notice that setting

$$g(x) = \ln\left(\frac{e^{\gamma x}-1}{e^{\gamma x}}\right) = \ln\left(e^{\gamma x}-1\right) - \gamma x \quad \Rightarrow \frac{dg}{dx} = \frac{\gamma e^{\gamma x}}{e^{\gamma x}-1} - \gamma = \frac{\gamma e^{\gamma x} - \gamma(e^{\gamma x}-1)}{e^{\gamma x}-1} = \frac{\gamma}{e^{\gamma x}-1}$$

and we have the formula:

$$\widetilde{F}_1(t) = \frac{t}{b} + \frac{e^{\gamma t}-1}{\gamma b}\int_t^b \frac{\gamma dx}{e^{\gamma x}-1} = \frac{t}{b} + \frac{e^{\gamma t}-1}{\gamma b}\left(g(b) - g(t)\right)$$

$$= \frac{t}{b} + \frac{e^{\gamma t}-1}{\gamma b}\left(\ln\left(\frac{e^{\gamma b}-1}{e^{\gamma t}-1}\right) - \gamma(b-t)\right)$$

Example 2 (COLA case): Recall that

$$F_2(\varphi,a,b;t) = F(t) = \begin{cases}\left(\frac{t}{a}\right)^{\varphi} & t \le a \\ 1 & a \le t \le b\end{cases}$$

then $\widetilde{F}_2(t) \ge F_2(t) = 1$ for $t \ge a$ and for $t \le a$:

$$\widetilde{F}_2(t) = F_2(t) + \left(e^{\gamma t}-1\right)\int_t^a \frac{f_2(x)}{e^{\gamma x}-1}dx = \left(\frac{t}{a}\right)^{\varphi} + \frac{\varphi\left(e^{\gamma t}-1\right)}{a^{\varphi}}\int_t^a \frac{x^{\varphi-1}}{e^{\gamma x}-1}dx.$$

Example 3 (COLA case): Recall that

$$f_3(t) = \begin{cases}\frac{2(a-t)}{a^2} & t \le a \\ 0 & a \le t \le b < \infty\end{cases}$$

$$F_3(a,b;t) = \begin{cases}\frac{t(2a-t)}{a^2} & t \le a \\ 1 & a \le t \le b\end{cases}$$

then $\widetilde{F}_3(t) \ge F_3(t) = 1$ for $t \ge a$ and for $t \le a$:

$$\widetilde{F}_3(t) = F_3(t) + \left(e^{\gamma t}-1\right)\int_t^a \frac{f_3(x)}{e^{\gamma x}-1}dx = F_3(t) + \frac{2\left(e^{\gamma t}-1\right)}{a^2}\int_t^a \frac{a-x}{e^{\gamma x}-1}dx$$

$$= \frac{1}{a^2}\left(t(2a-t) + 2a\left(\frac{e^{\gamma t}-1}{\gamma}\right)\int_t^a \frac{\gamma dx}{e^{\gamma x}-1} - 2\left(\frac{e^{\gamma t}-1}{\gamma^2}\right)\int_t^a \frac{\gamma^2 x}{e^{\gamma x}-1}dx\right)$$

The function $\text{dilog}(x) = \int_1^x \frac{\ln(t)}{1-t} dt$ is useful for evaluating $\tilde{F}_3(t)$ because of the integral formula:

$$\int \frac{x}{e^x - 1} dx = -\text{dilog}(e^x) - \frac{x^2}{2}.$$

Combining this formula with what was observed in Example 1(COLA), it can be verified that:

$$\tilde{F}_3(t) = \frac{1}{a^2}\left(t(2a-t) + 2a\left(\frac{e^{\gamma t}-1}{\gamma}\right)\int_t^a \frac{\gamma dx}{e^{\gamma x}-1} - 2\left(\frac{e^{\gamma t}-1}{\gamma^2}\right)\int_t^a \frac{\gamma^2 x}{e^{\gamma x}-1}dx \right)$$

$$= \frac{1}{a^2}\left(\begin{array}{l} t(2a-t) + 2a\left(\frac{e^{\gamma t}-1}{\gamma}\right)\left(\ln\left(\frac{e^{\gamma a}-1}{e^{\gamma t}-1}\right) - \gamma(a-t)\right) \\ \quad + 2\left(\frac{e^{\gamma t}-1}{\gamma^2}\right)\left(\text{dilog}(e^{\gamma a}) - \text{dilog}(e^{\gamma t}) + \frac{\gamma^2}{2}\left(a^2 - t^2\right)\right) \end{array} \right)$$

The following table provides values of the tail factor $\lambda = \tilde{F}_3(N)^{-1}$ at various values of δ, N and a; it provides some quantification of the sensitivity of the tail factor to inflation:

COLA Case			
δ	N	a	$\tilde{F}_3(N)^{-1}$
-0.05	10	40	1.222
-0.05	20	40	1.032
-0.05	30	40	1.002
0	10	40	1.323
0	20	40	1.06
0	30	40	1.006
+0.05	10	40	1.431
+0.05	20	40	1.094
+0.05	30	40	1.011

Findings-Step Case

In this section we replace the assumption of a flat payment pattern with the assumption that payments are at a constant rate β during the last unit of time (i.e. the interval $(x-1,x)$ prior to closure at x), and otherwise at the constant rate α. We also require that the ratio $\rho = \frac{\beta}{\alpha}$ is the same for all claims. Payments are still assumed to be made continuously over the interval from the time of loss, $t = 0$, to claim closure. Obviously, the "flat case" is just the special case $\alpha = \beta$ of this "step case". We assume $b>1$ in this step case section, as otherwise this would reduce to the flat case.

Under these assumptions, it is again straightforward—but messier still--to determine the PLDD $\tilde{F}(t)$, on the time interval $(0,b)$. Indeed, by our assumptions on the payment pattern, for any claim with closure at $x \le 1$, payments are at the rate $\beta = \beta_x = \frac{1}{x}$ and the payment pattern implies

$$G(x,t) = \begin{cases} \frac{t}{x} & t \le x \\ 1 & x \le t \end{cases}$$

while for $x \ge 1$:

$$1 = \int_0^{x-1} \alpha_x + \int_{x-1}^{x} \beta_x = \alpha_x (x-1) + \rho\alpha_x \Rightarrow \alpha_x = \frac{1}{x+\rho-1}$$

and we find that:

$$G(x,t) = \begin{cases} \alpha_x t = \dfrac{t}{x+\rho-1} & t \le x-1 \\ \alpha_x (x-1) + \beta_x (t-x+1) = 1 - \rho\left(\dfrac{x-t}{x+\rho-1}\right) & x-1 \le t \le x \\ 1 & x \le t \end{cases}$$

A straightforward verification, again using equation (*) and the fact that $f(x)=0$ for $x>b$, yields:

$$\tilde{F}(t) = \begin{cases} F(t+1) + \int_t^1 \dfrac{(t-x)f(x)}{x} dx + \rho \int_1^{t+1} \dfrac{(t-x)f(x)}{x+\rho-1} dx + t\int_{t+1}^{b} \dfrac{f(x)}{x+\rho-1} dx & t \le 1 \\ F(t+1) + \rho \int_t^{t+1} \dfrac{(t-x)f(x)}{x+\rho-1} dx + t\int_{t+1}^{b} \dfrac{f(x)}{x+\rho-1} dx & 1 \le t \end{cases}$$

or:

$$\tilde{F}(t) = F(t+1) + \delta(t)\int_t^1 \frac{(t-x)f(x)}{x}dx + \rho\int_{Max(1,t)}^{t+1} \frac{(t-x)f(x)}{x+\rho-1}dx + t\int_{t+1}^{b} \frac{f(x)}{x+\rho-1}dx$$

where δ is the characteristic function of the interval (0,1), i.e. $\delta(t) = 1$ on the interval (0,1) and is 0 elsewhere. Note that the function $\tilde{F}(t)$ is continuous, even though δ is not. Note too that the last integral in the formula vanishes when $t>b-1$ and in that case the upper limit of the middle integral can be shifted down to b—this observation is helpful when the functional form of $f(t)$ only behaves on $(0,b)$.

In the step case, we may regard the PLDD $\tilde{F}(t) = \tilde{F}(\rho,t)$ as a function of ρ and we have:

$$\frac{\partial \tilde{F}}{\partial \rho} = \frac{\partial}{\partial \rho}\left(F(t+1) + \delta(t)\int_t^1 \frac{f(x)}{x}dx + \rho\int_{Max(1,t)}^{t+1} \frac{(t-x)f(x)}{x+\rho-1}dx + t\int_{t+1}^{b} \frac{f(x)}{x+\rho-1}dx \right)$$

$$= \rho\int_{Max(1,t)}^{t+1} \frac{\partial}{\partial \rho}\left(\frac{1}{x+\rho-1}\right)(t-x)f(x)dx + \int_{Max(1,t)}^{t+1} \frac{(t-x)f(x)}{x+\rho-1}dx + t\int_{t+1}^{b} \frac{\partial}{\partial \rho}\left(\frac{1}{x+\rho-1}\right)f(x)dx$$

$$= \int_{Max(1,t)}^{t+1}\left(\rho \ln(x+\rho-1) + \frac{1}{x+\rho-1}\right)(t-x)f(x)dx + t\int_{t+1}^{b} \ln(x+\rho-1)f(x)dx.$$

As the following examples illustrate, the integral form for $\tilde{F}(t)$ may be preferable to some closed form expressions, especially when there is access to decent numerical integration software. In the examples, we set $\hat{t} = Max(1,t) = t + \delta(t)(1-t)$.

Example 1 (Step case): Recall that here $F_1(t) = \frac{t}{b}$, $f_1(t) = \frac{1}{b}$, $b < \infty$, then for $t \le b-1$:

$$\tilde{F}_1(t) = \frac{1}{b}\left(\begin{array}{c} t+1+\delta(t)(t-1+t\ln t) \\ +\rho\left((t+\rho-1)\ln\left(\frac{t+\rho}{\hat{t}+\rho-1}\right) + \hat{t} - t - 1\right) \\ +t\ln\left(\frac{b+\rho-1}{t+\rho}\right) \end{array} \right)$$

and for $t \ge b-1$:

$$\tilde{F}_1(t) = \frac{1}{b}\left(\begin{array}{c} t+1+\rho\hat{t}+\delta(t)(t-1+t\ln t) \\ +\rho\left((t+\rho-1)\ln\left(\frac{b+\rho-1}{\hat{t}+\rho-1}\right)\right) \end{array} \right) - \rho$$

Example 2 (Step case): Recall that

$$F_2(\varphi,a,b;t) = F(t) = \begin{cases} \left(\frac{t}{a}\right)^{\varphi} & t \le a \\ 1 & a \le t \le b \end{cases}$$

then $\tilde{F}_2(t) \ge F_2(t) = 1$ for $t \ge a$ and for $t \le a$. When φ is a positive integer >1 and $t \le a-1$, we have:

$$\begin{aligned}\tilde{F}_2(t) &= \left(\frac{t+1}{a}\right)^{\varphi} + \frac{\delta(t)}{a^{\varphi}(\varphi-1)}\left(\varphi(t-1) - t^{\varphi} + 1\right) \\ &+ \left(\frac{\varphi}{a^{\varphi}}\right)\sum_{i=1}^{\varphi-1}\frac{(1-\rho)^{\varphi-i-1}\left((\rho-1)(\rho+t)(t+1)^i - \rho(t+\rho-1)\hat{t}^i + ta^i\right)}{i} \\ &+ \left(\frac{\varphi}{a^{\varphi}}\right)\left(\rho(t+\rho-1)(1-\rho)^{\varphi}\ln\left(\frac{t+\rho}{\hat{t}+\rho-1}\right) + t(1-\rho)^{\varphi-1}\ln\left(\frac{a+\rho-1}{t+\rho}\right)\right)\end{aligned}$$

When φ is a positive integer and $t \ge a-1$, we have:

$$\begin{aligned}\tilde{F}_2(t) &= \left(\frac{t+1}{a}\right)^{\varphi} + \frac{\delta(t)}{a^{\varphi}(\varphi-1)}\left(\varphi(t-1) - t^{\varphi} + 1\right) \\ &+ \left(\frac{\varphi\rho}{a^{\varphi}}\right)\sum_{i=1}^{\varphi-1}\frac{t(1-\rho)^{\varphi-i-1}\left(a^i - \hat{t}^i\right)}{i} \\ &+ \left(\frac{\varphi\rho}{a^{\varphi}}\right)\left(\frac{\hat{t}^{\varphi} - a^{\varphi}}{\varphi}\right) \\ &+ \left(\frac{\varphi\rho}{a^{\varphi}}\right)\left((t+\rho-1)(1-\rho)^{\varphi}\ln\left(\frac{a+\rho-1}{\hat{t}+\rho-1}\right)\right)\end{aligned}$$

Example 3 (Step case): Recall that

$$f_3(t) = \begin{cases} \frac{2(a-t)}{a^2} & t \le a \\ 0 & a \le t \le b < \infty \end{cases}$$

$$F_3(a,b;t) = \begin{cases} \frac{t(2a-t)}{a^2} & t \le a \\ 1 & a \le t \le b \end{cases}$$

then $\tilde{F}_3(t) \ge F_3(t) = 1$ for $t \ge a$. When $t \le a-1$ we have:

$$\tilde{F}_3(t) = \frac{1}{a^2}\left(a^2 - \hat{t}^2 + 2(1-\rho-a-t)(a-\hat{t}) + 2(\rho+t-1)(a+\rho-1)\ln\left(\frac{t+\rho}{\hat{t}+\rho-1}\right)\right)$$

and when $t \ge a-1$:

$$\widetilde{F}_3(t) = \left(\frac{t+1}{a}\right)\left(2-\left(\frac{t+1}{a}\right)\right) + \frac{2\delta(t)}{a^2}\left(\frac{1}{2} - a - t - at\ln t + (a+t)t - \frac{t^2}{2}\right)$$
$$+\frac{\rho}{a^2}\left(a^2 - \hat{t}^2 + 2(1-\rho-a-t)(a-\hat{t}) + 2(\rho+t-1)(a+\rho-1)\ln\left(\frac{a+\rho-1}{\hat{t}+\rho-1}\right)\right)$$

The following table provides values of $\lambda = \widetilde{F}_3(N)^{-1}$ at various values of ρ, N and a; it provides some quantification of the sensitivity of the tail factor to a change of payment in the terminating year.

Step Case			
ρ	N	a	$\widetilde{F}_3(N)^{-1}$
1/2	10	40	1.307
1/2	20	40	1.056
1/2	30	40	1.005
1	10	40	1.323
1	20	40	1.06
1	30	40	1.006
2	10	40	1.354
2	20	40	1.068
2	30	40	1.008

References:

[1] Corro, Dan, *Annuity Densities with Application to Tail Development*, in preparation.

[2] Corro, Dan, *Determining the Change in Mean Duration Due to a Shift in the Hazard Rate Function*, CAS Forum, Winter 2001.

[3] Corro, Dan, *Fitting Beta Densities to Loss Data*, CAS Forum, Summer 2002.

[4] Gillam, William R.; Couret, Jose R. *Retrospective Rating: 1997 Excess Loss Factors*, PCAS LXXXIV, 1997, including discussion: Mahler, H. PCAS, LXXXV, 1998.

[5] Kellison, Stephen G., *Theory of Interest*, McGraw Hill/Irwin, 1991.

[6] Klugman, Stuart A.; Panjer Harry H.; and Willmot, Gordon E., *Loss Models*, Wiley Series in Probability and Statistics, John Wiley & Sons, Inc., 1998.

Financial Pricing Models for Property-Casualty Insurance Products: Modeling the Equity Flows

Sholom Feldblum, FCAS, FSA, MAAA,
and Neeza Thandi, FCAS, MAAA

Financial Pricing Models for Property-Casualty Insurance Products: Modeling the Equity Flows

by Sholom Feldblum and Neeza Thandi

This paper and its companion papers present the use of return on capital financial models to price property-casualty insurance products. This paper focuses on the cash flow and equity flow modeling that underlies the financial models. The companion papers complete the description of return on capital pricing models. The first two appendices to this paper – Appendix A on federal income taxes and Appendix B showing the workers' compensation pricing exhibits – apply to all the papers in this series.

FLOW OF FUNDS

Financial models to quantify the expected profitability of a business project consider the net present value of a series of cash flows or the internal rate of return of those cash flows. The models are used to set product prices and to measure business performance.

The cash flows represent the expected flow of funds to and from the suppliers of capital. The suppliers of capital are termed the owners, the investors, the shareholders (or stockholders), or the equityholders; we generally use the term "equityholders" in this paper. Although mutual insurance companies do not have stockholders or investors, and the "ownership" status of their policyholders is not clear, we assume that they face the same capital management constraints as a stock insurance company faces.

For non-regulated industries, the cash flows to and from the company are reasonable proxies for the equity flows to and from the equityholders.

1. A cash flow into the company provides shareholder dividends to the equityholders.
2. A cash flow out of the company necessitates a capital contribution by the equityholders.

The use of company cash flows as a proxy for the implied equity flows assumes that the company cash flows can be paid as dividends to equityholders. For regulated industries, such as insurance and other financial services, this assumption is not correct, because of statutory reserve requirements and risk-based capital requirements.

For pricing property-casualty insurance products, we explicitly examine the expected flow of funds to and from the suppliers of capital – or the implied equity flows – not the flow of funds to and from the company. To calculate the implied equity flows, we project the future cash flows, and we adjust for statutory requirements and federal income taxes.[1]

---

[1] When the meaning is clear, we sometimes refer to implied equity flows simply as equity flows.

Cash flow projections are easier for insurance contracts than for many other commercial products. Most insurance underwriting costs are variable, and the overall demand for insurance is relatively stable. In contrast, other commercial products often have large fixed cost components and fluctuating demand from year to year.

Cash flow estimation techniques are not unique to insurance, and actuaries are proficient at these estimation tasks. We assume here that future cash flows to and from the company have been properly estimated and that statutory and tax provisions have been accounted for. Our task is to determine the implied equity flows from which the net present value and the internal rate of return are calculated.

Cash Flows and Equity Flows

In most industries, cash flows to and from the company are reasonable proxies for the cash flows to and from the suppliers of capital. If a company invests $1,000 at time t = 0 and it receives $1,100 at time t = 1, the pricing model assumes that the company's owners provide the $1,000 in capital at time t = 0 and receive the $1,100 at time t = 1.

This is true for unregulated manufacturing or service enterprises and for some utilities. The accounting requirements for these industries are not directly relevant to return on capital pricing models. The models focus on projected cash flows and current income tax liabilities.[2]

Regulated financial institutions – life insurance companies, property-casualty insurance companies, depository institutions, and certain investment firms – are different. For these industries, the cash flows to and from the company do not necessarily reflect the cash flows to and from investors. The potential dividends to equityholders and the capital required from equityholders depend on the statutory funding requirements for loss reserves and on risk-based capital requirements, not just on the cash and similar assets held by the company.

For the major property-casualty insurance transactions – premium collections and loss payments – the two sets of cash flows generally have opposite signs. A premium collection, which is a cash inflow to the company, generally necessitates a capital contribution by the equityholders. A loss payment, which is a cash outflow from the company, generally allows a return of capital to the equityholders.

---

[2] Current income tax liabilities are the taxes assessed by the IRS. Accrued tax liabilities are the sum of the current tax liabilities and the deferred tax liabilities. If there are no expected changes in the tax rate, the accrued tax liabilities equals the book income (either GAAP or statutory) times the tax rate. We use a balance sheet orientation – not an income statement orientation – to evaluate the deferred tax assets and liabilities; see SFAS 109. We adjust the statutory deferred tax asset for its admitted portion; see SSAP No. 10.

Solvency monitoring by governmental authorities is the underlying rationale for the difference between the company cash flows and the implied equity flows. The flow of funds to suppliers of capital, the "equityholders," can be inferred from

- the company cash flows,
- capital requirements imposed by regulatory authorities,
- reserve requirements and other accounting regulations.

To distinguish between the company cash flows and the flow of funds to suppliers of capital, we refer to the former as *cash flows* and to the latter as *implied equity flows*. The financial community uses the term "free cash flow" instead of implied equity flow. Atkinson and Dallas [2000], chapter 11, use the term "distributable earnings" instead of implied equity flows.[3]

We summarize below the signs of the company cash flows and the implied equity flows for the four major types of property-casualty insurance transactions. A positive flow means an inflow to the company or to the equityholders. A negative flow means an outflow from the company or the equityholders.

	Company Cash flow	Implied Equity flow
Premiums collected	+	–
Losses paid	–	+
Expenses (including federal income taxes)[4]	–	–
Investment income received	+	+

The actual relationship between the cash flows and the implied equity flows is complex, and the chart above does not do full justice to this topic. The illustration below shows the intuition for the implied equity flows stemming from premium collection. The text of this paper, and the associated flowcharts, graphics, and tables, works through the cash flows and implied equity flows for a more complete illustration. Appendix B shows the full cash flows and implied equity flows for a workers' compensation pricing analysis, using a 50 year (200 quarter) return on capital pricing model.

Illustration: A policy with a premium of $1,000 is written and collected on December 31, 20XX. The agent's commission is 20%, and the capital requirements equal 25% of the written premium. We ignore for the moment the tax liability and the deferred tax asset.

---

3 The life insurance pricing model in Atkinson and Dallas [2000] parallels the property-casualty insurance pricing model in this paper. The differences between the models reflect the differences in reserve requirements and federal income tax liabilities between life insurance and property-casualty insurance.

4 For both life insurance and property-casualty insurance, deferred taxes are included with current tax liabilities. When Atkinson and Dallas wrote their textbook, statutory accounting did not recognize deferred tax assets and liabilities, so they restricted their treatment of deferred taxes to GAAP return on equity models.

- The net cash *inflow* from the policyholder to the company is $1,000 (premium) – $200 (commission) = $800 on December 31.
- The company receives $800 (net of commission), but it must hold a $1,000 unearned premium reserve and it must support the policy with $250 of surplus. The cash *outflow* from the equityholders is ($1,000 – $800) + $250 = $450.

Free Cash Flows and Implied Equity Flows

The distinction between company cash flows and implied equity flows is identical to the distinction between company cash flows and free cash flows. Financial analysts use free cash flows for return on capital pricing models. The use of company cash flows without consideration of changes in net working capital fails to take full account of invested capital.

We summarize below the distinction between non-regulated and regulated industries. For non-regulated manufacturing enterprises, the company cash flows are adjusted for (i) depreciation and amortization and (ii) required investment (or "capital expenditures"). This adjusted income minus the change in net working capital equals the free cash flows. The free cash flows are used to determine the net present value and the internal rate of return.

For regulated insurance enterprises, the statutory income equals the statutory cash flow adjusted for the capitalization and amortization of the unearned premium reserves and the loss reserves. The statutory income *minus the change in required capital* equals the implied equity flow. The implied equity flows determine the net present value and the internal rate of return. The change in net working capital for other industries is equivalent to the change in required capital for insurance enterprises.[5]

Early forms of cash flow pricing models presumed that the cash flows equal the accounting income adjusted for non-cash revenues and expenditures. This presumption is no longer used, because many cash flows do not result in revenues or expenditures. Cash is used to purchase material and supplies to produce goods and services. The purchase of supplies is the exchange of cash for a non-cash asset (inventory). There is no revenue or expenditure on the firm's income statement.

---

[5] Some analysts argue that the change in net working capital for manufacturing enterprises is a real cash flow item. It is determined by business and economic constraints, not by regulation, and it reflects the cash expenditures of the firm. In contrast, reserve requirements and risk-based capital requirements for insurance enterprises depend on regulatory mandate, not on business and economic constraints. The perspective in this paper is that a regulatory mandate in a regulated industry is a business constraint.

ILLUSTRATION: AIRCRAFT MANUFACTURING

The following illustration clarifies the difference between cash flows and free cash flows. A non-leveraged (all-equity financed) firm manufactures aircraft. The firm leases the factory, equipment, and work-force, so there are no capital expenditures.

- At time t=0, the firm purchases material and supplies for $10 million to produce an airplane.
- From time t=0 to time t=4, the firm manufactures the airplane, at a cost of $2 million a year in rent and wages.
- The firm sells the airplane one year later, at time t=5, for $25 million, after paying $1 million in storage costs and sales commissions in the last year.

To simplify the computations, we assume that rent and wages are paid at the end of the year.

There are no interest payments, debt payments, amortization, or depreciation of fixed assets for this firm. The earnings for the firm are

- –$2 million at times t=1, 2, 3, and 4 (production expenses), and
- +$14 million at time t=5, calculated as sales revenue of $25 – cost of goods sold of $10 – sales expenses of $1.

There are no non-cash revenues or expenditures in this illustration. At time t=0, when the firm purchased the material and supplies, $10 million in cash is exchanged for $10 million of inventory. The inventory changes in form over the next five years, but its accounting value remains $10 million, with no effect on the balance sheet or the income statement.

If the cost of equity capital is 15% per annum, the apparent NPV of this project is (in millions of dollars)

$$-\$2 / (1.15) - \$2 / (1.15)^2 - \$2 / (1.15)^3 - \$2 / (1.15)^4 + \$14 / (1.15)^5 = \$1.25 .$$

This analysis does not take into account the change in net working capital. At time t=0, there is a cash outflow from the equityholders equal to the increase in the net working capital of $10 million.[6] At time t=5 there is an additional cash inflow to the equityholders equal to the decrease in net working capital of $10 million. The corrected net present value of this project is (in millions of dollars)

$$-\$10 - \$2 / (1.15) - \$2 / (1.15)^2 - \$2 / (1.15)^3 - \$2 / (1.15)^4 + \$24 / (1.15)^5 = -\$3.78.$$

[6] Net working capital = inventory + accounts receivable + cash on hand – accounts payable.

Proper consideration of the timing of the cash flows turns the $1.25 million indicated profit into a $3.78 million indicated loss.

INSURANCE ANALOGY

The aircraft manufacturing example is elementary. We discussed it because early property-casualty insurance net present value models overstated the returns by mistaken timing of the implied equity flows. A property-casualty insurance enterprise holds no asset called "inventory," but there is an equivalent equity flow stemming from regulatory constraints.[7]

Suppose an insurance policy is written at time t=0, a loss occurs at time t=1, and the loss is settled for $100,000 at time t=5.

- At time t=1, there is a debit on the income statement of $100,000 for the incurred loss. This accounting debit is a non-cash expenditure; it does not affect the cash flows of the firm. $100,000 of policyholders' surplus is transferred to loss reserves on the balance sheet. There is no change in the cash account on the asset side of the balance sheet.
- At time t=5, there is a cash outflow of $100,000 along with an offsetting non-cash reserve reduction of $100,000.

The cash flows noted above are the firm's cash flows, not the implied equity flows (or the free cash flows). They are not the proper base for the IRR or the NPV calculations. The proper perspective is that at time t=1, the insurance company purchases a loss reserve for $100,000. The equityholders no longer have access to these funds.

The loss reserve is like the inventory. It is an inventory of money, instead of an inventory of goods. But this money has changed from free cash that equityholders can use to a stock of funds that is not accessible to the equityholders.

Similarly, when the insurance company collects premium, it uses the funds to purchase an unearned premium reserve, as though it were purchasing a premium inventory. The premium collected is not a free cash flow.

IMPLIED EQUITY FLOWS

The implied equity flows are the implicit flow of funds to and from the suppliers of capital. The implied equity flows may be determined directly from the assets held by the company to support the insurance operations. The required assets comprise three pieces:

---

[7] We use the term "implied equity flow" to emphasize that there is not – and there need not be – any actual flow of funds. The term "equityholders" refers to the firm's owners, whether they be common stock investors or other owners of a property-casualty insurance company.

- assets backing the (gross) unearned premium reserves
- assets backing the (full value) loss reserves
- assets backing policyholders' surplus

Because stockholder dividends from insurance enterprises are restricted by statutory accounting rules, statutory accounting determines the implied equity flows.

Illustration: A company begins operations on January 1, 20XX, and writes property-casualty insurance business during the year. The invested capital on December 31, 20XX, equals the sum of three components:

- the gross unearned premium reserves minus the present value of future losses and expenses stemming from unexpired policies
- the full value loss reserves minus the present value of future loss and loss adjustment expenses on claims that have already occurred
- policyholders' surplus to satisfy the NAIC's risk-based capital requirements and similar rating agency capital requirements.

If the company holds only financial assets, the fundamental equation linking implied equity flows and required assets is that

cash flows minus implied equity flows during the accounting period equal the change in required assets from the beginning to the end of the period.

This equation is not correct if the company holds non-financial assets, such as premium receivables and deferred tax assets. Beginning with the codification of statutory accounting in January 1, 2001, almost all companies hold substantial deferred tax assets, stemming from revenue offset and from IRS loss reserve discounting.

Illustration: An insurer writes workers' compensation policies during 20XX. At December 31, 20XX–1, the required assets are zero. During 20XX, it writes $100 million of business. During 20XX, the company collects all the premium, and it pays $25 million in expenses, $18 million in losses, and $5 million in federal income taxes. At year end, it has $50 million in unearned premium reserves and $45 million in full value loss reserves. The capital requirements at year end consist of $12 million in written premium risk charges and $15 million in reserving risk charges.

- The required assets at the end of the year are $50 million + $45 million + $12 million + $15 million = $122 million.
- The cash flows during the year are $100 million – $25 million – $18 million – $5 million = $52 million.

- The implied equity flow is \$52 million – (\$122 million – \$0) = –\$70 million.

The negative implied equity flow is a flow from the equityholders to the company. It represents an investment in the insurance operations by the equityholders.

Equity Flows in Practice

Some actuaries object that the equity flows are not real, based on the following reasoning:

> *We speak of an implied equity flow to fund the underwriting loss at policy inception. But there is no actual capital contribution when a policy is written. In contrast, the company cash flows used in other industries are actual transfers of cash.*

This objection is specious. The implied equity flows are real, though they are submerged under a multitude of policies and the other capital structure decisions of the company.

Illustration A: A reinsurer writes a \$100 million book of casualty excess-of-loss reinsurance on January 1, 20XX. The risk-based capital requirements are \$25 million for the written premium risk charge.

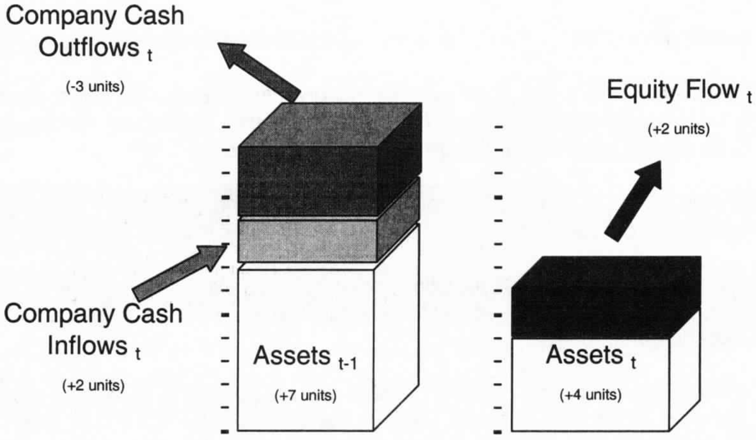
The pricing model uses an implied equity flow of \$25 million on January 1. There is, of course, no actual cash flow. But on December 31, 20XX–1, the reinsurer's book of 20XX–1 policies expired, and \$25 million of written premium risk capital was freed. The capital is transferred from one block of business to another block of business.

Illustration B: The reinsurer's premium volume increases from \$80 million in 20XX–1 to \$100 million in 20XX. The written premium risk charge increases from \$20 million to \$25 million, with a net capital contribution of \$5 million. There may be no actual cash flow corresponding to this implied capital contribution. But the reinsurer may have decided not to write other business because of capital constraints, or it may have decided not to pursue other financial activities, such as acquisitions. These are the real world reflections of the implied equity flow.

Illustration C: The reinsurer's premium volume decreases from \$120 million in 20XX–1 to \$100 million in 20XX. The written premium risk charge decreases from \$30 million to \$25 million, with a net capital contribution of –\$5 million. In this scenario, the reinsurer might pay greater stockholder dividends than it otherwise would have, it might buy back some stock, or it might use the capital to write other business or to engage in other financial activities.

The real world reflections of the implied equity flows may be slow, but eventually they are realized. The pricing model attributes these equity flows to the policies that require the capital.

Determination of the Equity Flow



Cash Flow Definition of Equity Flow:

Equity Flow $_t$ = -Asset Flow $_t$ + Company Cash Flow $_t$

Company Cash Flow $_t$ = Co Cash Inflow $_t$ - Co Cash Outflow $_t$

Asset Flow$_t$ = Assets$_{(t+1)}$ - Assets$_t$

Composition of Assets at time t is a function of business environment constraints:

Assets $_t$ = UEPR $_t$ + Loss Reserve $_t$ + Surplus $_t$

Income Statement Definition of Equity Flow:

Equity Flow $_t$ = Accting Net Income $_t$ - Δ (Accting Capital) $_t$

For the above Illustration:

Asset Flow $_t$ = (4-7) = -3

Company Cash Flow $_t$ = Co Cash Inflow $_t$ - Co Cash Outflow $_t$

= +2 - 3 = -1

$\therefore$ Equity Flow $_t$ = -(-3) + (-1) = 2

Insurance Transactions

The implied equity flows depend on the cash flows, statutory accounting rules, and capital requirements. The two illustrations below cover the major insurance transactions:

- premium writing
- premium collection
- loss incurral
- loss payment
- expense incurral and payment
- investment income
- federal income tax payments
- deferred tax assets and liabilities

The implied equity flow equals the statutory income minus the change in required capital (see also Robbin [1993; 1998], who uses the same perspective). We use statutory income, not GAAP income, since statutory income and capital requirements determine the funds that are available for distribution to owners.[8] The economic income (the NPV) to the equityholders from issuance of the insurance policy is the present value of the future implied equity flows.

An actual pricing model would use quarterly valuations for the lifetime of the book of business.[9] For this illustration, we use a simplified example with three years of semi-annual valuation periods. We provide item by item documentation, along with graphical depictions of the implied equity flows from premium and loss transactions.

Translating the cash flows, accounting requirements, and capital requirements into implied equity flows is the largest hurdle to proper use of the return on capital pricing models.[10]

---

[8] The American Academy of Actuaries *Standard of Practice No. 19, Actuarial Appraisals*, paragraph 5.2.1, emphasizes the centrality of statutory accounting for modeling insurance company cash flows and equity flows: "Distributable Earnings – For insurance companies, statutory earnings form the basis for determining distributable earnings, since the availability of dividends to owners is constrained by the amount of accumulated earnings and minimum capital and surplus requirements, both of which must be determined on a statutory accounting basis. Distributable earnings consist of statutory earnings, adjusted as appropriate to allow for the retention of a portion thereof or the release of a portion of prior accumulated earnings therein, in recognition of minimum capital and surplus levels necessary to support existing business."

[9] Appendix B shows the modeled equity flows for a workers' compensation block of business, using 50 years of quarterly valuation periods, or a total of 200 periods.

[10] Some older papers on property-casualty insurance pricing models have used company cash flows instead of equity flows for the return on capital models. Modeling the cash flows instead of the equity flows provides distorted rate indications. Since the signs of the premium and loss equity flows are the opposite of the signs of the premium and loss cash flows, use of company cash flows instead of implied equity flows transforms an investment project into a borrowing transaction. For an investment project, higher IRR's are

PREMIUM TRANSACTIONS

This paper is both a template for the practicing actuary and a teaching text for the actuarial student. We do not simply state the results; we show the calculations step-by-step, so that readers can replicate the procedures.

The illustration here is clear, and the computations are straight-forward. Nevertheless, there are many figures, and it is easy to lose track of the relationships. To keep the intuition clear, we divide the illustration into two pieces. The first piece deals with premiums, expenses, and the associated investment income, federal income taxes, capital requirements, and implied equity flows. The second piece adds losses and the associated equity flows.

The documentation in this paper has three components: (i) textual exposition, (ii) numerical exhibits, and (iii) graphics. Readers may find it helpful to trace the figures in the exhibits and the graphics as they proceed through the text.

ILLUSTRATION A: PREMIUMS

> *A company writes and collects a $1,000 annual premium on December 31, 20XX. Acquisition expenses of $250 are incurred and paid on that day. Maintenance and general expenses of $150 are incurred and paid evenly over the policy term.*
>
> *The pre-tax investment yield benchmark is an 8% per annum bond equivalent yield (semi-annual compounding). The marginal tax rate on both underwriting income and investment income is 35%.*
>
> *No losses are incurred. The capital requirements are based on the NAIC risk-based capital formula. For this scenario, the capital requirements are 25% of annual written premium plus 15% of loss reserves.*

We use the following modeling order:

a. underwriting cash flows: premiums, expenses, losses, and taxes on underwriting
b. statutory accounting entries: loss reserves and deferred tax assets
c. required surplus amounts: risk-based capital requirements
d. investment income on the investable assets
e. implied equity flows to and from the equityholders

The same sequence is used for the second half of the illustration as well, which deals with the loss cash flows and the associated components.

---

better; for a borrowing transaction, lower IRR's are better.

Valuation Dates

In theory, we should use continuous accounting changes, cash flows, and implied equity flows. Even if accounting entries are made at specified dates, such as the end of the year, insurance contracts are written throughout the year. An individual policy may have discrete accounting entries, but a policy year has continuous entries, cash flows, and equity flows.

In practice, we use quarterly valuation periods. Many readers find discrete entries simpler than continuous functions. Spreadsheet representations are also easier with discrete entries. The enhanced accuracy from continuous functions is outweighed by the added complexity.[11]

For the illustration here, we use semi-annual valuations, not quarterly valuations. This simplifies the exposition while retaining the structure of the analysis.

Premiums and Premium Receivables

A premium inflow of $1,000 occurs on December 31, 20XX. This illustration assumes full premium collection on the effective date. If the premium is not fully collected on December 31, 20XX, the company shows a premium receivable for the uncollected portion.

- If the premium receivable is an admitted asset, there is no effect on the implied equity flows.
- Any *non-admitted* premium receivable increases the implied equity outflow from the equityholders.
- The premium collection pattern affects the investment income cash flows, even if all premium receivables are admitted. The investment income cash flows affect the implied equity flows in subsequent periods.

For blocks of business with significant deferral of premium collection, such as large account workers' compensation policies, the actuary may estimate the expected non-admitted portion of the premiums receivable asset and increase the required capital contribution.[12]

Distribution Systems and Acquisition Expenses

Companies use a variety of distribution systems, such as independent agency, direct writing, salaried sales force, and mass marketing systems. Within each system, there are different

---

[11] Pricing models using continuous functions have often been proposed for life insurance contracts, though generally not for property-casualty insurance contracts.

[12] The statutory rules for the non-admitted portion of the premiums receivable are summarized in Feldblum [2002: SchP].

methods of premium billing. The modeling process for premium collection and acquisition expenses depends on the distribution system and the premium billing system.

- For independent agency companies, acquisition costs are primarily agents' commissions. These expenses are incurred on the policy writing date or on the premium collection date.
- For direct writers and for companies with salaried sales forces, acquisition costs include advertising expenses and fixed costs of the agency system. These additional expenses occur before the policy effective date.
- For commercial lines companies writing large accounts, acquisition costs include the costs of developing sales proposals and of soliciting business. These expenses also occur before the policy effective date.

The pricing actuary should use the policy distribution system and the premium billing system consistent with actual company practice.

The illustration in the text uses the independent agency distribution system, with which most readers are familiar. An acquisition expense outflow of $250 occurs on December 31, 20XX. The independent agency distribution system is particularly useful for the exposition in this paper, since it has a large initial underwriting loss, causing a large difference between the company cash flows and the implied equity flows.

General Expenses

A general expense outflow of $150 occurs evenly over the policy term. This can be modeled

- as a single $150 outflow on June 30, 20XX+1
- as three semi-annual outflows of $50 on each half year valuation date from December 31, 20XX, through December 31, 20XX+1.
- as five quarterly outflows of $30 on each quarterly valuation date from December 31, 20XX, through December 31, 20XX+1.

Common actuarial practice is to use quarterly cash flows and equity flows, at least for the first several years. For ease of exposition, this illustration uses a single expense payment at June 30, 20XX+1, not quarterly payments.

Many actuaries model some general expense costs at the policy effective date. For instance, the NAIC defines pre-paid acquisition expenses as commissions, other acquisition expenses, premium taxes, and one half of general expenses (see Feldblum [1997: IEE]). Alternatively, the general expenses incurred on the policy effective date may be included with acquisition costs. This illustration assumes that the $250 of pre-paid acquisition costs includes the general expenses that are incurred on the effective date of the policy.

TAX ON UNDERWRITING INCOME

Taxable underwriting income equals

written premium
- underwriting expenses
- 80% of the change in the gross unearned premium reserve
- paid losses
- the change in the discounted loss and loss adjustment expense reserves.

This may also be written as

statutory earned premium
+ 20% of the change in the statutory unearned premium reserve
- underwriting expenses
- statutory incurred losses
+ the change in the IRS loss and loss adjustment expense reserve discount.

Taxable underwriting income for 20XX is

$200 (income from revenue offset) – $250 (acquisition expenses) = –$50.

The tax outflow is a negative $17.50 (or a tax refund of $17.50). The "tax refund" does not rely on tax carrybacks or carry-forwards. The tax refund stemming from negative taxable income offsets tax liabilities stemming from other insurance contracts and from investment income. We chose an illustration with an acquisition expense greater than 20% of premium to emphasize that expected tax cash flows can be positive or negative.

The taxable premium income of $200 may be evaluated in two ways.

- As written premium minus 80% of the change in the unearned premium reserves = $1000 – 80% × $1000 = $200
- As statutory earned premium income plus 20% of the change in the unearned premium reserves = $0 + 20% × $1000 = $200.

The tax liability is 35% times the taxable income: 35% × ($200 – $250) = –$17.50.

Taxable underwriting income for 20XX+1 equals

$800 of taxable premium income – $150 of general expenses = $650.

The tax liability is $650 × 35% = $227.50. Written premium during the year is $0 and the unearned premium reserve declines from $1,000 to $0. We use the same two computation methods:

(i) $0 – 80% × (–$1,000) – $150 = $650
(ii) $1,000 + 20% × (–$1,000) – $150 = $650.

We use semi-annual valuations for this illustration, and we assume the tax on underwriting income is incurred evenly between the two halves of the year, or $227.50 / 2 = $113.75 in each half year. Alternatively, if all general expenses are assumed to be paid exactly on June 30, 20XX+1, the underwriting income is $400 – $150 = $250 in the first half of the year and $400 – $0 = $400 in the second half of the year.

For the exposition, we model the tax cash flows stemming from premium collection separately from the tax cash flows stemming from incurred losses. We do this to clarify the tax liabilities and the deferred tax assets; there is no qualitative difference in the cash flows.[13]

We have not yet included the investment income cash flows, since these depend on statutory accounting constraints and on the capital requirements.[14]

STATUTORY ACCOUNTING ENTRIES

The following statutory accounting entries are relevant for the implied equity flows:

---

[13] In contrast, Myers and Cohn distinguish between the taxes stemming from premium earning and those stemming from loss accruals because they use different capitalization rates to determine the present values of each. See Myers and Cohn [1987] as well as the discussion in Feldblum, [2003: *PCAS* d/d discussion].

[14] In the past, some casualty actuaries differentiated between investment income from policyholder supplied funds and investment income from equityholder supplied funds; cf Bailey [1967]. The investment income on policyholder supplied funds depends on the underwriting cash flows; the investment income on equityholder supplied funds depends also on the accounting entries and the capital requirements. The rationale was that policyholders were entitled to the investment income on their own funds but they were not entitled to the investment income on capital and surplus funds.

Although this distinction is not relevant to return on capital pricing models, it is useful for modeling the source of profits. If premiums are exactly adequate, the profit in the policyholder supplied funds (sometimes called policyholder supplied capital) is needed to fund the difference between the cost of equity capital and the after-tax investment yield on equityholder supplied capital. Myers and Cohn [1978] have a similar perspective, though they fail to take into account the equityholder supplied capital embedded in the statutory loss reserves and the gross unearned premium reserves.

- A unearned premium reserve of $1000 is set up on December 31, 20XX. It is amortized ratably as the insurance protection is provided. For most types of business, the amortization is even over the policy term.[15]

- Surplus of $250 is added to the balance sheet on December 31, 20XX, and it is removed on December 31, 20XX+1. The surplus requirement is *not* amortized over the policy term. This is a consequence of the NAIC risk-based capital formula and the corresponding rating agency formulas. The illustrations evaluate capital requirements by applying the risk-based capital charges to the written premium for the year and to the loss reserves at the valuation date.[16]

- A deferred tax asset of $200 × 35% = $70 stemming from the revenue offset provision is entered on the balance sheet on December 31, 20XX, and it is amortized over the course of the policy term. The full deferred tax asset from revenue offset is recognized on the statutory balance sheet, since it reverses within 12 months of the balance sheet date (for annual policies).[17]

IMPLIED EQUITY FLOW AT POLICY INCEPTION

We discuss the implied equity flows in more detail further below. The implied equity flows affect the investment income, and the investment income affects the implied equity flows. We show the implied equity flow at policy inception before discussing the investment income for the first half of the year to clarify the inter-relationship of these items.

The required assets of the company on December 31, 20XX, equal the $1,000 of unearned premium reserve plus the $250 of required surplus, or $1,250. The company holds statutory assets equal to the following:

---

[15] See SFAS 60 and SSAP No. 65. Exceptions occur for certain lines of business. (a) Workers' compensation premiums may be earned when billed for statutory accounting purposes, though not for tax purposes; see SSAP No. 53, "Property-Casualty Contracts – Premiums," paragraph 4 and IRS tax regulation 2001 FED ¶ 26,153, §1.832-4, sections (a)(4) and (a)(5). Companies which use different statutory accounting and tax accounting procedures for recording workers' compensation premium must use two sets of premium writing patterns in the pricing model. We do not show these scenarios in the exhibits. (b) Product warranty unearned premium reserves have more complex computations; see SSAP No. 65, "Property and Casualty Contracts," paragraphs 21-33, and Hayne [1999].

[16] Atkinson and Dallas [2000], chapter 8, use a slightly different procedure. They determine the risk-based capital requirements at year-end dates and they discount at the after-tax investment yield to the beginning of the year.

[17] See SFAS 109 for a general discussion of deferred tax assets and liabilities and SSAP No. 10, "Income Taxes," paragraph 10, for the statutory accounting rules on recognition of deferred tax assets. Appendix A of this paper reviews the post-codification statutory accounting rules for deferred tax assets and liabilities.

+ \$1000.00 in cash (written premium collected)
– \$ 250.00 in cash (acquisition costs paid)
+ \$ 17.50 in cash (tax refund)
+ \$ 70.00 (deferred tax asset, which is an admitted non-cash asset)
= \$ 837.50.

The remaining capital needed to fund the required assets is provided by equityholders: \$1250 – \$837.50 = \$412.50. We assume that all capital supplied by equityholders is investable.

INVESTMENT INCOME

The invested assets on December 31, 20XX, are \$750 (net premium) + \$412.50 (capital) + \$17.50 (tax refund) = \$1,180.00. The deferred tax asset of \$70 is not investable.

Equivalently, the total assets required are \$1,000 for the unearned premium reserve and \$250 in surplus, or \$1,250. Subtracting the \$70 deferred tax asset, which is not investable, gives \$1,180 of investable assets.

- For this illustration, we use an 8% per annum bond equivalent yield with semiannual evaluations. New money investment yields are often given in bond equivalent form.
- The investment income is received on June 30, 20XX+1, and December 31, 20XX+1, with a yield of 4% per half-year.
- The unearned premium reserve declines to \$500 on June 30, 20XX+1, and to \$0 on December 31, 20XX+1.
- The deferred tax asset declines to \$35 on June 30, 20XX+1, and to \$0 on December 31, 20XX+1. The change in the deferred tax asset reflects tax payments and refunds in 20XX+1.

The investment income during the period is the product of (i) the assets required at the beginning of the period minus the amount of non-investable (but admitted) assets and (ii) the investment yield during the time period.

Illustration: The investment income earned during the first half of 20XX+1 is 4% × \$1,180 = \$47.20. The investment income is assumed to be received on June 30, 20XX+1.

The assets required on July 1, 20XX+1, are \$500 of unearned premium reserve plus \$250 of required capital = \$750. The non-investable deferred tax asset is \$500 × 20% × 35% = \$35, and the investable assets are \$750 – \$35 = \$715. The investment income earned during the second half of 20XX+1 and received on Dec 31, 20XX+1, is \$715 × 4% = \$28.60.

Investment Income and Underwriting Expenses

For determining federal income taxes on underwriting income, we subtracted underwriting expenses. For determining investment income, we have *not* subtracted the underwriting expenses. At first glance, this appears incongruous, since money paid out as expenses is not available for investment. It might seem that the investable assets should also be reduced for underwriting expenses, just as the underwriting income is reduced for expenses.

The reasoning above is mistaken; it is mentioned here to clarify the equity flow modeling. The investment income is based on the investment yield times the investable assets. The investable assets are based on the statutory reserves and the required capital. Any expenses paid with policyholder supplied funds are replenished with equityholder supplied funds, since the company must hold the gross unearned premium reserve. Expense payments do not affect investable assets in the pricing model.

Rather, an expense outflow causes a federal income tax *inflow* and an implied equity *outflow*.

- The federal income tax *inflow* offsets 35% of the expense paid from the policy premium.
- The implied equity *outflow* offsets the other (1 – 35%) of the expense paid. The implied equity outflow is an investment in the insurance project, so it is modeled as a negative number. The equity outflow is an inflow to the company.[18]

The net change in the company's assets is $0: expense outflow – 35% federal income tax inflow – 65% implied equity flow into the company.

In practice, the amortization of the unearned premium reserves is offset by the accrual of loss reserves. The full pricing model considers premium writing patterns, premium collection patterns, loss accrual patterns, and loss payment patterns. To simplify the exposition, we have separated the premium section from the loss section. This highlights the equity inflows as the premium is earned and the equity outflows as losses are incurred. The numerical exhibits at the end of this paper show the combined premium and loss transactions.

In a companion paper ("Income Recognition and Performance Measurement"), we show net income under different accounting frameworks: statutory, GAAP, tax, fair value, net present value, and internal rate of return. No matter what accounting framework is used to measure income and management performance, the *statutory* accounting framework determines the capital contributions (cf. Atkinson and Dallas [2000], chapter 11).

---

[18] The premium is a pre-tax cash flow, so there is an offsetting tax liability or return. An equity flow is an after-tax flow, so there is no offsetting tax liability or return.

TAXES ON INVESTMENT INCOME

We examined earlier the federal income taxes on the net premium income. These are $113.75 each half year, payable on June 30, 20XX+1, and on December 31, 20XX+1; see page ?. Federal income taxes on investment income are (i) $47.20 × 35% = $16.52, paid on June 30, 20XX+1, and (ii) $28.60 × 35% = $10.01, paid on December 31, 20XX+1.

Some analysts presume that taxes are not paid until March or April of the following year. That is not correct. The federal income tax liability is computed for the tax year as whole, but payments to the U.S. Treasury are made quarterly in advance. For simplicity, this illustration shows payments on June 30, 20XX+1, and December 31, 20XX+1. The payment date for the tax liabilities is actually earlier than modeled here, not later.[19]

Although the taxes are paid earlier, the premium is also earned earlier, since it is earned evenly over the policy term. As long as the timing of the premium earning and the federal income taxes on the premium income are consistent, the model is not materially biased.

IMPLIED EQUITY FLOWS

The implied equity flow on any valuation date is (i) a capital distribution if it is positive or (ii) a capital contribution if it is negative. A capital distribution may be a stockholder dividend or a stock repurchase. Since common stocks are cash equivalents, a capital gain – even if unrealized – is also a capital distribution. Shareholders can sell some of their shares to produce a virtual dividend payment.[20]

---

[19] The method used in the text is equivalent to using an after-tax investment yield; cf. Atkinson and Dallas [2000], chapter 8. Some pricing models assume that tax liabilities for the current year are paid evenly during the year. The total investment income for the year is $47.20 + $28.60 = $75.80 and the tax is $75.80 x 35% = $26.53. If we assume that taxes are paid evenly between the two halves of the year, the tax payments at each semiannual valuation date are $26.53 ÷ 2 = $13.265.

The IRS allows taxpayers some leeway in the quarterly tax estimates, and companies differ in the timing of their tax payments. Modeling the precise tax payment stream is complex. It does not have a material effect on the rate indications, as long as reasonable assumptions are used.

The procedure in the text of the paper splits the tax on underwriting income evenly between the two halves of the year, but it computes the tax on investment income based on the investment income earned in each half of the year. It is difficult to quantify the amortization of the loss reserve discount between the two halves of the year, so we use an even spread. In contrast, using an after-tax investment yield is not difficult.

[20] The choice between paying dividends and allowing the capital to accumulate in the company depends on the investment opportunities of the company and the tax situation of the equityholders. This choice affects the personal tax liabilities of the investors and cost of holding capital. In practice, it is difficult to model personal income taxes, since they vary with the tax situation of the equityholders.

Illustration: The ABC Insurance Company has 10 million shares outstanding and a market value of $500 million on January 1, 20XX. The stock price is $50 per share. The company earns $50 million in 20XX, and it has a market value of $550 million on December 31, 20XX.

- *Scenario A:* The company pays a stockholder dividend of $5 per share on December 31, 20XX. The market value of the company declines to $500 million after the dividend. The stockholder dividend is the implied equity flow on December 31, 20XX.
- *Scenario B:* The company pays no stockholder dividend. The market value remains $550 million on December 31, 20XX. Since the common stock is a cash equivalent, which can be sold in the open market, the liquid assets of the owners increase by $5 per share. The capital accumulation is the implied equity flow on December 31, 20XX.

We may conceive of the implied equity flow from a balance sheet perspective or from an income statement perspective.

- *Balance sheet perspective:* At valuation dates (that is, between valuation periods), excess assets are distributed to equityholders, and insufficient assets are augmented by equityholder contributions. At the start of each valuation period, the held assets equal the required assets. The required assets are the sum of the liabilities and the required capital.

  During the valuation period, there are two types of changes in the balance sheet entries: (i) cash inflows and cash outflows affect investable assets, and (ii) non-cash increases and non-cash decreases affect non-cash assets and liabilities. The capital requirements may also change from one valuation date to the next valuation date.

  To determine the implied equity flow at any valuation date, we begin with the balance sheet entries at the start of the valuation period and we adjust them for both the cash inflows and outflows and the non-cash increases and decreases, including any change in capital requirements. The adjusted assets at the end of the period minus the required assets at the beginning of the period is the implied equity flow at the end of the period.

- *Income statement perspective:* The implied equity flow at valuation date "*t*" equals the statutory income during the period from "*t–1*" to "*t*" <u>minus</u> the change in required capital during this period. A direct charge or credit to surplus at the valuation date is treated as a component of statutory income.

The two definitions are equivalent. We show both methods to determine the implied equity flow on June 30, 20XX+1.

BALANCE SHEET PERSPECTIVE: The assets held at the beginning of the valuation period (on January 1, 20XX+1) are the required assets at that date. This equals $1250, which is the sum of the liabilities (the unearned premium reserve) and the required capital (the written premium risk charge). The $1250 consists of $1180 of investable assets and a $70 deferred tax asset.

The investable assets on the balance sheet are accumulated for investment income; the non-investable assets are capitalized or amortized.

- Cash inflow for investment income ($47.20 during the valuation period).
- Cash outflows for expenses ($150 in maintenance expenses) and tax accruals ($113.75 tax accrual for underwriting income; $16.52 tax accrual for investment income).[21]
- Amortization of the unearned premium reserve and of the deferred tax asset stemming from revenue offset ($500 amortization of the unearned premium reserve; $35 amortization of the deferred tax asset).

The loss reserves section of this illustration shows the capitalization of the incurred losses during the policy term and the capitalization and amortization of the deferred tax assets through the lifetime of the claims.

The accumulated assets at June 30, 20XX+1 equal the required assets at the beginning of the period adjusted for these cash flows and the non-cash increases and decreases:

Investable assets on December 31, 20XX	=	$1180.00
Investment income (cash inflow)	=	+$ 47.20
General expenses (cash outflow)	=	–$150.00
Federal income taxes on underwriting income (cash outflow)	=	–$113.75
Federal income taxes on investment income (cash outflow)	=	–$ 16.52
Deferred tax asset (non-cash asset)	=	+$ 35.00
Total	=	+$981.93

The required assets on June 30, 20XX+1, equal the statutory reserves plus the required capital. The unearned premium reserves are now $500 and the required capital remains $250, for a total of $750. The implied equity flow to the equityholders on June 30, 20XX+1, is $981.93 – $750.00 = $231.93.

INCOME STATEMENT PERSPECTIVE: The statutory income during the first half of 20XX+1 is shown below. Direct charges or credits to surplus, such as the change in the deferred tax asset, are included with statutory income.

---

[21] The tax payments are computed separately for underwriting income and investment income:

- underwriting income: $227.50 for the full year, or $113.75 for each half year, and
- investment income: $16.52 for the first half year and $10.01 for the second half year

Earned premium	=	+$500.00
Investment income	=	+$ 47.20
General expenses	=	–$150.00
Federal income taxes on underwriting income	=	–$113.75
Federal income taxes on investment income	=	–$ 16.52
Change in deferred tax asset	=	–$ 35.00
Total	=	+$231.93

There is no change in the required capital, so the implied equity flow is +$231.93. The two methods are alternative ways of describing the cash flows and balance sheet changes.

We calculate the implied equity flows for the second half of the year in the same two manners. For heuristic purposes, we show a third method below. A change in the implied equity flows between the first half of the year and the second half of the year stems from changes in statutory income or changes in capital requirements. We list these differences below.

a. General expense payments are $150 on June 30, 20XX+1, and $0 in the second half of the year.[22]
b. Investment income is $47.20 in the first half of the year and $28.60 in the second half of the year. The difference in the investment income reflects the difference in the investable assets.
c. The federal income tax on investment income is $16.52 in the first half of the year and $10.01 in the second half of the year.
d. Required capital declines to $0 on December 31, 20XX+1, since we have not yet included losses in the illustration.
e. The tax liability and the amortization of the deferred tax asset are spread evenly over the two halves of the year.

As an alternative to the full calculation, we adjust the implied equity flow from June 30, 20XX+1 with these differences to get the implied equity flow on December 31, 20XX+1:

Equity flow on June 30, 20XX+1	=	+$231.93
Difference in general expenses	=	+$150.00
Difference in investment income	=	+($28.60 – $47.20)
Difference in federal income taxes on investment income	=	–($10.01 – $16.52)
Difference in surplus change	=	+$250.00
Equity flow on December 31, 20XX+1	=	+$619.84

[22] This difference stems from the modeling assumptions. In truth, the maintenance expenses are incurred evenly over the year, but we assume a single payment date to simplify the modeling. For the federal income tax payments on underwriting income, we implicitly assumed that the maintenance expenses are incurred evenly over the course of the year, and we spread the federal income tax on underwriting income evenly over the year.

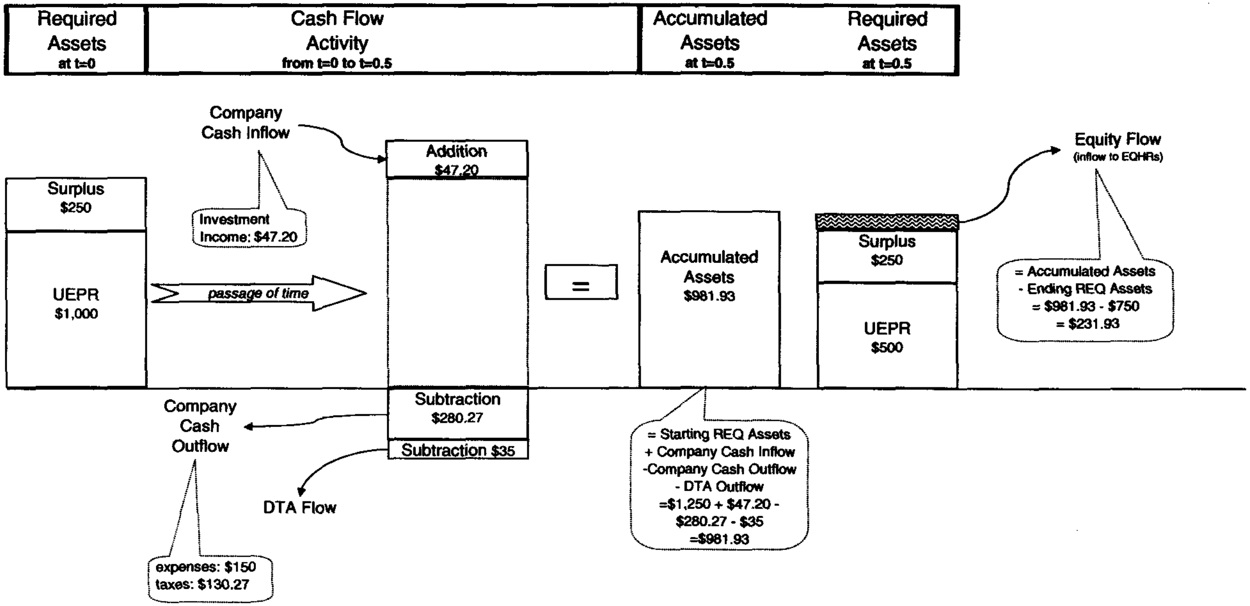
We show these adjustments to highlight the sources of the implied equity flows. This analysis is helpful for judging the reasonableness of the implied equity flows in each period. The actual pricing model performs separate calculations for each valuation date.[23]

The accompanying graphic shows the cash flow view of the implied equity flows. The following table shows the income statement view of the implied equity flows. These views are various perspectives of the same phenomenon.

---

[23] Instead of using values at the end of the period, some actuaries use average values for each entry and implied equity flows in the middle of the valuation period. To do this, we amortize the unearned premium reserve and the deferred tax asset evenly over the policy term. The invested assets decline from $1180 on December 31, 20XX, to $250 on December 31, 20XX+1. The average invested assets are ($1180 + $250) / 2 = $715.00, and the investment income is $715.00 × 8% = $57.20. This is less than the 20XX investment income of $47.20 + $28.60 = $75.80 in the illustration. The lower investment income is offset by the earlier implied equity flows in the alternative model, and there is no material change in the model results.

Cash Flow View of the Equity Flow



Income Statement View of Equity Flows

STATEMENT OF INCOME	0.00	0.50	1.00
Earned Premium		500	500
Undiscounted Incurred Loss		0	0
Paid Taxes		130.27	123.76
Undiscounted Incurred Expenses		150	0
Incurred Costs		280.27	123.76
Investment Income		47.2	28.6
Change in DTA		-35	-35
Net Income		231.93	369.84
CAPITAL and SURPLUS ACCOUNT			
Surplus, end of previous period		250	250
GAINS (and LOSSES) IN SURPLUS			
Net Income		231.93	369.84
Surplus Adjustments (Capital Contributions)		0	-250
Surplus, end of current quarter		250	0
Implied Equityflow		231.93	619.84

CASH FLOWS AND EQUITY FLOWS: PREMIUM AND EXPENSE TRANSACTIONS

The exposition above traces the cash flows and equations by type of transaction. We put the pieces together to show the cash flows and equity flows by valuation date. The four accompanying schematics show the cash flows and equity flows for *Illustration A: Premium and Expense Transactions*. The first schematic shows a summary of the cash flows and equity flows at the three valuation dates relevant to Illustration A. The next three schematics show more detailed information about the cash flows and equity flows at these valuation dates:

- t = 0.0, or December 31, 20XX–1
- t = 0.5, or June 30, 20XX
- t = 1.0, or December 31, 20XX

Each schematic shows transactions among the following seven nodes.

f. *Insurer:* All transactions pass through the insurance company and are taxed at a 35% rate. This causes the multiple layers of taxation of the profit margin in the policy premium.
g. *U/W:* All underwriting flows – premium collections, loss payments, and expense payments – are between the insurance company and the rectangle labeled *U/W*.
h. *IRS:* Tax payments are flows from the insurance company to the IRS; tax refunds are flows from the IRS to the insurance company.
i. *Assets:* To track the flow of funds, we imagine that there is a fiduciary handling the insurer's assets. All cash received by the insurer is sent to the rectangle labeled "assets."
j. *DTA:* The insurance company holds several types of non-cash assets, the most important of which is the deferred tax asset (DTA). There is no cash flow or equity flow underlying the non-cash assets. However, an increase in a non-cash asset causes an equal and offsetting reduction in the implied equity flows. An increase in a non-cash asset, such as a deferred tax asset, is shown as though it were a flow from the *DTA* rectangle to the *Assets* rectangle.
k. *FinMkts:* All cash assets are invested at the benchmark investment yield. The receipt of investment income by the insurance company is shown as a cash flow from the rectangle labeled *FinMkts* to the *Insurer*.
l. *Eqhr:* If the insurer needs more assets than it has from its underwriting and investment operations, the equityholders provide additional capital. If the insurer has excess assets, it distributes the excess capital to its equityholders.

ILLUSTRATION A (Case: No Losses) CASH FLOWS

	Time	0.0	0.5	1.0	1.5	2.0	2.5	3.0
	UW TRANSACTIONS							
(1)	Premium	1,000.00	0.00	0.00	0.00	0.00	0.00	0.00
(2)	Expense - Acquisition	250.00	0.00	0.00	0.00	0.00	0.00	0.00
(3)	Expense - General	0.00	150.00	0.00	0.00	0.00	0.00	0.00
(4)	Loss	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	OTHER							
	DTA							
(5)	Due to Revenue Offset	70.00	35.00	0.00	0.00	0.00	0.00	0.00
(6)	Due to Loss Reserve Discountin	0.00	0.00	0.00	0.00	0.00	0.00	0.00
(7)	TOTAL	70.00	35.00	0.00	0.00	0.00	0.00	0.00
	RESERVES							
(8)	UEPR	1,000.00	500.00	0.00				
(9)	Stat Loss Reserve		0.00	0.00	0.00	0.00	0.00	0.00
(10)	IRS Discount Factors			0.86		0.88		0.90
(11)	Tax Basis Loss Reserve			0.00		0.00		0.00
(12)	Surplus	250.00	250.00	0.00	0.00	0.00	0.00	0.00
(13)	**Total Assets**	1,250.00	750.00	0.00	0.00	0.00	0.00	0.00
(14)	**Investable Assets**	1,180.00	715.00	0.00	0.00	0.00	0.00	0.00
	TAXES							
(15)	Tax Basis UW Revenue	200.00		800.00	0.00	0.00	0.00	0.00
(16)	Tax Basis Expenses	250.00		150.00		0.00		0.00
(17)	Tax Basis Inc Loss	0.00		0.00		0.00		0.00
(18)	Tax Basis Inv Income	0.00	47.20	28.60	0.00	0.00	0.00	0.00
	Valuation of Taxes							
(19)	Tax on UW Income (Annual)	-17.50		227.50		0.00		0.00
(20)	Tax on UW Income (Semi-Annual Pa	-17.50	113.75	113.75	0.00	0.00	0.00	0.00
(21)	Tax on Invest Income	0.00	16.52	10.01	0.00	0.00	0.00	0.00
(22)	Total Semi-Annual Payment	-17.50	130.27	123.76	0.00	0.00	0.00	0.00
	CASH FLOWS							
(23)	Asset Flow	1,250.00	-500.00	-750.00	0.00	0.00	0.00	0.00
(24)	UW Flow	750.00	-150.00	0.00	0.00	0.00	0.00	0.00
(25)	Inv Inc Flow		47.20	28.60	0.00	0.00	0.00	0.00
(26)	Tax Flow	17.50	-130.27	-123.76	0.00	0.00	0.00	0.00
(27)	DTA Flow	70.00	-35.00	-35.00	0.00	0.00	0.00	0.00
(28)	Equityflow	-412.50	231.93	619.84	0.00	0.00	0.00	0.00

U/W Assumptions

Target Return on Capital = 12.0%
Invest Rate of Return = 8.0%
Premium = $1,000
Dollars of Ultimate Loss = $0
Combined Ratio = 40%
Loss Ratio = 0%

Surplus Assumptions

Premium Leverage Ratio = 25%
Reserve Leverage Ratio = 15%

Results

IRR on Equityflows (annual rate) = 136.8%

Economic Value Added = $360.08

Formulae For Exhibit 1A

CASE: No Losses

$(5)_t = 0.35 * 0.2 * (8)_t$
$(7)_t = (5)_t + (6)_t$

$(12)_t = 0.25 * WP$ for $t = 0, 0.5$
$(13)_t = (8)_t + (9)_t + (12)_t$
$(14)_t = (13)_t - (7)_t$

$(15)_t = (1)_t - [(8)_t - (8)_{t-1}]$
$(16)_t = (2)_t + (2)_{t-0.5} + (3)_t + (3)_{t-0.5}$
$(18)_t = (14)_{t-0.5}$ * interest rate
$(19)_t = [(15)_t - (16)_t - (17)_t] * 0.35$
$(20)_t = (19)_t$ for $t = 1, 2, 3$
$(20)_{t-0.5} = (19)_t$ for $t = 0.5, 1.5, 2.5$
$(21)_t = (18)_t * 0.35$

$(23)_t = (13)_t - (13)_{t-0.5}$
$(24)_t = (1)_t - (2)_t - (3)_t - (4)_t$
$(25)_t = (14)_{t-0.5}$ * interest rate
$(26)_t = (22)_t$
$(27)_t = (7)_t - (7)_{t-0.5}$
$(28)_t = -(23)_t + (24)_t + (25)_t + (26)_t$

Exhibit 1C

ILLUSTRATION A (Case: No Losses) Summary of Cash Flows

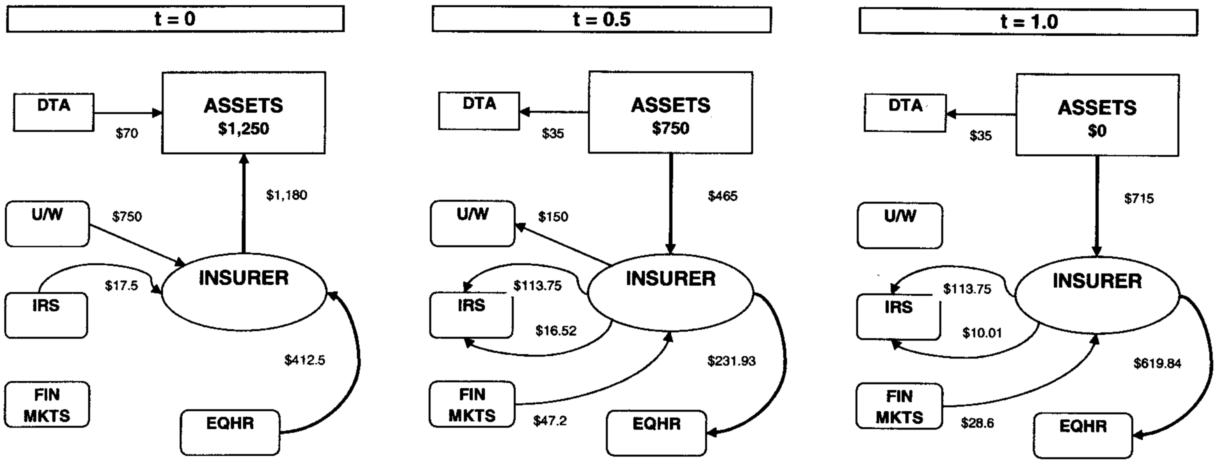
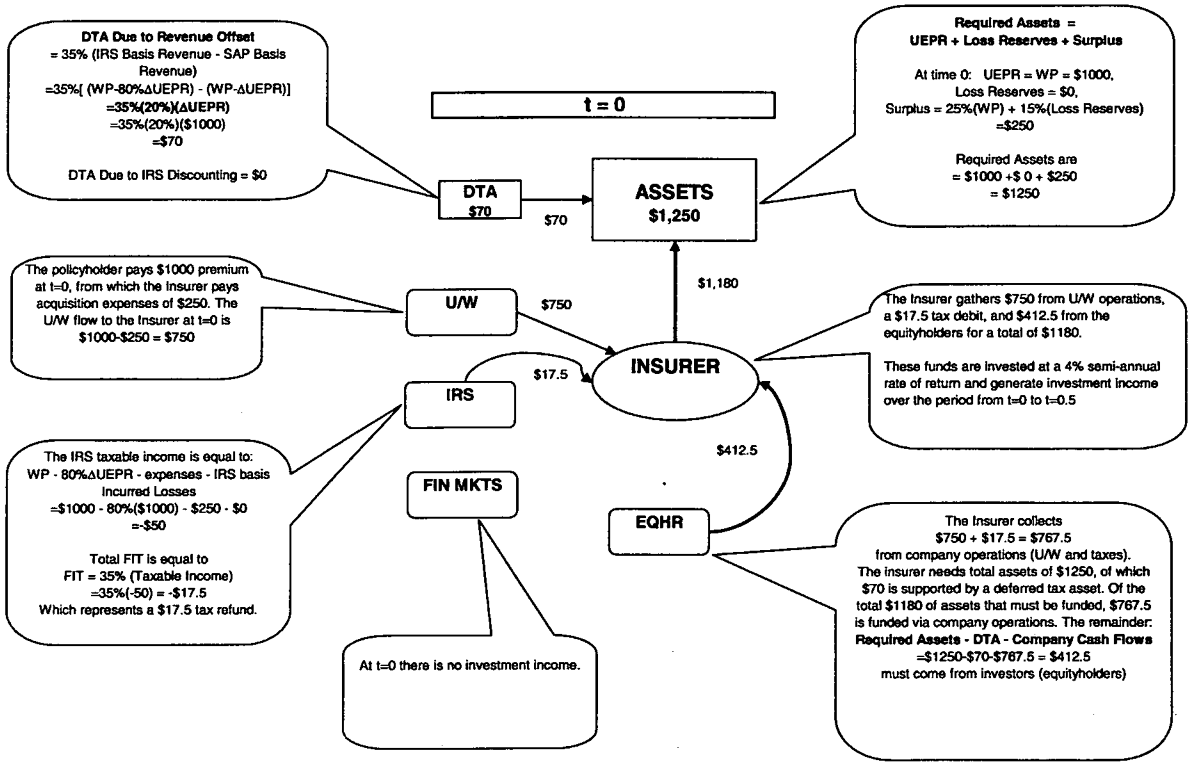


ILLUSTRATION A (No Losses Incurred)

Exhibit 1D



475

ILLUSTRATION A (No Losses Incurred)

Exhibit 1D

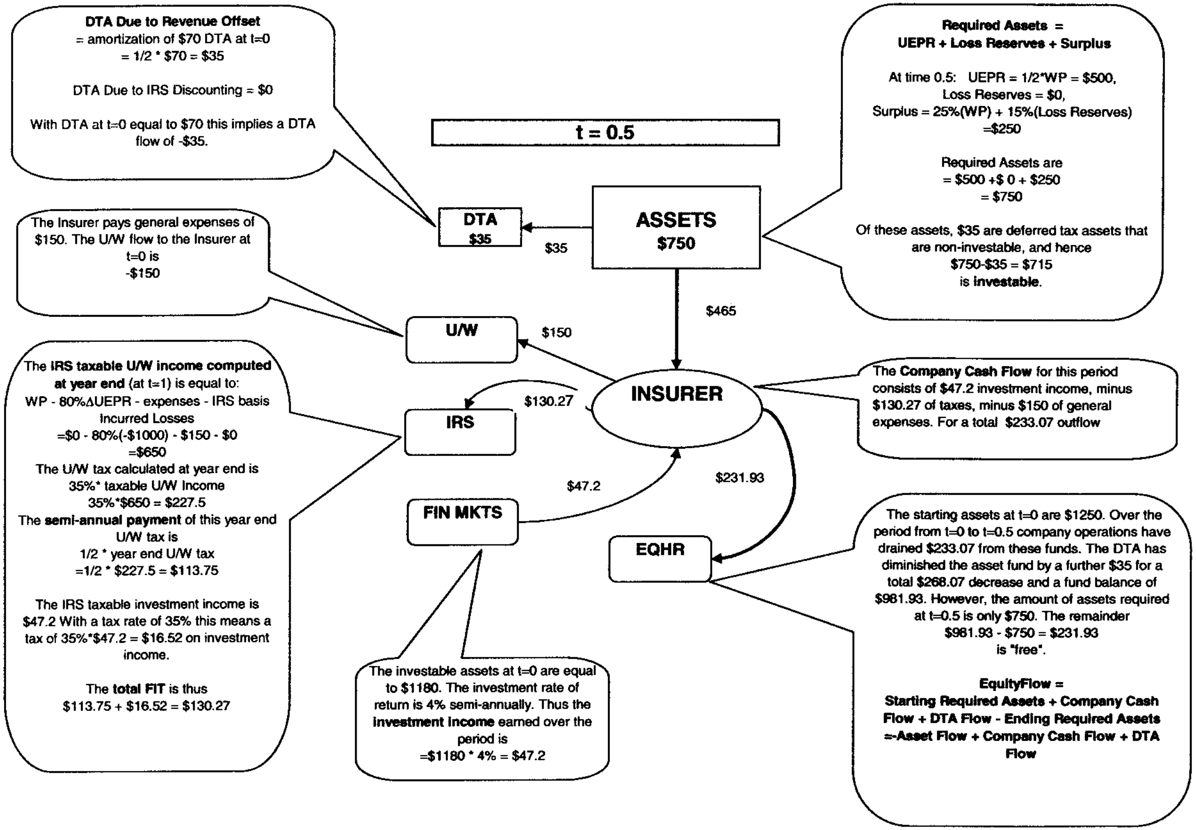


ILLUSTRATION A (No Losses Incurred)

Exhibit 1D

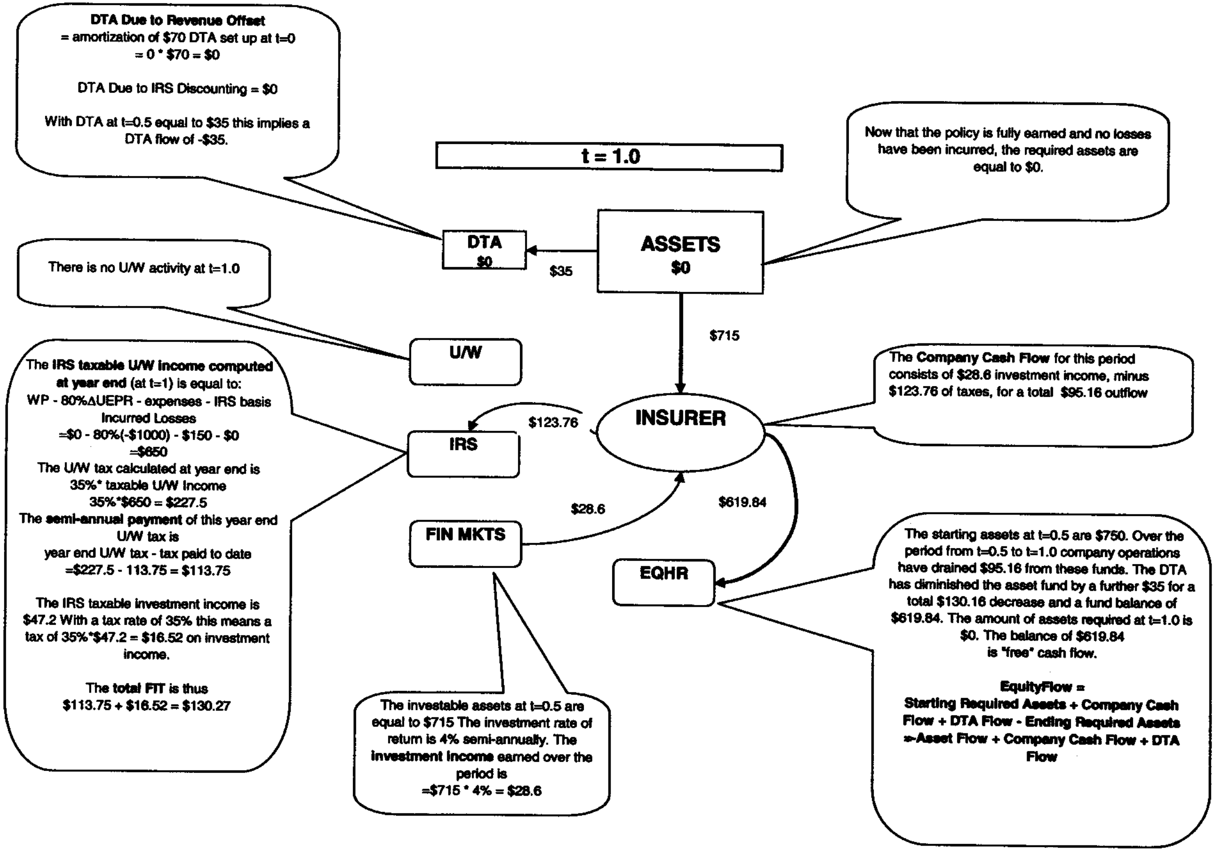


ILLUSTRATION B – LOSS TRANSACTIONS

We retain the scenario in the previous illustration, and we add losses incurred evenly during the policy term and paid over several years. To simplify the illustration, we model the losses as if there were two losses with ultimate values of $400 each occurring on June 30, 20XX+1, and December 31, 20XX+1. Both losses are paid on December 31, 20XX+3. Tax rates and capital requirements are the same as before.

There are several additional cash flows: a loss payment on December 31, 20XX+3 of $800, federal income tax payments or refunds for two and a half years between June 30, 20XX+1, and December 31, 20XX+3, and investment income over the same time period. There are non-cash changes to the balance sheet for loss reserves and for deferred tax assets stemming from IRS loss reserve discounting. There are changes in the capital requirements stemming from the reserving risk charge. The assumption of a single payment date simplifies the computations yet leaves enough detail to highlight the modeling principles.

Federal Income Taxes

We use the IRS loss reserve discount factors for computing taxable income and tax liabilities. The IRS provides loss reserve discount factors for each line of business and accident year.

The text of this paper presumes knowledge of IRS tax calculations and the post-codification statutory accounting rules for deferred tax assets and liabilities. The appendix to this paper contains a more complete exposition of the tax accounting rules.

For the illustration, we assume IRS loss reserve discount factors of 86%, 88%, and 90% for accident year 20XX+1 as of 12 months, 24 months, and 36 months, respectively. Since we are pricing prospectively, the actuary must estimate the future loss reserve discount factors based on federal mid-term rates and either industry or company loss payment patterns.

The offset to taxable income for 20XX+1 equals the tax rate times the change in the discounted loss reserve. The change in the discounted reserve for 20XX+1 is the discounted reserve itself, since losses first occur in 20XX+1. The computation for 20XX+1 is

$$35\% \times 86\% \times \$800 = \$240.80.$$

The offset to taxable income may be viewed either as

(i) 35% of the change in the IRS discounted losses or
(ii) 35% × (statutory incurred losses minus the change in the IRS loss reserve discount).

On December 31, 20XX+1, the IRS discounted reserves are 86% x \$800 = \$688.00, and the IRS loss reserve discount is (1 – 86%) x \$800 = \$112.00. Since there are no loss reserves on December 31, 20XX, these figures are also the year-to-year changes in these quantities.

We pro-rate the annual tax liability among the portions of the year to estimate tax payments. For the semi-annual valuation periods in this illustration, we assume tax offsets of \$120.40 at each valuation date: June 30, 20XX+1, and December 31, 20XX+1.

The tax rate times the change in discounted reserves during 20XX+2 is

$$35\% \times (88\% \times \$800 - 86\% \times \$800) = \$5.60.$$

The change in discounted reserves is an offset to taxable income, so the tax rate times the change in discounted reserves is the offset to the tax liability for December 31, 20XX+2. We split the offset of \$5.60 into equal halves of \$2.80 each for the semi-annual valuation periods.

The tax basis incurred loss in 20XX+3 is the paid loss plus the change in discounted reserves:

$$\$800 + (90\% \times \$0 - 88\% \times \$800) = \$96.00.$$

The offset to the tax liability on December 31, 20XX+3, equals

$$35\% \times (\$800 + 90\% \times \$0 - 88\% \times \$800) = \$33.60.$$

We split the \$33.60 offset into equal halves of \$16.80 each for semi-annual valuation periods.

Unless a policy is written on December 31 or January 1, there are two accident years for tax purposes. This complicates the pricing model, so we have chosen a December 31 effective date for the illustration.[24]

Cash Flow and Equity Flow Patterns

The illustration shows the occurrence of losses and the payment of losses in different years. The company cash flows for incurred losses show the following pattern:

- There are significant cash *inflows* stemming from the offset to taxable income on the dates the losses occur (June 30, 20XX+1, and December 31, 20XX+1). The cash inflow from the offset to taxable income precedes the cash outflow from the payment of losses.
- There is a large cash outflow on the date the loss is paid, or December 31, 20XX+3.

[24] The exhibits in Appendix B use an effective date of July 1, which is a proxy for the average effective date for policies written evenly through the year. Two accident years are used to evaluate the underwriting tax effects. This is the standard modeling technique for property-casualty insurance policies.

- The offset to taxable income stemming from the unwinding of the IRS loss reserve discount is balanced by the investment income on the assets backing the discounted reserves. The total cash inflow during the period equals the *pre-tax* investment income. The balancing is not perfect, since it depends on the accuracy of the loss reserve discount factors.
- The pricing model treats the federal income taxes on the investment income separately from the tax refund on the amortization of the discount in the IRS loss reserves. The offsetting cash flows are shown separately.
- The investment income on the capital supporting the reserves is not balanced by amortization of the interest discount in the reserves. The supporting capital comprises both the capital embedded in full value reserves and the capital held in surplus for the reserving risk charge.
- Changes in the deferred tax asset are not company cash flows.

The implied equity flows show a different pattern. The losses cause large equity outflows during 20XX+1 and a modest equity inflow on December 31, 20XX+3.

We assume initially that the company posts full value loss reserves on June 30, 20XX+1, and December 31, 20XX+1. If the coverage is priced adequately, the policyholder premium provides for the present value of the expected losses plus an amount to fund the cost of holding capital and the associated federal income taxes.[25] The equityholders fund the loss reserve with assets equal to

♦ the difference between the held reserve and the present value of the reserve
\+ ♦ the capital requirement for reserving risk
– ♦ the amount of capital provided by the policyholders.[26]

To distinguish the components of the loss reserves and the sources of capital, we conceive of the held reserve as the present value of future loss payments plus the capital embedded in the statutory held reserve.

The *capital requirement* is the explicit capital in the surplus account. For illustration, we assume a capital requirement equal to 15% of held loss reserves. This is consistent with current NAIC risk-based capital requirements.[27]

---

[25] On the cost of holding capital, see Feldblum and Thandi, "Target Return on Capital" and "Federal Income Taxes and the Cost of Holding Capital."

[26] Adjustments must be made for the tax refund and the admitted portion of the deferred tax asset stemming from IRS loss reserve discounting.

[27] The risk-based capital requirements for the long-tailed lines of business (except workers' compensation) are higher than 15%, and companies hold surplus about twice the risk-based capital requirements. The covariance adjustment in the risk-based capital formula reduces the effective risk charge by about 50%. The 15% factor understates the capital requirements for the long-tailed casualty lines of

The effects of the reserve valuation rate on the NPV and IRR calculations are examined in a separate paper (Feldblum and Thandi [2002], "Reserve Valuation Rates"). The loss reserve valuation rate is an accounting item; it does not affect the loss cash flows. However, it affects the loss reserve, the capital requirements, the tax payments, and the deferred tax asset, all of which affect the implied equity flows and the internal rate of return. For this paper, we assume that loss reserves are held at full value (undiscounted value).

Computations

On June 30, 20XX+1, the statutory loss reserves are $400 and the required capital is $400 × 15% = $60.00. On December 31, 20XX+1, the statutory loss reserves increase to $800 and the required capital increases to $800 × 15% = $120.00.

The federal income tax liability on December 31, 20XX+1, resulting from the incurred losses is –$240.80, computed as –35% × 86% × $800. The total deferred tax asset (DTA) is $5.60 + $33.60 = $39.20. This is the deferred tax asset on a GAAP balance sheet, as well as the deferred tax asset in column 1 of the statutory balance sheet (line 15 of page 2 in 2001). Of the $39.20 DTA, only $5.60 reverses within 12 months. This is the admitted portion on the 12/31/XX+1 statutory balance sheet.

We assume that the tax on underwriting income is paid (or the offset to taxable income is received) evenly over the year. For simplicity, we assume that half the deferred tax asset is accrued on June 30, 20XX+1, and the other half is accrued on December 31, 20XX+1.

To fund the incurred losses at June 30, 20XX+1, the policyholders and equityholders must provide assets equal to

the held reserves + the capital requirement – the tax refund – the deferred tax asset, or

$400 + $60 – $120.40 – $2.80 = $336.80.

The policyholder funds are provided by the policy premium. At policy inception, the money is transferred to the unearned premium reserves. Over the course of the year, the money in the

---

business and somewhat overstates the capital requirements for workers' compensation. The pricing model in the appendix uses actual factors or best estimates by line of business.

The illustration does not explicitly distinguish between the reserving risk charge applied to held reserves (R_5) and the asset risk charges applied to the assets backing the held reserves (R_1 and R_2). Because the marginal effect of a risk charge varies directly with the magnitude of the charges in its risk category, the marginal effect of the asset risk charges is only about 10% to 20% of the marginal effect of the reserving risk charges for the long-tailed lines of business; see Feldblum [1996: RBC]. One may conceive of the 15% capital requirement as a 13% to 14% reserving risk charge and a 1% to 2% asset risk charge.

unearned premium reserves is transferred to the loss reserves. The equityholder funded capital is a capital infusion at the time of the loss occurrence.

If the policy is adequately priced, the policyholder funds the fair value of the losses plus an amount to cover the cost of holding capital and the associated taxes. The equityholders provide the capital embedded in the reserves and the capital explicitly held in statutory surplus, minus the capital provided by the policyholders.

The assets needed to support the incurred losses at December 31, 20XX+1, are:

$800 + $120 – $240.80 – $5.60 = $673.60.

We have separated the premium transactions from the loss transactions in this illustration to highlight the relationships among the company cash flows and the implied equity flows.

- The premium transactions show the funds supplied at time t=0 by both policyholders and equityholders to support the unearned premium reserve and the initial underwriting loss. At times t=½ and t=1, the funds are distributed to the equityholders.
- The loss transactions show equityholder supplied funds used at times t=½ and t=1 to support the loss reserves. At time t=3, the remaining funds are returned to the equityholders.

In practice, loss reserves and paid losses gradually replace the unearned premium reserves. The policyholder supplied funds collected at time t=0 to support the unearned premium reserve are transferred to support the loss reserves as the premium is earned and the losses are incurred. The profit in the policyholder premium is transferred gradually to the equityholders over the same time period to provide the required return on the invested capital.

In the illustration, there is no change in the undiscounted reserves between December 31, 20XX+1, and December 31, 20XX+3. We use this simplistic scenario to clarify the cash flows and implied equity flows, without having to deal with changing loss reserves and paid losses at each valuation date. In practice, reserves run off gradually.

DEFERRED TAX ASSET: IRS LOSS RESERVE DISCOUNTING

The deferred tax assets are computed at year-end dates. To clarify the exposition, we begin this sub-section with annual valuation periods, not semi-annual periods. We then turn to semi-annual valuation periods to explain the calculation of the deferred tax assets at the mid-year valuation dates.

On December 31, 20XX+1, the offset to *statutory* income stemming from the incurred losses is $800. The federal income tax offset that would result from an offset to taxable income of $800 is 35% × $800 = $280.00.

- The actual offset to taxable income on December 31, 20XX+1, is 86% × $800 = $688.00.
- The offset to the federal income tax liability is 35% × 86% × $800 = $240.80.

The actual federal income tax liability is greater than the federal income tax liability that is implied by the statutory balance sheet. The difference is $280.00 – $240.80 = $39.20.

The difference of $39.20 is recouped by tax refunds in subsequent years. Since statutory accounting recognizes the full value loss reserve on the occurrence date, it should recognize the full tax offset on that date as well. Tax accounting defers the last $39.20 of the offset over the period during which the interest discount unwinds. Both GAAP and statutory accounting treat this as a receivable. It is shown as a deferred tax asset on the balance sheet.

GAAP financial statements recognize the full receivable if the firm expects to collect it. Statutory accounting admits only the portion of the deferred tax asset that is expected to reverse within 12 months of the statement date. To calculate the admitted portion of the deferred tax asset on the December 31, 20XX+1, statutory balance sheet, we estimate the portion of this deferred tax asset that remains on December 31, 20XX+2.

- On December 31, 20XX+2, the offset to taxable income is 88% × $800 = $704.00.
- The offset to the federal income tax liability is 35% × 88% × $800 = $246.40.
- The change in the federal income tax between December 31, 20XX+1, and December 31, 20XX+2, is $246.40 – $240.80 = $5.60.

Date	Statutory	Tax	Difference	Change
12/31/20XX+1	$800	$688	$112	—
12/31/20XX+2	$800	$704	$96	$16

35% × $16 = $5.60.

The expected change in the deferred tax asset from the current valuation date to the valuation date one year hence is the portion of the deferred tax asset that is recognized on the statutory balance sheet.

The appendix explains the calculation of the deferred tax asset when there are multiple loss payments during the year.

INVESTABLE ASSETS

We calculate the investable assets to determine the expected investment income. We show first the procedure for annual valuation periods, and we then extend the computations to semi-annual valuation periods.

Investable Assets, with annual valuation periods: The investable assets equal the total assets minus the admitted portions of any non-investable assets on the statutory balance sheet. The non-investable assets include agents' balances, earned but unbilled premiums, accrued retrospective premiums, and deferred tax assets.

The admitted portion of the DTA is $5.60 at December 31, 20XX+1, and $33.60 at December 31, 20XX+2. The investable assets are

- $920.00 – $5.60 = $914.40 during 20XX+2 and
- $920.00 – $33.60 = $886.40 during 20XX+3.[28]

Investable Assets, with semi-annual valuation periods: The deferred tax asset at December 31 of year X is set up evenly over the course of year X and declines linearly to zero over the course of year X+1. During year X+1, a new deferred tax asset is set up, which declines linearly to zero over the course of year X+2.[29] Each deferred tax asset follows an accrual and amortization pattern shaped like a carot ("/\"). Each carot is two years long.

The rationale for the two year up-down pattern is the assumption that losses are paid evenly during the calendar year. Before a loss is paid, there is a gross (GAAP) deferred tax asset associated with its reserve. The deferred tax asset is admitted on the statutory balance sheet only during the 12 month period immediately prior to its payment.

Illustration: The $5.60 deferred tax asset at December 31, 20XX+1, is accrued evenly over 20XX+1: $2.80 on June 30, 20XX+1, and the remainder on December 31, 20XX+1. It declines to $2.80 at June 30, 20XX+2, and to $0 by December 31, 20XX+2.

A new deferred tax asset of $33.60 is shown at the December 31, 20XX+2 valuation date. It accrues evenly over 20XX+2 ($16.80 on June 30, 20XX+2, and $16.80 on December 31, 20XX+2). It declines to $16.80 at June 30, 20XX+3, and to $0 on December 31, 20XX+3.

The total deferred tax asset on June 30, 20XX+2, is $2.80 + $16.80 = $19.60.

---

[28] The implication of this reasoning seems to be that if more of the DTA is not admitted, the investable assets increase. This raises the investment income, which lowers the need for underwriting income, thereby causing a smaller rate indication.

In fact, the investable assets increase only because the non-admitted DTA is replaced by equityholder supplied funds, which are investable. This raises the invested capital, and it more than offsets the higher investment income. The net result is to raise the rate indication, not to lower it.

[29] Since we are using discrete (semi-annual) functions, not continuous functions, "evenly over the course of the year" means one half on June 30 and the other half on December 31.

This modeling procedure interpolates between the deferred tax assets at December 31 of year X and December 31 of year X+1 to derive the deferred tax asset at June 30 of year X+1.

- The deferred tax asset at June 30, 20XX+1, is ($0.00 + $5.60)/2 = $2.80.
- The deferred tax asset at June 30, 20XX+2, is ($5.60 + $33.60)/2 = $19.60.
- The deferred tax asset at June 30, 20XX+3, is ($33.60 + $0)/2 = $16.80.

The change in the deferred tax asset is not itself a cash flow. However, the recognized portion of the deferred tax asset is an admitted statutory asset. A change in the admitted portion of the deferred tax asset causes an implied equity flow. Since the deferred tax asset is not investable but the capital contribution from the equityholders is investable, the change in the deferred tax asset also affects the investment income during the year. A decrease in the deferred tax asset causes an increase in the capital contributed by equityholders. This causes an increase in the expected investment income and an associated increase in the federal income taxes on this investment income.

The total assets held by the company are $920 throughout the two years 20XX+2 and 20XX+3. The investable assets are $914.40 at December 31, 20XX+1, $917.20 at June 30, 20XX+2, $886.40 at December 31, 20XX+2, and $903.20 at June 30, 20XX+2.

Period	Total Assets	Deferred Tax Asset	Investable Assets
1/1 – 6/30/XX+2	$920.00	$5.60	$914.40
7/1 – 12/31/XX+2	$920.00	$19.60	$900.40
1/1 – 6/30/XX+3	$920.00	$33.60	$886.40
7/1 – 12/31/XX+3	$920.00	$16.80	$903.20

The investment income during the first half of 20XX+2 equals the investable assets times the investment yield, or $914.40 × 4% = $36.58. We use the same computation for each semi-annual valuation period in 20XX+2 and 20XX+3:

Period	Investable Assets	Investment Yield	Investment Income
1/1 – 6/30/XX+2	$914.40	4% (half year)	$36.58
7/1 – 12/31/XX+2	$900.40	4% (half year)	$36.02
1/1 – 6/30/XX+3	$886.40	4% (half year)	$35.46
7/1 – 12/31/XX+3	$903.20	4% (half year)	$36.13

The implied equity flow at each valuation date equals the statutory income during the preceding period minus the change in capital requirements from the beginning to the end of the period. The components of statutory income during the first half of 20XX+2 are as follows:

- Investment income is $36.58.
- The tax on the investment income is 35% × $36.58 = $12.80.
- The change in the deferred tax asset is $19.60 – $5.60 = $14.00.
- The tax refund for the 20XX+2 amortization of the loss reserve is $5.60. We split the tax refund evenly over the year, giving a tax refund of $2.80 for the first half of 20XX+2.

The total statutory income is $36.58 – $12.80 + $14.00 + 2.80 = $40.58. There is no change in the required capital, so the implied equity flow is +$40.58.

Period	*Investment Income*	*Tax on Inv Income*	*Change in DTA*	*Tax on Reserve Amortization*	*Statutory Income*
1/XX+2	$36.58	$12.80	$14.00	–$2.80	$40.58
2/XX+2	$36.02	$12.61	$14.00	–$2.80	$40.21
1/XX+3	$35.46	$12.41	–$16.80	–$16.80	$23.05
2/XX+3	$36.13	$12.65	–$16.80	–$16.80	$23.48

For the first three half years in the table above, the implied equity flows equal the statutory income. On December 31, 20XX+3, the loss is paid, the loss reserve decreases to $0, and the required capital decreases from $120 to $0. The implied equity flow on December 31, 20XX+3 is

$23.48 – (– $120.00) = $143.48.

Cash Flows and Equity Flows: Loss Transactions

The exposition above traces the cash flows and equations by type of transaction. We put the pieces together to show the cash flows and equity flows by valuation date. The five accompanying schematics show the cash flows and equity flows for *Illustration B: Premium, Expense, and Loss Transactions.*

The first schematic shows a summary of the cash flows and equity flows at all seven valuation dates from t=0 through t=3. The subsequent four schematics show more detailed information about the cash flows and equity flows at four of these valuation dates:

- t = 0.5, or June 30, 20XX
- t = 1.0, or December 31, 20XX
- t = 2.5, or June 30, 20XX+2
- t = 3.0, or December 31, 20XX+2

At times t=0.5 and t=1.0, the losses are incurred. The schematics show the relevant cash flows and equity flows for setting up case reserves, added the reserving risk charge to the risk-based capital requirements, obtaining the tax offset for incurred losses, and setting up the deferred tax asset stemming from IRS loss reserve discounting.

At time t=1.5, 2.0, and 2.5, there are no accidents or loss payments. The cash flows and equity flows stem from changes in the IRS discounted reserves, leading to tax payments and changes in the deferred tax asset. In addition, there are investment income flows stemming from the investment return on the assets backing the loss reserves and the reserving risk charge of the risk-based capital requirements.

At time t=3.0, the losses are paid and the assets supporting the reserving risk charge are returned to the equityholders.

The schematics show transactions among the same seven nodes as used for the premium and expense transactions.

ILLUSTRATION B (Case: Losses Incurred) CASH FLOWS

Time	0.0	0.5	1.0	1.5	2.0	2.5	3.0
UW TRANSACTIONS							
(1) Premium	1,000.00	0.00	0.00	0.00	0.00	0.00	0.00
(2) Expense - Acquisition	250.00	0.00	0.00	0.00	0.00	0.00	0.00
(3) Expense - General	0.00	150.00	0.00	0.00	0.00	0.00	0.00
(4) Loss	0.00	0.00	0.00	0.00	0.00	0.00	800.00
OTHER							
DTA							
(5) Due to Revenue Offset	70.00	35.00	0.00	0.00	0.00	0.00	0.00
(6) Due to Loss Reserve Discountin	0.00	2.80	5.60	19.60	33.60	16.80	0.00
(7) TOTAL	70.00	37.80	5.60	19.60	33.60	16.80	0.00
RESERVES							
(8) UEPR	1,000.00	500.00	0.00				
(9) Stat Loss Reserve		400.00	800.00	800.00	800.00	800.00	0.00
(10) IRS Discount Factors			0.86		0.88		0.90
(11) Tax Basis Loss Reserve			688.00		704.00		0.00
(12) Surplus	250.00	310.00	120.00	120.00	120.00	120.00	0.00
(13) **Total Assets**	1,250.00	1,210.00	920.00	920.00	920.00	920.00	0.00
(14) **Investable Assets**	1,180.00	1,172.20	914.40	900.40	886.40	903.20	0.00
TAXES							
(15) Tax Basis UW Revenue	200.00		800.00	0.00	0.00	0.00	0.00
(16) Tax Basis Expenses	250.00		150.00		0.00		0.00
(17) Tax Basis Inc Loss	0.00		688.00		16.00		96.00
(18) Tax Basis Inv Income	0.00	47.20	46.89	36.58	36.02	35.46	36.13
Valuation of Taxes							
(19) Tax on UW Income (Annual)	-17.50		-13.30		-5.60		-33.60
(20) Tax on UW Income (Semi-Annual Pa	-17.50	-6.65	-6.65	-2.80	-2.80	-16.80	-16.80
(21) Tax on Invest Income	0.00	16.52	16.41	12.80	12.61	12.41	12.64
(22) Total Semi-Annual Payment	-17.50	9.87	9.76	10.00	9.81	-4.39	-4.16
CASH FLOWS							
(23) Asset Flow	1,250.00	-40.00	-290.00	0.00	0.00	0.00	-920.00
(24) UW Flow	750.00	-150.00	0.00	0.00	0.00	0.00	-800.00
(25) Inv Inc Flow		47.20	46.89	36.58	36.02	35.46	36.13
(26) Tax Flow	17.50	-9.87	-9.76	-10.00	-9.81	4.39	4.16
(27) DTA Flow	70.00	-32.20	-32.20	14.00	14.00	-16.80	-16.80
(28) Equityflow	-412.50	-104.87	294.93	40.57	40.21	23.05	143.48

U/W Assumptions

Target Return on Capital = 12.0%

Invest Rate of Return = 8.0%

Premium = $1,000

Dollars of Ultimate Loss = $800

Combined Ratio = 120.0%

Loss Ratio = 80%

Surplus Assumptions

Premium Leverage Ratio = 25%

Reserve Leverage Ratio = 15%

Results

IRR on Equityflows (annual rate) = 3.0%

Economic Value Added = -$62.49

Exhibit 2B

Formulae For Exhibit 1B

CASE: Losses Incurred

$(5)_t = 0.35 * 0.2 * (8)_t$
$(6)_t = 0.35 * \{ [(9)_t - (11)_t] - [(9)_{t+1} - (11)_{t+1}] \}$
$(7)_t = (5)_t + (6)_t$

$(11)_t = (9)_t * (10)_t$
$(12)_t = \begin{cases} 0.25 * WP + 0.15 * (9)t & \text{for } t = 0, 0.5 \\ 0.15 * (9)t & \text{for } t >= 1.0 \end{cases}$
$(13)_t = (8)_t + (9)_t + (12)_t$
$(14)_t = (13)_t - (7)_t$

$(15)_t = (1)_t - [(8)_t - (8)_{t-1}]$
$(16)_t = (2)_t + (2)_{t-0.5} + (3)_t + (3)_{t-0.5}$
$(17)_t = (4)_t + (4)_{t-0.5} + [(11)_t - (11)_{t-1}]$
$(18)_t = (14)_{t-0.5}$ * interest rate
$(19)_t = [(15)_t - (16)_t - (17)_t] * 0.35$
$(20)_t =$ $(19)_t$ for t = 1, 2, 3
$(20)_{t-0.5} =$ $(19)_t$ for t = 0.5, 1.5, 2.5
$(21)_t = (18)_t * 0.35$

$(23)_t = (13)_t - (13)_{t-0.5}$
$(24)_t = (1)_t - (2)_t - (3)_t - (4)_t$
$(25)_t = (14)_{t-0.5}$ * interest rate
$(26)_t = (22)_t$
$(27)_t = (7)_t - (7)_{t-0.5}$
$(28)_t = -(23)_t+(24)_t+(25)_t+(26)_t$

Exhibit 2C

ILLUSTRATION B (Case: Losses Incurred) Summary of Cash Flows

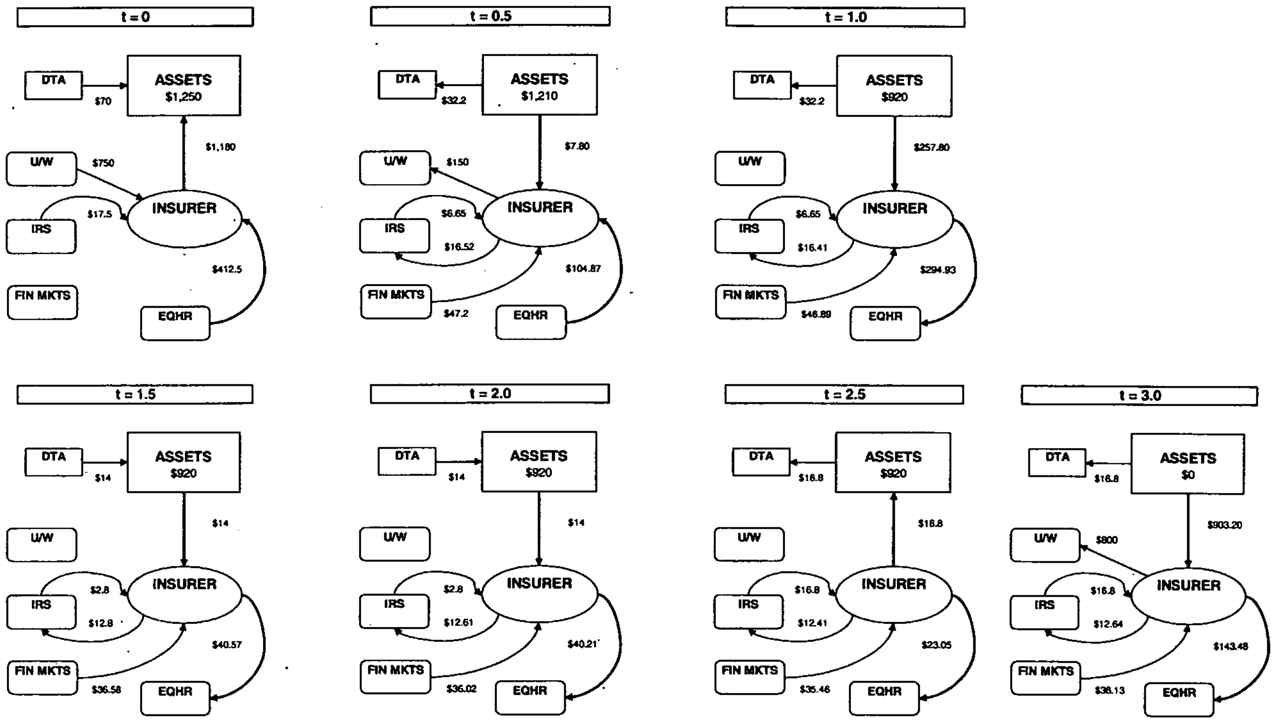


ILLUSTRATION B (Case Losses Incurred)

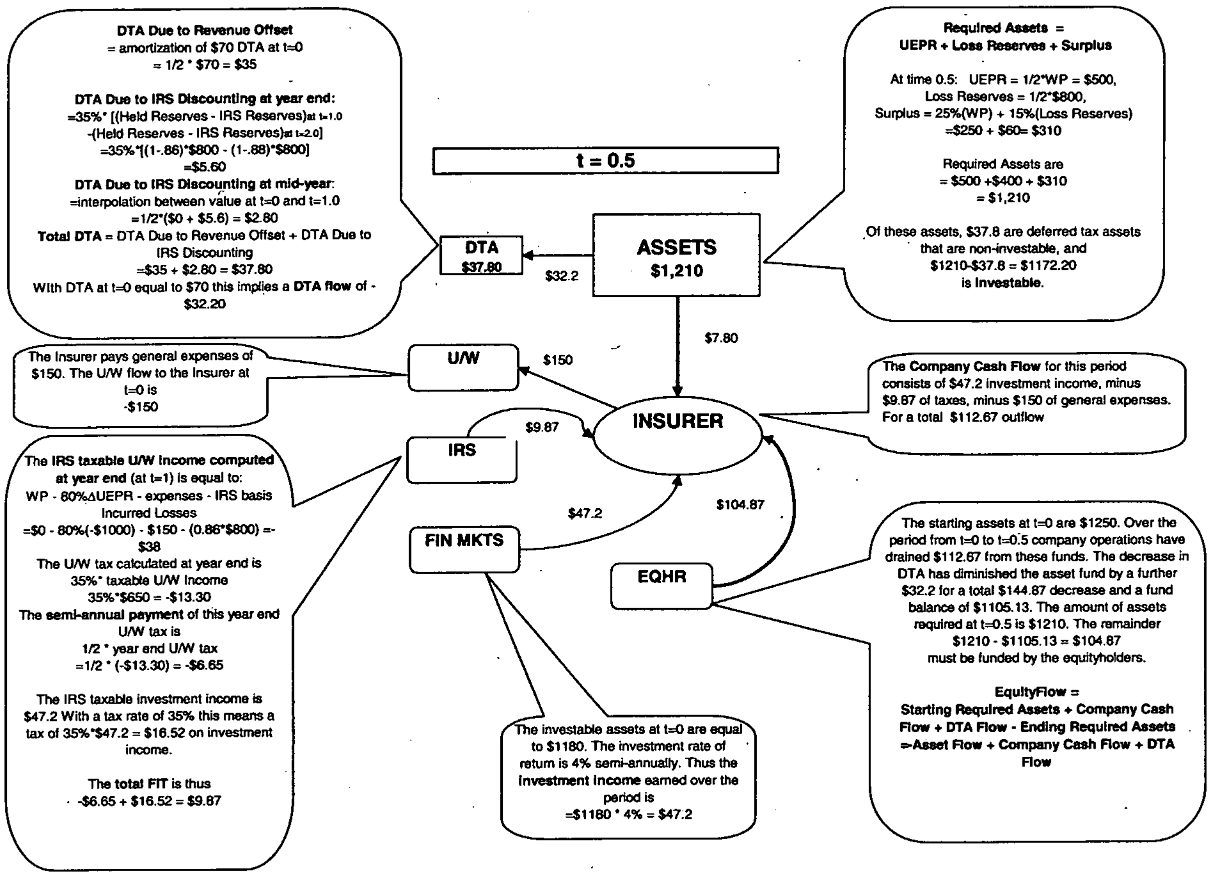


ILLUSTRATION B (Case Losses Incurred)

Exhibit 2D

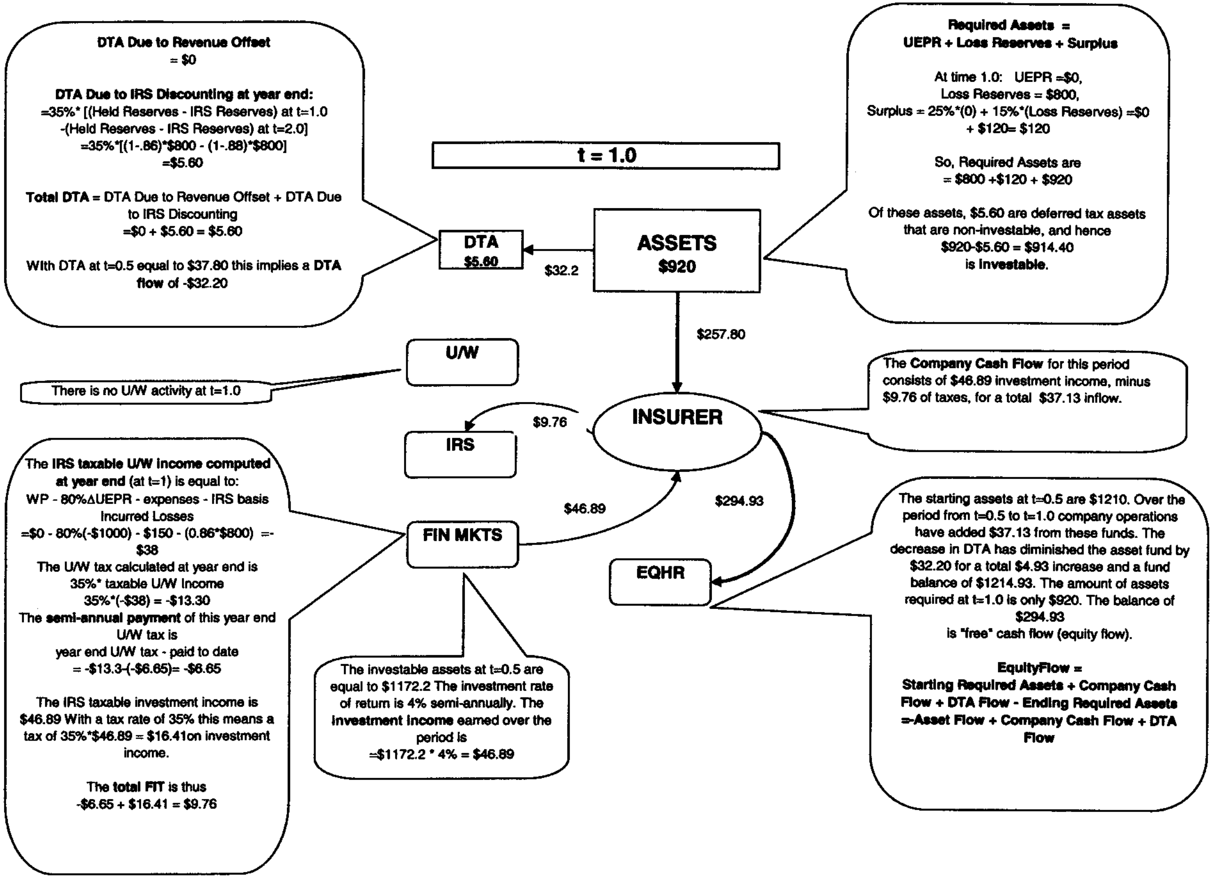


ILLUSTRATION B (Case Losses Incurred)

Exhibit 2D

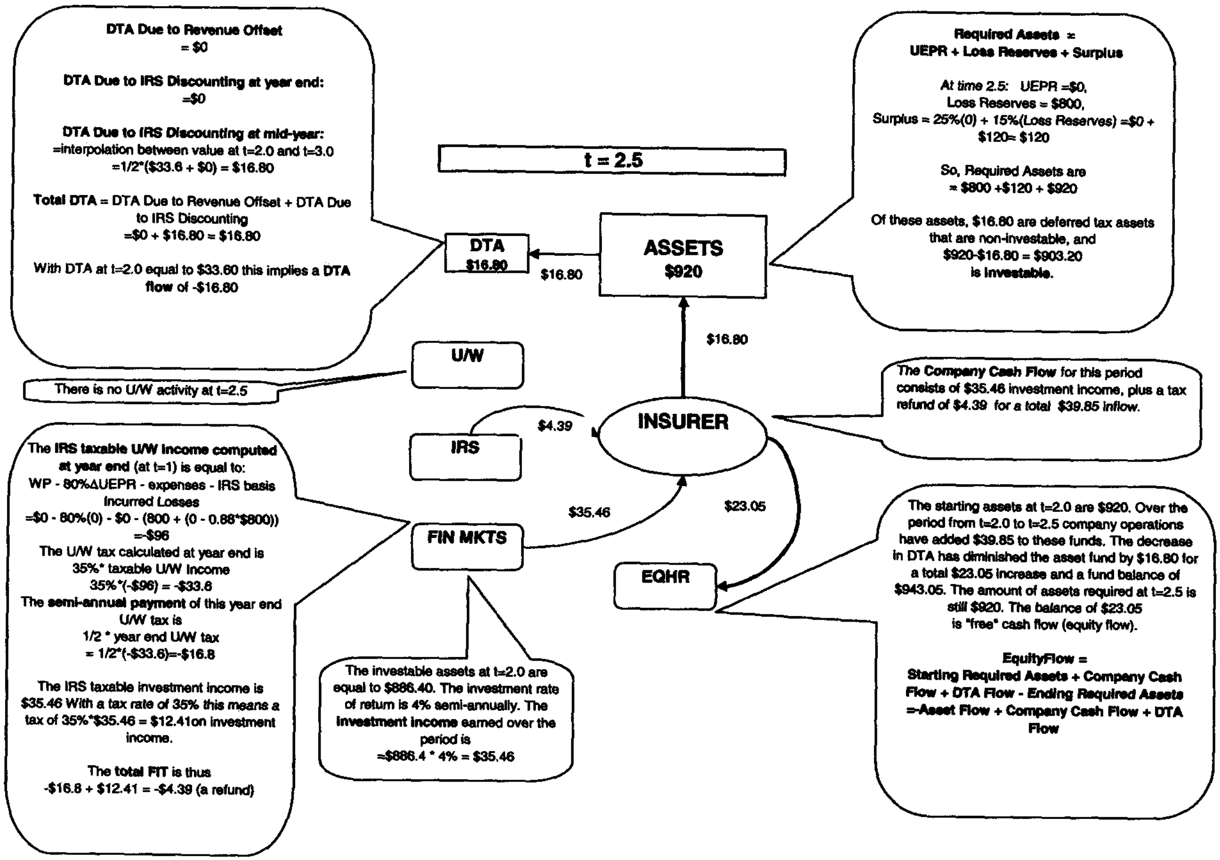
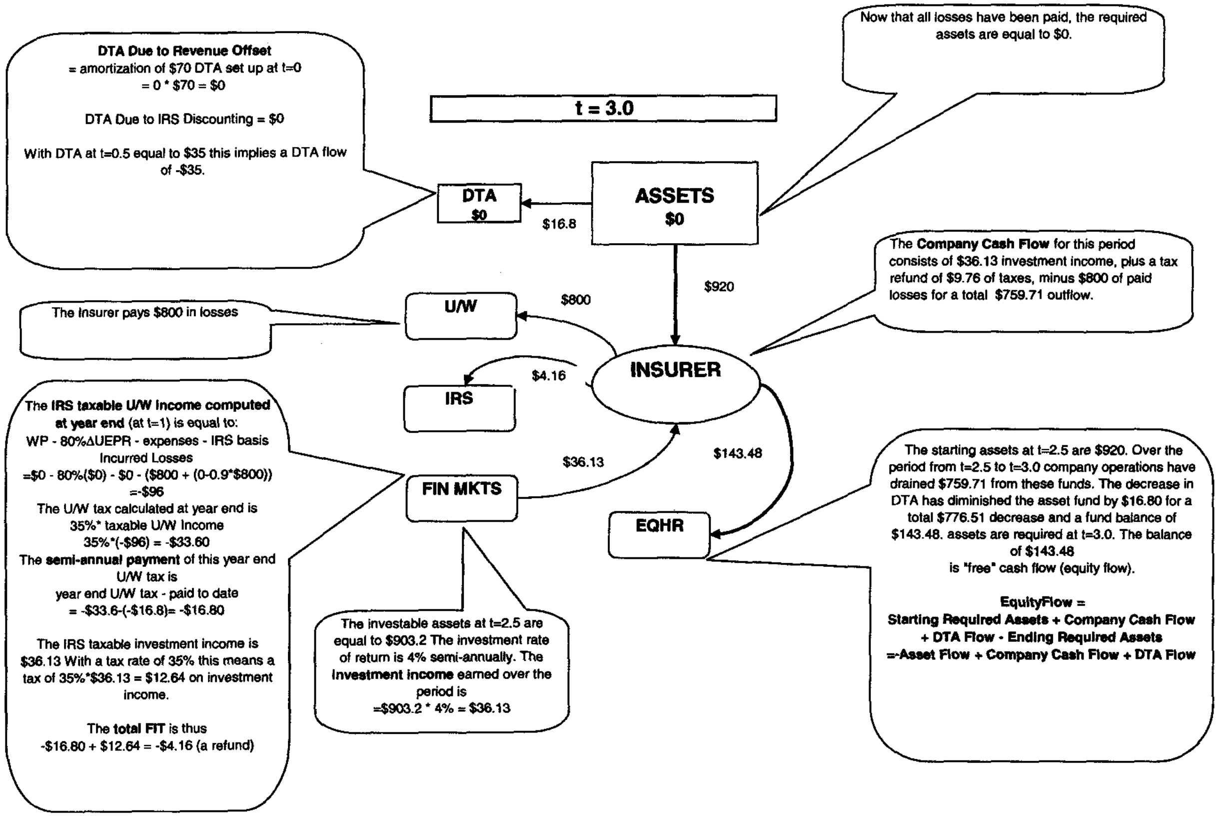


ILLUSTRATION B (Case Losses Incurred)

Exhibit 2D



Premiums and Losses

To complete the modeling of the implied equity flows, we overlay the premium transactions with the loss transactions. On June 30, 20XX+1, and December 31, 20XX+1, there are implied equity flows stemming from both premium and loss transactions.

JUNE 30, 20XX+1: The implied equity flow stemming from the earning of the premiums is +$231.91. This includes the effects of premium earnings, expense payments, federal income taxes, the takedown of one half of the deferred tax asset from revenue offset, and investment income. There is no change during the year in the capital requirements stemming from the written premium RBC charge.

The implied equity flow stemming from the occurrence of the first loss is –$336.80. This includes the effects of loss accrual, federal income taxes, the deferred tax asset stemming from IRS loss reserve discounting, and capital requirements for held loss reserves. The net implied equity flow on June 30, 20XX+1, is +$231.93 – $336.80 = $-104.87.

DECEMBER 31, 20XX+1: *The implied equity flow stemming from the earning of the premiums* is +$619.84. The implied equity flow stemming from the occurrence of the second loss is –$336.80.

This includes the same items as the implied equity flow on June 30, 20XX+1; it does not include the effect of investment income on the assets held to support the first loss, which was *incurred on June 30, 20XX+1. The investable assets supporting the first loss are* $400 of loss reserve plus 15% × $400 = $60 of supporting surplus minus $2.80 of deferred tax asset = $457.60. The after-tax investment income on these investable assets is $457.20 × 4% × (1–35%) = $11.89. The net implied equity flow on December 31, 20XX+1, is +$619.84 – $336.80 + $11.89 = $294.93 (premium flows, loss flows, and after-tax investment income on the assets supporting the loss reserves).

Internal Rate of Return and Net Present Value

The internal rate of return is the interest rate that sets the present value of the implied equity flows to zero. The table at the end of this paper shows the cash flows, statutory accounting entries, and implied equity flows for the illustration discussed above. The internal rate of return on the implied equity flows is the solution to the following equation:

$$0 = -412.50 - 104.87/(1+x)^1 + 294.93/(1+x)^2 + 40.57/(1+x)^3 + 40.21/(1+x)^4 + 23.05/(1+x)^5 + 143.48/(1+x)^6.$$

This solution is x = 1.485%, which is a 3.0% effective annual rate ($1.01485^2 = 1.030$).[30]

RATE FILINGS

Since the internal rate of return is less than the cost of capital, the policy generates a loss for the company, not a gain. We have not discussed the cost of capital, but it is at least equal to the investment yield of 8% per annum.

Some regulators presume that a positive internal rate of return implies a profit, even if the profit is not as great as the company desires. The National Council on Compensation Insurance (NCCI) was discomfited by this perception among some state rate regulators in the 1980's, when it filed advisory premium indications for its members.

Net present value models circumvent this misinterpretation. The NPV model shows a dollar gain or loss. An indicated IRR less than the cost of equity capital produces a dollar loss, and an indicated IRR greater than the cost of equity capital produces a dollar gain.

For performance measurement purposes, we use an EVA yardstick in addition to the IRR; see Feldblum and Thandi, "Income Recognition and Performance Measurement." Applying a net present value analysis to the implied equity flows (not to the company cash flows) is similar to an economic value added analysis. Both net present value and economic value added translate the implied return into a dollar amount, so that the gain or loss to the company is more readily understood.

Assumptions and Precision

These illustrations cover the major equity flows that affect financial pricing models for property-casualty insurance. Realistic pricing models have more entries, but they are not conceptually different.

Some readers might dismiss the analysis in the previous sections as needless precision. They presume that the modeling of tax cash flows and deferred tax assets and liabilities imposes excessive costs for little benefit.

This may have been true in past years, when computations were done with pencil and paper or with desk calculators. There is much arithmetic manipulation, but the principles are straightforward. The task of the pricing actuary is to construct the pricing model and to provide reasonable assumptions for the cash flows; the arithmetic is done by computer. Once the model is in place, the computation rules remain the same from year to year.

---

[30] By Descartes' rule, the maximum number of real solutions to this polynomial equation is equal to the number of sign reversals. Since there is only 1 sign reversal, the 1.485% solution is unique.

Time demands on casualty actuaries have hindered some companies from developing financially appropriate return on capital pricing models. There is a temptation to use rules of thumb or expedient short-cuts, such as

- traditional combined ratio targets with discounted loss ratios,
- after-tax investment yields instead of explicit modeling of federal income taxes,
- fair values of insurance costs instead of the cost of holding capital.

These short-cuts often lead to severe pricing errors in long-tailed lines of business.

Consumers' Perspective and Suppliers' Costs

We conclude this paper by comparing two perspectives on insurance pricing: the consumers' value perspective and the supplier's cost perspective.

Suppliers: The *equityholder's cost model* described in this paper determines the minimum price that investors demand to fund the insurance product. To optimize shareholders' return on invested capital, an insurer must focus on the amount of invested capital and the return on that capital.

Consumers: The *consumer's value model* focuses on the value received by the consumer from the insurance product. The rational consumer looks at the expected cash flows to and from the insurance company. Pricing an insurance product from the consumer's perspective focuses on expected loss payments for casualty products or expected benefit payments for life insurance products.

The expected loss costs gives the pure premium. By adding underwriting expenses and income taxes on the policy cash flows, and discounting at an appropriate interest rate, one determines the present value of benefits. This is the value of the product to the consumer. The consumer does not include the insurer's cost of capital in the value of the product.

In theory, the consumer's cost of capital should be included in the consumer's value perspective; see the following illustration. In practice, few consumers would set aside the requisite capital to fund the insurance risks.

Illustration: An employer might self-insure its workers' compensation exposures. The self-insured employer faces substantial process risk caused by random loss fluctuations. In theory, it should hold capital to guard against adverse loss fluctuations and to ensure payment of benefits to injured employees. In practice, the employer would pay benefits out of current cash flow, since it has no reserve requirements and no risk-based capital requirements.

The additional capital costs imposed by state regulation cause a discrepancy between the value to the consumer and the cost to the insurer. Even if the rational insurer and the rational consumer have the same expectations for the insurance cash flows, they calculate different prices for the insurance product. A common solution to this problem in many jurisdictions is an economically inefficient battle of wits among pricing actuaries, regulators, and consumer representatives at state rate hearings.

The magnitude of the discrepancy between the consumer's value perspective and the insurer's cost perspective is not always appreciated by regulatory authorities. This discrepancy was low at the beginning of the 20th century, when property-casualty products covered primarily the short-tailed property lines of business and capital requirements were low. The discrepancy has risen steadily through the 20th century, as casualty lines have increased, payment patterns have lengthened, and more stringent capital requirements have been imposed. The capital to assets ratio for property-casualty insurers is now many times higher than for life insurers, commercial banks, or other financial institutions. The costs of this high capital to asset ratio should be understood and properly weighed by state regulators.

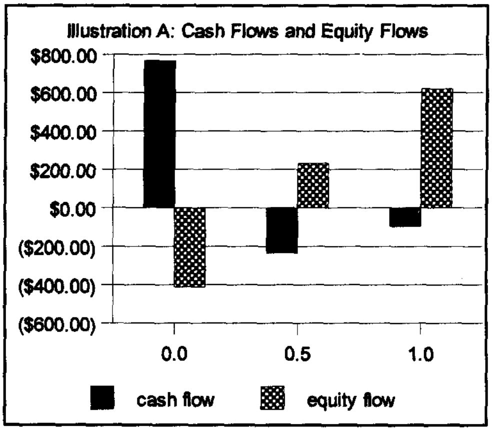
CASH FLOWS AND EQUITY FLOWS

The leitmotif of this paper is that the company cash flows for property-casualty insurance operations are not a suitable proxy for implied equity flows. The accompanying chart shows the cash flows and the associated equity flows for "Illustration A: Premium and Expenses."

Valuation Date	0.0	0.5	1.0
cash flow	$767.50	($233.07)	($95.16)
equity flow	($412.50)	$231.93	$619.94

The *premium collection at time* t=0 causes a large cash inflow to the company. The unearned premium reserves and risk-based capital requirements cause a large equity outflow at time t=0.

Expenses and tax payments cause cash outflows at times t=½ and t=1. As the premium is earned and the unearned *premium reserve* is taken down, there are equity inflows.



LOSS TRANSACTIONS

The loss transaction show a similar pattern: company cash inflows are associated with implied equity outflows, and company cash outflows are associated with implied equity inflows. The loss transactions begin at time t=½ and extend through time t=3.

Date	0.5	1.0	1.5	2.0	2.5	3.0
cash flow	$123.20	$135.09	$40.57	$40.21	$23.05	($776.52)
equity flow	($336.80)	($324.91)	$40.75	$40.21	$23.05	$143.48

Losses are incurred at times t=½ and t=1. The cash inflows at these dates stem from the federal income tax contra-liabilities, which are 35% of the present value of the incurred losses. The equity outflows stem from the case reserves and the risk-based capital requirements.

The cash flows and equity flows at times t = 1.5, 2.0, and 2.5, stem from investment income, tax payments, and changes in the deferred tax asset. The equity flows equal the cash flows.

At time t=3, the loss is paid, resulting in a large cash outflow. There is an equity inflow stemming from the takedown of the risk-based capital requirements.

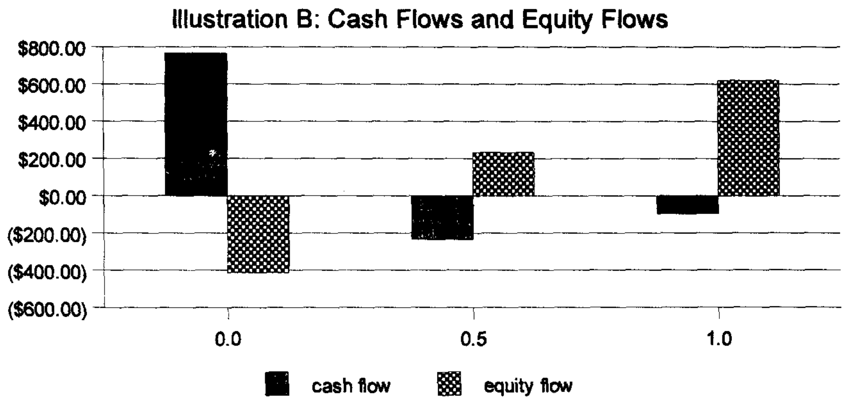
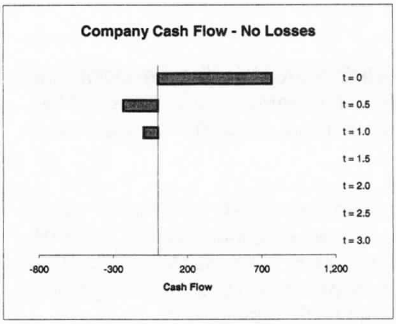
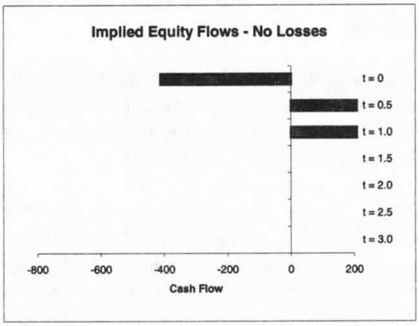


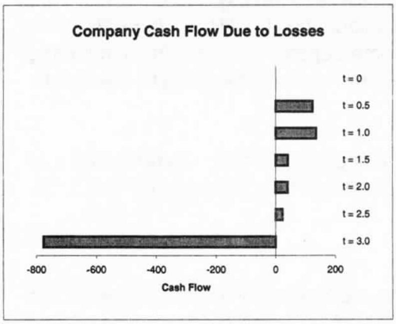
Exhibit 3

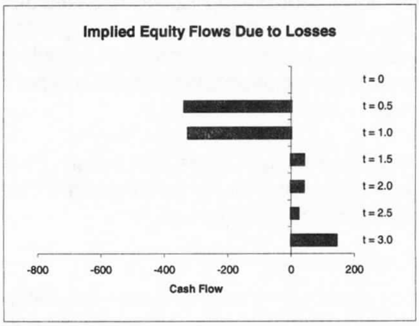
Company Cash Flows and Equity Flows - Case No Losses



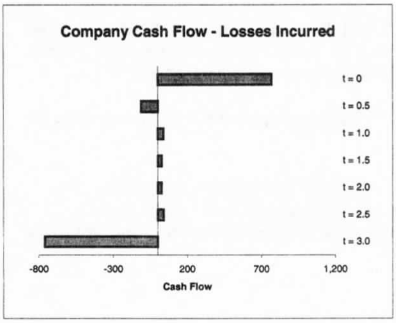


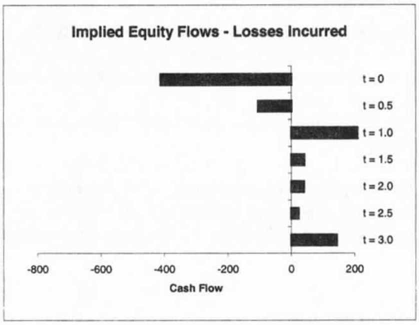
Company Cash Flows and Equity Flows Due to Losses





Company Cash Flows and Equity Flows - Case Losses Incurred





Appendix A: Federal Income Taxes and Deferred Tax Assets

INTRODUCTION

For prospective pricing, the actuary must estimate future IRS loss reserve discount factors and future deferred tax assets stemming from revenue offset and from loss reserve discounting. The post-codification changes to statutory accounting further complicate the consideration of tax effects in actuarial pricing models.

This appendix provide a tutorial through the relevant tax laws and regulations. It provides clear documentation of the IRS provisions concerning loss reserve discounting and the post-codification statutory accounting rules regarding deferred tax assets and liabilities.[31] To avoid repetition, this appendix covers the tax laws and regulations for the series of pricing model papers by S. Feldblum, N. Thandi, E. Schirmacher, and D. Schirmacher.[32]

DATA SOURCES

The equity flows in the pricing model are based on statutory accounting. For non-insurance companies, taxable income depends on general accounting [GAAP] statements. For property-casualty insurance companies, taxable income depends on statutory accounting statements. The exposition in the text of this paper determines taxable income from statutory income with several adjustments.

A. The tax computation begins with statutory pre-tax income from the Underwriting and Investment Exhibit of the Annual Statement: Part 1 for investment income and Parts 2, 2A, and 3 for underwriting income.[33]

---

[31] For a general treatment of federal income taxes relating to property-casualty insurance companies, with emphasis on items of particular concern to casualty actuaries, see Sarason, *et al.* [2003].

[32] For more extensive coverage of the tax aspects of insurance pricing, see Feldblum and Thandi, "Federal Income Taxes and the Cost of Holding Capital," Feldblum and D. Schirmacher, "Reinsurance Pricing and Capital Management," and Feldblum, "The Pricing of Commutations."

[33] See the Treasury regulations, 2001FED 26,153, §1.832-4(a)(1): "Gross income means the gross amount of income earned during the taxable year from interest, dividends, rents, and premium income, computed on the basis of the underwriting and investment exhibit of the annual statement."

The Internal Revenue Code lists numerous adjustments, of which the following are the most important for the pricing actuary:

1. The earlier incurral of the tax liability resulting from revenue offset and loss reserve discounting.
2. The effects of anticipated salvage and subrogation and the discounting provisions relating thereto.
3. The reduction of the tax liability resulting from municipal bond income and the dividends received deduction.

B. The additional tax liability resulting from the revenue offset provision is calculated from Part 2 of the Underwriting and Investment Exhibit.[34]
C. Schedule P, Part 1, is used to calculate the additional tax liability resulting from the IRS loss reserve discounting provision. Schedule P, Part 3, may be used to determine the non-admitted portion of the deferred tax asset stemming from the loss reserve discounting.

This appendix focuses on IRS loss reserve discounting and the admitted portion of the resulting statutory deferred tax asset.

Loss Reserve Discounting

For statutory financial statements, calendar year incurred losses equal the losses paid during the year plus the change in the full value loss reserves from the beginning of the year to the end of the year. For federal income tax purposes, the incurred losses during the tax year equal the losses paid during the year plus the change in the *discounted* loss reserves from the beginning of the year to the end of the year.

The determination of discounted loss reserves relies on Schedule P. For a prospective pricing model, the actuary must estimate (i) the discounted loss reserves, (ii) the amount of the discount, and (iii) the deferred tax assets stemming from loss reserve discounting and revenue offset. For retrospective analysis of product profitability, the actuary must consider also the effects of reserve changes on taxable income.[35]

The cost of capital is a major factor for the pricing of insurance contracts, and the double taxation of the investment income on capital funds is a significant component of this cost. The IRS loss reserve discounting provisions and the statutory deferred tax asset affect the cost of holding capital for insurers.[36]

The alternative minimum income tax provisions may also cause earlier incurral of the tax liability. Changes in the incurral dates of the tax liabilities lead to deferred tax assets and liabilities on the statutory balance sheet.

[34] The recognition of taxable revenue from earned but unbilled premiums and accrued retrospective premiums are affected by the January 2000 tax regulations and the statutory accounting codification changes effective on January 1, 2001. The illustration in the text of the paper does not consider earned but unbilled premiums or accrued retrospective premiums. The workers' compensation illustration in Appendix B considers the billing and collection pattern of the policy premiums.

[35] See E. Schirmacher and S. Feldblum, "Retrospective Analysis and Performance Measurement."

[36] See Feldblum and Thandi, [2002], "Federal Income Taxes and the Cost of Holding Capital."

INVESTMENT INCOME AND AMORTIZATION

For long-tailed lines of business, the statutory accounting rules cause an underwriting loss during the policy term when losses occur. After policy expiration, the investment income on the assets backing the loss reserves provide steady and positive net income. For tax accounting, the expected investment income on the assets backing the loss reserves offsets the expected amortization of the interest discount in the reserves. The underwriting gain or loss is realized during the policy term, with no expected net gain or loss in subsequent years.[37]

Complete (exact) offsetting depends on the following conditions:

- There are no implicit (undisclosed) discounts in the statutory loss reserves.
- The IRS discount rate equals the investment yield of the company.
- The IRS loss payment pattern equals the actual liquidation pattern for the block of business.
- The company holds fully discounted reserves, with disclosure of the amount of discount.

These conditions are not consistent with current statutory requirements, so complete offsetting is not expected. Nonetheless, they clarify the heuristic illustration below.[38]

Illustration: Offsetting

A one day policy is written on December 31, 20XX, for a net premium of $10,000. One loss occurs on December 31, 20XX, which is paid for $12,100 on December 31, 20XX+2. The term structure of interest rates is flat at 10% per annum. To simplify the illustration, we assume that the IRS loss payment pattern is the same as the actual loss payment pattern here.

In 20XX, statutory accounting shows an underwriting loss of $10,000 – $12,100 = $2,100. The $10,000 net premium is invested at 10% per annum. The investment income is $10,000 × 10% = $1,000 in 20XX+1 and $11,000 × 10% = $1,100 in 20XX+2. There is no underwriting gain or loss in 20XX+1 or 20XX+2, so these are the statutory income amounts.

If we assume a two year IRS loss payment pattern and a discount rate of 10% per annum, the discounted loss reserves are $12,100 / 1.100^2 = $10,000 at December 31, 20XX. Tax accounting shows no underwriting gain or loss in 20XX and a tax liability of $0 for 20XX.

In 20XX+1, investment income is $10,000 × 10% = $1,000. The discounted loss reserve on December 31, 20XX+1, is $12,100 / 1.100 = $11,000. The underwriting loss (or the offset

---

[37] Statutory, GAAP, and tax accounting are discussed in detail in Feldblum and Thandi [2002], "Income Recognition and Performance Measurement."

[38] For complete discussion of this subject, see Feldblum and Thandi, [2002], "Reserve Valuation Rates."

to underwriting income) for tax year 20XX+1 equals amortization of the interest discount on the loss reserves, or $11,000 – $10,000 = $1,000. The underwriting loss just offsets the investment income. The net taxable income is $0, and the tax liability is $0.

In 20XX+2, investment income is $11,000 × 10% = $1,100. The incurred loss offset to taxable underwriting income in 20XX+2 is the paid loss plus the change in the discounted loss reserve, or $12,100 (paid on December 31, 20XX+2) + $0 – $11,000 = –$1,100.

This is the amortization of the interest discount on the 12/31/20XX+1 reserve of $11,000. It offsets the investment income in 20XX+2. Taxable income is $0, and the tax liability is $0.[39]

DISCOUNTING PRINCIPLES

The discounted loss reserves are determined from three components:

- The undiscounted loss reserves, as shown in Schedule P, Part 1.
- The loss reserve discount rate, which is promulgated each year by the Treasury.
- The loss payment pattern by line of business, which is determined from Schedule P data.

The illustration below shows the concepts, though the details differ from the IRS computation.

Illustration: The December 31, 20XX, undiscounted loss reserves are $100 million. The loss reserve discount rate is 8% per annum. The $100 million of reserves will be paid in three parts: 50% on December 31, 20XX+1, 30% on December 31, 20XX+2, and 20% on December 31, 20XX+3.[40] The discounted loss reserves equal

$$\$100 \text{ million} \times (50\%/1.08 + 30\%/1.08^2 + 20\%/1.08^3) = \$100 \text{ million} \times 0.879 = \$87.9 \text{ million.}$$

Undiscounted Loss Reserves

The Treasury assumes that the loss reserves in Schedule P, Part 1, are undiscounted values. If discounted values are shown, the losses may be "grossed up" to undiscounted amounts before application of the IRS loss reserve discounting procedure. The "gross-up" is permitted only if the amount of the discount is disclosed in (or with) the Annual Statement.[41]

---

[39] Some insurance personnel speak of the post-1986 federal income tax incurral pattern as a "prepayment of taxes by the insurance industry." This is correct from a statutory or GAAP perspective. The IRS would take the opposite view; before 1986 the Treasury helped fund the conservative insurance accounting practices.

[40] This illustration is simplified. The actual tax procedure assumes mid-year payments and a longer loss payment pattern.

[41] See section 846(b)(2) of the Internal Revenue Code: "Adjustment If Losses Discounted on Annual Statement: If the amount of unpaid losses shown in the annual statement is determined on a discounted basis,

Illustration: Schedule P, Part 1, is gross of non-tabular discount and net of tabular discount.

- A company incurs $10,000,000 of accident year 20XX workers' compensation losses, including lifetime pension claim reserves with a tabular discount of $1,000,000.
- The IRS loss reserve discount factor for workers' compensation accident year 20XX reserves is 85%.

If the company does not disclose the tabular discount in the Annual Statement, the offset to taxable income is $10 million × 85% = $8.5 million. If the company does disclose the tabular discount in the Annual Statement, the offset to taxable income is ($10 million + $1 million) × 85% = $9.35 million. The difference in taxable income is $9.35 million – $8.5 million = $0.85 million, and the difference in the tax liability is $0.85 million × 35% = $297,500.

DISCLOSURE AND TIMING COSTS

Reserve discounting is a timing difference; it reverses in subsequent years. The cost to the company is the present value of the expected after-tax investment yield on this money.

Illustration: Suppose the pension reserves are paid (on average) twelve years after policy expiration, and the after-tax investment yield is 6% per annum. The cost to the company is

$$\$297,500 \times [\,(1.06^{12} - 1)\,/1.06^{12}] = \$297,500 \times 0.503 = \$149,651.61.$$ [42]

RESERVE VALUATION RATES

Several of the pricing papers focus on the inter-relationships among the reserve valuation rate, the cost of holding capital, the federal income tax liability, and the indicated premium. We expand the comments above for the application to these papers.

The pricing model separates the undiscounted loss reserves into two pieces:

- the true undiscounted loss reserves, and
- the valuation rate at which the company's reserves are booked.

---

and *the extent to which the losses were discounted can be determined on the basis of information disclosed on or with the annual statement*, the amount of the unpaid losses shall be determined without regard to any reduction attributable to such discounting." The required disclosure of *non-tabular* discounts by accident year and by line of business is provided in columns 34 (losses) and 35 (loss adjustment expenses) of Schedule P, Part 1. The required disclosure of *tabular* discounts is shown in note 28 (in the 2001 Annual Statement) to the financial statements, "Discounting of Liabilities for Unpaid Losses or Unpaid Loss Adjustment Expenses."

[42] Because of the statutory deferred tax asset and the capital requirements imposed on insurance companies, the actual cost to equityholders is somewhat different; see Feldblum and Thandi [2002], "Federal Income Taxes and the Cost of Holding Capital," for a full discussion.

If the company holds full value loss reserves, the valuation rate is 0%. A pricing model used by a regulatory agency would use full value loss reserves. Similarly, a pricing model used for a rate filing would use full value loss reserves. A pricing model used for competitive pricing purposes should use the valuation rate implicit in the reserves held by the company.

Illustration: The December 31, 20XX, undiscounted loss reserves are $100 million. The loss reserve discount rate is 8% per annum. The $100 million of reserves will be paid in three parts: 50% on December 31, 20XX+1, 30% on December 31, 20XX+2, and 20% on December 31, 20XX+3. The company values its held reserves at an implicit 5% discount rate.

The reserves on the company's balance sheet are

$$\$100 \text{ million} \times (50\%/1.05 + 30\%/1.05^2 + 20\%/1.05^3) = \$100 \text{ million} \times 0.921 = \$92.1 \text{ million.}$$

If the company's implicit loss reserve discount is not disclosed in its Annual Statement, the IRS treats the held reserves as undiscounted loss reserves. The discounted loss reserves for tax purposes equal

$$\$92.1 \text{ million} \times (50\%/1.08 + 30\%/1.08^2 + 20\%/1.08^3) = \$92.1 \text{ million} \times 0.879 = \$80.96 \text{ million.}$$

The implicit discounting lowers the lowers the tax basis loss reserves. This is a timing effect, not a permanent effect, since the losses actually paid do not depend on the valuation rate used by the company. The company loses the investment income on the early incidence of the federal income tax liability. This loss may be offset by the capital management benefits of having less equityholder supplied funds tied up in full value loss reserves. The capital management benefit depends on the cost of holding capital and on possible rating agency revaluations of the indicated reserves. For a complete treatment of these items, see Feldblum and Thandi, "Reserve Valuation Rates" and Feldblum and Schirmacher, "Reinsurance Pricing and Capital Management."

LIMITATION

The IRS is concerned that a company might claim such a large discount for its statutory loss reserves that the discounted tax-basis loss reserves would be greater than the Annual Statement loss reserves, thereby reducing the tax liability by means of discounting instead of increasing the tax liability. To prevent this, the discounted IRS loss reserves may not be greater than the loss reserves shown in the Annual Statement.[43]

Statutory accounting allows only limited discounting: tabular discounts and exceptional cases of non-tabular discounts. For tabular discounts, most companies use conservative interest rates, such as 3.5% or 4% per annum. For non-tabular discounts, the permissible discount rate for statutory accounting is rarely greater than the discount rate used for IRS loss reserve discounting; see SSAP No. 65 on "Property and Casualty Contracts," paragraph 12.

In most cases, the statutory loss reserves are lower than the IRS discounted loss reserves. The workers' compensation "prior years" row (Part 1D) is an exception. These reserves are primarily indemnity reserves for lifetime pension cases, and many companies use tabular discounts. For this row, the "composite discount factor" used in the IRS discounting calculations assumes (on average) three more years of payment, whereas the pension cases in these reserves may have (on average) a future expected lifetime of 10 to 20 years.

ILLUSTRATION: THE LIMITATION

The workers' compensation prior years row shows unpaid losses and loss adjustment expenses of $30 million. In the Notes to the Financial Statements, the company reports a $10 million tabular discount for these claims. The IRS composite discount factor applicable to these reserves is 90%.

Without the limitation discussed above, the gross loss reserves are $30 million + $10 million = $40 million. The IRS discounted loss reserves are 90% x $40 million = $36 million. Since this exceeds the $30 million of statutory loss reserves, the IRS discounted loss reserves are capped at $30 million.

This rule has a significant effect on the pricing of workers' compensation commutations for permanent total disability cases; see Feldblum [2002], "The Pricing of Commutations," for a full discussion. The rule also affects the pricing of workers' compensation large dollar

[43] See the Internal Revenue Code §846(a)(3): "In no event shall the amount of the discounted unpaid losses with respect to any line of business attributable to any accident year exceed the aggregate amount of unpaid losses with respect to such line of business for such accident year included on the annual statement."

deductible business and workers' compensation excess coverage, since these two types of contract cover primarily long term disability cases.

Discount Rate

The discount rate varies by accident year. For each accident year, the discount rate is the 60 *month moving average of the federal mid-term rates ending on the December 1 preceding the* accident year. This rate is frozen and applies to that accident year's losses in all future calendar years. In tax parlance, the discount rate is "vintaged." The federal mid-term rate is the average rate on Treasury securities with 3 to 9 years remaining maturity.[44]

The federal mid-term rate is promulgated by the Treasury each month.[45] The 60 month moving average for an accident year can be determined once the last federal mid-term rate has been announced.

Illustration: The loss reserve discounting rate for accident year 20X9 is the 60 month average of the federal mid-term rates from January 1, 20X4, through December 1, 20X8. It can be computed in December 20X8, before the inception of accident year 20X9, so that companies can effectively determine their tax strategies during 20X9.

Yield Projections

The market values of future cash flows are based on the current term structure of interest rates. The date that the liability was incurred is not relevant. In contrast, the IRS bases the discount rate on the incurral year of the liability. The rationale is that the insurance company uses the premium cash flows from the policy to purchase fixed-income securities to fund the future loss payments. The yield on the fixed-income securities is determined at the date of purchase.

---

[44] See section 846(c)(2) of the Internal Revenue Code: "Determination of Annual Rate: The annual rate determined by the Secretary under this paragraph for any calendar year shall be a rate equal to the average of the applicable Federal mid-term rates (as defined in section 1274(d) but based on annual compounding) effective as of the beginning of each of the calendar months in the test period. The test period is the most recent 60-calendar-month period ending before the beginning of the calendar year for which the determination is made."

The federal mid-term rates are expressed as bond equivalent yields, since bond coupons are paid semi-annually in the United States. (A bond equivalent yield is a yield with semi-annual compounding.) The IRS loss reserve discounting procedure uses annual compounding, since it assumes that losses are paid in mid-year (i.e., once a year). The bond equivalent yields are converted to effective annual yields before averaging, using the formula $r_a = (1 + r_s/2)^2 - 1$, where r_a is the effective annual yield and r_s is the bond equivalent yield with semi-annual compounding. If the bond equivalent yield is 8% per annum, the equivalent effective annual rate is $(1 + 0.08/2)^2 - 1 = 8.16\%$.

[45] The yield among mid-term securities varies with the remaining maturity, in accordance with the term structure of interest rates. More recently issued securities tend to have slightly lower yields, since they are more marketable. The Secretary of the Treasury selects an appropriate average rate.

If the duration of the assets backing the reserves matches the duration of the loss liabilities, the losses will be paid from the coupon income and the principal repayment from these securities. The yield during the accident year is the relevant investment yield throughout the life of the policies.[46]

For prospective pricing, the actuary must project about two years of future yields. The pricing requirements are most easily seen by illustration.

Illustration – Projected Yields: The pricing actuary is setting rates for policies effective from July 1, 20XX, through June 30, 20XX+1. The losses on these policies extend from July 1, 20XX, through June 30, 20XX+2, since the last policy written under the new rates is effective on June 30, 20XX+1, and remains in effect through June 30, 20XX+2. The losses stemming from policies written under the new rates relate to accident years 20XX, 20XX+1, and 20XX+2.

To project loss reserve discount factors, the pricing actuary must estimate federal mid-term rates through December 1, 20XX+1. If the rate analysis is done in the last quarter of 20XX–1, the actuary must project rates from the last quarter of 20XX–1 through the end of 20XX+1.

The rate projection is generally done in one of two ways:

- The most recent monthly mid-term rate may be repeated for all future months. Alternatively, the average of the most recent three or six monthly mid-term rate may be repeated for all future months.
- The projected rates may be set equal to the current forward rates for the corresponding time period.

The second method is favored by many investment analysts. We explain by means of an illustration.

Illustration: Projecting Treasury Rates: The term structure of interest rates on January 1, 20XX, is upward sloping, as shown in the table below.

---

[46] Whether a moving average rate or the current rate is a better predictor for future rates is an open question. Accountants often prefer average rates, on the assumption that the most recent monthly figure may be abnormally high or low. Some financial analysts presume that interest rates revert towards a long-term mean, and a 60 month moving average may be a better reflection of this mean. Other analysts presume that interest rates form a random walk, and the present term structure of interest rates is the best reflection of expected future rates. The dominant view is that the current rate is a better estimator of the rate during the next 12 months than the 60 month moving average is; see Dr Jonathan Benjamini and S. Feldblum, *Dynamic Financial Analysis: a Primer for the Practicing Actuary* [2002].

Term	Spot Rate	Term	Spot Rate
1 year	5.00%	*5 years*	6.90%
2 years	6.00%	*6 years*	7.00%
3 years	6.40%	*7 years*	7.10%
4 years	6.70%	*8 years*	7.10%

Table AppA.1: *Term Structure of Interest Rates*

To project the five year spot rates for 20XX, 20XX+1, 20XX+2, and 20XX+3, we use one of two methods. The first method assumes that the five year spot rate remains unchanged at 6.9%. The second method determines the forward rates for the appropriate periods.

- The five year spot rate on Jan 1, 20XX+1, is estimated as $(1.070^6 / 1.05)^{1/5} - 1 = 7.40\%$.
- The five year spot rate on Jan 1, 20XX+2, is estimated as $(1.071^7 / 1.06^2)^{1/5} - 1 = 7.54\%$.
- The five year spot rate on Jan 1, 20XX+3, is estimated as $(1.071^8 / 1.064^3)^{1/5} - 1 = 7.52\%$.

This method of using forward rates to project future spot rates relies on the pure expectations hypothesis for the term structure of interest rates. Most financial analysts do not subscribe to a pure expectations hypothesis.

Loss Payment Pattern

The IRS determines the expected loss payment pattern by line of business from Schedule P, Part 1. To compute the tax liability, the accountant may use the loss reserve discount factors promulgated by the Treasury. For a financial pricing model, the actuary must estimate the future loss reserve discount factors a year or two in advance. The actuary should understand the relationship of the IRS loss reserve discount factors to actuarially determined loss reserve discount factors to estimate the tax effects on policy pricing; see the discussion below.

Illustration: An actuary is pricing a claim commutation. For a permanent total disability case, the actuarially determined loss reserve discount factors over the next twenty years rise from 70% to 100%, and the IRS loss reserve discount factors are level at about 92%. The IRS discounted reserves are greater than the fair value reserves, leading to a tax credit and a lower commutation price than if actuarial loss reserve discount factors were used; see Feldblum, "The Pricing of Commutations."

Illustration: After a period of falling interest rates, the IRS loss reserve discount factors provide a larger discount than is financially warranted. After a period of rising interest rates, the IRS loss reserve discount factors provide a smaller discount is financially warranted.

Data Grouping

The IRS loss reserve discount factors are determined by Schedule P line of business. The actual loss payment pattern for the business being priced, which is used in the pricing model to determine the implied equity flows, is not relevant for the IRS loss reserve discount factors. The illustrations in this appendix all use Schedule P loss triangles.

Illustration: The average lag between premium collection and loss payment for large dollar deductible workers' compensation business with a $500,000 deductible may be twenty years. The average lag for first dollar workers' compensation business in a state with limited duration permanent total disability cases and no cost of living adjustments may be four years. The IRS loss reserve discount factors are the same for the two sets of business, since both use the Schedule P, Part 1D (workers' compensation) factors.

Deriving Payment Patterns

We determine two sets of loss payment patterns: one for IRS loss reserve discount factors and one for statutory accounting deferred tax assets. Both loss payment patterns are derived from historical loss liquidation patterns. If there were no random fluctuations in the loss payment pattern for any accident year and no systematic changes in the payment pattern over the past ten years, the observed liquidation pattern of the oldest accident year recorded in Schedule P would be sufficient, as illustrated below.

Illustration: We are computing the loss payment pattern for the 20X9 accident year reserves. Suppose that Schedule P, Part 3, shows the following pattern for accident year 20X0:

Exhibit AppA.2: 20X9 Schedule P, Part 3 ($000,000)

Part 3	20X0	20X1	20X2	20X3	20X4	20X5	20X6	20X7	20X8	20X9
20X0	103	226	294	334	363	384	398	412	422	433

In addition, suppose the ultimate incurred losses for accident year 20X0 are $486 million. This estimate may be taken from Schedule P, Part 2, or it may be derived from an actuarial loss reserve projection; see Feldblum [2002: SchP] for estimation procedures.

Schedule P, Part 3, shows cumulative paid losses. The first differences between each adjoining set of figures is the incurred loss paid losses in each 12 month period. The ratio of these incurred loss paid loss figures to the estimated ultimate incurred losses is the percentage of ultimate losses paid in each 12 month period, as shown below.

Exhibit AppA.3: Incremental Loss Payment Pattern from Accident Year 20X0 ($000,000)

Part 3	20X0	20X1	20X2	20X3	20X4	20X5	20X6	20X7	20X8	20X9
1. 20X0	$103	$226	$294	$334	$363	$384	$398	$412	$422	$433
2. percent	0.212	0.465	0.605	0.687	0.747	0.790	0.819	0.848	0.868	0.891
3. incremental	0.212	0.253	0.140	0.082	0.060	0.043	0.029	0.029	0.021	0.023

- Row 1: The row labeled "20X0" shows the cumulative dollars (in millions) of accident year 20X0 losses paid by December 31 of each calendar year from 20X0 through 20X9.
- Row 2: The row labeled "percent" shows the cumulative percentages of accident year 20X0 ultimate losses paid by December 31 of the calendar year in each column.
- Row 3: The row labeled "incremental" shows the incremental percentages of accident year 20X0 ultimate losses paid in each calendar year.

The final row in the table above tells us that 21.2% of an accident year's incurred losses are paid during the accident year, another 25.3% are paid in the 12 months following the accident year, 14.0% are paid in the subsequent 12 months, and so forth. The remaining 10.9% [= 100% – 89.1%] are paid more than 10 years after the inception of the accident year.

The illustration above shows the logic underlying the estimated loss payment pattern. Because of the volatility of loss payments and possible systematic changes over the years, as noted below, we make several adjustments.

- Settlement of large losses may distort the payment pattern in any one accident year.
- The loss payment pattern does not reflect any changes in the intervening nine years.
- This method ignores the information embedded in the observed liquidation of accident years 20X1 through 20X8.

Recent Data

To use the most recent data, we examine the dollars paid in calendar year 20X9 divided by the total incurred losses for each accident year. Assume that the paid loss development illustration shows the following associated figures from Schedule P, Parts 2 and 3.[47]

Exhibit AppA.4: Loss Payment Pattern from Successive Accident Years ($000,000)
(Data from Schedule P, Parts 2 and 3, from the 20X9 Annual Statement)

Accident Year (1)	*Cum Paid by 20X8 (2)*	*Cum Paid by 20X9 (3)*	*Paid in 20X9 (4)*	*Ultimate Losses (5)*	*Percentage Paid (6)*
20X0	$422	$433	$11	$486	2.26%
20X1	$442	$454	$12	$520	2.31%
20X2	$391	$403	$12	$475	2.53%
20X3	$416	$434	$18	$522	3.45%
20X4	$504	$534	$30	$667	4.50%
20X5	$490	$542	$52	$707	7.36%
20X6	$463	$546	$83	$787	10.55%
20X7	$353	$485	$132	$802	16.46%
20X8	$152	$406	$254	$866	29.33%
20X9		$156	$156	$898	17.37%

The columns show the following figures:

Column (2): Cumulative dollars of loss paid through December 31, 20X8 (from Part 3),
Column (3): Cumulative dollars of loss paid through December 31, 20X9 (from Part 3).
Column (4): Incremental dollars of loss paid in 20X9 (= column (2) minus column (1)).
Column (5): Incurred losses (from Part 2).
Column (6): Incremental dollars of loss paid as a percentage of incurred losses (row 3 / row 4).

Consider the row for accident year 20X4:

Column 2: $504,000 has been paid by 12/31/20X8, or 60 months since inception of the accident year.

[47] The accident years are shown along the horizontal axis of the table. In the exhibits used for the paid loss chain ladder development method, the accident years are shown along the vertical axis.

Column 3: $534,000 has been paid by 12/31/20X9, or 72 months since inception of the accident year.
Column 4: $30,000 has been paid between 60 months and 72 months.
Column 5: The total accident year 20X4 incurred losses are $667,000.
Column 6: 4.5% (or $30,000 / $667,000) of the incurred losses are paid between 60 months and 72 months since inception of the accident year.

The loss payment pattern in the table above is theoretically sound, though both the IRS and common actuarial practice use slightly different methods.

- This procedure uses figures from Schedule P, Part 3, which shows cumulative paid losses at the current valuation date and the previous valuation date. The IRS used figures from Part 1, perhaps because Part 1 is an audited exhibit whereas Part 3 is not.
- This procedure uses incurred loss losses from a single accident year for each payment period. Common actuarial practice is to use averages from three or more years.

Incremental Percentages and Cumulative Differences

For the lines of business with ten year exhibits, the IRS makes one additional change. The procedure outlined above uses the incremental paid loss percentages in each accident year to estimate the percentage of losses paid in each time interval. The IRS uses the difference in the cumulative paid loss percentages between successive accident years.

Exhibit AppA.5: Loss Payment Pattern Between Accident Years ($000,000)
(Data from Schedule P, Parts 2 and 3, from the 20X9 Annual Statement)

AccYr	20X0	20X1	20X2	20X3	20X4	20X5	20X6	20X7	20X8	20X9
Row 1	$433	$454	$403	$434	$534	$542	$546	$485	$406	$156
Row 2	$486	$520	$475	$522	$667	$707	$787	$802	$866	$898
Row 3	89.1%	87.3%	84.8%	83.1%	80.1%	76.7%	69.4%	60.5%	46.9%	17.4%
Row 4	1.8%	2.5%	1.7%	3.1%	3.4%	7.3%	8.9%	13.6%	29.5%	17.4%

- Row (1) shows the cumulative paid losses at December 31, 20X9, for each accident year.
- Row (2) shows the incurred losses at December 31, 20X9, for each accident year.
- Row (3) shows the ratio of cumulative paid losses to incurred losses.
- Row (4) shows the differences in successive ratios. For accident year 20X9, nothing is paid before calendar year 20X9, so 17.4% of incurred losses are paid in the first 12 months. For losses paid between 12 months and 24 months, we reason as follows.

- ✓ From the 20X8 accident year, we infer that 46.9% of incurred losses are paid by 24 months since inception of the accident year.
- ✓ From the 20X9 accident year, we infer that 17.4% of incurred losses are paid by 12 months since inception of the accident year.
- ✓ This implies that 46.9% – 17.4% = 29.5% of incurred losses are paid between 12 months and 24 months since inception of the accident year.

The figures in row (4) sum to 89.1%. This is the ratio of cumulative paid losses to incurred losses for accident year 20X0.

The illustration above uses figures from Schedule P, Parts 2 and 3. The IRS actually uses figures from Schedule P, Part 1, which include more loss adjustment expenses.

- The Part 1 figures used by the IRS include all loss adjustment expenses.[48]
- The Part 3 figures include only defense and cost containment expenses.

IRS RATIONALE

We summarize the computations as follows:

1. For each accident year in Schedule P, Part 1, we calculate the cumulative paid losses at the current valuation date as a percentage of the incurred losses for that accident year.
2. We take the difference between successive accident years to determine the expected percentage of incurred losses paid in each 12 month interval.
3. We use this procedure for the ten accident years shown in Part 1. If the cumulative paid losses for the oldest year equal 100% of the incurred losses, we stop here. If the cumulative paid losses for the oldest year are less than 100% of the incurred losses, we extend the loss payment pattern for additional years, as described below.

The cumulative paid losses as of the current valuation date are shown in Part 1, column 11, "total net paid." The incurred losses at the current valuation date are shown in column 28, "total losses and loss expense incurred."

---

[48] See section 846(f)(2) of the Internal Revenue Code: *The term "unpaid losses" includes any unpaid loss adjustment expenses shown on the annual statement.*

Illustration A: No Extension of Payments

Although the concepts are straight-forward, the implementation is complex. We explain the details in this appendix with three illustrations. To *understand* the text of this paper, the reader need not know all the material in this appendix. To *implement* the pricing model described in this paper, the reader must be familiar with the tax rules and regulations.

We proceed incrementally in this appendix. For most casualty lines of business, the IRS loss payment pattern extend up to a maximum of 16 years. We show first the procedure for a line of business with no extension, so the loss payment pattern ceases in the eleventh year. The next illustration shows the extension through the sixteenth year.

The pricing actuary will normally estimate loss reserve discount factors for two or three future accident years. The loss payment pattern can be derived only for the first of these accident years if the company uses its own pattern. If the company uses the industry pattern, the estimation procedure depends on the particular accident years in relation to the redetermination year. For certain items, there is no good way to estimate the required figures. We explain the choices available to the pricing actuary after the illustrations.

The ABC Insurance Company elected to use its own loss payment pattern in the 2007 determination year. This election applies to accident years 2009 through 2013.

It is now July 1, 2010, and the pricing actuary is estimating premium rates for policy year 2011. Losses from policies written in policy year 2011 fall into accident years 2011 and 2012. We estimate IRS loss reserve discount factors for accident year 2011. After completing this illustration, we explain the possible methods of estimating loss reserve discount factors for accident year 2012.

The following figures are taken from ABC's 2009 Annual Statement, Schedule P, Part 1. The procedure shown here is applicable to any ten year line of business. These include all the casualty lines for which financial pricing models are commonly used.

Exhibit AppA.6: Casualty Line of Business Paid and Incurred Losses

Accident Year	Losses + LAE Paid	Losses + LAE Incurred
Prior	250,000	250,000
2000	270,000	275,500
2001	300,000	316,000
2002	320,000	348,000
2003	340,000	386,500
2004	350,000	421,500
2005	370,000	480,500
2006	380,000	550,500
2007	360,000	610,000
2008	330,000	687,500
2009	200,000	571,500

Discount Rate

The discount rate used for the loss reserve discounting procedure is the 60 month rolling average of the federal mid-term rate, from January 2006 through December 2010. Since it is now July 1, 2010, only 54 months are available; the last six months must be estimated.

The actuary has two alternatives for estimating future interest rates, as discussed above.

- Repeat the most recent federal mid-term rate for the remaining months, or use an average of recent federal mid-term rates for the remaining months.
- Determine the federal mid-term rates implied by the current term structure of interest rates.

For this illustration, we assume that the estimated 60 month moving average of the federal mid-term rates is 7% per annum.

Determination Year and Company Election

If the company uses its own data to determine the loss payment pattern, the data are updated each year. If the company uses industry data – that is, if the company uses Treasury factors determined from industry-wide data reported in Best's Aggregates and Averages, the data are updated in determination years (or re-determination years).

Determination years end in a "2" or a "7," and they use aggregate industry data for statement dates ending in a "0" or a "5."

- For determination year 20X2, data as of December 31, 20X0 are used.

- For determination year 20X7, data as of December 31, 20X5 are used.

Once every five years (determination years), the company makes an election to use either the loss reserve discount factors developed by the Treasury, which are based on industry aggregate data or its own loss reserve discount factors, which are based on its own data

The election is made with the company's tax filing for the determination year. It applies to that year and to the succeeding four years. If the company elects to use its own payment patterns, it uses data that are available before the beginning of each tax year. These are the data from two years earlier.

In this illustration, the company made an election with its 2007 tax filing to use its own data. The election applies to the 2007 through 2011 accident years.

PROJECTIONS

If the company uses the industry factors, the loss payment pattern is known until the next determination year. The pricing actuary must estimate only future federal mid-term rates, as discussed earlier.

At determination years, the loss payment pattern changes significantly. There is no simple method of projecting the future Schedule P data. The pricing actuary should use a loss payment pattern based on the actuarial projection, not the IRS projection.

If the company uses its own Schedule P data, the loss reserve discount factors can generally be estimated for one additional accident year only. The factors for the subsequent years should be based on actuarial projection techniques, not the IRS projection technique.

Illustration A: In July 2008, the pricing actuary is estimating loss reserve discount factors for accident years 2009, 2010, and 2011. The industry loss payment patterns were determined in 2007 for accident years 2007 through 2011. No future estimates are needed.

If the company uses its own loss payment patterns, the accuracy of the projection differs by accident year.

- Accident year 2009 uses loss payment patterns based on the 2007 Schedule P, which is available by July 2008. The only projection needed is for federal mid-term rates from July 2008 through December 2008.
- Accident year 2010 uses loss payment patterns based on the 2008 Schedule P entries, which the company's reserving actuary may be able to estimate.
- Accident year 2011 uses loss payment patterns based on the 2009 Schedule P. A projection of Schedule P entries 18 months in advance is unlikely to yield usable figures.

Illustration B: In July 2010, the pricing actuary is estimating loss reserve discount factors for accident years 2011, 2012, and 2013. The industry loss payment patterns are determined in 2012 for accident years 2012 through 2016.

- Accident year 2011 loss payment patterns are based on industry-wide 2005 Schedule P figures, which are available by the summer of 2006.
- Accident year 2012 and 2013 loss payment patterns are based on industry-wide 2010 Schedule P figures, which are not available until the summer of 2011. The IRS loss reserve discount factors are highly sensitive to random loss fluctuations. Attempting to project industry-wide Schedule P figures would not yield accurate discount factors. Instead, the actuary should estimate actuarial loss payment patterns through the tenth year based on the projection techniques discussed below. For years 11 through 16, the actuary should use the IRS procedure for extending the loss payment pattern, using the actuarial estimate for the percentage of losses paid in the tenth year.

One might presume that continuing the industry-wide loss payment patterns for accident years 2007 through 2011 for accident years 2012 and 2013 is a reasonable solution when the loss payment patterns for accident years 2012 and 2013 can not be readily estimated. This is not correct, since the loss payment patterns for accident years 2007 through 2011 are highly sensitive to the random loss fluctuations embedded in the 2005 Schedule P entries. The correct approach is to estimate an actuarial loss payment pattern for the line of business, not to repeat the previous IRS loss payment patterns.[49]

PROSPECTIVE PRICING AND RETROSPECTIVE ANALYSIS

The sensitivity of the IRS loss reserve discount factors to random loss fluctuations means that the actual discount factors may be quite different from the projected discount factors. In these situations, the actual discount factors should be used for retrospective analysis, and the variance should be ascribed to estimation error; see Feldblum [2002: Source of Earnings].

Illustration: In July 2010, the pricing actuary is estimating loss reserve discount factors for accident years 2011, 2012, and 2013. The company uses the industry-wide loss payment patterns. The actuary used actuarial projections for the accident year 2012 and 2013 loss payment patterns, as recommended above.

---

[49] The pricing actuary may be tempted to rely on the tax department's projections of future loss reserve discount factors, particularly if the tax department says that they can estimate future loss reserve discount factors. The tax department generally means that they can estimate the factors for accident year 20XX in December 20XX–1 instead of waiting for the official promulgation of the factors by the Secretary of the Treasury in the latter half of 20XX. This is not a "projection"; this is simply an independent computation of the factors. The tax department has no need to project loss reserve discount factors for future years. Many tax accountants would consider the actuary's request for future accident year factors as a misunderstanding of the vintaging provisions in the tax law.

For the retrospective analysis of policy profitability performed in calendar years 2012 and subsequent, the actuary should use the actual factors promulgated by the Treasury, not the original projections. The change from projected to actual is no different for loss reserve discount factors than it is for the loss trend factors in Feldblum [2002: SOE].

VINTAGING

The computed loss reserve discount factors are used for accident year 2011 only. The discount factors for previous accident years at every future valuation date have already been determined and frozen. In tax parlance, they are vintaged. They are not subsequently revised.

Illustration: We determine between 11 and 15 discount factors for accident year 2011. The first ten discount factors are used at valuation dates December 31, 2011, December 31, 2012, December 31, 2013, through December 31, 2020. The final one to five development factors are used at subsequent valuation dates. The development factors are combined into a composite development factor for the prior years row for valuation dates 2021, 2022, and subsequent. The discount factors all use the same discount rate and loss payment pattern. *The chart below shows the discount factors and the applicable valuation dates.*

Exhibit AppA.7: Valuation Dates for Loss Reserve Discount Factors

Discount Factor	Accident Year	Individual / Composite	Tax Year (Valuation Date)
12 mos	2011	individual	2011
24 mos	2011	individual	2012
...	...	...	...
120 mos	2011	individual	2020
132 mos	prior	composite	2021
144 mos	prior	composite	2022

The first ten discount factors apply to accident year 2011 only. They are used at valuation dates between 12 months and 120 months from inception of the accident year, corresponding to tax years 2011 through 2020. For subsequent valuation dates, the discount factor for accident year 2011 is combined with discount factors for other accident years to form a composite discount factor.

Discounting Sequence

The loss reserve discount factor computation can be divided into three steps.

A. Calculate the nominal (undiscounted) amounts for cumulative percentages paid, incremental percentages paid, and percentages unpaid.
B. Calculate the adjustments for long-tailed lines of business showing adjusted incremental percentages paid, long-tail extension of payments, and adjusted percentages unpaid.
C. Apply the appropriate discount rate to obtain the discounted percentages unpaid, and loss reserve discount factors.

UNDISCOUNTED PERCENTAGES

The loss reserve discount factors for this illustration are calculated in Exhibit AppA.1. Column 2 shows the cumulative net paid losses and loss adjustment expenses by accident year at the current statement date. Column 3 shows the incurred net losses and loss adjustment expenses by accident year at the current statement date. These entries include paid losses and loss adjustment expenses, case reserves, and bulk reserves.

Column 4 shows the cumulative percentage paid from inception of the accident year to the current statement date, or column 2 divided by column 3. For accident year 2009, the percentage is \$200,000 / \$571,500 = 35.00%. For accident year 2008, the percentage is \$330,000 / \$687,500 = 48.00%.

Assumed Incremental Percentage Paid

Column 5 shows the expected incremental percentage paid in each 12 month period. These entries are the first differences of the series in the previous column:

- For accident year 2009, the cumulative percentage paid at 12 months since inception of the accident year is 35.00%. For the most recent accident year, the incremental percentage paid equals the cumulative percentage paid.
- For accident year 2008, the cumulative percentage paid at 12 months since inception of the accident year is 48.00%. This implies that 48.00% – 35.00% = 13.00% of incurred losses are paid between 12 months and 24 months since inception of the accident year.

Schedule P shows 10 accident years of data, from which we estimate 10 twelve-month intervals of expected loss payments. If any losses remain unpaid at the end of 10 years – that is, if the cumulative paid losses for the oldest accident year does not equal the incurred losses for that accident year – we assume that all these losses are paid in the eleventh year, with the following limitation.

The amount assumed to be paid in the eleventh year is capped by the amount assumed to be paid in the tenth year. The excess amount is assumed to be paid in the twelfth year, but it is also capped at the same limit. The remaining excess is assumed to be paid in the thirteenth year, and so forth. We continue in this fashion through the fifteenth year. The remaining excess is assumed to be paid in the sixteenth year, with no limit. The next illustration (other liability) shows the computation of an extended loss payment pattern. We defer further explanation of the procedure for that illustration.

The Schedule P entries for the "prior years" row are not used in the computation of the loss reserve discount factors. The reserves and payments in this row relate to various accident years. A "composite" discount factor is used to determine the discounted loss reserves for the prior rows in Schedule P; see the discussion below.

In this illustration, the cumulative percentage paid for the ninth year (2001) is 94.94%, and the cumulative percentage paid for the tenth year (2000) is 98.00%. (The "n^{th}" year here means the "n^{th}" year working backwards from the current valuation date.)[50] The amount assumed to be paid from the end of the ninth year to the end of the tenth year is 98.00% – 94.94% = 3.06%. The amount still unpaid after 10 years is 100.00% – 98.00% = 2.00%. Since 2.00% is less than 3.06%, the full 2.00% is assumed to be paid in the eleventh year. No losses are assumed to be paid after 11 years.

Several of the commercial casualty lines of business have loss payment patterns extending beyond ten years; this is especially true for workers' compensation, other liability, products liability, and medical malpractice. For these lines of business, we don't expect the cumulative paid losses at the end of the tenth year to equal the incurred losses for that year.[51] The next illustration shows the adjustments used for these long-tailed lines of business.

DISCOUNTING COMPUTATIONS

Column 6 shows the percentage of losses unpaid at the end of the accident year, which equals the complement of the cumulative percentage of losses paid. For accident year 2009 in the illustration, the cumulative percentage of losses paid is 35.00%, and the percentage of losses unpaid at the end of the accident year is 100% – 35.00% = 65.00%.

---

[50] We estimate the amounts to be paid in future calendar years by looking at old accident years. The difference in the cumulative percentages paid between the n^{th} past accident year and the $(n+1)^{rst}$ past accident year is the percentage assumed to be paid between the end of the n^{th} calendar year from inception of the accident year to the end of the $(n+1)^{rst}$ calendar year from inception of the accident year. The n^{th} accident year working backwards from the most recent accident year corresponds to the n^{th} calendar year working forewards from the current statement date.

[51] The IRS computation of the loss reserve discount factors for all years is heavily influenced by the Schedule P entries for the ninth oldest accident year and the tenth oldest accident year. By random loss fluctuations, any long-tailed line of business may have an 11 year loss payment pattern one year and a 16 year loss payment pattern the next year.

Column 7 shows the discounted percentage of losses unpaid at the end of the accident year. To compute these figures, we assume that all losses are paid at mid-year. We may use either an iterative method, working backwards from the oldest accident year, or a formula method.[52]

Iterative Method

Two percent of the incurred losses are assumed to be paid in the eleventh year, labeled "AY + 10" in the exhibit. We assume that they are paid in mid-year. With a 7.0% discount rate, the discounted value of these losses at the preceding December 31 is $2\%/(1.070)^{0.5} = 1.93\%$.

Going backwards in accident years corresponds to going forwards in calendar years. The "current accident year" in this Schedule P exhibit is 2009, though the computed loss payment pattern is used for accident year 2011, not accident year 2009. The current valuation date for accident year 2011 for which this discount factor applies is December 31, 2011. Accident year AY+1 corresponds to calendar year 2011+1 = 2012. Accident year AY+10 corresponds to calendar year 2011 + 10 = 2021.[53]

To determine the discounted percentage of losses unpaid at the end of the ninth year, we combine two pieces:

i. The percentage of losses assumed to be paid in the tenth year – which are assumed to be paid at mid-year – discounted for half a year to the end of the ninth year.
ii. The discounted percentage of losses unpaid at the end of the tenth year, discounted for an additional year to the end of the ninth year.

In the illustration, the two pieces are as follows.

i. 3.07% of accident year 2011 losses are assumed to be paid in the middle of the tenth year, or July 1, 2020. They are discounted for half a year to December 31, 2019: $3.07\%/1.070^{0.5} = 2.97\%$.
ii. The discounted percentage of accident year 2011 losses unpaid at the end of the tenth year (December 31, 2020) is discounted for a full year: $1.93\%/1.070 = 1.80\%$.

The sum of 2.97% and 1.80% is 4.77%. We continue in this fashion for all accident years. This is the iterative method.

---

[52] The assumption that all losses are paid at mid-year is a proxy for an even distribution of paid losses during the year. In truth, losses are paid (on average) earlier than the middle of the year, particularly for losses paid in the 2 or 3 years following the inception of the accident year. The IRS procedure provides a slightly longer discount period than is warranted. This reduces the offset to taxable income and increases the income tax liability. This bias is offset by the shorter payment patterns implicit in the IRS extension past ten years.

[53] For an excellent explanation of this technique, see Salzmann [1984], who uses a similar version to develop a reserving method for allocated loss adjustment expenses.

Formula Method

Alternatively, formulas may be used for each year. The formula for the 2009 accident year in the Schedule P exhibit, which corresponds to accident year 2011 valued at December 31, 2011, is

$$(13.00\% \div 1.07^{0.5}) + (11.02\% \div 1.07^{1.5}) + \ldots + (3.07\% \div 1.07^{8.5}) + (2.00\% \div 1.07^{9.5}) = 52.26\%.$$

LOSS RESERVE DISCOUNT FACTORS

Column 8 shows the loss reserve discount factors used in the tax calculation. These factors are the discounted percentage of unpaid losses at the end of each year divided by the undiscounted percentage of unpaid losses at the end of that year. For accident year 2009, the loss reserve discount factor is 52.26% / 65.00% = 80.3944%. This corresponds to the loss reserve discount factor for accident year 2011 valued at December 31, 2011. If the accident year 2011 undiscounted reserves at December 31, 2011, are \$450,000, the corresponding discounted reserves are \$450,000 × 80.3944% = \$361,775.

The loss reserve discount factor in the preceding row, 81.6659%, is applied to the accident year 2011 reserves on December 31, 2012, not to the reserves of any other accident year. If the 2012 Schedule P reserves for accident year 2011 are \$350,000, the 2012 discounted reserves for accident year 2011 are \$350,000 × 81.6659% = \$281,380.

ILLUSTRATION B – LONG-TAILED EXTENSION OF PAYMENTS

In actual practice, the lines of business for which financial pricing models are appropriate will probably have extended loss payment patterns. The illustration below shows the procedure for extending the loss payment pattern beyond the eleventh year.

The following figures are taken from the 2009 Annual Statement, Schedule P, Part 1H (other liability), of a company that has elected to use its own loss payment pattern for computing discounted reserves for accident year 2011.

Exhibit AppA.8: Paid and Incurred Losses

Accident Year	Losses + LAE Paid	Losses + LAE Incurred
Prior	235,000	250,000
2000	50,000	55,500
2001	55,000	62,000
2002	60,000	70,000
2003	65,000	80,000
2004	70,000	96,000
2005	65,000	103,000
2006	60,000	115,000
2007	50,000	125,000
2008	35,000	140,000
2009	15,000	180,000

The 60 month rolling average of the federal mid-term rate, from January 2006 through December 2010, is 7.0% per annum.

Extension of Payments

The loss reserve discount factors are used for accident year 2011 only. In this illustration, we determine 15 separate loss reserve discount factors. The first ten discount factors are used for valuation dates December 31, 2011, through December 31, 2020. The 11th through the 15th discount factors are used at valuation dates December 31, 2021, through December 31, 2025 as part of the composite discount factor for accident years more than 10 years old.

CAPPING

The amount assumed to be paid in the eleventh year is capped by the amount assumed to be paid in the tenth year. In this illustration, 90.09% – 88.71% = 1.38% of incurred losses are assumed to be paid in the tenth year. The amount remaining unpaid after 10 years is

100.00% – 90.09% = 9.91% of the incurred losses. Only 1.38% is assumed to be paid in the eleventh year. The remaining 9.91% – 1.38% = 8.53% is assumed to be unpaid at the end of the eleventh year.

The 1.38% *cap affects the subsequent years as well. The amount assumed to be paid in each* of the five years immediately following the tenth year is the lesser of (i) the amount unpaid at the end of the previous year and (ii) the 1.38% cap. We show first an illustration with a loss payment pattern that does not extend through the 16th year before returning to the other liability illustration here.

Illustration: Suppose that the IRS loss reserve discounting procedure indicates that 90.90% is paid within 10 years and 88.10% is paid within nine years. This implies that 90.90% – 88.10% = 2.80% is paid in the tenth year. The amounts assumed to be paid in the 11th, 12th, and 13th years are also 2.80%. Only 9.10% – 3 × 2.8% = 0.70% remains unpaid after thirteen years. This is the amount assumed to be paid in the 14th year.

Whatever remains after 15 years is assumed to be paid in the 16th year, even if it exceeds the 1.38% cap.

Illustration: For illustration B, 9.91% – 5 × 1.38% = 3.01% remains unpaid after 15 years, so 3.01% is assumed to be paid in the sixteenth year.[54]

EXTENDED DEVELOPMENT

We begin the computation of the discounted percentages unpaid at the December 31 preceding the final loss payment. For this (other liability) illustration, the loss payment pattern extends through 16 years, so we begin the computation of the discounted percentage unpaid with the end of the fifteenth year.

3.01% of the accident year 2011 incurred losses are assumed to be paid in the middle of the 16th year, or July 1, 2026. The discounted loss reserve at the end of the 15th year (or December 31, 2025) is 3.01% / $1.070^{0.5}$ = 2.91%.

---

[54] See the Internal Revenue Code §§ 846(d)(3)(C) and (D), "Special rule for certain long-tail lines": In the case of any long-tail line of business, the period taken into account shall be extended (but not by more than 5 years), and the amount of losses which would have been treated as paid in the 10th year after the accident year shall be treated as paid in such 10th year and each subsequent year in an amount equal to the amount of the losses treated as paid in the 9th year after the accident year (or, if lesser, the portion of the unpaid losses not theretofore taken into account). To the extent such unpaid losses have not been treated as paid before the last year of the extension, they shall be treated as paid in such last year. The term "long-tail line of business" means any line of business if the amount of losses which would be treated as paid in the 10th year after the accident year exceeds the losses treated as paid in the 9th year after the accident year.

The discounted percentage unpaid at the end of the 14th year equals the sum of (i) the 2.91% discounted percentage unpaid at the end of the 15th year discounted for an additional full year and (ii) the 1.38% of the incurred losses assumed to be paid on July 1 of the 15th year discounted for half a year. This is $2.91\% / 1.070 + 1.38\% / 1.070^{0.5} = 4.05\%$. (The 0.01 percentage point difference from the figure in the exhibit is a rounding discrepancy.)

Alternatively, we calculate each discounted percentage unpaid by formula. For the 2011 valuation date for the 2011 accident year, the discounted percentage unpaid equals

$$(16.67\% \div 1.07^{0.5}) + (15.00\% \div 1.07^{1.5}) + \ldots + (1.38\% \div 1.07^{13.5}) + (3.01\% \div 1.07^{14.5}) = 70.87\%.$$

Patterns

The pricing model discussed in this paper and its companion papers serves two functions:

- It determines the premium rates needed to provide a target return on capital, and
- It shows the pattern of income recognition under alternative accounting systems.

The unwinding of the interest discount on the loss reserves affects the pattern of income recognition. The income recognition pattern is compared for six accounting systems in Feldblum and Thandi [2002], "Income Recognition and Performance Measurement." The discussion here shows the expected pattern of IRS loss reserve discount factors.

In Illustration B, the loss reserve discount factors are similar for the ten accident years that are separately reported in Schedule P, ranging from 77% to 80%. Some actuaries presume that loss reserve discount factors should be lowest (i.e., furthest below unity) at inception and should increase towards unity as the reserves become more mature. This presumption is that the amount of the discount as a percentage of the remaining reserves is greatest at early maturities and declines to zero at later maturities.

This presumption is correct for the true discount factor for an individual loss. Suppose a loss occurs on July 1, 20X1, and it will be paid on July 1, 20X9. The amount of the discount is greatest on December 31, 20X1, and it declines steadily thereafter.

This presumption is not correct for an accident year. If loss payments follow an exponential decay, as modeled by McClenahan [1975] and Butsic [1981], the loss reserve discount factor remains relatively constant as long as some claims remain unpaid. The expected discount factor depends on the rate of decay and the discount rate, not on the development period. As Butsic [1981] shows, if the loss payments follow an exponential decay, the average remaining time to settlement is constant over the lifetime of the reserves.[55]

The loss reserve discount factors in Illustration B increase steadily in the final six years, from 80% to about 97%. This is caused by the IRS assumption of a constant percentage of incurred losses paid in each development period during the extended part of the loss payment pattern, instead of the declining percentage of incurred losses assumed by an exponential decay pattern.[56] For instance, Illustration B uses a 1.38% figure for each

[55] For workers' compensation, the decay is slower than exponential. Temporary total claims dominate the early payments; most of these claims are settled within a year or two. Permanent partial disability and permanent total disability claims dominate the reserves for mature years. These claims may remain open for 30 or 40 years. The loss payment pattern is rapid initially but it is very slow by ten years of maturity.

[56] The exponential decay assumes that a constant percentage of the remaining reserves (not of the total incurred losses) is paid in each development period.

development period. The assumption of a final lump sum payment in the last year, whether or not the payment pattern is extended, augments the upward trend in the loss reserve discount factors for mature periods.

Composite Discount Factors

The loss reserve discount factors calculated above are applied to the unpaid losses for the appropriate accident year. Schedule P shows loss reserves by accident year only for the ten most recent years, to which ten separate loss reserve discount factors are applied. The 11th through 15th loss reserve discount factors are applied to the reserves in the Schedule P prior years row, which is not divided into the component accident years.

The IRS loss reserve discounting procedure assumes that all losses are paid no later than the 16th year. The prior years row in Schedule P contain losses that will be paid in the 12th through the 16th year, which use the loss reserve discount factors for years AY+11 through AY+15. A composite discount factor is formed from the five individual discount factors for application to the prior years row.

Each discount factor is the ratio of discounted reserves to undiscounted reserves for a given accident year at a given valuation date. For instance, the "tenth" accident year 2010 discount factor for AY+10 represents the discounted reserves for accident year 2010 at December 31, 2020, divided by the undiscounted reserves for accident year 2101 at December 31, 2020. This discount factor is computed in tax year 2010, not in tax year 2020.

We explain the calculation of the composite discount factor by illustration.

Illustration: Composite Discount Factors

For tax year 2019, Schedule P shows ten individual accident years: 2010 through 2019. Previous accident years – 2009 and prior – are grouped in the prior years row. Since the IRS loss reserve discounting procedure assumes that all losses are paid by the 16th year, we assume that the loss reserves in the prior years row represent losses from accident year 2005 through 2009.

We form a composite discount factor based on the following discount factors:

- Accident year 2005 discount factor for a valuation date 15 years after inception of year.
- Accident year 2006 discount factor for a valuation date 14 years after inception of year.
- Accident year 2007 discount factor for a valuation date 13 years after inception of year.
- Accident year 2008 discount factor for a valuation date 12 years after inception of year.
- Accident year 2009 discount factor for a valuation date 11 years after inception of year.

Some of these loss reserve discount factors use the same loss payment pattern. However, they all use different discount rates, and they are computed in separate years.

Suppose these five loss reserve discount factors are as shown below:

Exhibit AppA.11: Composite Discount Factor

Accident Year (1)	*Valuation Date (2)*	*Undiscounted Reserve (3)*	*Discounted Reserve (4)*	*Discount Factor (5)*
2005	AY + 15	5.0%	4.8%	96.9%
2006	AY + 14	7.2%	6.8%	93.9%
2007	AY + 13	9.1%	8.3%	91.0%
2008	AY + 12	11.7%	10.3%	88.2%
2009	AY + 11	13.3%	11.4%	85.4%
Total	prior years row	46.3%	41.6%	89.8%

The calculation of the individual discount factors is explained earlier. Each discount factor in column 5 is the ratio of the discounted reserves in column 4 to the undiscounted reserves in column 3. The reserve figures in columns 3 and 4 are expressed as percentages of the corresponding year's incurred losses. We compute the total of the five percentages for the discounted reserves and the undiscounted reserves. We divided these totals to obtain the composite discount factor for the prior years row.[57]

PROSPECTIVE PRICING

For the financial pricing model, we must project loss reserve discount factors until all the reserves run off. This time frame is the actual run-off date of the reserves, not the sixteen years assumed by the IRS. We explain the projection process by means of an illustration.

Illustration: The actuary is setting premium rates for policies written in 2005. For this block of business, reserves remaining 20 years after inception of the policy year are not material.

The losses on this block of business fall into accident years 2005 and 2006. Separate loss reserve discount factors are determined for each accident year. We examine here the factors for accident year 2005. The procedure for accident year 2006 is analogous, though it requires more estimation.

[57] Using a simple average to obtain the "total" row assumes that each year has the same volume of incurred losses. It might seem better to weight the discount factors by the actual percentage of incurred losses by accident year in the prior years row. However, the IRS bases the loss reserve discounting procedure on information contained in the Annual Statement. The distribution of the prior years row reserves by accident year is not found in the Annual Statement.

For valuation dates 12/31, 2005 through 2014, the discount factors are the factors for accident year 2005. For valuation date 12/31/2015, the remaining accident year 2005 reserves appear in the prior years row in the 2015 Schedule P. The discount factor applied to this row is the composite factor determined from the following individual factors:

- accident year 2005 at 11 years
- accident year 2004 at 12 years
- accident year 2003 at 13 years
- accident year 2002 at 14 years
- accident year 2001 at 15 years

The factors for accident years 2001 through 2004 factors are available to the pricing actuary. No additional estimation is required for the composite factor beyond what is required for other accident year 2005 loss reserve discount factors.

For valuation date December 31, 2016, the remaining accident year 2005 reserves appear in the policy years row in the 2016 Schedule P. The loss reserve discount factor applied to this row is the composite factor determined from the following individual factors:

- accident year 2006 at 11 years
- accident year 2005 at 12 years
- accident year 2004 at 13 years
- accident year 2003 at 14 years
- accident year 2002 at 15 years

The pricing actuary has not calculated any accident year 2006 loss reserve discount factors. The random component of the discount factors for the eleventh through the fifteenth years is so great that the projection of these discount factors for future accident years is not feasible.

Instead, the pricing actuary should compile average discount factors for the eleventh through the fifteenth years. Since the random component of these factors is so great, a long-term average should be used. We assume here that the actuary uses ten year averages.

The average loss reserve discount factor for the eleventh year is the average of the eleventh year factors for the ten accident years 1996 through 2005. (If the 2005 factor is only a projection, the actuary may use a ten year average for accident years 1995 through 2004.) This average loss reserve discount factor would be combined with the actual factors for accident years 2002 through 2005 to form the composite factor for accident year 2005 at December 31, 2016.

We continue this process for all subsequent valuation dates. By the December 31, 2020, all five loss reserve discount factors used for the composite discount factor are averages. This same composite factor is used for all subsequent valuation dates.

Revenue Offset

Proper treatment of the tax liability and the deferred tax asset stemming from revenue offset has a material effect on the indicated premiums. This section explains the revenue offset provision in the 1986 Tax Reform Act.

For other industries, sales constitute revenues for income tax purposes. Similarly, premium due is the taxable revenue (as well as the statutory and GAAP revenue) for life insurance companies. For property-casualty insurance companies, earned premium is the revenue for both statutory and taxable income, not written premium or collected premium.

For the statutory income statement, earned premium equals written premium minus the change in the unearned premium reserves. For taxable income, earned premium equals written premium minus 80% of the change in the unearned premium reserves.[58] [59]

- A change in written premium with no change in earned premium does not affect statutory income.
- A change in written premium with no change in earned premium affects the unearned premium reserve and changes the tax liability by means of the revenue offset provision.

Statutory and taxable income also differ in the treatment of accrued retrospective premiums. The statutory vs tax treatment of accrued retrospective premiums is important for the pricing of commercial casualty lines of business, such as workers' compensation and general liability. The pricing model must incorporate the both the statutory and the tax accounting treatment for this asset, as well as the resulting deferred tax asset. This subject is discussed in Feldblum [2002: Schedule P], and it is not repeated here.

The statutory (full expensing) vs GAAP (deferred policy acquisition cost) vs tax (revenue offset) treatment of policy acquisition costs is an important component of the financial pricing model discussed in this paper; see especially Feldblum and Thandi, "Income Recognition and Performance Measurement."

---

[58] See the Treasury regulations, 2001 FED 26,153, §1.832-4(a)(3): "The determination of premiums earned on insurance contracts during the taxable year begins with the insurance company's gross premiums written on insurance contracts during the taxable year, reduced by return premiums and premiums paid for reinsurance. This amount is increased by 80 percent of the unearned premiums on insurance contracts at the end of the preceding taxable year, and is decreased by 80 percent of the unearned premiums on insurance contracts at the end of the current taxable year."

[59] Life insurance companies and annuity writers are subject to a DAC-tax that is identical in concept though more complex than the property-casualty tax provision explained here; see Atkinson and Dallas [2000], chapter 9.

ILLUSTRATION: SINGLE POLICY

An insurer writes a policy with a $10,000 written premium on December 31, 20XX, and it pays $2,000 in agents' commissions on that day. Losses of $8,000 are incurred and paid evenly through the policy term. There are no other expenses or losses on this policy. We assume that losses are paid when they are incurred so that we need not deal with IRS loss reserve discounting.

The unearned premium reserve for this policy is $0 on January 1, 20XX, and $10,000 on December 31, 20XX. The change in the unearned premium reserve during the year is $10,000. The earned premium in 20XX is $10,000 of written premium minus the $10,000 change in the unearned premium reserve, or $0. Expenses during 20XX are $2,000, and statutory income during 20XX is –$2,000. Without revenue offset, the federal income tax liability would be 35% × –$2,000 = –$700, or a $700 tax refund.

The unearned premium reserve on December 31, 20XX+1, is $0. The change in the unearned premium reserve during 20XX+1 is –$10,000. The earned premium in 20XX+1 is $0 of written premium minus the –$10,000 change in the unearned premium reserve, or $0 – (–$10,000) = +$10,000. Losses of $8,000 are incurred and paid in 20XX+1. The statutory income is $10,000 – $8,000 = $2,000. The tax liability (ignoring revenue offset) would be 35% × $2,000 = $700.

Statutory accounting recognizes a loss at policy inception and a gradual profit during the remainder of the policy lifetime, thereby preventing companies from recognizing income until it has been fully earned.[60]

Were there no revenue offset provision in the tax code, the U.S. Treasury would fund part of the initial underwriting loss at policy inception. The illustration above shows a tax refund of $700 in 20XX and a tax liability of $700 in 20XX+1. Before 1987, statutory accounting helped the insurance industry defer its tax liabilities. Steady growth (in nominal dollar terms) led to persistent deferral of tax liabilities.

Direct and Indirect Methods

---

[60] Some analysts see a conservative bend in statutory accounting's write-off of pre-paid acquisition costs when they are incurred, particularly in comparison with GAAP's capitalization and amortization of the deferred policy acquisition cost asset. This is not quite correct. Statutory accounting is correct from a tangible asset perspective, since the prepaid acquisition costs may be incurred whether or not the company retains the policy. International accounting standards follow statutory accounting on this issue. GAAP capitalizes an "imaginary" asset called DPAC to match revenues and expenses and show a better portrayal of the company's profitability. However, statutory accounting is unduly conservative in its double treatment of underwriting expenses: once when they are incurred and a second time in the gross unearned premium reserves. See Yoheved and Sarason [2003] for further discussion of GAAP and statutory accounting of property-casualty insurance companies.

The Tax Reform Act of 1986 introduced the revenue offset provision of the Internal Revenue Code. The provision may be stated in two equivalent ways. These two perspectives are used in the two fashions of computing taxable income and the federal income tax liability, which are termed here the "direct method" and the "indirect method." The direct method is easier to understand; the indirect method is the method actually used in the Internal Revenue Code for computing taxable income.

1. *Direct method:* The taxable earned premium equals the taxable written premium minus 80% of the change in the unearned premium reserve. This may be stated as "only 80% of the change in the unearned premium reserve is an offset to taxable income."
2. *Indirect method:* Twenty percent of the change in the unearned premium reserve is an *addition* to statutory income for computing taxable income.

We can use either method for the illustration.

Direct method: The taxable earned premium in 20XX equals the taxable written premium minus 80% of the change in the unearned premium reserve, or $10,000 – 80% × ($10,000 – $0) = $2,000 in 20XX. Agents' commissions are $2,000 on December 31, 20XX. Taxable income is $2,000 – $2,000 = $0, and the tax liability is $0.

In 20XX+1, the taxable earned premium equals $0 – 80% × ($0 – $10,000) = $8,000. The losses incurred and paid in 20XX+1 are $8,000. The taxable income is $8,000 – $8,000 = $0, and the tax liability is $0.

Indirect method: Twenty percent of the change in the unearned premium reserve in 20XX is 20% × ($10,000 – $0) = $2,000. The statutory income in 20XX is –$2,000. Taxable income is –$2,000 + $2,000 = $0, and the tax liability is $0.

In 20XX+1, twenty percent of the change in the unearned premium reserve is 20% × ($0 – $10,000) = –$2,000. The statutory income in 20XX+1 is +$2,000. The taxable income is +2,000 – $2,000 = $0, and the tax liability is $0.

ILLUSTRATION B: TWO YEARS

An insurer writes a policy with a $10,000 written premium on July 1, 20XX, and it pays $2,000 in agents' commissions on that day. Losses of $8,000 are incurred evenly over the policy term, and they are paid when they are incurred. On July 1, 20XX+1, the insurer renews the policy for a written premium of $15,000, and it pays $3,000 in agents' commissions on that day. Losses of $12,000 are incurred evenly over the policy term, and they are paid when they are incurred. There are no other expenses on these policies.

Illustration B shows the importance of computing the *change* in the unearned premium reserve during the year. The statutory unearned premium reserve equals $0 on December

31, 20XX–1, $5,000 on December 31, 20XX, $7,500 on December 31, 20XX+1, and $0 on December 31, 20XX+2.

CALENDAR YEAR 20XX

Statutory earned premium is $10,000 written premium minus the ($5,000 – $0) = $5,000 change in the unearned premium reserve; the earned premium is $5,000. Expenses are $2,000, and incurred losses are $4,000. The statutory income in 20XX is $5,000 – $1,000 – $4,000 = –$1,000. There are two methods to calculate the taxable income.

i. *Direct method:* The taxable earned premium is taxable written premium minus 80% of the change in the unearned premium reserve, or $10,000 – 80% × ($5,000 – $0) = $6,000. The taxable income is $6,000 – $2,000 – $4,000 = $0, and the tax liability is $0.
ii. *Indirect method:* Twenty percent of the change in the unearned premium reserve is 20% × ($5,000 – $0) = $1,000. The statutory income in 20XX is –$1,000. The taxable income is –$1,000 + $1,000 = $0, and the tax liability is $0.

CALENDAR YEAR 20XX+1

Statutory earned premium is $15,000 written premium minus the ($7,500 – $5,000) = $2,500 change in the unearned premium reserve; the earned premium is $12,500. Expenses incurred and paid on January 1, 20XX+1, are $3,000, and incurred losses during the year are $4,000 (first six months) + $6,000 (latter six months) = $10,000. The statutory income is $12,500 – $3,000 – $10,000 = –$500. There are two methods to calculate taxable income.

i. *Direct method:* The taxable earned premium is the taxable written premium minus 80% of the change in the unearned premium reserve, or $15,000 – 80% × ($7,500 – $5,000) = $13,000. Expenses and losses are the same as for statutory income. The taxable income is $13,000 – $3,000 – $10,000 = $0, and the tax liability is $0.
ii. *Indirect method:* Twenty percent of the change in the unearned premium reserve is 20% × ($7,500 – $5,000) = $500. The statutory income in 20XX+1 is –$500. The taxable income is –$500 + $500 = $0, and the tax liability is $0.

CALENDAR YEAR 20XX+2

Statutory earned premium is $0 written premium minus the ($0 – $7,500) = –$7,500 change in the unearned premium reserve, or $7,500. Expenses incurred in 20XX+2 are $0, and incurred losses during the year are $6,000. Statutory income is $7,500 – $6,000 = $1,500.

There are two methods to calculate the taxable income.

i. *Direct method:* The taxable earned premium is \$0 – 80% × (\$0 – \$7,500) = \$6,000. The taxable income is \$6,000 – \$6,000 = \$0, and the tax liability is \$0.
ii. *Indirect method:* Twenty percent of the change in the unearned premium reserve is 20% × (\$0 – \$7,500) = –\$1,500. The statutory income in 20XX+2 is \$1,500. The taxable income is \$1,500 + –\$500 = \$0, and the tax liability is \$0.

Deferred Tax Assets

The computation of the admitted portion of the deferred tax asset stemming from IRS loss reserve discounting is based on two items:

- the loss reserve discount factors by accident year and by line of business for the current valuation date and for the valuation date 12 months hence, and
- the company's loss payment pattern by line of business.

The IRS loss payment pattern is used to compute the loss reserve discount factors. The actuary's estimated loss payment pattern is used to compute the admitted portion of the deferred tax asset.

Of all the changes in the NAIC's codification project, the deferred tax asset stemming from IRS loss reserve discounting has the greatest effect on policy pricing and company valuation. The deferred tax assets stemming from revenue offset and loss reserve discounting are used extensively in this paper and in all its companion papers. We present first the requisite background explanations of deferred tax assets and liabilities, and we illustrate the computation of the deferred tax assets relevant for policy pricing.

CURRENT TAXES VS DEFERRED TAXES

There are two ways of accounting for federal income taxes:

- The incurred tax liability is the tax liability actually incurred by the taxpayer, based on the provisions of the Internal Revenue Code, or
- The accrued tax liability is the tax liability implied by the company's balance sheet, whether GAAP or statutory.

Current taxes are the incurred tax liability. The current year's change to the deferred tax asset or liability is the difference between the incurred tax liability and the accrued tax liability.[61] The change to the deferred tax asset or liability is a direct charge or credit to surplus shown on line 24 of the NAIC Annual Statement. As a direct charge and credit to surplus, it has the same effect on the implied equity flows as though it flowed through the income statement.[62]

---

[61] This definition uses a retrospective computation. SFAS 109 requires a prospective computation, which may be different if the tax rate changes or if there are other changes in tax regulations. For simplicity, we use the retrospective viewpoint at first. We explain the prospective viewpoint further below.

[62] Direct charge and credit to surplus are not included in after-tax net income for standard accounting statements, both statutory and GAAP. The implied equity flows depend on the statutory balance sheet entries, not the income statement entries.

Before 2001, insurers could not admit any deferred tax asset or liabilities on the statutory balance sheet. In contrast, GAAP recognizes deferred tax assets and liabilities if they are expected to be realized; see SFAS 109. With the implementation of codification in 2001, statutory accounting recognizes deferred tax liabilities and a portion of deferred tax assets.

Permanent Differences and Timing Differences

Tax accounting differentiates between permanent differences and timing differences, as defined below.

- *Permanent differences* are differences that do not reverse in later accounting periods. The tax exemption for municipal bond interest is a permanent difference.
- *Timing differences* are differences that reverse in later accounting periods. The revenue offset provision creates a timing difference between statutory income and taxable income.

An alternative perspective is to view permanent differences as differences in the tax rates applicable to different sources of income; see Feldblum and Thandi, "Income Recognition and Performance Measurement." For property-casualty insurers, both corporate bond income and municipal bond income are taxable income, but the former has a 35% tax rate and the latter has a 5.25% tax rate; see Feldblum and Thandi, "Investment Yields."

Income Statement vs Balance Sheet

It is tempting to define timing differences as differences in the timing of income between the book income statement (i.e., GAAP or statutory) and the tax income statement. This is not correct.

> *Timing differences are differences between the tax income statement and the income statement implied by the GAAP or statutory balance sheet.*[63]

UNREALIZED CAPITAL GAINS AND LOSSES

For each accounting year, we compute the difference between the book value and the cost of the financial asset. The change in this difference from the previous year to the current year is the unrealized capital gain or loss. For common stocks, the book value is the market value.

---

[63] This definition is particularly relevant to the deferred tax liabilities and assets stemming from unrealized capital gains and losses. For the deferred tax assets stemming from revenue offset and loss reserve discounting, we could use the difference between statutory income and taxable income.

Unrealized capital gains and losses are admitted on the statutory (as well as GAAP) balance sheet, though they do not flow through the income statement. They are direct charges and credits to surplus, not a portion of net income.

For tax purposes, capital gains and losses are not part of income until they are realized.

- Unrealized capital gains increase the book value of common stocks on the statutory balance sheet. There is no incurred tax liability in the current tax year. Instead, the reporting company shows a deferred tax liability.
- Similarly, unrealized capital losses decrease the book value of common stocks on the statutory balance sheet. There is no tax refund in the current tax year. Instead, the reporting company shows a deferred tax asset.

Illustration

ABC Insurance Co buys common stock for $50 million on December 31, 20XX.

- On December 31, 20XX+1, the common stock are worth $40 million;
- On December 31, 20XX+2, the common stock are worth $60 million; and
- On December 31, 20XX+3, the common stock are worth $80 million.

The federal income tax rate is 35%. On December 31, 20XX+3, the ABC Insurance Company sells the common stock. We calculate the following accounting entries:

- The unrealized capital gains and losses in years 20XX+1, 20XX+2, and 20XX+3.
- The realized capital gains and losses in years 20XX+1, 20XX+2, and 20XX+3.
- The deferred tax assets and liabilities in years 20XX+1, 20XX+2, and 20XX+3.

Tax year 20XX+1

The market value of the stock has decreased by $10 million. The stock has not been sold yet, so the capital loss is unrealized. There are no realized capital gains and losses.

- On December 31, 20XX, book value – cost = $50 million – $50 million = $0.
- On Dec 31, 20XX+1, book value – cost = $40 million – $50 million = –$10 million.
- The unrealized capital gain or loss = –$10 million – $0 million = –$10 million.

The current balance sheet shows a decline of $10 million. When the stocks are sold, ABC Insurance Company will have an income loss of only $6.5 million, since the capital loss can offset other capital gains, and the company's tax liability will be reduced by $3.5 million. There is a $3.5 million deferred tax asset on the 20XX+1 balance sheet.

Tax year 20XX+2

The stock prices have increased. The unrealized capital gain is the *change* in the difference between book value and cost of the stocks. The unrealized capital gain for 20XX+2 is $20 million. The realized capital gain is again zero, since the stocks have not been sold.

- On December 31, 20XX+1, book value – cost = $40 million – $50 million = –$10 million.
- On December 31, 20XX+2, book value – cost = $60 million – $50 million = +$10 million.
- The unrealized capital gain or loss = +$10 million – (–$10 million) = +$20 million.

The company's balance sheet is $20 million stronger than it was a year ago. However, if the stocks were sold now, the company would realize a gain of only $13 million, since $7 million would go to taxes. The *change* in the deferred tax assets and liabilities is a credit of $7 million. Since we began with a deferred tax asset (a debit) of $3.5 million, we now have a deferred tax liability (a credit) of $3.5 million.

Tax year 20XX+3

The company sells the stock. The difference between market value and cost of the stocks is now $0 (since there are no more stocks on the balance sheet), so the unrealized capital gain is –$10 million.

- On December 31, 20XX+2, book value – cost = $60 million – $50 million = +$10 million.
- On December 31, 20XX+3, book value – cost = $0 million – $0 million = $0 million.
- The unrealized capital gain or loss = $0 million – ($10 million) = –$10 million.

The realized capital gain, which is defined as the sale price minus the purchase price, is +$30 million. The deferred tax assets and liabilities are now zero.[64]

Prospective Pricing

To see the effects of deferred tax assets and liabilities on determining a benchmark investment yield for policy pricing, we consider an illustration of bond and stock returns.

---

[64] Unrealized capital gains and losses give rise to deferred tax liabilities and assets, respectively. Realized capital gains and losses affect current taxes; they do not give rise to deferred tax assets and liabilities. An exception stems from the rule that capital losses can offset capital gains but not operating gains.

If capital losses exceed capital gains, the company may carry forward the unused capital losses. The tax rate times the unused capital loss is a deferred tax asset, not a deduction in current tax liabilities.

Capital losses can be carried forward a limited number of years. If during these years the company has not realized sufficient capital gains to offset all the capital losses, the remaining capital losses expire unused, and the deferred tax asset is removed.

Illustration: An insurer invests in bonds yielding 10% per annum and in common stocks that pay no dividends but that are expected to increase in market value by 10% per annum. We examine the effects of each on the benchmark investment yield for a financial pricing model. To keep the illustration clear, we assume that the insurer begins with $100 million of each type of security.

Single Year: During the first year, the bonds yield $10 million. The federal income tax liability is $3.5 million, and the increase in statutory surplus is $6.5 million. The after-tax investment yield for the purposes of the pricing model is 6.5% per annum.

During the first year, the expected change in the common stock value is +$10 million. The expected deferred tax liability is $3.5 million, so the expected change in the statutory balance sheet is a $6.5 million increase in surplus. For the single year scenario, the after-tax investment yield is 6.5% for the common stocks.

For a multi-period scenario, the yields on the bonds and the common stocks are not identical, even when deferred tax assets and liabilities are considered. We show this with a two year scenario. We assume that the investment income in reinvested in the same securities.

Two Years: The bond portfolio yields investment income of $10 million the first year. Of this amount, $3.5 million is paid to the U.S. Treasury, and $6.5 million is reinvested in the bond portfolio. During the second year, the bond portfolio yields investment income of $106.5 million × 10% = $10.65 million. Of this amount, $10.65 million × 35% = $3.7275 million is paid to the U.S. Treasury, and $10.65 million × 65% = $6.9225 million is reinvested in the bond portfolio. The total in the bond portfolio after two years is $106.5 million + $6.9225 = $113.4225. This is a 6.5% annual yield, since $1.065^2 = 1.134225$.

The common stock portfolio appreciates to $110 million the first year. Nothing is paid to the U.S. Treasury, and the company sets up a $3.5 million deferred tax liability on its balance sheet. During the second year, the common stock portfolio appreciates to $121 million. Nothing is yet paid to the U.S. Treasury, and the company increases the deferred tax liability to $21 million × 35% = $7.35 million. The net common stock asset is $121 million – $7.35 million = $113.65 million. This is a 6.607% annual yield, since $1.06607^2 = 1.1365$.

The effect of the tax deferral on the effective investment yield is small for a short holding period. The effect is material for longer holding periods.[65] Formulas for calculating the effective investment yield for different securities are shown in Feldblum and Thandi, "Investment Yields."

---

[65] The effective holding period is generally estimated as the reciprocal of the turnover rate. If 5% of the common stocks are sold each year, the effective holding period is 1 ÷ 5% = 20 years. This approximation is not exact, but the difference from the exact result is insignificant.

We sum up the gist of this discussion as follows:

- For calculating after-tax net income, a deferred tax liability is not the same as an actual tax liability.
- For calculating implied equity flows, a deferred tax liability is the same as an actual tax liability.
- For calculating the investment yield, a deferred tax liability is slightly different from an actual tax liability. The magnitude of the difference depends on the effective holding period of the securities.

Statutory Recognition of Deferred Tax Assets

All deferred tax liabilities are recognized on the statutory balance sheet. For most deferred tax assets, the admitted statutory portion equals the entire asset, and statutory accounting is the same as GAAP.[66] In certain instances, only a portion of the deferred tax assets are recognized on the statutory balance sheet. This applies particularly to the deferred tax asset stemming from IRS loss reserve discounting for medium- and long-tailed lines of business.

SSAP No. 10, "Income Taxes," paragraph 10, says:

Gross DTAs shall be admitted in an amount equal to the sum of:

a Federal income taxes paid in prior years that can be recovered through loss carrybacks for existing temporary differences that reverse by the end of the subsequent calendar year;

b The lesser of:

  i. The amount of gross DTAs, after the application of paragraph 10 a., expected to be realized within one year of the balance sheet date; or

  ii. Ten percent of statutory capital and surplus as required to be shown on the statutory balance sheet of the reporting entity for its most recently filed statement with the domiciliary state commissioner adjusted to exclude any net DTAs, EDP equipment and operating system software and any net positive goodwill; and

---

[66] There are two potential differences between GAAP and statutory accounting even when the full deferred tax asset passes the 12 month test:

- Some companies use a valuation allowance on the GAAP balance sheet for deferred tax assets and liabilities that may not reverse.
- Some companies use fair values, or discounted values, for deferred tax assets and liabilities that may not reverse for many years.

c. *The amount of gross DTAs, after application of paragraphs 10 a. and 10 b., that can be offset against existing gross DTLs.*

A gross deferred tax asset is admissible if it will reverse within one year, as required by paragraph (a) and by paragraph (b.i).

The limitation of 10% of surplus in paragraph (b.ii) is applicable for some companies, depending on the circumstances of the company's business and capital. Actuaries estimating the admitted portion of the deferred tax asset for these companies must take this limitation into account.

This is one of the rare instances where the implied equity flows depend not just on the book of business being priced but on all operations of the company. It would be rare for the deferred tax asset stemming from a single policy year to exceed 10% of statutory surplus, but is may occur that the total deferred tax asset of the company exceeds 10% of its surplus.

Illustration: The ABC Insurance Company writes annual policies with effective dates spread evenly through the year. Its premium to surplus ratio is 2 to 1, and its reserves to surplus ratio is 4 to 1. Its average loss reserve discount factor is 80%, and 30% of its deferred tax asset from loss reserve discount will reverse within 12 months. We work out its gross deferred tax asset and the portion admitted on the statutory balance sheet.

To simplify the computations, we use numbers instead of algebraic variables. We assume that statutory surplus is $100 million. Any other figure would work just as well.

The premium to surplus ratio is 2 to 1, so annual premium is $200 million. Since it uses annual policies with effective dates spread evenly through the year, the unearned premium reserves are $200 million × 50% = $100 million. The deferred tax asset stemming from revenue offset is $100 million × 20% × 35% = $7 million.

The reserves to surplus ratio is 4 to 1, so the undiscounted loss reserves are $400 million. The average loss reserve discount factor is 80%, so the discounted reserves are $400 million × 80% = $320 million. The gross deferred tax asset stemming from IRS loss reserve discounting is ($400 million – $320 million) × 35% = $28 million. The portion admitted on the statutory balance sheet is $28 million × 30% = $8.4 million.

The total statutory deferred tax asset is $7 million + $8.4 million = $15.4 million. This is limited to 10% of statutory surplus. Ten percent of statutory surplus is $10 million, so an additional $5.4 million is not admitted.

This illustration is reasonable, but it does not reflect the average company. Various changes in the scenario would reduce or eliminate the effect of the "10% of surplus" restriction.

- A lower premium to surplus ratio would reduce the effect of this restriction. The U.S. insurance industry as a whole has a premium to surplus ratio of about 1 to 1. With a 1 to 1 premium to surplus ratio in the illustration above, the 10% of surplus restriction has no effect on the admitted portion of the deferred tax asset.
- A shorter policy term would reduce the deferred tax asset stemming from revenue offset. With six month policies, the unearned premium reserve would be only half the size and the deferred tax asset stemming from revenue offset would be only half the size.
- The property lines of business do not have slowly liquidating reserves that would generate a significant deferred tax asset from loss reserve discounting.
- Companies with effective dates clustered around January 1 have much lower unearned premium reserves at the end of the year, and therefore have lower deferred tax assets stemming from revenue offset.

Most companies do not have deferred tax assets that will reverse in the coming year and that exceed 10% of policyholders' surplus. The deferred tax asset stemming from revenue offset is usually about 1% to 5% of statutory surplus. The deferred tax asset stemming from IRS loss reserve discounting is larger for companies that predominate in the long-tailed casualty lines of business, but most of this deferred tax asset does not reverse within one year. For companies with low surplus, this restriction is important.

The offsetting against existing gross deferred tax liabilities mentioned in paragraph (c) is relevant for companies with large unrealized capital gains from common stock holdings. The actuary should take this provision into account when quantifying the admitted portion of the deferred tax asset.

Common stock that has suffered an unrealized capital loss may be sold within the next 12 months to realize the tax benefits. A literal reading of the SSAP would permit the recognition of the deferred tax asset only if the company expects to realize the capital loss during the coming calendar year. In practice, most auditors do not require an explicit company expectation to realize the loss in order to admit the deferred tax asset.

Revenue Offset

The deferred tax asset stemming from revenue offset is similar to the deferred tax asset stemming from loss reserve discounting. For annual policies, the entire deferred tax asset will reverse during the coming year, and it is fully admitted on the statutory balance sheet.

BACKGROUND

All acquisition expenses flow through the statutory income statement when they are incurred. No deferred policy acquisition cost (DPAC) asset is entered on the statutory balance sheet.

On GAAP financial statements, acquisition expenses are capitalized on the balance sheet and amortized through the income statement over the term of the policy. The DPAC asset depends on the actual expenses incurred by the company.

For tax purposes, 20% of the written premium is treated as acquisition expenses that are capitalized and amortized over the term of the policy.[67] More precisely, the revenue offset provision defines the taxable earned premium.

- Statutory earned premium equals written premium minus the change in the unearned premium reserves.
- Taxable earned premium equals written premium minus 80% of the change in the unearned premium reserves.

ILLUSTRATION: DPAC OF 20%

An annual policy with a premium of $1,000 and acquisition expenses of $200 is written on December 31, 20XX.

- The statutory balance sheet shows a loss of $200. The written premium of $1,000 is offset by the unearned premium reserve of $1,000, and the incurred acquisition cost of $200 flows through the income statement.
- For tax purposes, the $1,000 written premium is offset by only $800 of unearned premium reserves, leaving a $200 gain. This $200 gain combined with the $200 acquisition cost yields a $0 net gain or loss.

The income implied by the statutory balance sheet – taxable income = –$200 – $0 = –$200.

In 20XX+1, statutory earned premium is $1000, since the entire unearned premium reserve is taken down over the course of the year. The taxable income is $800, since only 80% of the change in the unearned premium reserve is considered. For 20XX+1, the income implied by the statutory balance sheet – taxable income equals $1000 – $800 = $200.

At the end of 20XX+1, the statutory balance sheet equals the implied tax balance sheet. Both show net cash received of $1000 – $200, or the written premium minus the acquisition expense. The temporary balance sheet difference at December 31, 20XX fully reverses by December 31, 20XX+1.

At December 31, 20XX, taxable income is $200 greater than the income implied by the statutory balance sheet. The tax liability for 20XX is 35% × $200 = $70 greater than the tax liability that would be determined from the statutory balance sheet. Since the $70 difference

---

[67] Life and health insurers and annuity writers have a similar "DAC-tax."

will reverse over the coming 12 months, it is recognized as a deferred tax asset on the statutory balance sheet.

The deferred tax asset on the statutory balance sheet does not depend on the amount of actual acquisition expenses. In contrast, the deferred tax asset on the GAAP balance sheet depends on the size of the GAAP deferred policy acquisition cost asset relative to the 20% assumption in the revenue offset provision.

ILLUSTRATION: DPAC OTHER THAN 20%

A company writes and collects a $1000 annual premium on December 31, 20XX. Acquisition expenses of $250 are incurred (and paid) on December 31, 20XX. The marginal tax rate on underwriting income is 35%. All acquisition costs are deferrable under GAAP.

Taxable underwriting income for 20XX is $200 (taxable premium income from revenue offset) – $250 (acquisition expenses) = –$50. The tax outflow is a negative $17.50 (or a tax refund of $17.50).[68]

The taxable premium income may be evaluated in either of two ways.

- Taxable earned premium = written premium minus 80% of the change in the unearned premium reserves = $1000 – 80% x $1000 = $200.
- Taxable earned premium = statutory earned premium plus 20% of the change in the unearned premium reserves = $0 + 20% x $1000 = $200.

The tax liability is 35% times the taxable income: 35% x ($200 – $250) = –$17.50.

Taxable underwriting income for 20XX+1 equals $800 of taxable premium income. The tax outflow is $800 x 35% = $280.00. Written premium during the year is $0 and the unearned premium reserve declines from $1000 to $0. We use the same two computation methods: (i) $0 – 80% x (–$1000) = $800, or (ii) $1000 + 20% x (–$1000) = $800.

A deferred tax asset of $70 stemming from the revenue offset provision is entered on the balance sheet on December 31, 20XX, and it is amortized over the course of the policy term. The full deferred tax asset from revenue offset is recognized on the statutory balance sheet, since it reverses within 12 months of the balance sheet date (for annual policies).

On GAAP financial statements, the book income for 20XX is $1000 – $0 = $1000, since all acquisition expenses are capitalized. The taxable income is –$50 (as above), and the tax

---

[68] The tax refund stemming from negative taxable income offsets tax liabilities stemming from positive taxable income on other insurance contracts. There is no need to presume tax carrybacks or carryforwards.

liability is –$17.50 (i.e., a refund). GAAP shows a deferred tax liability (not an asset) of $17.50, exactly offsetting the tax refund.

The text of this paper uses an acquisition expense greater than 20% of premium in order to show the differing treatments under GAAP and statutory accounting. The same illustration is carried through to the companion papers.

LOSS RESERVE DISCOUNTING

The deferred tax asset stemming from loss reserve discounting is the most difficult component of the financial pricing model for some readers. The cause of this difficulty lies in the training and experience of many North American actuaries. The actuarial aspects of the pricing model are covered in the examination syllabus, and many actuaries have experience with cash flow analysis at work. Tax accounting is not emphasized on the actuarial syllabus, and few actuaries deal with deferred tax assets in their jobs.

This situation is unfortunate. The concepts involved in tax accounting are not difficult, and they significantly affect policy pricing. There is perhaps no better example than the deferred tax asset stemming from loss reserve discounting. There is no good documentation of the procedures (other than here), and many actuaries never have the opportunity to master the calculations. In truth, the procedure is straight-forward; it takes no more than fifteen or twenty minutes to learn how it is done.

The statutory incurred losses are the paid losses plus the change in the undiscounted loss reserves. The taxable incurred losses are the paid losses plus the change in the *discounted* loss reserves. The difference between statutory and taxable incurred losses is a timing difference. The change in the deferred tax asset is 35% of this difference.

Illustration: A policy is issued on January 1, 20XX, for a premium of $1000 and expenses of $200. Losses of $800 are incurred in 20XX, of which half are paid in 20XX and half are paid in 20XX+1. The IRS loss reserve discount factor at the 12 month valuation is 90%. For simplicity, we assume that the companies earns no investment income.

- The statutory incurred losses in 20XX are $400 of paid losses plus $400 of loss reserve change = $800. Statutory income is $1000 – $200 – $800 = $0. The accrued taxes on income of $0 is $0.
- The taxable incurred losses in 20XX are $400 of paid losses plus $360 of change in discounted loss reserves = $760. Taxable income is $1000 – $200 – $760 = $40. The tax liability on $40 is $14.

The difference between the income implied by the statutory balance sheet and taxable income is $0 – $14 = –$14. The gross deferred tax asset is $14.

Only the portion of the deferred tax asset that reverse within 12 months is admitted on the statutory balance sheet. We examine the statutory income and taxable income for 20XX+1.

- The statutory incurred losses in 20XX+1 are \$400 of paid losses plus –\$400 of loss reserve change = \$0. There is no premium or expense in 20XX+1, so statutory income is \$0. The accrued taxes on income of \$0 is \$0.
- The taxable incurred losses in 20XX+1 are \$400 of paid losses plus –\$360 of change in discounted loss reserves = \$40. There is no premium or expense in 20XX+1, so taxable income is \$0 – \$40 = –\$40. The tax liability is 35% × (–\$40) = –\$14.

The full difference between statutory and taxable income reverses in 20XX+1, so the full deferred tax asset of \$14 is admitted on the statutory balance sheet.

Twelve Month Reversal

We present the formula for computing the admitted portion of the deferred tax asset stemming from loss reserve discounting. The computations are done separately by line of business and by accident year.

Illustration: For accident year 20XX in a given line of business, the loss reserve discount factors are Z_1 at December 31, 20YY, and Z_2 at December 31, 20YY+1. Let "R" be the held loss reserves at December 31, 20YY. Let "P" be the percentage of accident year 20XX reserves that will be paid during calendar year 20XX.

- At December 31, 20YY, the difference between statutory and taxable income for accident year 20XX is $R \times (1 - Z_1)$. The gross deferred tax asset is $35\% \times R \times (1 - Z_1)$.
- At December 31, 20YY+1, the difference between statutory and taxable income for accident year 20XX is $R \times (1 - P) \times (1 - Z_2)$. The gross deferred tax asset is $35\% \times R \times (1 - P) \times (1 - Z_2)$.
- The admitted portion of the deferred tax asset on the statutory balance sheet at December 31, 20YY is $35\% \times R \times [(1 - Z_1) - (1 - P) \times (1 - Z_2)]$.

The value of "P" depends on the actuary's best estimate of the loss payment pattern. It is not the same as the IRS loss payment pattern. To estimate the pattern, we must derive actuarially justified discount factors.

Actuarial Discount Factors

The percentage of losses expected to be paid by each valuation date is the reciprocal of the paid loss development factor.[69] This is a standard actuarial procedure, not peculiar to tax accounting. We show an illustration here for the benefit of readers who have not dealt with reserve liquidation patterns. The illustration uses the same data as the previous illustrations of the IRS loss payment patterns.[70]

The illustration assumes the actuary is determining the deferred tax asset stemming from loss reserve discounting for accident year 20X9 workers' compensation business. Since the amount of the loss reserve discount depends on Schedule P data and the discount is based on Schedule P lines of business, it makes sense to determine the loss liquidation pattern from Schedule P data for Schedule P lines of business.[71]

If the characteristics of the book of business are changing, or if the book of business can be separated into components with different loss liquidation patterns, the pricing actuary should use separate analyses for each component. This is important for accurate policy pricing.

Illustration: An insurer writes two types of workers' compensation coverage:

- First dollar coverage for small and medium size insureds.
- Large dollar deductible policies for large accounts.

Both blocks of business are included in the company's workers' compensation line of business for Schedule P purposes. Both blocks of business have the same loss reserve discount factors for tax purposes. The two blocks of business have different deferred tax assets stemming from the IRS loss reserve discounting, since the actual loss liquidation pattern differs by type of policy. First dollar coverage pays out rapidly. Large dollar deductible (also termed "high deductible") coverage pays out very slowly, since the insurer's payments begin after a large deductible has been pierced. With a deductible of $500,000, which is common for large ("national") accounts, the insurer's payments may begin years after the accident date.

One may be tempted to think that the loss liquidation pattern for calculating the deferred tax asset should be the same as the loss payment pattern for calculating the loss reserve discount factors. This is not correct. The loss reserve discount factors are tax factors. The deferred tax assets are statutory accounting figures. In this illustration, we use the same raw data for the deferred tax assets as we used for the loss reserve discount factors, but the loss

---

69 See Feldblum [2002: SB] for a full discussion of this topic.

70 A more comprehensive discussion of this illustration may be found in Feldblum [2002: Schedule P].

liquidation pattern for the deferred tax assets is not the same as the loss payment pattern for the loss reserve discount factors.

Exhibit AppA.12 shows the Schedule P, Part 3D entries as they would appear in the 20X9 Schedule P for accident years 20X0 through 20X9 for the loss reserve discount factors estimated above.[72]

Exhibit AppA.12: 20X9 Schedule P, Part 3D ($000)

Part 3	20X0	20X1	20X2	20X3	20X4	20X5	20X6	20X7	20X8	20X9
20X0	103	226	294	334	363	384	398	412	422	433
20X1		111	238	309	356	387	409	428	442	454
20X2			108	221	286	328	354	375	391	403
20X3				111	238	311	357	392	416	434
20X4					135	299	394	458	504	534
20X5						146	314	418	490	542
20X6							159	343	463	546
20X7								146	353	485
20X8									152	406
20X9										156

Paid Loss Link Ratios

We determine the paid loss link ratios from these data. We use these link ratios to calculate the loss liquidation pattern, not to calculate the indicated reserves. Even if the company determines its reserve indications by another reserving method, it would use the procedure described here to determine the loss liquidation pattern.

Paid loss link ratios are the ratios of

i *cumulative paid losses for a specific accident year at a given valuation date to*
ii cumulative paid losses for the same accident year at a valuation date one year earlier.

For instance, the paid loss link ratio from two years to three years of development for accident year 20X6 is $463,000 divided by $343,000, or 1.350. The complete set of link ratios is shown in the table below.

[72] These data are based on actual Schedule P entries for a large commercial lines insurer that was acquired by a peer company in the mid-1990's. The figures have been disguised, and the accident years have been changed.

Exhibit AppA.13: 20X9 Schedule P, Paid Loss Link Ratios

	1 to 2	2 to 3	3 to 4	4 to 5	5 to 6	6 to 7	7 to 8	8 to 9	9 – 10
20X0	2.194	1.301	1.136	1.087	1.058	1.036	1.035	1.024	1.026
20X1	2.144	1.298	1.152	1.087	1.057	1.046	1.033	1.027	
20X2	2.046	1.294	1.147	1.079	1.059	1.043	1.031		
20X3	2.153	1.301	1.148	1.098	1.061	1.043			
20X4	2.215	1.318	1.162	1.100	1.060				
20X5	2.151	1.331	1.172	1.105					
20X6	2.157	1.350	1.179						
20X7	2.418	1.374							
20X8	2.671								

The row labels are accident years; the column captions are development intervals. The caption "2 to 3" means from two years of development to three years of development. We have rotated the triangle, turning the diagonals in Exhibit 3.4 into the columns in Exhibit 3.5.

No link ratio is calculated for the 20X9 accident year, since there is only one valuation. No link ratios are shown for the prior row, since the claims in this row stem from different accident years. For the prior years row, the time since inception of the accident year varies by claim.

We determine averages of the most recent three and the most recent five link ratios, and we select prospective factors from the historical figures and expectations about future conditions. In this illustration, the selected link ratios lie between the three and five year averages.[73]

[73] For the averaging procedures most suitable to reserving analyses, see Feldblum [2002: Schedule P] and Feldblum [2002: Stanard-Bühlmann reserving method].

Exhibit AppA.14: Paid Loss Development (dollars in thousands)

	1 to 2	2 to 3	3 to 4	4 to 5	5 to 6	6 to 7	7 to 8	8 to 9	9–10
Averages									
3 year	2.415	1.352	1.171	1.102	1.060	1.044	1.033		
5 year	2.322	1.335	1.162	1.094	1.059				
Select	2.350	1.340	1.170	1.100	1.060	1.040	1.030	1.030	1.020
Cumulative	4.835	2.057	1.535	1.312	1.193	1.125	1.082	1.051	1.020
Pd to Date	$156	$406	$485	$546	$542	$534	$434	$403	$454
Developed	$754	$835	$746	$716	$647	$601	$470	$423	$463
Ultimate	$830	$919	$819	$788	$711	$661	$517	$466	$509
Reserve	$674	$513	$334	$242	$169	$127	$83	$63	$55

PAID LOSS DEVELOPMENT FACTORS

The *cumulative* link ratios, or paid loss development factors, are the cumulative products of the appropriate link ratios (age-to-age factors) in adjacent columns. For instance, the cumulative link ratio from seven to ten years, or 1.082, is the product of 1.030, 1.030, and 1.020, which are the link ratios from seven to eight, eight to nine, and nine to ten years.

We incorporate a paid loss tail factor of +10%, which is not derived from the Schedule P data.

Illustration: The 1 year to 10 years cumulative paid loss development factor is 4.835. The 1 year to ultimate paid loss development factor is $4.835 \times 1.100 = 5.319$.

Exhibit AppA.15: Paid Loss Development Test of Reserve Adequacy

	1 yr	2 yrs	3 yrs	4 yrs	5 yrs	6 yrs	7 yrs	8 yrs	9 yrs
Pd LDF's	4.835	2.057	1.535	1.312	1.193	1.125	1.082	1.051	1.020
LDF w/ tail	5.319	2.263	1.689	1.443	1.312	1.238	1.190	1.156	1.122
Reciprocal	18.8%	44.2%	59.2%	69.3%	76.2%	80.8%	84.0%	86.5%	89.1%
Incr'tl Pd %	18.8%	25.4%	15.0%	10.1%	6.9%	4.6%	3.2%	2.5%	2.6%

The rows in the table are described below.

- The "Pd LDF's" are the paid loss development factors from each development date to 10 years of maturity, derived from Schedule P, Part 3, data. The paid loss development factor from 1 year to 10 years of maturity is 4.835.
- The "LDF w/ tail" is the paid loss development factors from each development date to ultimate, using a tail factor of +10%. The paid loss development factor from 1 year to ultimate is 5.319.
- The "Reciprocal" of the paid loss development factor to ultimate shows the percentage of losses paid by the development date. The cumulative losses paid by 1 year after the inception of the accident year is 1/5.319 = 18.8% of ultimate paid losses.
- The "Incr'tl Pd %" is the incremental paid losses during each development period as a percentage of ultimate paid losses. The losses paid between 1 year and 2 years after inception of the accident year are 44.2% – 18.8% = 25.4% of ultimate paid losses.

Loss Reserve Discounting

For GAAP financial statements, the deferred tax asset from loss reserve discounting is treated in the same fashion as the deferred tax asset from revenue offset. Both are fully recognized on the balance sheet.

ILLUSTRATION

In the other liability loss reserve discounting illustration in this paper, the accident year 2009 loss reserves for statutory and GAAP balance sheets on December 31, 2009 are $180,000 – $15,000 = $165,000. The corresponding discounted tax basis loss reserves are

$165,000 × 77.8022% = $128,373.63.

The difference between the GAAP loss reserves and the tax basis loss reserves is

$165,000.00 – $128,373.63 = $36,626.37.

The addition to taxable income stemming from loss reserve discounting for accident year 2009 at December 31, 2009 is $36,626.27 × 35% = $12,819.23. This is the deferred tax asset on the GAAP balance sheet.

The admitted portion of the deferred tax asset on the statutory balance sheet depends on the portion of the loss reserve that will still be unpaid in one year's time. This is an actuarial estimate; it is not the IRS provision used in the loss reserve discounting calculation. We may estimate this amount from Schedule P, Part 3, as discussed earlier.

Suppose the projected paid loss link ratios for other liability are 8.000 at 12 months of development and 5.000 at 24 months of development.

- At 12 months of development, 1/8.000 = 12.5% of incurred losses have been paid and 1 – 1/8.000 = 87.5% of incurred losses are still unpaid.
- At 24 months of development, 1/5.000 = 20.0% of incurred losses have been paid and 1 – 1/5.000 = 80.0% of incurred losses are still unpaid.

We expect 80.0% / 87.5% = 91.428571% of the December 31, 2009, accident year 2009 loss reserves to remain unpaid at December 31, 2010. This amount is \$165,000 × 91.4285714% = \$150,857.14. The expected IRS discounted reserves at December 31, 2010 equal this amount times the IRS loss reserve discount factor for accident year 2009 at 24 months of development, or 78.7611% in the other liability illustration:

\$150,857.14 × 78.7611% = \$118,816.75.

Implicit Discounting

Some companies implicitly discount reserves for long-tailed lines of business. Implicit discounting means that the company consciously holds less than full value loss reserves (for capital management purposes), not that the company mis-estimates the reserve indication.

One might be tempted to think that the amount of the implicit reserve discount should be taken into consideration when calculating the deferred tax asset. This is not correct. The deferred tax asset must be calculated as if the company held full value loss reserves.[74]

Illustration: An insurer expects to pay a loss for \$100,000 in three years. The IRS loss reserve discount factor for this line of business and accident year is 80% for the current valuation date and 85% for the valuation date 12 months hence.

- The (gross) deferred tax asset on the GAAP financial statements is 35% × \$100,000 × (1 – 80%) = \$7,000.
- The (net admitted) deferred tax asset on the statutory financial statements is 35% × \$100,000 × (85% – 80%) = \$1,750.

If the insurer implicitly discounts reserves at 5% per annum, its held reserves are \$100,000 / 1.05^3 = \$86,383.76, and its tax basis reserves are 80% × \$100,000 / 1.05^3 = \$69,107.01. Its expected held reserves one year hence are \$100,000 / 1.05^2 = \$90,702.95, and its expected tax basis reserves at that time are 85% × \$100,000 / 1.05^2 = \$77,097.51.

One might think that the gross (GAAP) and net admitted (statutory) deferred tax assets should be computed as follows:

- Gross (GAAP): 35% × (\$100,000 – \$69,107.01) = \$10,812.55.

---

[74] For more complete discussion of this topic, see Feldblum and Thandi, "Reserve Valuation Rates."

- Net admitted (statutory): 35% × ($77,097.51 – $69,107.01) = $2,796.68.

This is not correct. If the company shows a reserve of $86,383.76 on its statutory financial statements, it must treat that reserve at though it were a full value loss reserve for calculating the deferred tax asset. The appropriate calculations are as follows:

- The (gross) deferred tax asset on the GAAP financial statements is 35% × $86,383.76 × (1 – 80%) = $6,046.86.
- The (net admitted) deferred tax asset on the statutory financial statements is 35% × $86,383.76 × (85% – 80%) = $1,511.72.

FEDERAL INCOME TAXES

The previous sections of this appendix explain the federal income tax regulations that are most relevant for property-casualty insurance pricing. This section reviews the tax effects of the accounting transactions in the two illustrations in this paper.

Tables Tx.1 and Tx.2 show the components of the federal income tax liability and the deferred tax assets at each valuation date.

Table Tx.1, "Tax Decomposition," shows the two components of the current tax liability: the tax on underwriting income and the tax on investment income. Each component is divided into two parts: the part stemming from premium revenue and the part stemming from losses and expenses.

Tax Basis

The set of rows under the caption "tax basis" shows the tax basis earned premium, incurred expenses, and incurred losses. The tax basis earned premium is the statutory earned premium adjusted for the revenue offset provision. It is defined either as

written premium – 80% × the change in the unearned premium reserves

or as

statutory earned premium + 20% × the change in the unearned premium reserves

Illustration: For t=0, the written premium is $1000, the statutory earned premium is zero and the unearned premium reserves are $1000. The tax basis earned premium is computed either as $1000 – 80% × ($1000 – $0) = $200 or as $0 + 20% × ($1000 – $0) = $200.

For t=1, the written premium is zero, the statutory earned premium is $1000 and the unearned premium reserves are zero. The tax basis earned premium is computed either as $0 – 80% × ($0 – $1000) = $800 or as $1000 + 20% × ($0 – $1000) = $800.

The tax basis expenses are the same as the statutory expenses. The tax basis incurred loss is the statutory incurred loss adjusted for loss reserve discounting. It is defined either as

paid loss + the change in the discounted reserves

or as

statutory incurred loss – the change in the reserve discount

Illustration: The paid losses are $800 at time t=3.0 and zero before then. The assumed IRS discount factors are 86% at time t=1.0, 88% at time t=2.0, and 90% at time t=3.0. The statutory incurred loss is $800 at time t=1.0 and zero at other valuation dates.

- At time t=1.0, the tax basis incurred loss is computed either as \$0 + (\$800 × 86% – \$0) = \$688 or as \$800 – [\$800 × (1 – 86%) – \$0] = \$688.
- At time t=2.0, the tax basis incurred loss is computed either as \$0 + (\$800 × 88% – \$800 × 86%) = \$16 or as \$0 – [\$800 × (1 – 88%) – \$800 × (1 – 86%)] = \$16.
- At time t=3.0, the tax basis incurred loss is computed either as \$800 + (\$0 × 90% – \$800 × 88%) = \$96 or as \$0 – [\$0 × (1 – 90%) – \$800 × (1 – 88%)] = \$96.

Tax Basis Investment Income

The illustration in the text assumes that all investable assets are fully taxable. If the company holds municipal bonds or common stocks, the yield on the security is adjusted to a pre-tax equivalent yield; see Feldblum and Thandi, "Investment Yields."

The investment income at any valuation date equals investable assets at the previous valuation date times the benchmark investment yield. In the illustration, the benchmark investment yield is an 8% per annum bond equivalent yield, or 4% each half-year.

The investable assets at any valuation date equal the required assets at that date minus the non-cash assets. The only non-cash asset in this illustration is the deferred tax asset.

The tax on underwriting income uses the tax basis earned premium and incurred loss. The tax on investment income stemming from premiums and losses uses the statutory earned premium and incurred loss, with adjustments for the deferred tax asset admitted on the statutory balance sheet.

Illustration: At valuation date t=2.0, or December 31, 20XX+2, the statutory incurred loss is zero and the tax basis incurred loss is \$16.00 (see above). The investment income is based on the statutory loss reserves of \$800, the statutory capital requirements of \$120, and the deferred tax asset of ½ × (\$5.60 + \$33.60) = \$19.60 admitted on the statutory balance sheet. The investment income for the second half of 20XX+2 is

$$8\% \times \tfrac{1}{2} \times (\$800 + \$120 - \$19.60) = \$36.02.$$

TAX LIABILITY

The rows under the caption "Tax on" show the current tax liability by valuation date. The company is in the regular tax environment, with a 35% marginal tax rate. The current tax liability for each cell is 35% of the corresponding figures in the rows under the "Tax Basis" caption.

Illustration: The tax on "underwriting revenue minus expenses" for valuation date t=1.0 is

$$35\% \times (\$800 - \$150) = \$227.50.$$

The rows under the caption "Tax Due to" combine the underwriting income and the investment income for the premium and loss components of the tax. The tax on underwriting income is spread equally across the two halves of each year.

Illustration: The "tax due to writing of policy" on valuation date t=½ is computed as

$$(½ \times \$227.50) + \$16.52 = \$130.27.$$

DEFERRED TAX ASSETS

Table Tx.2 calculates the deferred tax assets on both a statutory and a GAAP basis. The statutory basis deferred tax assets are used for calculating the implied equity flows. The GAAP basis deferred tax assets are used for calculating the GAAP income recognition pattern in Feldblum and Thandi, "Income Recognition and Performance Measurement."

The deferred tax assets for the illustration are computed on December 31 of each year. The deferred tax assets at the June 30 valuation dates in the illustrations are computed by interpolation between the surrounding December 31 dates; see the text of this paper.

The rows under the caption "Tax Basis" show the tax basis premium, expenses, and losses and the federal income tax liability on each. The entries in this table are the same as the entries in Table Tx.1.

The rows under the caption "Statutory Basis" show the corresponding statutory basis premium, expenses, and losses.

- The tax basis premiums are adjusted for revenue offset; the statutory basis premiums are not.
- The expenses are the same for statutory and tax basis entries.
- The statutory incurred losses use full value loss reserves; the tax basis incurred losses use discounted loss reserves.

The rows under the caption "GAAP Basis" show the corresponding GAAP basis premiums, expense, and losses.

- The GAAP premiums equal the statutory premiums.
- The GAAP expenses are adjusted for deferred policy acquisition costs.
- The GAAP losses equal the statutory losses.

The exhibit assumes that all $250 of acquisition expenses are deferrable under GAAP. These expenses are capitalized at time t=0 and amortized over the policy term. Since the exhibit shows only the December 31 valuation dates, it shows a –$250 DPAC expense at time t=0 and a +$250 DPAC expenses at time t=1. The $250 DPAC expenses is spread evenly over the two halves of the year.

The rows under the caption "Statutory DTA Flow" show the deferred tax assets on the statutory balance sheet. The change in the *gross* deferred tax asset is the tax rate times the difference between taxable income and the income implied by the statutory balance sheet.

Illustration: At time t=0, the taxable income is +$200 from revenue offset and –$250 from incurred expenses, for a total of –$50. The income implied by the statutory balance sheet is $0 from premium and –$250 from incurred expenses. The difference is –$50 – (–$250) = +$200. The deferred tax asset is 35% × $200 = $70.

At time t=2, the taxable income from unwinding of the loss reserve discount is –$16.00. The income implied by the statutory balance sheet is zero. The difference is –$16.00 – $0 = –$16.00. The negative difference implies a reduction in the deferred tax asset of 35% × $16.00 = $5.60, from $39.20 at time t=1 to $33.60 at time t=2.

The net admitted deferred tax asset on the statutory balance sheet is the amount of the gross deferred tax asset that will reverse within 12 months.

Illustration: The entire $70 DTA at time t=0 stemming from revenue offset reverses over the policy term, so all $70 is admitted on the statutory balance sheet.

Illustration: The gross DTA stemming from IRS loss reserve discounting are $39.20 at time t=1 and $33.60 at time t=2. Since only $5.60 of this DTA reverses during year 2, only $5.60 is admitted on the statutory balance sheet at time t=1.

GAAP Deferred Tax Asset

The year to year change in the GAAP deferred tax asset is the tax rate times the difference between taxable income and the income implied by the GAAP balance sheet. There are no admissibility constraints for the GAAP DTA.

Illustration: At time t=0, the taxable income is –$50. Since no premium has been earned or expenses incurred on the GAAP balance sheet at time t=0, the GAAP income is zero. The difference is –$50 – $0 = –$50. The change in the GAAP deferred tax asset is 35% × –$50 = –$17.50. This is shown as a deferred tax liability on the GAAP balance sheet.

Exhibit 4

TAX DECOMPOSITION

	t = 0	t = 0.5	t = 1.0	t = 1.5	t = 2.0	t = 2.5	t = 3.0
Tax Basis							
(1) UW Revenue	200.00		800.00		0.00		0.00
(2) Expenses	250.00		150.00		0.00		0.00
(3) Inc Loss	0.00		688.00		16.00		96.00
(4) Inv Income							
(4a) on funds due to (Rev - Exp)	0.00	47.20	28.60	0.00	0.00	0.00	0.00
(4b) on funds due to Incurral of L	0.00	0.00	18.29	36.58	36.02	35.46	36.13
(4c) Total Inv Income	0.00	47.20	46.89	36.58	36.02	35.46	36.13
Tax on							
(5) U/W Revenue - Expenses	-17.50		227.50		0.00		0.00
(6) Inc Loss	0.00		-240.80		-5.60		-33.60
(7) Inv Inc							
(7a) on funds due to (Rev - Exp)	0.00	16.52	10.01	0.00	0.00	0.00	0.00
(7b) on funds due to Incurral of L	0.00	0.00	6.40	12.80	12.61	12.41	12.64
Tax Due to							
(8) Writing of Policy (semi-ann paym	-17.50	130.27	123.76	0.00	0.00	0.00	0.00
(9) Incurral of Loss (semi-ann payment)		-120.40	-114.00	10.00	9.81	-4.39	-4.16
(10) Total Tax (semi-ann payment)	-17.50	9.87	9.76	10.00	9.81	-4.39	-4.16

FORMULAS

(5) = 0.35 * [(1) - (2)]
(6) = -0.35 * (3)
(7a) = 0.35 * (4a)
(7b) = 0.35 * (4b)

(8) t = $(7a)_t + 0.5 * [(5)_{t+0.5}]$, for t = 0.5, 1.5, 2.5
(8) t = $(7a)_t + 0.5 * [(5)_t]$, for t = 0.0, 1.0, 2.0, 3.0

$(9)_t = (7b)_t + 0.5 * [(6)_{t+0.5}]$, for t = 0.5, 1.5, 2.5
$(9)_t = (7b)_t + 0.5 * [(6)_t]$, for t = 0.0, 1.0, 2.0, 3.0

(10) t = $(7a)_t + (7b)_t + 0.5 * [(5)_{t+0.5} + (6)_{t+0.5}]$, for t = 0.5, 1.5, 2.5
(10) t = $(7a)_t + (7b)_t + 0.5 * [(5)_t + (6)_t]$, for t = 0.0, 1.0, 2.0, 3.0

Exhibit 5

Deferred Tax Asset

	t = 0	t = 1.0	t = 2.0	t = 3.0
(1) **Tax Basis**				
a. Revenue	200.00	800.00	0.00	0.00
b. Expense	250.00	150.00	0.00	0.00
c. Incurred Loss	0.00	688.00	16.00	96.00
d. Tax due to Revenue	70.00	280.00	0.00	0.00
e. Tax due to Expense	-87.50	-52.50	0.00	0.00
f. Tax due to Losses	0.00	-240.80	-5.60	-33.60
g. Tax on U/W Total	-17.50	-13.30	-5.60	-33.60
(2) **Statutory Basis**				
a. Revenue	0.00	1,000.00	0.00	0.00
b. Expense	250.00	150.00	0.00	0.00
c. Incurred Loss	0.00	800.00	0.00	0.00
d. Tax due to Revenue	0.00	350.00	0.00	0.00
e. Tax due to Expense	-87.50	-52.50	0.00	0.00
f. Tax due to Losses	0.00	-280.00	0.00	0.00
g. Tax on U/W Total	-87.50	17.50	0.00	0.00
(3) **GAAP Basis**				
a. Revenue	0.00	1,000.00	0.00	0.00
b. Expense	250.00	150.00	0.00	0.00
b' DPAC	250.00	0.00	0.00	0.00
c. Incurred Loss	0.00	800.00	0.00	0.00
d. Tax due to Revenue	0.00	350.00	0.00	0.00
e. Tax due to (Expense - Δ DPA(	0.00	-140.00	0.00	0.00
f. Tax due to Losses	0.00	-280.00	0.00	0.00
g. Tax on U/W Total	0.00	-70.00	0.00	0.00
(4) **Statutory DTA Flow**				
a. due to Revenue Offset	70.00	-70.00	0	0
b. due to Loss Reserve Discount	0	39.20	-5.6	-33.6
c. DTA Flow w/out Reversal	70.00	-30.80	-5.60	-33.60
d. DTA w/out Reversal	70.00	39.20	33.60	0.00
e. DTA w/ Reversal	70.00	5.60	33.60	0.00
(5) **GAAP DTA Flow**				
a. due to Revenue Offset				
b. due to Loss Reserve Discounting				
c. DTA Flow	-17.50	56.70	-5.60	-33.60
d. DTA	-17.50	39.20	33.60	0.00

FORMULAE

$(4a)_t = (2d)_t - (1d)_t$

$(4b)_t = (2f)_t - (1f)_t$

$(4c) = (4a) + (4b)$

$(4d)_{t+1} = (4d)_t + (4c)_{t+1}$

$(4e)_0 = (4d)_0$

$(4e)_t = -(4b)_{t+1}$ for $t>0$

Statutory DTA w/ Reversal is that portion of the DTA that reverses in the year. Hence, it is by definition the negative of the DTA flow w/out reversal that occurs in the subsequent period.

Exhibit AppA.16: illustration A Loss Reserve Discount Factors

Accident Year (1)	Paid Loss + LAE (2)	Incurred Loss + LAE (3)	Cumulative Paid/Incurred Ratio (4)	Incremental Paid/Incurred Ratio (5)	Undiscounted Percentage Unpaid (6)	Discounted Percentage Unpaid (7)	Loss Reserve Discount Factor (8)
AY + 15							
AY + 14							
AY + 13							
AY + 12							
AY + 11							
AY + 10				2.00%	100.00%	0.00%	
2000	$270,000	$275,500	98.00%	3.07%	2.00%	1.93%	96.6735%
2001	$300,000	$316,000	94.94%	2.98%	5.06%	4.77%	94.1800%
2002	$320,000	$348,000	91.95%	3.99%	8.05%	7.34%	91.2271%
2003	$340,000	$386,500	87.97%	4.93%	12.03%	10.71%	89.0399%
2004	$350,000	$421,500	83.04%	6.03%	16.96%	14.78%	87.1281%
2005	$370,000	$480,500	77.00%	7.98%	23.00%	19.65%	85.4281%
2006	$380,000	$550,500	69.03%	10.01%	30.97%	26.07%	84.1740%
2007	$360,000	$610,000	59.02%	11.02%	40.98%	34.04%	83.0660%
2008	$330,000	$687,500	48.00%	13.00%	52.00%	42.47%	81.6659%
2009	$200,000	$571,500	35.00%	35.00%	65.00%	52.26%	80.3944%

Exhibit AppA.17: Illustration B Loss Reserve Discount Factors

Accident Year (1)	Paid Loss + LAE (2)	Incurred Loss + LAE (3)	Cumulative Paid/Incurred Ratio (4)	Incremental Paid/Incurred Ratio (5)	Undiscounted Percentage Unpaid (6)	Discounted Percentage Unpaid (7)	Loss Reserve Discount Factor (8)
AY + 15			100.00%	3.01%	0.00%	0.00%	
AY + 14			96.99%	1.38%	3.01%	2.91%	96.6736%
AY + 13			95.61%	1.38%	4.39%	4.05%	92.3385%
AY + 12			94.23%	1.38%	5.77%	5.12%	88.7803%
AY + 11			92.85%	1.38%	7.15%	6.12%	85.6177%
AY + 10	$235,000	$250,000	91.47%	1.38%	8.53%	7.06%	82.7122%
2000	$50,000	$55,500	90.09%	1.38%	9.91%	7.93%	79.9988%
2001	$55,000	$62,000	88.71%	3.00%	11.29%	8.74%	77.4439%
2002	$60,000	$70,000	85.71%	4.46%	14.29%	11.07%	77.4718%
2003	$65,000	$80,000	81.25%	8.33%	18.75%	14.66%	78.1822%
2004	$70,000	$96,000	72.92%	9.81%	27.08%	21.76%	80.3309%
2005	$65,000	$103,000	63.11%	10.93%	36.89%	29.82%	80.8185%
2006	$60,000	$115,000	52.17%	12.17%	47.83%	38.44%	80.3644%
2007	$50,000	$125,000	40.00%	15.00%	60.00%	47.69%	79.4828%
2008	$35,000	$140,000	25.00%	16.67%	75.00%	59.07%	78.7611%
2009	$15,000	$180,000	8.33%	8.33%	91.67%	71.32%	77.8022%

Appendix B: Workers' Compensation Pricing Exhibits

This appendix describes the implementation of an Excel© based version of the Equity Flow Pricing Model by tracing the steps required to price a fully insured workers' compensation policy (the details of which are described below). The model generates the equity flows associated with the policy being priced. The indicated premium is determined by setting the IRR on the equity flows equal to the target cost of capital. The solution is found by running the goal seek algorithm in Excel©.

To generated the implied equity flows, the model first calculates the company cash flows in the following three categories:

- U/W cash flows
- Investment income cash flows
- Federal income tax flows ("+" denotes a refund; "–" denotes a payment)

The required assets are prescribed by statutory reserve requirements and capital requirements. The asset flow is defined as the change in required assets. The inplied equity flow is calculated by the equation below:

Equity Flow = U/W Flow + Investment Income Flow + Tax Flow – Asset Flow

A positive equity flow denotes a distribution of earnings, or a flow of cash from the insurer to the equityholders, and a negative equity flow denotes a capital contribution, or a payment by the equityholders to the insurer.

Illustration

The illustrative workers' compensation policy is effective on July 1, 20XX. The model assumes effective dates at the inception of a quarter. The July 1 effective date serves as a proxy for a policy year 20XX book of business. The effective date affects the federal income tax calculations and the capital requirements.

The following policy costs serve as inputs to the model:

- The ultimate loss & ALAE,
- The acquisition expenses as a percentage of written premium,
- The general expenses as a percentage of written premium,
- The ULAE as a percentage of ultimate loss & ALAE,
- The policyholder dividends as a percentage of written premium,

The following collection/payment patterns are additional inputs to the model:

- The premium collection pattern,
- The loss & ALAE accident quarter payment pattern,
- The ULAE accident quarter payment pattern,
- The policyholder dividend payment pattern.

The model uses quarterly valuations and assumes all accounting and cash flow activity occur at quarter end.

The model requires the following parameter inputs:

- The benchmark investment yield on invested assets,
- The effective tax rate for both investments and U/W income, which is set at 35% (unless otherwise indicated),
- The surplus leverage ratios (which determine the capital requirements),
- The IRS loss & LAE reserve discount factors.
- The target return on capital.

The effective tax rate and the IRS loss reserve discount factors are either known figures or they are estimated by the pricing actuary. The target return on capital is a discretionary figure that is chosen by company management. The surplus leverage ratios and the benchmark investment yield are a mix of empirical data (such as actual investment yields, risk-based capital requirements, and rating agency capital formulas) and discretionary management choice.

Both the policy characteristics and the parameters are described in more detail in the section on pricing assumptions.

Exhibits

The accompanying exhibits show the quarterly valuation of the various items in the model. Only the first 10 years of valuations are shown. The model which produced these exhibits shows 50 years of quarterly valuations.

Exhibit 1 summarizes the pricing assumptions and shows the pricing results.

Exhibit 2 shows the assumed cash flow patterns (pattern of loss payment, premium collection, policyholder dividend payment, IRS loss reserve discount factors, and tabular discount factors applicable to pension indemnity cases).

Exhibit 3 shows the cash flows and balance sheet items needed to generate the implied equity flow associated with the workers' compensation policy.

Exhibit 4 shows the determination of paid losses and loss adjustment expenses.

Exhibit 5 summarizes the held assets and resulting investment income.

Exhibit 6 shows the Federal Income Tax calculation.

Exhibit 7 summarizes the relevant cash flows and calculates the equity flow.

The exhibits use two conventions for expressing time. The first expresses time in absolute terms: time 0 stands for Jan. 1, 20XX and time 3.50 stands for July 1, 20XX+3. The second convention marks time relative to the policy inception date.

In exhibits 3 through 7 the time column represents absolute time. In exhibit 2 the payment and collection patterns are expressed relative to the age of the policy. The time column for the cash flow patterns is labeled "Age". An age of zero refers to the policy inception date of July 1, 20XX. An age of 1.25 refers to April 1, 20XX+1, which is 1¼ years after policy inception. The exception to this dating convention in exhibit 2 is the IRS discount factors. The discount factors are applicable to specific calendar periods and the time column for these factors represents absolute time (i.e. calendar date). See Appendix A for explanation of the IRS loss reserve discount factors.

Pricing Assumptions

Policy Characteristics

The policy is a fully insured workers' compensation policy effective on July 1.

Premium is collected in the pattern specified in Exhibit 2. The indicated premium, denoted as WP (written premium), is determined by setting the internal rate of return (IRR) of the implied equity flows equal to the target return on capital (TROC) of 12%. The premium of $1,374 shown in all the exhibits is determined by the goal seek algorithm in Excel©. The dollars of premium collected shown in column (4) of Exhibit 3 are calculated as WP x Premium Collection Pattern.

Illustration: From exhibit 2, 18% of the charged premium is collected at policy inception (time 0.5), 17.1% is collected one quarter later (time 0.75), 21.8% is collected two quarters later (time 1.0), 25.3% is collected three quarters later (time 1.25), 7.2% is collected four quarters later (time 1.5), 6.9% is collected five quarters later (time 1.75), etc.

Since the written premium is $1,374, the collection pattern implies that the dollars of premium collected (column (4) exhibit 3) are equal to

1,374 x 18.0% = $247	at time 0.50 (0 years after policy inception)
1,374 x 17.1% = $235	at time 0.75
1,374 x 21.8% = $300	at time 1.00
1,374 x 25.3% = $348	at time 1.25

1,374 × 7.2% = $ 99	at time 1.50
1,374 × 6.9% = $ 95	at time 1.75

The Premium Receivable (column (5) of Exhibit 3) is defined to be
Written Premium – Cumulative Premium Collected.

Illustration: The premium receivable is

$1,374 – $247	= $1,127	at time 0.50
$1,374 – ($247+$235)	= $ 892	at time 0.75
$1,374 – ($247+$235+$300)	= $ 892	at time 1.0

Ultimate Loss & ALAE is $1000, with the Accident Quarter (AQR) loss payment pattern as shown in Exhibit 2. The determination of the dollars of loss payments, given the AQR pattern, as well as the calculation of loss reserves, is described below.

ULAE is 7.2% of Ultimate Loss & ALAE, with a payment pattern as shown in Exhibit 2. The determination of the ULAE payments and reserves is discussed below.

Acquisition expenses are 17.9% of written premium and are paid at policy inception. For the illustration, the acquisition expenses are $1,374 × 17.9% = $246 as shown in column (6) of Exhibit 3.

General expenses are 7.7% of written premium and are assumed to be paid at policy inception. In the illustration, general expenses are $1,374 × 7.7% = $105. Some general expenses actually occur both before and after policy inception. For this illustration we make the simplifying assumption that we can approximate this payment pattern with a single payment at policy inception.

Other Assumptions

We assume an effective annual rate of return on investments of 8% (or $(1+0.08)^{0.25} - 1$ = 1.94277% per quarter). Investment income is earned on all investable assets, which exclude, in particular, the premium receivable asset.

The Federal Income Tax Rate of 35% applies to both underwriting income and investment income.

Surplus is allocated as 43.7% of written premium and is held only for the policy term. Surplus could also be allocated in proportion to loss reserves and held until all losses are paid, if this deemed a more appropriate allocation method.

Loss & ALAE Payments

Because the model uses quarterly cash flows, it requires a quarterly valuation of accident quarter losses. If we needed to determine just the payment of losses on a single policy we could use the loss payment pattern for the accident year as a whole. This would reproduce the payment pattern for an individual policy. It would not suffice modeling the equity flows in the model, since paid and unpaid losses segregated by accident year are needed to estimate the federal income tax liabilities.

The tax basis discounted loss reserves are the product of the IRS loss reserve discount factor and the held loss & LAE reserves. The discount factor varies by line of business, accident year, and age of the accident year.[1] The appropriate discount factor depends on the year in which the losses occurred. A segregation of losses into the years in which they occur is necessary to determine taxable incurred losses. A description of the segregation procedure follows below.

The policy term spans two accident years, which we term AYR1 and AYR2. For a policy with an effective date during year 20XX, AYR1 extends from January 1, 20XX to December 31, 20XX, and AYR2 extends from January 1, 20XX+1 to December 31, 20XX+1. The portion of the policy term that spans each accident year is given in the table below:

Policy Effective Date	AYR1	AYR2
Jan 1, 20XX	Jan 1, 20XX --> Dec. 31, 20XX	-----
Apr 1, 20XX	Apr 1, 20XX --> Dec. 31, 20XX	Jan 1, 20XX+1 --> May 31, 20XX+1
July 1, 20XX	July 1, 20XX --> Dec. 31, 20XX	Jan 1, 20XX+1 --> June 30, 20XX+1
Oct 1, 20XX	Oct 1, 20XX --> Dec. 31, 20XX	Jan 1, 20XX+1 --> Sept 30, 20XX+1

We assume that losses are incurred evenly over the policy term. We use accident quarter loss and LAE payment patterns to segregate losses into AYR1 and AYR2. The illustration uses an annual workers' compensation policy with an effective date at the beginning of a calendar quarter, so the policy term spans four accident quarters. We assume that one quarter of ultimate losses are incurred each quarter, and that the loss & ALAE payment pattern is the same for each accident quarter.

Given the paid losses by accident quarter, we calculate the accident year losses. The terms accident year 1 and accident year 2 refer to accident years measured from January 1 to December 31. The term AQR1 (or accident quarter 1) refers to the first accident quarter in a given policy.

AYR1 (or AYR2) paid losses consist of the sum of the losses paid in each accident quarter that falls into AYR1 (or AYR2). The classification of accident quarter losses into accident years depends on the policy effective date. The chart below shows the classification of accident quarter losses into accident years:

Policy Effective Date	AYR1	AYR2

[1] The age of the accident year is the valuation date and the inception of the accident year. Accident year 20XX valued at December 31, 20XX+2, is accident year 20XX aged 36 months.

Jan 1, 20XX	(AQR1 + AQR2 + AQR3 + AQR4)	-----
April 1, 20XX	(AQR1 + AQR2 + AQR3)	(AQR4)
July 1, 20XX	(AQR1 + AQR2)	(AQR3 + AQR4)
Oct 1, 20XX	(AQR1)	(AQR2 + AQR3 + AQR4)

Illustration: A portion of the accident quarter loss & ALAE pattern is reproduced below. The pattern is from inception of the policy. For example, accident quarter 3 means the third quarter of the policy term, and accident quarter 5 means the first quarter after expiration of the policy.

Time (from Policy Inception)	AQR Payment Pattern
0.00	0.0000
0.25	0.0480
0.50	0.1210
0.75	0.0819
1.00	0.0622
1.25	0.0543

This pattern and the assumption that $250 (¼ of ultimate losses of $1,000) are incurred in each accident quarter implies the following payment pattern by accident year for a policy inception date of July 1, 20XX:

Policy Effective July 1, 20XX

	AYR1		AYR2		→ AYR1	AYR2
Time	AQR1	AQR2	AQR3	AQR4		
0.00	-	-	-	-		
0.25	-	-	-	-		
0.50	250 * 0.0000 = 0	-	-	-	0 + 0 = 0	
0.75	250* 0.0480 = 12	250 * 0.0000 = 0	-	-	12 + 0 = 12	
1.00	250* 0.1210 = 30	250* 0.0480 = 12	250 * 0.0000 = 0	-	30 + 12 = 42	0 + 0 = 0
1.25	250* 0.0819 = 20	250* 0.1210 = 30	250* 0.0480 = 12	250 * 0.0000 = 0	20 + 30 = 51	12 + 0 = 12
1.50	250* 0.0622 = 16	250* 0.0819 = 20	250* 0.1210 = 30	250* 0.0480 = 12	16 + 20 = 36	30 + 12 = 42
1.75	250* 0.0543 = 13	250* 0.0622 = 16	250* 0.0819 = 20	250* 0.1210 = 30	13 + 16 = 29	20 + 30 = 51

(time 0 refers to Jan 1, 20XX)

Figure 4

A policy with a September 1, 20XX inception date would have the following loss & ALAE payment by accident year:

Policy Effective Sept 1, 20XX

AYR1	AYR 2	→ AYR1	AYR2

Time	AQR1	AQR2	AQR3	AQR4		
0.00	-	-	-	-		
0.25	-	-	-	-		
0.50	-	-	-	-		
0.75	250 * 0.0000 = 0	-	-	-	0	
1.00	250* 0.0480 = 12	250 * 0.0000 = 0	-	-	12	0 + 0 + 0 = 0
1.25	250* 0.1210 = 30	250* 0.0480 = 12	250 * 0.0000 = 0	-	30	12 + 0 + 0 = 12
1.50	250* 0.0819 = 20	250* 0.1210 = 30	250* 0.0480 = 12	250 * 0.0000 = 0	20	30 + 12 + 0 = 42
1.75	250* 0.0622 = 16	250* 0.0819 = 20	250* 0.1210 = 30	250* 0.0480 = 12	16	20 + 30 + 12= 63

(time 0 refers to Jan 1, 20XX) **Figure 5**

Loss & ALAE Reserves (Nominal)

Nominal reserves are statutorily mandated reserves in the absence of discounts. Held reserves and tabular discount are defined in terms of Nominal Reserves.

Once the policy is fully earned the nominal reserves of each accident year are the sum of the unpaid losses of the component accident quarters.

Illustration: For a policy effective at time 0.5 (July 1, 20XX), the AYR1 nominal reserve at time 1.75 (Sept. 1, 20XX+1) equal the unpaid losses on AYR1. The unpaid losses are the AYR1 ultimate losses minus the AYR1 losses paid to date.

- The AYR1 ultimate loss equals ½ x $1,000 = $500.
- The cumulative losses paid by time 1.75 is $(12+42+51+36+29) = $170; see column (5) of exhibit 4 or Figure 4 above.
- The AYR1 unpaid loss at time 1.75 is $500 – $170 = $330; see column (15) of exhibt 4.

- The AYR2 nominal reserve at time 1.75 is the AYR2 ultimate loss minus the AYR2 losses paid to date.
- The AYR2 ultimate loss is ½ x $1,000 = $500.
- The cumulative losses paid by time 1.75 are $(12+42+51) = $105; see column (5) of exhibit 4 or Figure 4 above.
- The AYR2 unpaid loss at time 1.75 is $500 - $105 = $395; see column (15) of exhibt 4.

If a policy is not fully earned at a given valuation date, the recognition of nominal reserves by accident year depends on the policy effective date and the earning of the policy.

Illustration: If the policy effective date is time 0, all losses attributable to AQR1 have occurred by time 0.25. AQR1 losses represent ¼ of all losses attributable to AYR1. The AYR1 nominal reserves at time 0.25 are equal to ¼ of ultimate AYR 1 Losses minus AYR1 Losses Paid to Date.

If the policy effective date is time 0.5, then at time 0.75, all losses attributable to AQR1 have occurred . AQR1 losses represent ½ of all losses attributable to AYR1 → ½ of all losses attributable to AYR1 have occurred. Thus the AYR1 nominal reserves at time 0.75 is equal to ½ of ultimate AYR 1 Losses minus AYR1 Losses Paid to Date).

The AYR1 ultimate losses that are recognized at time T equal the AYR1 ultimate losses times a factor that represents the proportion of the total AYR1 losses that have occurred, less the cumulative amount of losses that have been paid to date. Figure 1 below presents a schematic of the determination of the AYR1 factor. AYR2 is handled in the same fashion. Figure 2 presents the determination of the corresponding AYR2 reserve recognition pattern.

The total amount of losses attributable to AYR1 is equal to the total ultimate losses times the proportion of the policy term that falls in AYR1. The amount of losses attributable to AYR2 is equal to the total ultimate losses times the proportion of the policy terms that falls in AYR2.

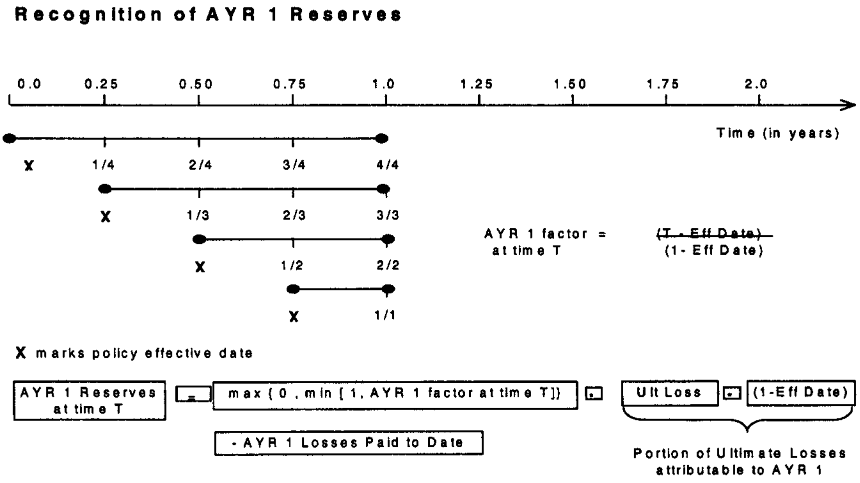


Figure 6

Illustration: We determine the AYR1 nominal reserve valued at time 0.75 for a policy effective at time 0.5.
The AYR1 ultimate loss is ½ × $1,000 = $500.
The AYR1 earnings factor at time 0.75 is the portion of AYR1 losses that have occurred by time 0.75, which equals

$$\frac{0.75-0.5}{1-0.5}=\frac{0.25}{0.5}=0.5$$

The cumulative losses paid by time 0.75 is \$0 + \$12 = \$12; see column (5) of exhibit 4 or Figure 4 above.
The AYR1 nominal reserve at time 0.75 is equal to

AYR1 factor × AYR1 ultimate loss – AYR1 paid to date
= 0.5 × (\$500) – \$12 = \$238

(See column (15) of exhibit 4).

The more complicated formula for the AYR1 reserves at time T, as shown in Figure 6, applies whether or not the policy is fully earned.

Recognition of AYR 2 Reserves

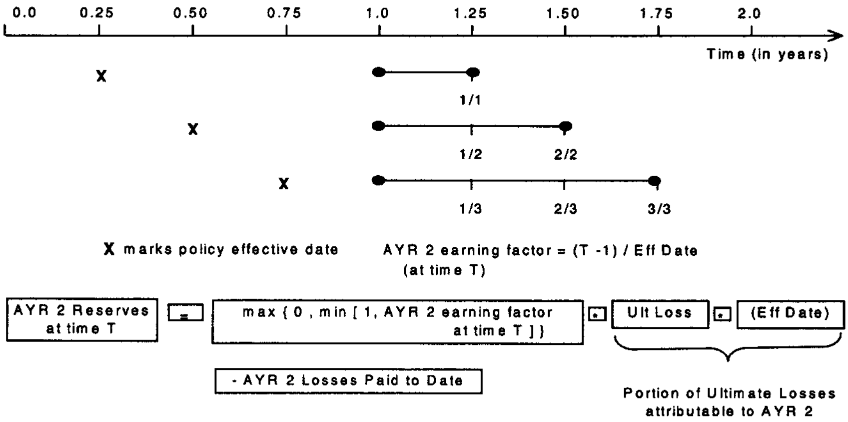


Figure 7

Illustration: We determine the AYR2 nominal reserve valued at time 1.5 for a policy effective at time 0.5.
The AYR2 ultimate loss is ½ × \$1,000 = \$500.
The AYR2 earnings factor at time 1.5 is the portion of AYR2 losses that have occurred by time 1.25, or

$$\frac{1.5-1}{0.5}=\frac{0.5}{0.5}=1.0$$

The cumulative losses paid by time 1.5 are \$0+\$12+\$42 = \$54; see column (6) of exhibit 4 or Figure 4 above.
The AYR2 nominal reserve at time 1.5 is equal to

AYR2 factor × AYR2 ultimate loss – AYR2 paid to date
= 1.0 × (\$500) – \$54 = \$446

(See column (16) of exhibit 4).

Nominal ULAE Reserves

The consideration of nominal ULAE reserves is analogous to nominal loss & ALAE reserves. Instead of using accident quarter loss & ALAE payment patterns we use accident quarter ULAE payment patterns. The ultimate ULAE is estimated as ultimate loss & ALAE × the ULAE ratio. With these modifications, the exposition in the preceding section applies to ULAE as well.

Tabular Discount

Statutory accounting rules allow insurers to discount WC pension indemnity cases. The tabular discounts by line of business and by accident year are disclosed in the notes to the financial statements.[2] To evaluate the tax basis reserves, we need the dollar amount of tabular discount on a given accident year's loss reserves. This dollar amount will vary by age of the accident year.

The dollar amount of tabular discount is calculated by applying the ratio

$$\frac{\text{tabular discount}}{\text{nominal reserves}}$$

to the nominal reserves calculated by the model. This ratio can be based on an analysis of

$$\frac{\text{actual tabular discount}}{\text{actual nominal reserves}}$$

by accident year, using company data.

[2] The tabular discount may also be derived from a comparison of Schedule P, Part 1, which is net of tabular discount, with Schedule P, Part 2, which is gross of tabular discount; see Feldblum [2002: Schedule P].

Held Loss & ALAE Reserves

Held reserves are the sum of nominal loss & ALAE reserves and nominal ULAE reserves, times the level of reserve adequacy, minus the amount of tabular discount:[3]

$$\text{AYR X Held}_{\text{Loss + LAE}} = \text{Res Adeq} * (\text{AYR X Nominal Loss \& ALAE} + \text{AYR X Nominal ULAE}) - \text{Tabular Discount}$$

Illustration: For our WC policy we assume that the level of reserve adequacy is 100%. We confirm that the AYR1 held reserve is $246 at time 0.75 (as shown in column (25) of Exhibit 4). From column (15) the nominal loss & ALAE nominal reserve at t=0.75 is $238; the nominal ULAE reserve from column (18) is $18; from column (22) the tabular discount is $10. Thus the AYR1 Held Reserves = 1.0 * ($238 + $18) – $10 = $246.
The AYR1 held reserve at time 1.75 is $337: the nominal loss & ALAE reserve is $330; the nominal ULAE reserve is $26; the tabular discount is $20.
The AYR2 held reserve at time 1.75 is $411: the nominal loss & ALAE reserve is $395; the nominal ULAE reserve is $32; the tabular discount is $16.

Assets

Required Surplus

Surplus is held only for the policy term in our illustration. Surplus is held to cover unforeseen contingencies and to maintain an acceptable level of risk.

Most pricing model use either a premium to surplus leverage ratio or a reserves to surplus leverage ratio or both. The pricing model described here supports not just prospective pricing but also an economic value added performance measurement system. Using a premium to surplus leverage ratio for the surplus assumptions allows a more responsive performance measurement system, which is more likely to be accepted by company personnel. See Kelly [2002] for further discussion.

Illustration: The premium leverage ratio is 43.7%, and he written premium is $1,374. The required surplus for the full policy term (time 0.5 until time 1.5) is 43.7% x $1374 = $601. For time 1.5 and subsequent (once the policy is fully earned) the required surplus is zero.

Total Reserve

The total reserve is the sum of the unearned premium reserve and the held loss & LAE reserves.

[3] The level of reserve adequacy is discussed in Feldblum and Thandi [2002], "Reserve Valuation Rates."

Illustration: At time 0.5 (policy inception) the UEPR is equal to the written premium which is equal to $1,374. Held loss & LAE reserves are equal to 0 → Total Reserves = $1,374
At time 0.75 the UEPR = ¾ * (1,374) = $1,031. The held loss & LAE reserves = $246 → Total Reserves = $1,031 + $246 = $1,277.
At time 1.75 the UEPR = $0. The held loss & LAE reserves = $748 → Total Reserves = $0 + $748 = $748

Required Assets

The assets needed to support the policy (the Required Assets in column (1) of Exhibit 5) equal the total reserves (sum of columns (2) and (3) of Exhibit 5) plus the required surplus (column (4) of Exhibit 5):

$$\text{Required Assets} = \text{Total Reserves} + \text{Required Surplus}$$

Illustration: At time 0.5, Required Assets = $601 + $1,374= $1,975
At time 0.75, Required Assets = $601 + $1,277= $1,878
At time 1.5, Required Assets = $0 + $833 = $ 833
At time 1.75, Required Assets = $0 + $748 = $ 748

Income Producing Assets

Not all of the assets held by the company to support the policy generate investment income. The premium receivable (column (5) of Exhibit 3) and the deferred tax asset (column (34) of Exhibit 3) are non-income producing assets:

$$\text{Income Producing Assets} = \text{Required Assets} - \text{Premium Receivable} - \text{DTA}$$

Illustration: At time 0.5 required assets = $1,975, premium receivable = $1,127, and the DTA = $96. Thus Income Producing Assets (column (18) Exhibit 3) = $1,975-$1,127 – $96 = $752.
At time 0.75 the required assets = $1,877, premium receivable = $892, and the DTA = $96. Thus Income Producing Assets = $1,975-$1,127 – $96 = $752.
At time 1.5 the required assets = $833, premium receivable = $145, and the DTA = $9. Thus Income Producing Assets = $833-$145– $9 = $679.
At time 1.75 the required assets = $748, premium receivable = $50, and the DTA = $96. Thus Income Producing Assets (column (18) Exhibit 3) = $1,975-$1,127 – $96 = $752.

Investment Income

The Investment Income earned over a quarter is the product of the quarterly effective investment rate of return times the amount of income producing assets held at the beginning of the quarter:

$$\text{Invest Inc}_{@\,\text{time T}} = \text{Qtlry Invest ROR} * \text{Investible Assets}_{@\text{time T-1}}$$

Illustration: Invest Income at time 0.75 = 1.9427% * (Investable Assets at time 0.5) = 1.9427% * $752 = $15 (Investable Assets shown in column (18) of Exhibit 3, Investment Income shown in column (20) of Exhibit 3).
Invest Income at time 1.75 = 1.9427% * (Investable Assets at time 1.5) = 1.9427% * $679 = $13.

Taxes

IRS Discounted Reserves

The tax basis (discounted) reserves are the produce of the held reserves, gross of any tabular discount, and the IRS loss reserve discount factor. The discount factor varies by line of business, by accident year, and by age of the accident year.

The calculation of the loss reserve discount factors is described in Appendix A, along with explanation of the estimation procedures required of the pricing actuary. For this appendix, we take the IRS discount factors as given. We use the accident year 2000 discount factors as the factors for "AYR1" and the accident year 2001 factors for AYR2. The formula for IRS discounted reserves is

$$\text{IRS Discounted Reserves} = \text{IRS Discount Factor} * (\text{Held Reserves} + \text{Tabular Discount})$$

Illustration: For the accounting year ending at time 1.0, the AYR1 Held Reserves are $460, the AYR1 tabular discount is $20, and the IRS discount factor is 0.8194 (from columns (25), (22) and (28), respectively, of Exhibit 4). Thus the IRS Discounted Loss & LAE reserves at time 1.0 is equal to 0.8194 * ($460 + $20) = $393.

For the accounting year ending at time 2.0, the AYR1 Held Reserves are $308, the AYR1 tabular discount is $19, and the IRS discount factor is 0.8027 (from columns (25), (22) and (28), respectively, of Exhibit 4). Thus the IRS Discounted Loss & LAE reserves at time 2.0 is equal to 0.8027 * ($308 + $19) = $263.
For AYR2 at time 2.0 the Held reserves are $372, the tabular discount is $16, and the discount factor is 0.8214 (columns (26), (23), and (29) respectively of Exhibit 4). Thus the AYR2 IRS Discounted Reserve is 0.8214*(372+16)=$319.
Thus the total IRS Discounted Reserve at time 2.0 is $263 + $319 = $582.

Taxable U/W Income

The taxable U/W income over an accounting year is

$$\text{Written Premium} - 0.8 * \Delta\text{UEPR} - \text{Paid Expenses} - [\text{ Paid Losses} + \Delta\text{IRS Disc Reserves}]$$

where all activity is over the relevant accounting year.

Illustration: For the accounting year ending at time 1.0 the written premium is $1374 (column (1) Exhibit 3), the change in the UEPR is $687 (column (2) Exhibit 3), the expenses paid are $351 (column (9) Exhibit 3), the paid loss & ALAE is $(42+12)=$64 (column (10) Exhibit 3), the paid ULAE is $1 (column (11) Exhibit 3), and the change in the IRS discounted reserves is $393-$0 =$393 (column (16) Exhibit 3). Thus the Taxable U/W income is 1374-0.8*(687)-351-(42+12+1 +393) = $24.
For the accounting year ending at time 2.0 the written premium is $0, the change in the UEPR is -$687, the expenses paid are $12, the paid loss & ALAE is $(63+78+80+62)=$282, the paid ULAE is $(2+4+5+6)= $16, and the change in the IRS discounted reserves is $582-$393= $289. Thus the Taxable U/W income is 0-0.8*(-687)-12-(282+17 +289) = $49.

The paragraph above determines the annual federal income taxes on U/W income. In practice, taxes are paid quarterly. The taxpayer (the insurance company) projects its annual U/W accounting income and pays one quarter of that amount in each calendar quarter. We estimate the taxes for the two accident years for a given policy, and we spread the tax payments over the quarters in which the policy is effective.

Illustration: U/W Tax at t=0.75 = ½ × (24) × (35%) = $4.20
U/W Tax at t=1.0 = ½ × (24) × (35%) = $4.20
U/W Tax at t=1.25 = ½ × (49) × (35%) = $8.58
U/W Tax at t=1.50 = ½ × (49) × (35%) = $8.58
(See columns (30) and (31) of Exhibit 3.)

Tax on Investment Income

The tax on investment income is paid quarterly as investment income is earned.

Illustration: Qtrly Tax on Inv Inc at t = 0.75 = Qtrly Inv Inc at t = 0.75 * 35%
= 15 * 35% = $5

Qtrly Tax on Inv Inc at t = 1.75 = Qtrly Inv Inc at t = 1.75 * 35%
= 13 * 35% = $5

Column (36) of Exhibit 3.

Total Tax

The total federal income tax paid each quarter is equal to the sum of the quarterly tax on U/W income and the quarterly tax on investment income.

Illustration: At t=0.75 the total FIT = $4 + $5 = $9
At t=1.75 the total FIT = $4 + $5 = $9

See column (37) of exhibit 3.

Deferred Tax Asset

The calculation of the deferred tax asset is described in Appendix A. We trace the calculation of the deferred tax asset for the workers' compensation policy.

There are two components to the DTA: the portion due to the Revenue Offset; and the portion due to IRS Discounting of Loss&LAE Reserves.

The DTA due to the Revenue Offset is equal to

$$35\% * 20\% * \Delta UEPR$$

Illustration: The DTA due to Revenue Offset at t=0.5 is equal to 35% * 20% * $1,374 = $96.
The DTA due to Revenue Offset at t=0.75 = 35% * 20% * ($1,031) = $72
The DTA due to Revenue Offset at t=1.75 = 35% * 20% * ($0) = $0.

(See column (32) of exhibit 3).

The DTA due to IRS Discounting at the end of Accounting Year X is equal to

35% * [(AYR1 Held Loss Reserve $_{\text{at time X}}$ – AYR1 IRS Loss Reserve $_{\text{at time X}}$) -
(AYR1 Held Loss Reserve $_{\text{at time X+1}}$ – AYR1 IRS Loss Reserve $_{\text{at time X+1}}$)]
+ 35% * [(AYR2 Held Loss Reserve $_{\text{at time X}}$ – AYR2 IRS Loss Reserve $_{\text{at time X}}$) -
(AYR2 Held Loss Reserve $_{\text{at time X+1}}$ – AYR2 IRS Loss Reserve $_{\text{at time X+1}}$)]

The amount in each square bracket is the amount that reverses over the next twelve months, following statutory accounting rules; see Appendix A.

Illustration: At time 2.0 the Held Loss&LAE Reserve for AYR1 is $308, for AYR2 is $372 (columns (25), (26) of Exhibit 4); the IRS Loss Reserves for AYR1 are $263, for AYR2 are $319 (columns (30), (31) of Exhibit 4). At time 3.0 the Held Loss&LAE Reserve for AYR1 is $223, for AYR2 is $262; the IRS Loss Reserves for AYR1 are $191, for AYR2 are $225.Thus the DTA due to IRS Discounting at time 2.0 is

35% * [(308-263) – (223-191)] + 35% * [(372-319)-(262-225)] = $10

The DTA due to IRS Discounting for any time other than and Accounting Year end is an interpolation year end DTAs (due to IRS discounting).

Illustration: At time 1.0 the DTA due to IRS Discounting is $7, at time 2.0 the DTA due to IRS Discounting is $10, at time 3.0 it is $5. Hence the DTA due to IRS Discounting is $4 at time 0.75 (interpolation between $0 and $7)

$10 at time 1.75 (interpolation between $7 and $10) (rounded to nearest whole dollar). See column (33) of exhibit 3.

Cash Flows

The relevant cash flows for determining the Equity Flow are described below.

U/W Cash Flow

The underwriting cash flow is defined as

U/W Cash Flow = WP – Paid Expenses – Paid Loss & LAE
Column (38) exhibit 3 = column (1) – column (9) – [column (10) + column (11)]

Illustration: At t=0.5 U/W CF = 1374 – 351 – 0 = $1023
At t=0.75 U/W CF = 0 – 0 – 12 = -$12
At t=1.75 U/W CF = 0 – 12 – 85 = -$97
(See column (38) Exhibit 3).

Investment Income Flow

The investment income cash flow is defined at the quarterly investment income earned each quarter. The calculation is described above

Illustration: At t=0.5 the investment income is $0
At t=0.75 the investment income is $15.
At t=1.75 the investment income is $13.
(See column (39) of Exhibit 3).

Tax Flow

The Tax Cash Flow is defined at the negative (to denote a flow from the company) of the federal income taxes paid that quarter. The calculation of this flow item is described above.

Illustration: At time t=0.5 FIT = 0 → Tax Flow = 0
At time 0.75 FIT = $9 → Tax Flow = -$9
At time 1.75 FIT = $9 → Tax Flow = -$9

(See column (40) Exhibit 3).

DTA Flow

The DTA Flow is defined as the change in the DTA asset over a calendar quarter.

Illustration: At t=0.5, DTA Flow = $DTA_{t=0.5}$ = $96
At t=0.75, DTA Flow = $DTA_{t=0.75} - DTA_{t=0.5}$ = $76 - $96 = -$20
At t=1.75, DTA Flow = $DTA_{t=1.75} - DTA_{t=1.5}$ = $10 - $9 = $1
(See column (41) Exhibit 3)

Asset Flow

The asset flow is defined as the change in the required assets. The composition and calculation of the required assets are described above.

Illustration: At t=0.5 Asset Flow = $Assets_{t=0.5}$ = $1975
At t=0.75 Asset Flow = $Assets_{t=0.75} - Assets_{t=0.5}$ = $1877 - $1975 = -$98
At t=1.75 Asset Flow = $Assets_{t=1.75} - Assets_{t=1.5}$ = $748 - $833 = -$85
(See column (42) of Exhbit 3).

Equity Flow

To compute the Equity Flow at each quarter we use the cash flow definition:

Equity Flow = - Asset Flow + U/W Flow + Investment Income Flow + FIT Flow + DTA Flow

Recall that we use the convention that a positive equity flow denotes a flow of cash from the insurer to the equityholders, and a negative a payment by the equityholders to the insurer.

Illustration: At t=0.5 Equity Flow = -$1,975 + $1,023 + $0 + $96 = -$856
At t=0.75 Equity Flow = -(-$98) – $12 + $15 – $9 – $20 = $70
At t=1.75 Equity Flow = -(-$85) – $97 + $13 – $9 + $1 = -$7.

SUMMARY OF ASSUMPTIONS AND RESULTS FOR WC Fully Insured Policy

I. UNDERWRITING ASSUMPTIONS

A) Policy Costs	
Expense Ratio (as % WP)	25.6%
Dividend Ratio (as % WP)	5.7%
ULAE Ratio (as % of Loss&ALAE)	7.2%
Ultimate Loss & ALAE	1,000
B) Cash Flow Patterns	
Disc Loss&ALAE to Undisc	73.0%
Duration of Losses (in yrs)	4.3
Disc Premium to Undisc	95.3%
C) Average Effective Date	0.5
D) Level of Reserve Adequacy	
Held to Nominal Reserves	100.0%

II. FINANCE ASSUMPTIONS

A) Investment Rate of Return	
On all Investable Assets	8.0%
B) Federal Income Taxes	
Tax Rate on U/W Income	35.0%
Tax on Investment Income	35.0%
C) Target Return on Capital	
Post-Tax Return	12.0%

III. RISK (SURPLUS) ASSUMPTIONS

Reserve Leverage Ratio	0.0%
Premium Leverage Ratio	43.7%

IV. PRICING RESULTS

A) Premium		
Nominal Premium	1,374	
Discounted Premium	1,309	
B) Summary of Costs		
Disc Loss & LAE	784	
Disc Expense (incl PHR Dividends)	416	
Disc Taxes	67	
	1,267	
	Nominal	Discounted
C) Ratios	(% of Nominal Prem)	(% of Disc Prem)
Loss & ALAE Ratio	72.8%	55.8%
ULAE Ratio	5.2%	4.1%
Expense Ratio (incl PHR Dividends	31.3%	31.8%
Combined Ratio	109.2%	91.6%

V. PROFITABILITY

A) Equity Charge	Nominal	Discounted
	107.48	42.56
B) IRR on Equity Flows	12.0%	

Exhibit 2

AGE	Accident Quarter Loss & ALAE Payout Pattern	ULAE Payout Pattern	Dividend Payout Pattern	Premium Collection Pattern	Tabular Discount	CYR	IRS Discount Factors	
0.00	0.00%	0.00%	0.00%	18.0%	3.6%	0.00		
0.25	4.80%	1.36%	0.00%	17.1%	4.1%	0.25		
0.50	12.10%	4.89%	0.00%	21.8%	4.5%	0.50		
0.75	8.19%	7.09%	0.00%	25.3%	5.0%	0.75		
1.00	6.22%	8.62%	0.00%	7.2%	5.5%	1.00	0.819398	0.819398
1.25	5.43%	8.46%	15.00%	6.9%	5.9%	1.25		
1.50	5.05%	6.78%	0.00%	0.3%	6.4%	1.50		
1.75	4.40%	5.43%	28.33%	0.3%	6.6%	1.75		
2.00	4.09%	5.02%	0.00%	0.3%	6.8%	2.00	0.802722	0.809786
2.25	3.54%	4.33%	0.00%	0.2%	7.0%	2.25		
2.50	3.21%	4.07%	0.00%	0.2%	7.2%	2.50		
2.75	2.77%	3.65%	28.33%	0.2%	7.9%	2.75		
3.00	2.58%	3.17%	0.00%	0.2%	8.5%	3.00	0.797466	0.804824
3.25	2.18%	2.80%	0.00%	0.1%	9.2%	3.25		
3.50	1.90%	2.56%	0.00%	0.1%	9.8%	3.50		
3.75	1.77%	2.19%	28.33%	0.1%	10.5%	3.75		
4.00	1.58%	2.12%	0.00%	0.1%	11.1%	4.00	0.754828	0.764156
4.25	1.42%	1.77%	0.00%	0.1%	11.7%	4.25		
4.50	1.31%	1.59%	0.00%	0.1%	12.3%	4.50		
4.75	1.15%	1.51%	0.00%	0.1%	12.5%	4.75		
5.00	1.10%	1.28%	0.00%	0.1%	12.7%	5.00	0.733432	0.714034
5.25	0.95%	1.30%	0.00%	0.1%	12.9%	5.25		
5.50	1.01%	1.04%	0.00%	0.1%	13.1%	5.50		
5.75	0.81%	0.99%	0.00%	0.1%	13.7%	5.75		
6.00	0.82%	0.92%	0.00%	0.1%	14.3%	6.00	0.706716	0.678684
6.25	0.65%	0.91%	0.00%	0.0%	14.9%	6.25		
6.50	0.56%	0.84%	0.00%	0.0%	15.5%	6.50		
6.75	0.57%	0.66%	0.00%	0.0%	15.7%	6.75		
7.00	0.54%	0.65%	0.00%	0.0%	15.9%	7.00	0.693485	0.665723
7.25	0.51%	0.60%	0.00%	0.0%	16.1%	7.25		
7.50	0.51%	0.62%	0.00%	0.0%	16.3%	7.50		
7.75	0.45%	0.53%	0.00%	0.0%	16.9%	7.75		
8.00	0.44%	0.51%	0.00%	0.0%	17.6%	8.00	0.666403	0.674000
8.25	0.40%	0.50%	0.00%	0.0%	18.3%	8.25		
8.50	0.35%	0.45%	0.00%	0.0%	19.0%	8.50		
8.75	0.39%	0.40%	0.00%	0.0%	20.0%	8.75		
9.00	0.34%	0.41%	0.00%	0.0%	21.0%	9.00	0.697093	0.706858
9.25	0.33%	0.36%	0.00%	0.0%	22.0%	9.25		
9.50	0.26%	0.36%	0.00%	0.0%	23.0%	9.50		
9.75	0.29%	0.34%	0.00%	0.0%	23.0%	9.75		
10.00	0.26%	0.37%	0.00%	0.0%	22.9%	10.00	0.693861	0.742613

Exhibit 3

	PREMIUM					EXPENSES				LOSS & LOSS ADJUSTMENT EXPENSES						
Year Ending	WP	UEPR	EP	Collected Premium	Premium Receivable	Acq	Maint.	PHR Dividends	Total	Paid Loss & ALAE	Paid ULAE	Nominal Loss&ALAE (Req Reserve)	Nominal ULAE (Req Reserve)	Tabular Discount	Held Reserve	IRS Disounted Reserve
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0.25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0.50	1,374	1,374	0	247	1,127	246	105	0	351	0	0	0	0	0	0	
0.75	0	1,031	344	235	892	0	0	0	0	12	0	238	18	10	246	
1.00	0	687	344	300	592	0	0	0	0	42	1	446	34	20	460	393
1.25	0	344	344	348	244	0	0	0	0	63	2	633	50	20	663	
1.50	0	0	344	99	145	0	0	0	0	78	4	805	64	36	833	
1.75	0	0	0	95	50	0	0	12	12	80	5	725	59	36	748	
2.00	0	0	0	5	45	0	0	0	0	62	6	663	53	36	680	582
2.25	0	0	0	5	40	0	0	22	22	53	5	610	48	35	623	
2.50	0	0	0	5	36	0	0	0	0	47	5	562	43	34	572	
2.75	0	0	0	3	33	0	0	0	0	43	4	520	39	33	526	
3.00	0	0	0	3	30	0	0	0	0	38	3	482	36	33	485	416
3.25	0	0	0	3	27	0	0	22	22	34	3	448	33	32	448	
3.50	0	0	0	3	25	0	0	0	0	30	3	417	30	32	416	
3.75	0	0	0	2	23	0	0	0	0	27	2	391	28	31	387	
4.00	0	0	0	2	21	0	0	0	0	24	2	367	26	31	362	307
4.25	0	0	0	2	19	0	0	22	22	21	2	346	24	31	338	
4.50	0	0	0	2	18	0	0	0	0	19	2	327	22	32	317	
4.75	0	0	0	1	17	0	0	0	0	17	2	311	20	32	299	
5.00	0	0	0	1	16	0	0	0	0	15	1	295	19	33	282	236
5.25	0	0	0	1	14	0	0	0	0	14	1	282	18	32	267	
5.50	0	0	0	1	13	0	0	0	0	12	1	269	17	32	254	
5.75	0	0	0	1	13	0	0	0	0	11	1	258	16	32	242	
6.00	0	0	0	1	12	0	0	0	0	11	1	248	15	31	231	186
6.25	0	0	0	1	11	0	0	0	0	10	1	238	14	31	221	
6.50	0	0	0	1	10	0	0	0	0	9	1	229	13	31	211	
6.75	0	0	0	0	10	0	0	0	0	8	1	221	13	31	203	
7.00	0	0	0	0	9	0	0	0	0	7	1	214	12	30	195	155
7.25	0	0	0	0	9	0	0	0	0	7	1	207	11	30	188	
7.50	0	0	0	0	8	0	0	0	0	6	1	201	11	30	182	
7.75	0	0	0	0	8	0	0	0	0	5	0	196	10	30	176	
8.00	0	0	0	0	8	0	0	0	0	5	0	190	10	30	170	133
8.25	0	0	0	0	7	0	0	0	0	5	0	185	9	30	165	
8.50	0	0	0	0	7	0	0	0	0	5	0	181	9	30	159	
8.75	0	0	0	0	7	0	0	0	0	4	0	176	9	30	155	
9.00	0	0	0	0	6	0	0	0	0	4	0	172	8	30	150	124
9.25	0	0	0	0	6	0	0	0	0	4	0	168	8	31	145	
9.50	0	0	0	0	6	0	0	0	0	4	0	164	8	32	140	
9.75	0	0	0	0	6	0	0	0	0	4	0	161	7	32	136	
10.00	0	0	0	0	6	0	0	0	0	3	0	158	7	33	132	115

Exhibit 3

Year Ending	ASSETS				CAPITAL						
	Total	Investible	Non-Investible	Pre-Tax Investment Income	Surplus Capital	PHR Funded Capital	EQHR Funded Capital	TOTAL	Contributed Capital	Net Income	Value Added
	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)	(25)	(26)	(27)
0.00	0	0	0	0	0	0	0	0	0	0	0
0.25	0	0	0	0	0	0	0	0	0	0	0
0.50	1,975	752	1,223	0	601	107	856	963	856	0	0
0.75	1,877	910	967	15	601	80	810	890	-46	25	0
1.00	1,748	1,101	647	18	601	56	760	817	-50	23	0
1.25	1,608	1,332	276	21	601	39	721	759	-40	22	0
1.50	833	679	154	26	0	28	60	88	-661	21	0
1.75	748	689	59	13	0	25	68	94	9	2	0
2.00	680	625	55	13	0	24	65	89	-3	2	0
2.25	623	573	49	12	0	23	79	102	13	2	0
2.50	572	528	43	11	0	22	71	93	-8	2	0
2.75	526	487	39	10	0	21	64	85	-7	2	0
3.00	485	450	35	9	0	20	57	77	-7	2	0
3.25	448	417	32	9	0	20	72	92	16	2	0
3.50	416	387	29	8	0	19	67	86	-6	2	0
3.75	387	361	26	8	0	18	61	80	-5	2	0
4.00	362	337	24	7	0	18	56	74	-5	2	0
4.25	338	316	22	7	0	17	73	90	17	2	0
4.50	317	298	20	6	0	16	69	85	-4	2	0
4.75	299	281	18	6	0	16	65	80	-4	2	0
5.00	282	266	16	5	0	15	61	76	-4	2	0
5.25	267	252	15	5	0	14	59	74	-2	2	0
5.50	254	240	14	5	0	14	58	71	-2	2	0
5.75	242	228	14	5	0	13	56	69	-2	2	0
6.00	231	218	13	4	0	13	55	67	-1	2	0
6.25	221	208	13	4	0	12	53	66	-1	2	0
6.50	211	200	12	4	0	12	52	64	-1	2	0
6.75	203	192	11	4	0	11	51	62	-1	1	0
7.00	195	184	11	4	0	11	50	61	-1	1	0
7.25	188	177	11	4	0	10	48	58	-2	1	0
7.50	182	171	11	3	0	10	46	56	-2	1	0
7.75	176	165	11	3	0	10	44	54	-2	1	0
8.00	170	159	11	3	0	9	43	52	-2	1	0
8.25	165	154	11	3	0	9	41	50	-2	1	0
8.50	159	149	11	3	0	9	39	48	-2	1	0
8.75	155	144	10	3	0	8	38	46	-2	1	0
9.00	150	140	10	3	0	8	36	44	-2	1	0
9.25	145	135	10	3	0	8	34	41	-2	1	0
9.50	140	130	10	3	0	7	31	39	-2	1	0
9.75	136	126	10	3	0	7	29	36	-2	1	0
10.00	132	121	10	2	0	7	27	34	-2	1	0

Exhibit 3

INCOME TAX

Year Ending	UW Income: Statutory UW Inc (28)	UW Income: Taxable UW Inc (29)	UW Income: Year End Tax on UW Inc (30)	UW Income: Qtrly Tax on UW Inc (31)	Deferred Tax Asset: DTA due to Revenue Offset (32)	Deferred Tax Asset: DTA due to IRS Disc (33)	Deferred Tax Asset: DTA (34)	Investment Income: Qtrly Invest Income (35)	Investment Income: Qtrly Tax on Total II (36)	Tax Total (37)
0.00										
0.25	0			0	0	0	0	0	0	0
0.50	0			0	96	0	96	0	0	0
0.75	0			4	72	4	76	15	5	9
1.00	-180	24	9	4	48	7	56	18	6	10
1.25	0			4	24	8	32	21	7	12
1.50	0			4	0	9	9	26	9	13
1.75	0			4	0	10	10	13	5	9
2.00	155	49	17	4		10	10	13	5	9
2.25	0			-5		9	9	12	4	-1
2.50	0			-5		8	8	11	4	-1
2.75	0			-5		6	6	10	4	-1
3.00	-25	-55	-19	-5		5	5	9	3	-1
3.25	0			-3		4	4	9	3	0
3.50	0			-3		4	4	8	3	0
3.75	0			-3		3	3	8	3	-1
4.00	-24	-38	-13	-3		3	3	7	2	-1
4.25	0			-3		2	2	7	2	0
4.50	0			-3		2	2	6	2	0
4.75	0			-3		1	1	6	2	-1
5.00	-21	-29	-10	-3		1	1	5	2	-1
5.25	0			0		1	1	5	2	2
5.50	0			0		1	1	5	2	1
5.75	0			0		1	1	5	2	1
6.00	-1	-3	-1	0		1	1	4	2	1
6.25	0			0		1	1	4	1	1
6.50	0			0		1	1	4	1	1
6.75	0			0		1	1	4	1	1
7.00	-1	-5	-2	0		1	1	4	1	1
7.25	0			0		2	2	4	1	1
7.50	0			0		2	2	3	1	1
7.75	0			0		3	3	3	1	1
8.00	0	-4	-1	0		4	4	3	1	1
8.25	0			-1		4	4	3	1	0
8.50	0			-1		4	4	3	1	0
8.75	0			-1		4	4	3	1	0
9.00	0	-10	-4	-1		4	4	3	1	0
9.25	0			-1		4	4	3	1	0
9.50	0			-1		4	4	3	1	0
9.75	0			-1		4	4	3	1	0
10.00	3	-7	-3	-1		5	5	2	1	0

CASH FLOW

Year Ending	UW (38)	Invest Income (39)	FIT (40)	Deferred Assets (41)	Asset (42)	Post-tax Equity (44)
0.00	0	0	0	0	0	0
0.25	0	0	0	0	0	0
0.50	1,023	0	0	96	1,975	-856
0.75	-12	15	-9	-20	-98	70
1.00	-43	18	-10	-20	-130	73
1.25	-65	21	-12	-23	-140	61
1.50	-82	26	-13	-23	-775	682
1.75	-97	13	-9	1	-85	-7
2.00	-68	13	-9	1	-68	5
2.25	-80	12	1	-1	-57	-12
2.50	-52	11	1	-1	-51	10
2.75	-47	10	1	-1	-46	9
3.00	-41	9	1	-1	-41	9
3.25	-59	9	0	0	-37	-14
3.50	-33	8	0	0	-33	8
3.75	-29	8	1	0	-29	7
4.00	-26	7	1	0	-25	7
4.25	-45	7	0	-1	-23	-16
4.50	-20	6	0	-1	-21	6
4.75	-18	6	1	-1	-19	6
5.00	-17	5	1	-1	-17	6
5.25	-15	5	-2	0	-15	3
5.50	-14	5	-1	0	-13	3
5.75	-12	5	-1	0	-12	3
6.00	-11	4	-1	0	-11	3
6.25	-10	4	-1	0	-10	3
6.50	-10	4	-1	0	-9	3
6.75	-9	4	-1	0	-9	3
7.00	-8	4	-1	0	-8	3
7.25	-7	4	-1	1	-7	3
7.50	-6	3	-1	1	-6	3
7.75	-6	3	-1	1	-6	3
8.00	-6	3	-1	1	-6	3
8.25	-5	3	0	0	-5	3
8.50	-5	3	0	0	-5	3
8.75	-5	3	0	0	-5	3
9.00	-4	3	0	0	-5	3
9.25	-4	3	0	0	-5	3
9.50	-4	3	0	0	-5	3
9.75	-4	3	0	0	-4	3
10.00	-4	2	0	0	-4	3

LOSS & LAE PAYMENTS

Year Ending	Paid Loss & ALAE							Paid ULAE						
	AQtr 1	AQtr 2	AQtr 3	AQtr 4	AYR 1	AYR 2	TOTAL	AQtr 1	AQtr 2	AQtr 3	AQtr 4	AYR 1	AYR 2	TOTAL
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.25	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.50	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.75	12	0	0	0	12	0	12	0	0	0	0	0	0	0
1.00	30	12	0	0	42	0	42	1	0	0	0	1	0	1
1.25	20	30	12	0	51	12	63	1	1	0	0	2	0	2
1.50	16	20	30	12	36	42	78	2	1	1	0	3	1	4
1.75	14	16	20	30	29	51	80	2	2	1	1	3	2	5
2.00	13	14	16	20	26	36	62	1	2	2	1	3	3	6
2.25	11	13	14	16	24	29	53	1	1	2	2	2	3	5
2.50	10	11	13	14	21	26	47	1	1	1	2	2	3	5
2.75	9	10	11	13	19	24	43	1	1	1	1	2	2	4
3.00	8	9	10	11	17	21	38	1	1	1	1	2	2	3
3.25	7	8	9	10	15	19	34	1	1	1	1	1	2	3
3.50	6	7	8	9	13	17	30	1	1	1	1	1	2	3
3.75	5	6	7	8	12	15	27	1	1	1	1	1	1	2
4.00	5	5	6	7	10	13	24	0	1	1	1	1	1	2
4.25	4	5	5	6	9	12	21	0	0	1	1	1	1	2
4.50	4	4	5	5	8	10	19	0	0	0	1	1	1	2
4.75	4	4	4	5	7	9	17	0	0	0	0	1	1	2
5.00	3	4	4	4	7	8	15	0	0	0	0	1	1	1
5.25	3	3	4	4	6	7	14	0	0	0	0	1	1	1
5.50	3	3	3	4	6	7	12	0	0	0	0	0	1	1
5.75	2	3	3	3	5	6	11	0	0	0	0	0	1	1
6.00	3	2	3	3	5	6	11	0	0	0	0	0	0	1
6.25	2	3	2	3	5	5	10	0	0	0	0	0	0	1
6.50	2	2	3	2	4	5	9	0	0	0	0	0	0	1
6.75	2	2	2	3	4	5	8	0	0	0	0	0	0	1
7.00	1	2	2	2	3	4	7	0	0	0	0	0	0	1
7.25	1	1	2	2	3	4	7	0	0	0	0	0	0	1
7.50	1	1	1	2	3	3	6	0	0	0	0	0	0	1
7.75	1	1	1	1	3	3	5	0	0	0	0	0	0	0
8.00	1	1	1	1	3	3	5	0	0	0	0	0	0	0
8.25	1	1	1	1	2	3	5	0	0	0	0	0	0	0
8.50	1	1	1	1	2	3	5	0	0	0	0	0	0	0
8.75	1	1	1	1	2	2	4	0	0	0	0	0	0	0
9.00	1	1	1	1	2	2	4	0	0	0	0	0	0	0
9.25	1	1	1	1	2	2	4	0	0	0	0	0	0	0
9.50	1	1	1	1	2	2	4	0	0	0	0	0	0	0
9.75	1	1	1	1	2	2	4	0	0	0	0	0	0	0
10.00	1	1	1	1	1	2	3	0	0	0	0	0	0	0

LOSS & LAE RESERVES

Year Ending	Nominal Loss & ALAE Reserves			Nominal ULAE Reserve			Tabular Discount				Held Reserves			IRS Discounted Reserves				
	AYR 1	AYR 2	TOTAL	AYR 1	AYR 2	TOTAL	Discount Factor	AYR 1	AYR 2	Total	AYR 1	AYR 2	TOTAL	IRS Disc Factor AYR 1	IRS Disc Factor AYR 2	AYR 1	AYR 2	TOTAL
	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)	(25)	(26)	(27)	(28)	(29)	(30)	(31)	(32)
0.00	0		0	0		0		0		0	0		0					
0.25	0		0	0		0	0	0		0	0		0					
0.50	0		0	0		0	0.036	0		0	0		0					
0.75	238		238	18		18	0.041	10		10	246		246					
1.00	446		446	34		34	0.045	20		20	460		460	0.8194		393		393
1.25	395	238	633	32	18	50	0.050	20	0	20	408	256	663					
1.50	359	446	805	29	34	64	0.055	20	16	36	369	464	833					
1.75	330	395	725	26	32	59	0.059	20	16	36	337	411	748					
2.00	304	359	663	24	29	53	0.064	19	16	36	308	372	680	0.8027	0.8214	263	319	582
2.25	280	330	610	21	26	48	0.066	18	16	35	283	340	623					
2.50	259	304	562	20	24	43	0.068	18	17	34	261	311	572					
2.75	240	280	520	18	21	39	0.070	17	17	33	241	285	526					
3.00	223	259	482	16	20	36	0.072	16	17	33	223	262	485	0.7975	0.8098	191	225	416
3.25	208	240	448	15	18	33	0.079	16	16	32	207	242	448					
3.50	195	223	417	14	16	30	0.085	17	15	32	192	224	416					
3.75	183	208	391	13	15	28	0.092	17	15	31	179	208	387					
4.00	172	195	367	12	14	26	0.098	17	14	31	167	194	362	0.7548	0.8048	139	168	307
4.25	163	183	346	11	13	24	0.105	17	14	31	157	181	338					
4.50	155	172	327	10	12	22	0.111	17	15	32	148	170	317					
4.75	147	163	311	10	11	20	0.117	17	15	32	140	159	299					
5.00	141	155	295	9	10	19	0.123	17	15	33	132	150	282	0.7334	0.7642	110	126	236
5.25	134	147	282	8	10	18	0.125	17	15	32	126	141	267					
5.50	129	141	269	8	9	17	0.127	16	16	32	120	134	254					
5.75	124	134	258	7	8	16	0.129	16	16	32	115	127	242					
6.00	119	129	248	7	8	15	0.131	16	16	31	110	121	231	0.7067	0.7140	89	98	186
6.25	114	124	238	7	7	14	0.137	16	15	31	105	116	221					
6.50	110	119	229	6	7	13	0.143	16	15	31	101	111	211					
6.75	106	114	221	6	7	13	0.149	16	15	31	97	106	203					
7.00	103	110	214	6	6	12	0.155	16	14	30	93	102	195	0.6935	0.6787	76	79	155
7.25	101	106	207	5	6	11	0.157	16	15	30	90	98	188					
7.50	98	103	201	5	6	11	0.159	16	15	30	87	94	182					
7.75	95	101	196	5	5	10	0.161	15	15	30	85	91	176					
8.00	93	98	190	5	5	10	0.163	15	15	30	82	88	170	0.6664	0.6657	65	69	133
8.25	90	95	185	4	5	9	0.169	15	15	30	79	85	165					
8.50	88	93	181	4	5	9	0.176	16	15	30	77	83	159					
8.75	86	90	176	4	4	9	0.183	16	14	30	74	80	155					
9.00	84	88	172	4	4	8	0.190	16	14	30	72	78	150	0.6971	0.6740	61	62	124
9.25	82	86	168	4	4	8	0.200	16	15	31	70	75	145					
9.50	80	84	164	4	4	8	0.210	17	15	32	67	73	140					
9.75	79	82	161	4	4	7	0.220	17	15	32	65	71	136					
10.00	77	80	158	3	4	7	0.230	18	15	33	63	69	132	0.6939	0.7069	56	59	115

ASSETS

Year Ending	Total Assets	UEPR	Held Loss & LAE Reserves	Required Surplus	Income Producing	Non-Income Producing	II on Total Assets
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
0.00	0	0	0	0	0	0	0
0.25	0	0	0	0	0	0	0
0.50	1,975	1,374	0	601	752	1,223	0
0.75	1,877	1,031	246	601	910	967	15
1.00	1,748	687	460	601	1,101	647	18
1.25	1,608	344	663	601	1,332	276	21
1.50	833	0	833	0	679	154	26
1.75	748	0	748	0	689	59	13
2.00	680	0	680	0	625	55	13
2.25	623	0	623	0	573	49	12
2.50	572	0	572	0	528	43	11
2.75	526	0	526	0	487	39	10
3.00	485	0	485	0	450	35	9
3.25	448	0	448	0	417	32	9
3.50	416	0	416	0	387	29	8
3.75	387	0	387	0	361	26	8
4.00	362	0	362	0	337	24	7
4.25	338	0	338	0	316	22	7
4.50	317	0	317	0	298	20	6
4.75	299	0	299	0	281	18	6
5.00	282	0	282	0	266	16	5
5.25	267	0	267	0	252	15	5
5.50	254	0	254	0	240	14	5
5.75	242	0	242	0	228	14	5
6.00	231	0	231	0	218	13	4
6.25	221	0	221	0	208	13	4
6.50	211	0	211	0	200	12	4
6.75	203	0	203	0	192	11	4
7.00	195	0	195	0	184	11	4
7.25	188	0	188	0	177	11	4
7.50	182	0	182	0	171	11	3
7.75	176	0	176	0	165	11	3
8.00	170	0	170	0	159	11	3
8.25	165	0	165	0	154	11	3
8.50	159	0	159	0	149	11	3
8.75	155	0	155	0	144	10	3
9.00	150	0	150	0	140	10	3
9.25	145	0	145	0	135	10	3
9.50	140	0	140	0	130	10	3
9.75	136	0	136	0	126	10	3
10.00	132	0	132	0	121	10	2

Exhibit 6

INCOME TAX

Year Ending	UW Income: Statutory UW Income	UW Income: Taxable UW Income	UW Income: Year End Tax on UW Income	UW Income: Qtrly Tax on UW Income	Deferred Tax Asset: DTA: Revenue Offset	Deferred Tax Asset: DTA: IRS Discounting	Deferred Tax Asset: DTA: TOTAL	Investment Income: Qtrly Investment Income	Investment Income: Qtrly Tax on Inv Inc	FIT: Total
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
0.00										
0.25				0	0	0	0	0	0	0
0.50				0	96	0	96	0	0	0
0.75				4	72	4	76	15	5	9
1.00	-180	24	9	4	48	7	56	18	6	10
1.25				4	24	8	32	21	7	12
1.50				4	0	9	9	26	9	13
1.75				4	0	10	10	13	5	9
2.00	155	49	17	4		10	10	13	5	9
2.25				-5		9	9	12	4	-1
2.50				-5		8	8	11	4	-1
2.75				-5		6	6	10	4	-1
3.00	-25	-55	-19	-5		5	5	9	3	-1
3.25				-3		4	4	9	3	0
3.50				-3		4	4	8	3	0
3.75				-3		3	3	8	3	-1
4.00	-24	-38	-13	-3		3	3	7	2	-1
4.25				-3		2	2	7	2	0
4.50				-3		2	2	6	2	0
4.75				-3		1	1	6	2	-1
5.00	-21	-29	-10	-3		1	1	5	2	-1
5.25				0		1	1	5	2	2
5.50				0		1	1	5	2	1
5.75				0		1	1	5	2	1
6.00	-1	-3	-1	0		1	1	4	2	1
6.25				0		1	1	4	1	1
6.50				0		1	1	4	1	1
6.75				0		1	1	4	1	1
7.00	-1	-5	-2	0		1	1	4	1	1
7.25				0		2	2	4	1	1
7.50				0		2	2	3	1	1
7.75				0		3	3	3	1	1
8.00	0	-4	-1	0		4	4	3	1	1
8.25				-1		4	4	3	1	0
8.50				-1		4	4	3	1	0
8.75				-1		4	4	3	1	0
9.00	0	-10	-4	-1		4	4	3	1	0
9.25				-1		4	4	3	1	0
9.50				-1		4	4	3	1	0
9.75				-1		4	4	3	1	0
10.00	3	-7	-3	-1		5	5	2	1	0

Financial Pricing Models for Property-Casualty Insurance Products: The Target Return on Capital

Sholom Feldblum, FCAS, FSA, MAAA,
and Neeza Thandi, FCAS, MAAA

Financial Pricing Models for Property-Casualty Insurance Products: The Target Return on Capital

by Sholom Feldblum and Neeza Thandi[1]

INTRODUCTION

The target return on capital is the cost of capital for the insurance enterprise, or the return demanded by suppliers of capital. This paper describes the major considerations in selecting the target return on capital.

A financial pricing model determines the premium rate such that the insurer achieves a target return on capital. The pricing model may take either of two forms:

- A net present value model discounts the projected equity flows at the cost of capital.
- An IRR model compares the internal rate of return implied by the project's equity flows with the cost of capital.

The structure of the pricing model and most of the pricing assumptions are based on the characteristics of the insurance environment and of the line of business. In contrast, the cost of capital is not easily quantified. It is often selected by the insurer's management, based on recommendations by the financial, actuarial, and underwriting departments.

Profitability in the property-casualty insurance industry cycle between hard markets, when returns are high, and soft markets, when returns are low.[1] The target return on capital selected by the company's management may vary with the phases of the underwriting cycle.

Illustration: The selected long-term target return on capital may be 700 basis points above the risk-free interest rate on 90-day Treasury bills. The company may add up to 300 basis points during the profitable phases of the underwriting cycle, and it may subtract up to 300 basis points during the unprofitable phases of the underwriting cycle.

DEBT AND EQUITY CAPITAL

Pricing models for other industries use a weighted average cost of equity capital and of debt capital. The weights depend on the company's intended capital structure.

---

[1] We are indebted to Karl Goring for helpful review of this paper.

Illustration: A firm can issue long-term debt at an 8% yield. Its common stock is priced in the market to yield 13% per annum. We determine the cost of capital.

Suppose the company's target capital structure is 40% debt and 60% equity. The coupon payments on long-term debt are tax deductible. If the marginal tax rate is 35%, the after-tax interest payments are 8% × (1 – 35%) = 5.20%. Stockholder dividends are paid with after-tax funds. The weighted average cost of capital is

$$40\% \times 5.20\% + 60\% \times 13\% = 9.88\%.$$

Neither the market yield on the company's common stock nor the market yield on its long-term debt are choices of the company. They depend on investors' perceptions of the risk of the company and the volatility of its securities. This paper does not deal with the reasons for the different returns on equity capital and debt capital; this is a financial issue, not a pricing issue.

The Need for Cash

Long-term debt provides the cash needed to fund research and development, build plants, and purchase equipment. Accounting equity is not necessarily an operating requirement. A firm may operate with low or even negative capital. A high debt to equity ratio may raise the cost of debt capital, but it is not an absolute impediment to corporate operations.

Insurers – both property-casualty insurance companies and life insurance companies – have little or no long-term debt. Insurers have more than sufficient cash for their operations, since they receive premiums well before they pay losses and other benefits. Insurers need statutory surplus to operate. Long-term debt provides cash, but it does not enhance statutory surplus and it can not satisfy capital requirements.

Illustration: Leveraged buy-outs (LBO's) illustrate the workings of a firm financed primarily with debt and with little accounting equity. LBO's often provide strong management incentives, and they have succeeded in several industries.

There is no such thing as an insurance company LBO. The low invested capital in the LBO would trigger failure of the risk-based capital requirements and possible liquidation or rehabilitation of the company by state solvency regulators.

For insurance pricing models, we must quantify the cost of equity capital. We need not quantify the cost of debt capital, and we need not deal with capital structure. The following sections of this paper consider several methods which are commonly used to quantify the cost of equity capital.[2]

MARKET BENCHMARK

The standard benchmark for the cost of equity capital is the average rate of return for publicly traded stock companies. The S&P 500 and the Russell 2000 are commonly used benchmarks for the rate of return in the U.S. for large companies and small companies, respectively.

The nominal rate of return varies with interest rates and inflation rates. Common practice is to treat the cost of equity capital as the risk-free interest rate plus a market risk premium. The risk-free interest rate is the yield on Treasury securities. The market risk premium is often assumed to remain fairly constant from year to year.

Illustration: The market risk premium may be estimated from historical experience to be about 7 percentage points above the risk-free yields on 90 day Treasury bills. If the current yield on short term Treasury bills is 5% per annum, the benchmark cost of equity capital is 12% per annum.

For the benchmark cost of equity capital, we use the average market risk premium for publicly traded companies. Even this simple benchmark involves subjective judgment in several areas. We mention three topics: (i) duration of the risk-free interest rate, (ii) multiplicative vs additive models, and (iii) stability of the market risk premium.

1. *Duration:* The market risk premium depends on the duration of the risk-free interest rate. If the average spread between 90 day Treasury bills and 30 year Treasury bonds is 250 basis points, a 90 day Treasury bill rate might have a market risk premium of 800 basis points and a long-term Treasury bond rate might have a market risk premium of 550 basis points. The cost of equity capital at any time depends on the shape of the yield curve.

   Illustration: The benchmark cost of equity capital might be estimated either as (i) the 90-day Treasury bill rate plus 800 basis points or as (ii) the 30-year Treasury bond rate plus 550 basis points. When the term structure of interest rates is upward sloping with a spread of 250 basis points between the high and low ends, the two formulas give the same cost of equity capital. If the yield curve is inverted one year, with an 8.5% Treasury bill rate and an 8% long bond rate, method (i) gives a cost of equity capital of 16.5% and method (ii) gives a cost of equity capital of 13.5%.

2. *Model:* It is unclear whether a multiplicative model or an additive model should be used. A multiplicative model uses the risk-free interest rate times a constant, whereas an additive model uses the risk-free interest rate plus a constant.

   Illustration: An additive model may estimate the cost of equity capital as the Treasury bill rate + 800 basis points. A multiplicative model may estimate the cost of equity capital as (1 + the Treasury bill rate) × (1.075) –1.

When the Treasury bill rate is 7% per annum, the two methods give the same cost of equity capital, since $0.07 + 0.08 = 15\%$ and $(1.07 \times 1.075) - 1 = 15\%$. When the Treasury bill rate is 10% per annum, the second method gives a slightly higher cost of equity capital.

3. *Stability:* It is not clear whether the market risk premium is stable over the years. In the late 1990's, some analysts argued for a lower market risk premium, as more investors become comfortable with stock market volatility.

These are financial issues, not insurance pricing issues. They are hotly debated and unresolved; we do not address them further. To use the market benchmark, the pricing actuary selects a duration for the risk-free interest rate, a market risk premium, and the type of model (additive or multiplicative). Reasonableness of assumptions and consistency of application are the key attributes of good pricing.

Illustration: A common benchmark is the yield on short term Treasury bills plus a 7 percentage point market risk premium. With a 5% yield on short term Treasury securities, the cost of equity capital is 12% per annum. These figures reflect the investment environment in 2002.

RISK

The risk of the project affects the required return. Investors seek to maximize their returns for a given level of risk, or to minimize their risk for a given return.

There is no consensus on the level of insurance risk versus the level of risk in other industries. We review several common perspectives on this issue.

Finance: Some financial analysts consider insurance enterprises to be less risky than the average company, implying that a lower cost of capital may be used in the pricing analysis. This view is generally based on a CAPM analysis (see below), which shows an average beta for property-casualty insurance companies of 85% to 90%. The economic rationale for the low beta value is that insurers have little up-front capital expenditures and most of their expenses are variable costs, thereby lessening their business risk.

Illustration: Pharmaceutical companies invest billions of dollars in extensive research to development new medications. Automobile manufacturers invest billions of dollars in plants and equipment to development new automobiles. Insurers do not have these up-front capital requirements.

Actuaries: Some actuaries perceive property-casualty insurance companies as more risky than other enterprises, for two reasons.

- Insurers don't know their costs until after the policy has been sold.
- The loss severity distribution in some lines of business is highly dispersed.

These two types of risk are of questionable relevance for selecting a target return on capital. For most lines of insurance, such as private passenger automobile or workers' compensation, these characteristics have little effect on business risk or earnings volatility. Even for lines of business where these two characteristics are significant, such as general liability, the risk is diversifiable for shareholders. Modern portfolio theory assumes that diversifiable risk does not receive any additional return.

Underwriting cycles: The property-casualty insurance industry has distinct profitability cycles, generally called underwriting cycles. Some past studies, such as the Arthur D. Little studies in the 1960's and the early 1970's, examined the standard deviation of the insurance industry's profitability versus that of other industries to justify a higher rate of return for insurers.

The effect of profitability cycles on the target return on capital depends on their severity, regularity, and correlation with general business cycles.[3] The property-casualty underwriting cycles may be stronger than the cycles in some other industries, but they are also more regular, mitigating the risk to investors. Financial analysts often presume that the market takes into account expected profitability cycle. If the underwriting cycles in the property-casualty insurance industry are not correlated with profitability cycles in other industries, a CAPM analysis would not imply a higher capitalization rate for the insurance industry.

Catastrophes: Some insurance industry personnel speak of the above average risks that they face from natural catastrophes (hurricanes, earthquakes) and from man-made catastrophic exposures (terrorism, asbestos, and environmental liabilities). These are indeed unusual risks, though the relative size of the risks in the insurance industry versus those in other industries is hard to judge.[4]

Longevity: Some analysts see the longevity of the insurance industry and the persistency of many companies as evidence that the level of risk in this industry is low. The slow rate of innovation in the insurance industry and the high customer loyalty reduces the business risk for insurers.

We do not attempt to resolve these issues. We discuss below the most common methods of quantifying the target return on capital, without claiming that any method is necessarily correct. We do not assume that the insurance industry faces higher or lower risk than other industries.

RETURN FACTOR MODELS

Several mathematical models have been developed to quantify the cost of equity capital for particular industries or firms. Generally, a return factor model is used, such as the Capital Asset Pricing Model or Arbitrage Pricing Theory.

A return factor model with N factors says that the expected return for security s in period t is

$$E(r_s) = \beta_{1s} \times F_{1t} + \beta_{2s} \times F_{2t} + \ldots \beta_{Ns} \times F_{Nt}$$

The factors $\{F_{1t}, F_{2t}, \ldots F_{Nt}\}$ depend on the time period *t* but not on the particular security. The beta coefficients $\{\beta_{1s}, \beta_{2s}, \ldots, \beta_{Ns}\}$ depend on the security *s* but not on the time period.

We explain this formula by the CAPM, which is a two factor model.

- F_{1t} is the risk-free interest rate in period *t*, and β_{1s} is unity for all securities.
- F_{2t} is the market risk premium, and β_{2s} is the market beta for security *s*.

THE CAPM

In the 1960's and 1970's, the CAPM was commonly accepted among many financial analysts, and the CAPM perspective on the cost of equity capital remains a predominant view. Recent studies of market anomalies have cast doubt on the empirical validity of the CAPM.[5] The CAPM is still widely used for its simplicity, but it has lost some of its former luster.

The acceptance of the CAPM by pricing actuaries and insurance company managers varies. Other return factor models used in securities valuation, such as Arbitrage Pricing Theory, have had negligible effect on actuarial pricing models.[6]

The CAPM says that a security's expected return depends on its systematic risk. Systematic risk is risk that cannot be eliminated by diversification; diversifiable risk is not compensated by additional return. Algebraically, the CAPM says that the expected return on a security equals

$$E[r_s] = r_f + \beta_s \times (E[r_m] - r_f)$$

where

r_s	is the return on the security
r_f	is the risk-free interest rate
$E[r_s]$	is the expected return on the security
$E[r_m]$	is the expected overall return on the market of risky securities
$E[r_m] - r_f$	is the market risk premium

$\beta_s = \text{cov}(r_m, r_s)/\text{var}(r_m) = \text{corr}(r_m, r_s) \times$ standard deviation (r_s) / standard deviation (r_m).

In this equation, r_s and $\tilde{r}_m$ are random variables. $E[r_s]$ and $E[r_m]$ are scalars; they are the expectations of these random variables.

The rationale for the consideration of systematic risk but not diversifiable risk is compelling. Suppose that the expected return of a security were based on total risk, not on systematic risk only. An arbitrageur might purchase securities with high specific (diversifiable) risk, combine them in a mutual fund with low specific risk, and sell these low risk shares of the mutual fund to other investors.

Illustration: Suppose securities $\{s_1, s_2, \ldots, s_N\}$ have high but uncorrelated risks. The price of each security is the present value of its future cash flows at the capitalization rate for that security. If the market were to base the capitalization rate on the total risk of the security, each security would have a high price to earnings ratio. A mutual fund composed of these securities would have a similarly high price to earnings ratio, even though its total risk is reduced by the addition of uncorrelated random variables. Based on its lower capitalization rate, the mutual fund could be sold at a higher price. This would lead to arbitrage profits.

Market Returns and Security Returns

The CAPM derives the expected return on specific securities from the risk-free interest rate and the expected overall market return. Most applications of the CAPM consider the market risk premium, or the expected market return minus the risk free interest rate, as a known value.

The expected excess return on a specific security, or the expected return on that security minus the risk-free interest rate, is a function of its beta. Since the beta equals corr(r_m,r_s) × standard deviation (r_s) / standard deviation (r_m), this excess return is proportional to

i. the standard deviation of the security's returns, and
ii. the correlation of the security's returns with the overall market return.

Intuitively, these two relationships imply that

i. The greater the specific security's standard deviation, the more uncertainty is inherent in that security and the greater must be the return to the investor.
ii. The stronger the correlation of a specific security with the overall market, the less risk-reduction is available from diversification and the greater must be the return to the investor.

CAPM and Insurance Returns

If three conditions are true, the CAPM enables us to derive an estimate of the cost of equity capital for insurance companies. Specifically

- If the CAPM formula is valid,
- if betas for property-casualty insurance enterprises can be reasonably estimated, and
- if these betas are stable from year to year,

we can derive the expected return for insurance companies. The expected return, or the capitalization rate, is the cost of equity capital.

Betas for individual securities in any industry are not easy to determine, since the random fluctuation of common stock returns provides unstable estimates for the standard deviations and for the correlations with the market.[7] The common practice in the investment community is to assume that the betas for firms in the same industry are similar, and to use the industry's beta as a proxy for the betas of individual securities. The assumptions that firms within an industry have similar betas is questionable in any industry. For the property-casualty insurance industry, there are two reasons why the assumption of a common beta is dubious.

Asset risks: Firms have different investment portfolios. A firm with a more aggressive investment portfolio should have a higher beta for its own equity.

An insurer that invests only in high grade corporate bonds and in Treasury securities might have an investment portfolio with an overall beta close to zero. An insurer that invests half of its assets in common stocks and venture capital might have an overall beta closer to unity. The systematic risk in the investment portfolio translates into a leveraged systematic risk for the insurer's own equity.

Illustration: Suppose an insurer has a four to one assets to capital ratio; assets are \$400 million and capital is \$100 million. The insurer's investment portfolio has a beta of unity. We assume that the liabilities of the insurer, which equal \$400 million – \$100 million = \$300 million, are not correlated with overall market returns.

If the overall stock market return increases by 1 percentage point, the insurer's assets increase by \$4 million. Since the liabilities are uncorrelated with the overall market returns, the insurer's capital increases by \$4 million as well, for a 4% increase.

In sum, the composition of the investment portfolio has a leveraged effect on the systematic risk of the insurer's equity. An industry-wide beta for individual firms gives biased results.

Underwriting risks: Many actuaries assume that risk and expected return vary significantly by line of business. Empirical estimation of betas by line of business requires data from publicly traded monoline insurers; such data are not available. There have been sporadic studies of the systematic risk for certain lines of business, such as workers' compensation and private passenger automobile, but there are no conclusions that are broadly accepted.

Insurance Betas

Betas for the property-casualty insurance industry as a whole are estimated by investment firms. Values between 85% and 95% have been used. With a risk-free interest rate between 5% and 6% and a market risk premium of 7 to 8 percentage points, the CAPM estimate for the cost of equity capital is between $5\% + 90\% \times 7\% = 11.3\%$ and $6\% + 95\% \times 8\% = 13.6\%$.

We have limited our comments on the CAPM to general statements. Our objective is not to advocate or to criticize the CAPM. In our own work, we examine the cost of capital implied by the CAPM as well as other financial valuation models, in combination with our judgment on the competitiveness of the insurance market in each state and line of business. We may summarize the implications of the CAPM as "the cost of capital for the property-casualty insurance industry *may be* a percentage point or so below the overall market average." More specific conclusions are hard to justify.

HISTORICAL EXPERIENCE

Historical returns are sometimes used to estimate the cost of capital. If there are no impediments to capital flows or to marketplace competition, the long-run observed return on capital should not deviate much from the required return on capital.

- If the actual return on capital exceeds the return required by investors, additional capital should flow into the industry and the actual return should decline.
- If the actual return is less than the return required by investors, capital should leave the industry and the actual return on the remaining capital should rise.[8]

In practice, capital flows in the insurance industry are not frictionless:

- Much capital is held by mutual insurance companies, who have less incentive to return excess capital to their owners.
- Much capital is tied up in full value loss reserves.
- Managers of insurance companies may seek to hold more capital than is economically efficient to avoid surplus drains during adverse scenarios.

Historical returns on *statutory surplus* are available from industry publications, such as Best's *Aggregates and Averages*. They have been used at times by state regulators to set ratemaking targets, but they are rarely used by market analysts.[9] Neither GAAP nor statutory book values reflect the invested capital for property-casualty insurance companies.[10]

Illustration: Suppose an insurer writes a $100 million block of workers' compensation large dollar deductible business on January 1, 20XX. Expenses equal to 30% of premium are paid at policy inception. Loss reserves with a nominal value of $150 million and a present value of $70 million are held on December 31, 20XX. No losses are paid during the year. The investment yield is 8% per annum and capital requirements are 15% of written premium and 10% of held reserves.

- The required statutory surplus on December 31, 20XX, is 15% × $100 million + 10% × $150 million = $30 million.[11]
- The invested capital on December 31, 20XX, is the statutory surplus plus the capital embedded in the undiscounted loss reserves, or $30 million + $80 million = $110 million.[12]

The pricing model bases the premium rate on the target return on capital. The return on surplus and the return on equity are not suitable proxies for the return on capital.

The nominal rate of return varies with inflation. If inflation is 4% per annum, investors might be satisfied with a 12% annual return. If inflation is 15% per annum, the 12% annual return would be inadequate. Historical averages may be converted into real dollar terms by subtracting an adjustment for monetary inflation.[13]

We note several problems with basing the cost of capital on historical returns.

- Invested capital versus statutory surplus
- Calendar year investment income
- Portfolio yields versus new money yields
- Allocation of surplus by line of business
- Stock market fluctuations
- Possible over-capitalization of the insurance industry

INVESTED CAPITAL

Invested capital equals statutory surplus plus the capital embedded in the gross unearned premium reserves and the full value loss reserves to statutory surplus. The returns on statutory surplus are biased proxies for the return on capital. For a company holding full value loss reserves in the long-tailed lines of business, statutory surplus may be only half of invested capital. A 12% return on statutory surplus may be equal to a 6% return on invested capital.

The observed returns on surplus are not comparable across lines of business. Homeowners has little capital embedded in loss reserves, and the return on surplus is similar to the return on invested capital. For workers' compensation, the return on surplus may be twice as great as the return on invested capital.

Inter-company differences further hinder the interpretation of industry results. Expense ratios differ between direct writers and independent agency companies; industry results may not be appropriate for a specific insurer. Differing reserve adequacy levels by company are hard to measure, and they distort inter-company comparisons.[14]

CALENDAR YEAR INVESTMENT INCOME

The return on surplus calculated from Best's *Aggregates and Averages* uses the investment income from current calendar year reserves, not the expected investment income from future reserves on the current year's writings. The figures are distorted by growth or decline in the volume of business.

Illustration: Suppose losses in a block of business are 75% of gross premium, the losses are paid (on average) four years after they occur, and the investment yield is 10% per annum. In a steady-state,$1 million of premium would generate about $3 million in loss reserves.

- If policies are written evenly during the year, the average policy effective date is July 1 and the average date of loss is December 31.
- Total losses are three quarters of gross written premium.
- Since losses are paid four years after they occur (on average), loss reserves are four times annual incurred losses or three times annual written premium.

The pre-tax investment income on the assets backing the loss reserves is 10% × $3 million = $0.3 million, or 30% of the gross written premium.

Even if the investment income does not grown in exposure counts, it grows with monetary inflation. If inflation is 10% per annum and the company's book of business is growing with inflation, the reserves are $0.75 million × $(1 + 1.100^{-1} + 1.100^{-2} + 1.100^{-3})$ = $2,615,139.

The expected investment income from the current year's book of business is $0.75 million for four years. The present value at a 10% discount rate is $0.75 million × $(1 + 1.100^{-1} + 1.100^{-2} + 1.100^{-3})$ = $2,615,139. We summarize the steady state illustration as follows.

> *If the company's book of business grows with monetary inflation but there is no change in the overall exposures, and if the inflation rate equals the discount rate, then the present value of the investment income expected on the current year's book of business equals the calendar year investment income.*[15]

If the insurer has been growing, it holds less reserves than it would hold in a steady state. The investment income in the current calendar year is less than the present value of the investment income on the future reserves. The operating profit, or 1 minus the operating ratio, understates the true profitability.

The effects of this distortion are clearest when an insurer enters a new line of long-tailed business. There are no existing reserves, so there is no investment income on the assets backing previous years' reserves. The statutory operating ratio may be low or negative even for adequately priced business.

Illustration: Suppose an insurer commences operations by writing a $100 million block of workers' compensation large dollar deductible business on January 1, 20XX. Expenses equal to 30% of premium are paid at policy inception. Loss reserves with a nominal value of $150 million and a present value of $70 million are held on December 31, 20XX.

Were there no surplus requirements and no need to hold undiscounted loss reserves, the block of business would be profitable. The statutory operating gain, however, is the premium minus losses and expense plus the calendar year investment income. Since the investment income received during the calendar year is small, the statutory operating gain is negative.

The opposite distortion occurs when business volume is declining, as would be the case when a company switches from first dollar workers' compensation policies to large dollar deductible policies. The statutory operating ratio may overstate the true profitability of the company.

PORTFOLIO YIELDS

The operating income figures in statutory exhibits, such as the Insurance Expense Exhibit, and in most rating agency reports, such as Best's *Aggregates and Averages*, use portfolio yields, not market yields. If interest rates have been declining, the statutory exhibits overstate the present value of the investment income expected in the future and overstate the return.

Illustration: Suppose the insurer holds 10% coupon bonds valued at par on January 1, 20XX. On that day, interest rates on comparable bonds rise to 11% per annum. The bonds have a duration of four years on December 31, 20XX. We contrast the statutory yield, the GAAP yield, and the market yield.

- The statutory asset yield in 20XX is 10% per annum. The bonds are held at amortized cost. Neither the current market interest rates nor the change in the market value of the bonds affect the statutory investment yield.
- The GAAP asset yield in 20XX is the 10% coupon rate plus the change in the market value during the year. The market value change is about $-1 \times 4 \times 1\% = -4\%$ of the bond's market value before the change. The GAAP asset yield in 20XX is $10\% - 4\% = 6\%$.[16]
- The new money rate in 20XX is 11% per annum. The coupon rate of the bonds held by the insurer are not relevant to a financial pricing model.

The new money interest rate may vary considerably over the bond's life. The book yield from trade industry reports is a biased proxy for the new money interest rate.

SURPLUS BY LINE

The surplus figures in published reports are for all lines of business combined. Rates of return by line of business can not be observed. Best's does not allocate surplus by line of business, so it does not compute a return on surplus by line of business.

Return on statutory surplus by line of business can not be estimated indirectly. A. M. Best's *Aggregates and Averages* groups companies by category, such as personal lines predominating companies or commercial lines predominating companies. The company category is sometimes used as a proxy for the line of business, though this proxy is too crude for a financial pricing model.

EQUITY RETURNS

Property-casualty insurers hold considerable amounts of equity in their investment portfolios – common stocks, venture capital, and real estate. Fluctuations in equity markets affect the observed returns for property-casualty insurers. During the 1990's, the long bull market in common stocks raised the observed returns for insurance companies, and the stock market decline in 2000-2002 reduced the observed returns for insurance companies.

Random loss fluctuations can have a similar effect. The favorable weather during the latter half of the 1990's and the sparsity of natural catastrophes raised observed returns. The damages from the World Trade Center incident in 2001 reduced observed returns. The magnitude of these fluctuations offsets the value that might be gleaned from historical experience.

ADEQUACY OF RETURNS

The cost of capital reflects the return needed to induce investors to supply capital to insurance enterprises. For industries that are over- or under-capitalized, observed returns are not good proxies for required returns.

Returns for property-casualty insurance companies have been lower than the returns in other financial industries. Two perspectives are often heard:

- The property-casualty insurance industry has less systematic risk than the average industry, and the returns are proper.
- The lower than average returns stem from the competitive nature of the property-casualty insurance industry, not from the level of systematic risk.

The latter perspective is commonly associated with Michael Porter's writings on competitive strategy. The ease of entry into the insurance market, the hundreds of insurance companies, and the possible overcapitalization of the industry account for the lower than average returns, regardless of systematic risk.[17]

Flexible Pricing

The target return on capital is not a rigid figure. Some insurers select both a desired return on capital and a minimum return on capital. Management incentives and marketplace structure affect the target return on capital used in a financial pricing model.

Illustration: Suppose the average cost of equity capital for publicly traded stock companies is 14% per annum. Based on a CAPM analysis, the current state of the underwriting cycle, and a perceived over-capitalization of the industry, management believes that the expected return for property-casualty insurance companies is 12% per annum. The investment yield on a conservative investment grade fixed income portfolio is 8% per annum.

The company may price its policies with a 14% target return on capital, and allow its underwriters to give premium credits if necessary. It may use a 12% cost of capital to measure management performance. If the expected return on the business is less than 8% per annum, the company may curtail its writings in that market.

RISK ADJUSTED RETURNS

Actuarial standards relentlessly advise the actuary to use risk adjusted returns, risk adjusted discount rates, or risk adjusted yields. Rarely is there a coherent explanation of the risk adjustment to be used.[18]

Some actuaries assume that the IRR target return, or the NPV discount rate, should vary by line of business, depending on the risk inherent in the book of business. This inherent risk is sometimes assumed to exist without being rigorously quantified. Sometimes the risk of a line of business is assumed to be proportional to the duration of the liabilities. This is not consistent with the low risk of long-tailed fixed annuities or the high risk of short-tailed property insurance.[19]

Much of the actuarial literature on risk loads measures the process risk of individual policies, not the pricing risk in the book of business. More recently, some actuaries have tried to measure the expected process risk of an insurer's portfolio of risks. Although this is more meaningful than the process risk of an individual policy, there is little reason to assume that the expected process risk of an insurer's portfolio of risks is a proxy for the risk that is relevant to the target return on capital.[20]

Several consulting firms provide risk measures that purport to quantify correlations and covariances among lines of business. In many cases, these correlations and covariances reflect the white noise of random loss fluctuations.

Illustration: The ABC Consulting Firm quantifies the covariances of loss ratios by line from industry-wide Schedule P figures. These covariances stem primarily from random statistical correlations generated by the white noise of loss fluctuations.

Meaningful estimates of systematic risk by line of business have been impossible to attain. Even the CAPM adherents who propose risk adjustments based on modern portfolio theory have been forced to rely on round-about estimates. Underwriting betas and betas of losses are derived from asset betas and equity betas, using data for all lines of business combined. Sampling error and white noise obscure any information these derivations might have.[21]

RATES OF RETURN AND CAPITAL REQUIREMENTS

In theory, for any two lines of business, the line with the greater systematic risk should have a higher profit margin in competitive markets. There are two ways to conceive of this.

1. Both lines have the same capital requirements per dollar of written premium, but the riskier line requires a higher return on the invested capital.
2. The riskier line requires more capital per dollar of written premium, but neither line has a higher required return on the invested capital.

The first view was dominant before the advent of risk-based capital requirements in the early 1990's; the latter view is more common since then. We review the history of actuarial thinking on this topic, along with the related issues of capital structure.

Before 1992, capital requirements were based on rules of thumb, which were not related to the risk in each line of business. The capital requirements were based on overall leverage ratios, which were the same for most lines of business.

- State regulators often used a two to one premium to surplus leverage ratio (the revised "Kenny rule").
- The NCCI has used a three to one reserves to surplus leverage ratio in its IRR pricing model, reflecting the average leverage ratio for workers' compensation insurers.[22]

If the capital requirements per dollar of written premium do not differ by line of business, the target return on capital should depend on the risk in each line. The actuarial analysis of leverage ratios is similar to the financial analysis of capital structure, to which we now turn.

CAPITAL STRUCTURE

Some analysts have proposed viewing insurance reserves as debt capital and statutory surplus as equity capital; see Ferrari [1968], Bailey [1969], and Balcarek [1969]. The optimal leverage ratio for a property-casualty insurer is analogous to the optimal capital structure for a manufacturing concern.

Illustration: A bank receives money from depositors, which it lends to borrowers. The depositors are creditors of the bank (like bond-holders), and the interest paid on deposits times (1 – the corporate tax rate) is the after-tax cost of debt capital. The shareholders of the bank also contribute capital; their expected return is the cost of equity capital.

By writing policies, an insurer receive funds from policyholders. The underwriting loss of an insurer as a percentage of loss and unearned premium reserves is the implicit interest rate paid to policyholders for the use of their funds. This is the implicit cost of debt capital.

Beyond the Miller and Modigliani propositions, modern finance lacks an accepted theory for capital structure. The dominant interpretation of the Miller and Modigliani propositions has two implications.

- Capital structure does not matter in a tax-free world.
- Under the U.S. tax system, debt is often preferable to equity, since bond interest payments are tax deductible whereas stockholder dividends are not.

Illustration: A manufacturing concern needs $15 million in fixed assets and $5 million in net working capital. Corporate bonds can be issued at an 8% coupon rate. Equityholders expect a 12% per annum after-tax return.

- If the company is financed with debt only, it needs a $400,000 pre-tax return to meet its coupon payments.
- If the company is financed with equity only, it needs a $600,000 after-tax return, or a $600,000 / (1 – 35%) = $923,077 pre-tax return.

The implication for insurance is that equityholders should prefer higher leverage ratios wherever permitted by regulation.[23] This conclusion is belied by the one to one premium to surplus leverage ratio that now prevails in the property-casualty industry, despite the high cost of holding capital and the lack of regulation mandating this level. Property-casualty insurers have the highest capital to asset ratios for any financial intermediary: about 40% for property-casualty insurers, but less than 10% for life insurers. Modern financial theory has not been of much aid in explaining empirical leverage ratios for insurance or in recommending optimal leverage ratios for policy pricing.[24]

ACTUARIAL ANALYSES OF REQUIRED CAPITAL

Some actuaries have used to probability of ruin analyses to determine capital requirements.[25] The required capital was the capital needed so that the probability of ruin was below a given threshold. Butsic [1994] and Hodes, *et al.* [1999] extend the theory by looking at the expected policyholder deficit instead of the probability of ruin.[26]

These analyses, which often use financial analysis (DFA) models of an insurer's operations, suffer from two detriments.

1. Many actuaries contend that financial modeling using the probability of ruin or the expected policyholder deficit can suggest relative capital requirements among blocks of business. They are less useful for determining the absolute dollars of capital.[27]

2. A common perception is that DFA analyses can determine capital requirements if one first selects a probability of ruin or an expected policyholder deficit ratio. For example, if one selects a 1% probability of ruin or a 2% expected policyholder deficit, a DFA analysis can determine the capital needed to meet these requirements.

   This ascribes too much predictive power to DFA analysis. Solvency risks depend on variables that actuaries have not succeeding in quantifying, such as underwriting cycles, marketplace competition, regulatory actions, and unexpected catastrophes. The standard probability of ruin analyses, which focus on loss frequency and loss severity distributions, are of little relevance to these solvency risks. The events that have lead to most property-casualty insolvencies in recent years, such as Hurricane Andrew, asbestos claims, environmental exposures, or the September 2001 World Trade Center incident, are not amenable to standard loss frequency and loss severity analyses.

Risk-Based Capital

With the advent of risk-based capital requirements, the focus of actuarial work has shifted. Instead of improvising theoretical capital requirements, actuaries now address the actual capital requirements imposed by the NAIC or the rating agencies. The hypothesized relation between the required return on capital and various actuarial or financial measures – such as probability of ruin, expected policyholder deficit, process risk, or tail value at risk – are of limited relevance for policy pricing.

To determine the capital requirements for the financial pricing model, we use the actual capital requirements from the NAIC risk-based capital requirements and from the similar rating agency formulas. These requirements affect the capital that companies must hold to avoid regulatory intervention in their operations and to maintain their desired ratings.[28]

Cost of Holding Capital

The cost of holding capital connects the target return on capital and the indicated premium rate. Yet a problem with terminology has plagued many discussions of this topic. To clarify the terms, we differentiate between the *cost of capital* and the *cost of holding capital.*

- The *cost of capital* is the return on capital demanded by the equity-holders or other suppliers of capital to the firm. For a manufacturing enterprise, the cost of capital may be

8% for long-term debt, 13% for retained earnings, and somewhat higher for a new stock issue.[29] Insurance enterprises rarely have long-term debt. The cost of capital for insurers is the cost of internal equity (retained earnings). To highlight this attribute of insurance enterprises, we use the term "cost of *equity* capital" in our papers on financial pricing models.

- The *cost of holding capital* is the amount that equity-holders would lose by providing capital to the insurance enterprise were they not compensated by a profit margin in the policyholder premium. At a minimum, the cost of holding capital is the cost of double taxation. Investors supplying capital to an insurance enterprise are taxed twice on the investment income on capital funds.

Illustration – Double Taxation: Suppose insurance regulation requires investors to contribute $100 million to support the writing of insurance policies. The opportunity cost of this capital is the amount that the equity-holders would receive if they invested the $100 million elsewhere; this is the cost of capital. The cost of holding capital is the difference between this cost and the return received by investment through the insurance company.

Suppose the equity-holders would otherwise invest this $100 million in bonds with an investment yield of 10%. The insurance enterprise could invest the $100 million in the same bonds and receive the same investment yield.

If the equity-holders invest the $100 million in 10% coupon taxable bonds, they pay personal income taxes on the $10 million return. If the insurer makes the same investment, it pays $3.5 million of corporate income taxes before returning the remaining investment income to the equity-holders. The equity-holders pay personal income taxes on the $6.5 million that they receive from the insurance company.

The cost of holding this capital stemming from double taxation is the difference in the taxes incurred between (i) direct investment of capital and (ii) investment of capital through an insurance company.

- The taxes paid on direct investment of capital = *investment yield × personal tax rate.*

- The taxes paid on investment of capital through an insurance company =

  investment yield × [corporate tax rate + (1 – corporate tax rate) × personal tax rate]

- The difference between these two is

investment yield ×
[corporate tax rate + (1 – corporate tax rate) × personal tax rate – personal tax rate]
= investment yield × corporate tax rate × (1 – personal tax rate)

This is the after-tax difference to the equityholder. The difference before personal income taxes is the investment yield × the corporate tax rate.

Illustration: If the investment yield is 10% per annum, the corporate tax rate is 35%, and the average personal tax rate is 30%, the cost of holding capital is

$$10\% \times [35\% + (1 - 35\%) \times 30\% - 30\%] =$$
$$10\% \times 35\% \times (1 - 30\%) = 0.0245, \text{ or } 2.45\%.$$

The equityholders pay an additional 2.45% of the yield on their capital to the taxing authorities. This is the after-tax loss to the equityholders. The loss before personal income taxes is 10% × 35% = 3.5%.[30] To induce investors to fund the insurance enterprise, the 3.5% of lost yield must be paid by the policyholders, not the equityholders.

If the policyholders paid this money directly to the equityholders, this would be the full cost of holding capital. In practice, there are no direct transactions between the policyholders and the equityholders. Instead, the policyholders pay this money as part of the policy premium, and the insurance company remits the money to the equityholders. This introduces another layer of taxation, since the policy premium is pre-tax and the compensation to the equityholders is post-tax. The additional margin in the policy premium, as a percentage of the investment yield on equityholder supplied capital, is

$$\text{investment yield} \times \text{corporate tax rate} / (1 - \text{corporate tax rate}) =$$
$$\text{investment yield} \times 35\% / (1 - 35\%) = \text{investment yield} \times 53.85\%$$

The double taxation affect invested capital, whereas the money paid by policyholders is a margin on premium. This margin is *capital × investment yield × 53.85% / premium*

There are other potential costs to holding capital, which are subject to considerable debate in the financial community.[31] A common actuarial argument is that the cost of holding capital is the difference between the cost of equity capital and the after-tax investment yield of the insurance company. This perspective underlies the pricing model in Atkinson and Dallas [2000] as well as the pricing model in this paper.

Illustration: Suppose the cost of equity capital is 12% per annum, but the insurance enterprise invests in 8% Treasury securities. The cost of double taxation is 35% × 8% = 2.8%. The additional cost of holding capital stemming from the conservative investments of the insurance company is 12% – 8% = 4%. The total cost of holding capital is 2.8% + 4% = 6.8%. This is the amount that policyholders must pay to the equityholders to induce them to fund the insurance operations. Since the policyholders pay this money indirectly through the profit margin in the premium, which is taxed as underwriting income, the additional premium is 6.8%

/ (1–35%) = 10.46%.[32] Since the premium is paid at policy inception, the profit margin is 10.46% / 1.08 = 9.69%.

This implies that with an 8% investment yield and a 400 basis point spread between the target return on capital and the investment yield, the policyholders pay 9.69% of equityholder supplied capital to compensate for the indirect investment of their funds.[33]

CONCLUSION

The target return on capital is a somewhat discretionary assumption that drives any financial pricing model. There are diverse views on selecting the target return on capital, and we do not pretend to declare any of them correct. This paper reviews the considerations that the pricing actuary should take into account when selecting the target return.

Appendix: Investment Tax Rates and Double Taxation

The discussion of double taxation in the text of this paper does not fully reflect some adjustments that Miller and other have made to the theory. This appendix provides a brief synopsis for the interested reader.

MILLER'S TAX ADJUSTMENT

In 1963, Merton Miller qualified the tax advantage of debt financing; see also Myers (1999) and Miller (1997). Miller surmised that the double taxation of equity financing may be partially offset by the higher personal tax rates on interest income than on long-term capital gains. The following illustration explains this offset.

Illustration: Investors can receive 10% per annum interest on Treasury bonds, on which they pay personal income taxes. Assume that the investors have high personal income tax rates of 36%, so they receive $64 in after-tax income from $1,000 of invested capital. Alternatively, they can invest their capital in a property-casualty insurance company, which purchases Treasury bonds. The insurance company pays $35 in corporate income taxes on the interest income from the Treasury bonds. The investors receive the remainder as long-term capital gains (not as dividends), on which they pay a 20% marginal tax rate.[34] With a 10% average stock turnover rate, the effective tax rate on long-term capital gains is about 18%. The net return to the investors from the insurance company is $65 × (1 – 18%) = $53.30. The cost of double taxation is $64.00 – $53.30 = $10.70.

Miller deals with the gain to the company from debt financing, which is analogous to double taxation but viewed from the company's perspective.[35] He expresses the gain as

$$G_L = (1 - [(1 - t_C)(1 - t_{PS})]/(1 - t_{PB})] \times B_L$$

where t_C is the corporate tax rate

t_{PS} is the personal tax rate on stock capital gains

t_{pb} is the personal tax rate on bond coupon interest.[2]

If the personal tax rate on capital gains is low enough and the personal tax rate on coupon interest is high enough, the gain from debt financing disappears and the cost of double taxation is zero.

The exact cost of double taxation is unclear. Even if Miller's adjustment is correct, the current tax structure in the U.S. causes only a small reduction in the cost of double taxation. The cost is probably substantial, but there is disagreement on its exact size. It may depend on the tax brackets of individual investors and the form in which the investors receive income from the insurance company.

References

Appel, David and James Gerofsky, "Regulating Competition: The Case of Workers Compensation Insurance," *Journal of Insurance Regulation*, Volume 3, No. 4 (June 1985), pages 408-425.

Berger, Lawrence A., "A Model of the Underwriting Cycle in the Property and Casualty Insurance Industry," *Journal of Risk and Insurance*, Volume 55, No. 2 (June 1988), pages 398-306.

Blume, Marshall, "Betas and Their Regression Tendencies," *Journal of Finance*, X, No. 3 (June 1975), pp. 785-795.

Burns, Arthur F., and Wesley C. Mitchell, *Measuring Business Cycles* (New York: National Bureau of Economic Research, 1946).

Cummins, J. David and J. Francois Outreville, "An International Analysis of Underwriting Cycles in Property-Liability Insurance," *Journal of Risk and Insurance*, Volume LIV, Number 2 (June 1987), pages 246-252; reprinted in Georges Dionne and Scott E. Harrington (eds.), *Foundations of Insurance Economics: Readings in Economics and Finance* (Boston: Kluwer Academic Publishers, 1992), pages 609-626.

Cummins, J. David, "Multi-Period Discounted Cash Flow Ratemaking Models in Property-Liability Insurance," *Journal of Risk and Insurance*, Volume 57, No. 1 (March 1990), pages 79-109.

Cummins, J. David, and Scott E. Harrington, "Property-Liability Insurance Rate Regulation: Estimation of Underwriting Betas Using Quarterly Profit Data," The *Journal of Risk and Insurance*, Volume 52, No. 1 (March 1985), pages 16-43.

Cummins, J. David, and Scott E. Harrington, "The Relationship Between Risk and Return: Evidence for Property-Liability Insurance Stocks," *Journal of Risk and Insurance*, Volume 55, No. 1 (March 1988), pages 15-31.

---

[2] See also Megginson (1997), who provides a general review of this subject.

Cummins, J. David, and Joan Lamm-Tennant, "Capital Structure and the Cost of Equity Capital in the Property-Liability Insurance Industry," *Insurance Mathematics and Economics*, Volume 15 (1994), pages 187-201. (Philadelphia: National Council on Compensation Insurance, November 1991).

Cummins, J. David, Scott E. Harrington, and Robert W. Klein (eds.), *Cycles and Crises in Property/Casualty Insurance: Causes and Implications for Public Policy* (National Association of Insurance Commissioners, 1991).

D'Arcy, Stephen P. and Neil A. Doherty, *The Financial Theory of Pricing Property-Liability Insurance Contracts* (Homewood, IL: Richard D. Irwin, Inc., 1988).

D'Arcy, Stephen P., and Michael A. Dyer, "Ratemaking: A Financial Economics Approach," *Proceedings of the Casualty Actuarial Society*, Volume 84 (1997), pages 301-390.

Daykin, C. D., and G. B. Hey, "Applications of a Simulation Model of a General Insurance Company," *Third International Conference on Insurance Finance and Solvency* (Rotterdam, May 1991).

Daykin, Chris D., Teivo Pentikäinen, and M. Pesonen, *Practical Risk Theory for Actuaries*, First Edition (Chapman and Hall, 1994).

Daykin, C. D., G. D. Bernstein, S. M. Coutts, E. R. F. Devitt, G. B. Hey, D. I. W. Reynolds, and P. D. Smith, "Assessing the Solvency and Financial Strength of a General Insurance Company," *Journal of the Institute of Actuaries*, Volume 114, Part 2 (1987), pages 227-310; discussions by C. J. W. Czapiewski, pages 311-312; J. P. Ryan, pages 312-313; S. Benjamin, pages 313-315; J. Plymen, pages 315-316; P. S. Carroll, pages 316-317; T. Pentikainen, pages 317-318; T. Muir, pages 318-319; J. D. F. Dickson, page 320; W. M. Abbott, pages 320-321; M. H. Field, pages 321-322; P. N. S. Clark, pages 322-323; C. D. Daykin, pages 323-325.

Derrig, Richard A., "The Use of Investment Income in Massachusetts Private Passenger Automobile and Workers' Compensation Ratemaking," in J. David Cummins and Scott E. Harrington (eds.), *Fair Rate of Return in Property-Liability Insurance* (Boston: Kluwer*Nijhoff Publishing, 1987), pages 119-146.

Fairley, William, "Investment Income and Profit Margins in Property-Liability Insurance: Theory and Empirical Results," *The Bell Journal of Economics* 10 (Spring 1979) pages 192-210, reprinted in J. David Cummins and Scott E. Harrington (eds.), *Fair Rate of Return in Property-Liability Insurance* (Boston: Kluwer*Nijhoff Publishing, 1987), pages 1-26.

Fama, Eugene F., and Kenneth R. French, "The Cross-Section of Expected Stock Returns," *Journal of Finance*, 1992, pages 427-466.

Fama, Eugene F., and Kenneth R. French, "Common Risk Factors in the Returns on Stocks and Bonds," *Journal of Financial Economics*, 1993, pages 3-56.

Feldblum, Sholom, "Pricing Insurance Policies: The Internal Rate of Return Model," Second Edition (Casualty Actuarial Society Examination 9A Study Note, May 1992).

Feldblum and Thandi, "Income Recognition and Performance Measurement" [2003B]

Feldblum, Sholom, "Direct Charges and Credits to Surplus" (CAS Part 7 Examination Study Note, 1995).

Feldblum, Sholom, Discussion of Thomas Kozik's "Underwriting Betas: The Shadows of Ghosts," *Proceedings of the Casualty Actuarial Society*, Volume 83 (1996), 648-656.

Feldblum and Thandi, "Modeling the Equity Flows" [2003A]

Feldblum and Thandi, "Federal Income Taxes and the Cost of Holding Capital," [2003D]

Feldblum and Thandi, "Benchmark Investment Yield" [2003G]

Feldblum and Thandi, "Surplus Allocation" [2003J]

Feldblum, Sholom, "Underwriting Cycles and Business Strategies," *Proceedings of the Casualty Actuarial Society*, Volume 88 (2001), pages 175-235.

Ferrari, J. Robert, "A Theoretical Portfolio Selection Approach to Insuring Property and Liability Lines," *Proceedings of the Casualty Actuarial Society*, Volume 59 (1967), pages 33-54; discussion by Martin Bondy, pages 55-57; Robert A. Rennie, pages 57-59; Matthew Rodermund, pages 59-64; LeRoy J. Simon, pages 64-66; author's review of discussions, pages 66-69.

Gogol, Daniel F., "Pricing to Optimize an Insurer's Risk-Return Relation," *Proceedings of the Casualty Actuarial Society*, Volume 83 (1996), pages 41-74.

Harris, Milton, and Artur Raviv, "The Theory of Capital Structure," *Journal of Finance*, Volume 46 (March 1991), pages 297-355.

Harry DeAngelo and Ronald W. Masulis, "Optimal Capital Structure under Corporate and Personal Income Taxation," *Journal of Financial Economics*, Volume 8 (1980) pages 3-29

Hartwig, Robert P., William J. Kahley, Tanya E. Restrepo, and Ronald C. Retterath, "Workers Compensation and Economic Cycles: A longitudinal Approach," *Proceedings of the Casualty Actuarial Society*, Volume 84 (1997), pages 660-700.

Hartwig, Robert P., William J. Kahley, and Tanya E. Restrepo, " Workers Compensation Loss Ratios and the Business Cycle," NCCI Digest, Volume 9, No. 2 (December 1994), pages 1-13.

Hodes, Douglas M., Sholom Feldblum, and Gary Blumsohn, "Workers' Compensation Reserve Uncertainty," *Proceedings of the CAS*, Volume 86 (1999), pages 263-392.

Jeffrey, Robert H., "Tax Considerations in Investing," in *The Portable MBA in Investment*, edited by Peter L. Bernstein (John Wiley & Sons: 1995).

Joskow, Paul L., "Cartels, Competition, and Regulation in the Property-Liability Insurance Industry," *The Bell Journal of Economics and Management Science*, Volume 4, No. 2 (Autumn 1973) pages 375-427; reprinted in Georges Dionne and Scott E. Harrington (eds.), *Foundations of Insurance Economics: Readings in Economics and Finance* (Boston: Kluwer Academic Publishers, 1992), pages 468-520.

Kahley, William J., and Leigh H. Halliwell, "The NCCI Internal Rate of Return and Cost of Capital Models," *NCCI Digest*, Volume 7, Issue 4 (December 1992), pages 35-50.

Kozik, Thomas J., "Underwriting Betas – The Shadows of Ghosts," *Proceedings of the Casualty Actuarial Society*, Volume 81, Numbers 154 and 155 (1994), pages 303-329.

Kreps, Rodney E., "Reinsurer Risk Loads from Marginal Surplus Requirements," *Proceedings of the Casualty Actuarial Society*, Volume 77 (1990), pages 196-203.

Lamm-Tennant, Joan, and Mary A. Weiss, "International Insurance Cycles: Rational Expectations/Institutional Intervention," *Journal of Risk and Insurance*, Volume 64, Number 3 (1997), pages 415-439.

Mahler, Howard C., "The Myers-Cohn Pricing Model, A Practical Application," *Proceedings of the CAS*, Volume 85 (1998), pages 689-774.

McGee, Robert T., "The Cycle in Property/Casualty Insurance," *Federal Reserve Bank of New York Quarterly Review*, Autumn 1986, pages 22-30.

McNulty, James J., Tony D. Yeh, William S. Schulze, and Michael H.Lubatkin, "What's Your Real Cost of Capital," Harvard Business Review, October 2002, pages 114-121.

Megginson, W. L., Corporate Finance Theory (Old Tappan, NJ: Addison-Wesley 1997)

Meyers, Glenn G., "The Competitive Market Equilibrium Risk Load Formula for Increased Limits Ratemaking," *Proceedings of the Casualty Actuarial Society*, Volume 78 (1991), pages 163-200.

Meyers, Glenn G., "An Analysis of the Capital Structure of an Insurance Company," *Proceedings of the Casualty Actuarial Society*, Volume 76 (1989), pages 147-170.

Meyers, Glenn G., "The Competitive Market Equilibrium Risk Load Formula for Catastrophe Ratemaking," *Proceedings of the Casualty Actuarial Society*, Volume 83 (1996), pages 563-600; discussion by James E. Gant, pages 601-610.

Meyers, Glenn G., "An Analysis of the Capital Structure of an Insurance Company," *Proceedings of the Casualty Actuarial*

Miccolis, Robert S., "An Investigation of Methods, Assumptions, and Risk Modeling for the Valuation of Property/Casualty Insurance Companies," *Financial Analysis of Insurance Companies* (Casualty Actuarial Society 1987 Discussion Paper Program), pages 281-321.

Miccolis, Robert S., "On the Theory of Increased Limits and Excess of Loss Pricing," *Proceedings of the Casualty Actuarial Society*, Volume 64 (1977) page 27; discussion by Sheldon Rosenberg, page 60.

Miller, Merton H., "The Corporate Income Tax and Corporate Financial Policies," in *Stabilization Policies*, The Commission on Money and Credit, Prentice Hall, Inc., New Jersey (1963), pages 381-470.

Miller, Merton H., "Debt and Taxes," *Journal of Finance*, Volume 32, No. 2 (May 1977), pages 261-275.

Miller, Merton H., "The Modigliani-Miller Propositions after Thirty Years," *Journal of Economic Perspectives* (Fall 1988); reprinted in Daniel H. Chew, Jr. (ed.), *The New Corporate Finance: Where Theory Meets Practice*, Second Edition (Boston: Irwin McGraw-Hill, 1999), pages 192-204.

Modigliani, Franco, and Merton Miller, "The Cost of Capital, Corporation Finance, and the Theory of Investment," *American Economic Review*, Volume 48 (June 1958) pages 261-297.

Modigliani, Franco, and Merton Miller, "Corporate Income Taxes and the Cost of Capital," *American Economic Review*, Volume 53 (June 1963) pages 433-443.

Myers, Stewart C., "The Search for Optimal Capital Structure," in Daniel H. Chew, Jr. (ed.), *The New Corporate Finance: Where Theory Meets Practice*, Second Edition (Boston: Irwin McGraw-Hill, 1999), pages 205-213.

Myers, Stuart C., "The Capital Structure Puzzle," *Journal of Finance*, Volume 39 (July 1984) pages 575-592.

Pentikäinen, Teivo, and Jukka Rantala, *Solvency of Insurers*, Two Volumes (Helsinki: Insurance Publishing Co., 1982).

Pentikäinen, Teivo, Heikki Bonsdorff, Martti Pesonen, Jukka Rantala, and Matti Ruohonen, *Insurance Solvency and Financial Strength* (Helsinki, Finland: Finnish Insurance Training and Publishing Co., 1989).

Roll, Richard, "A Critique of the Asset Pricing Theory's Tests - Part I: On the Past and Potential Testability of the Theory," *Journal of Financial Economics*, Volume 4 (1977) pages 129-176.

Roth, Richard J., Jr., "Analysis of Surplus and Rate of Return Without Using Leverage Ratios," *Insurer Financial Solvency* (Casualty Actuarial Society 1992 Discussion Paper Program), Volume I, pages 439-464.

Stewart, Barbara D., "Profit Cycles in Property-Liability Insurance," in John D. Long and Everett D. Randall (eds.), *Issues in Insurance*, Volume 1, Third Edition (Malvern, PA: The American Institute for Property and Liability Underwriters, 1984), pp. 273-334.

Stiglitz, Joseph E., "A Reexamination of the Modigliani-Miller Theorem," *American Economic Review*, Volume 54 (December 1969) pages 784-793

Titman, Sheridan, and Roberto Wessels, "The Determinants of Optimal Capital Choice," *Journal of Finance*, Volume 43 (March 1988) pages 1-19.

Vasicek, Oldrich A., "A Note on Using Cross-Sectional Information in Bayesian Estimation of Security Betas," *Journal of Finance*, Volume 8, No. 5 (December 1973), pages 1233-1239.

Venezian, Emilio C., "Ratemaking Methods and Profit Cycles in Property and Liability Insurance," *Journal of Risk and Insurance*, Volume LII, Number 3 (September 1985) pages 477-500.

Damodaran, Aswath, *Investment Valuation* (New York, John Wiley and Sons, 1996).

Higgins, Robert C., *Analysis for Financial Management*, 6th edition (Boston: Irwin, McGraw Hill, 2001).

1 Hard markets and soft markets are insurance industry terms for the high profitability and low profitability phases of the underwriting cycle. On underwriting cycles in the U.S., see Berger [1988], Cummins, Harrington, and Klein [1991], McGee [1986], Stewart [1984], and Venezian [1985]. On underwriting cycles in Europe, see Daykin, Pentikäinen, and Pesonen, [1994], Cummins and Outreville [1987], and Lamm-Tennant and Weiss [1997]. On general business cycles in the U.S., with an emphasis on commodity cycles, see Sterman [2001]. For a pricing perspective on underwriting cycles, see Feldblum [2001].

2 This paper discusses considerations and methods; it does not present hard recommendations. For efficient corporate management, employees need definite targets, even if they are subjectively determined.

[3] Most analyst see little correlation of the insurance underwriting cycle with general business cycles, though this is disputed by some. For workers' compensation, loss experience is inversely correlated with general business cycles; see Hartwig, Kahley, and Restrepo [1994], Hartwig, Kahley, Restrepo, and Retterath [1997], and Feldblum [2003; wcr].

[4] The efficiency of the reinsurance marketplace further reduces the severity of catastrophic exposures. Well managed companies with sound reinsurance arrangements have weathered most natural catastrophes.

[5] See especially Fama and French [1992; 1993]. The theoretical underpinnings of the CAPM have also been weakened by the criticisms raised by Roll [1977].

[6] CAPM-based target returns and capitalization rates were used in Fairley's pricing model and in the Myers/Cohn discounted cash flow pricing model. These models were developed for the state mandated rating system in Massachusetts; neither of these models has been used in competitive markets. See Fairley [1987], Derrig [1987], Myers and Cohn [1987], D'Arcy and Doherty [1988], D'Arcy and Dyer [1997], Cummins [1990], Mahler [1998], Feldblum [2003: Fyf; 2003; mc]. On the estimation of underwriting betas, see Cummins and Harrington [1985], Kozik [1995], and Feldblum [1996].

[7] In statistical terms, both the sampling error of the historical observations and the subsequent changes in betas are too great for useable results. See Blume [1975] and Vasicek [1973] for theoretical discussions, and McNulty, Yeh, Schulze, and Lubatkin [2002] for practical problems with CAPM estimates.

[8] The relationship between the return on capital and the premium rate is similar. The price of a product depends on its supply, and its supply depends (in part) on the amount of capital available for its production. An increase in the available capital enlarges supply and lowers product prices. For property-casualty insurance, the relationship between the amount of capital and the supply of coverage is more tenuous. Prices fluctuate with the phases of the underwriting cycle, despite an ample supply of capital.

[9] California, Texas, New Jersey, and several other states use return on surplus measures in their rate approval process. See Roth [1992] for one regulator's perspective on this issue.

[10] The differences among statutory surplus, GAAP equity, and invested capital are discussed in Feldblum and Thandi, [2003A; 2003B]; see also Feldblum [1995].

[11] GAAP equity is larger by the amount of the deferred tax asset stemming from loss reserve discounting that is not recognized on the statutory balance sheet; see Feldblum and Thandi [2003D; 2003A, Appendix A].

[12] Federal income taxes paid increases invested capital, and the statutorily admitted portion of the deferred tax asset lowers it; for simplicity, we don't model the tax liability or the deferred tax asset in this paper.

[13] This is similar to the CAPM decomposition into the risk-free interest rate and a market risk premium. The use of the inflation rate versus the risk-free interest rate is not material; either adjustment may be used.

[14] Random loss fluctuations and the vicissitudes of the underwriting cycle render most published conclusions highly suspect. Some authors have used long time intervals to estimate the historical relationship between insurance industry returns and the risk-free interest rate. Even over long time intervals, this relationship has not been stable.

[15] This is the rationale for using calendar year investment income as a proxy for discounted reserves; see Feldblum [1997: IEE].

[16] We assume that the bonds are categorized as available for sale, not held to maturity; see SFAS 115.

[17] See Joskow [1973] on the possible over-capitalization of the property-casualty insurance industry.

[18] Actuarial Standards Board, "Actuarial Standard of Practice No. 19: Actuarial Appraisals" (October 1991), page 2, Definition 2.7, says: "Risk-Adjusted Rate of Return – An expected or target annual return to the investor that includes a risk-free return that compensates the investor for the use of the funds (recognizing anticipated inflation so as to maintain the real value of those funds), plus a risk premium above the risk-free rate that compensates the investor for the risk that actual returns will deviate from expected. The size of the risk premium varies with the degree of risk associated with the returns." This standard is too vague to guide actuarial practice.

[19] See Hodes, *et al.*, [1999] for an analysis of the risk inherent in long-tailed workers' compensation reserves.

[20] For the actuarial risk load literature, see Miccolis [1977], Meyers [1991; 1996], Feldblum [1993], Kreps [1990], Gogol [1996].

[21] See Kozik [1995]. The long quest to quantify underwriting betas may one day be recorded as an embarrassing interlude in the development of financial pricing models for property-casualty insurance.

[22] See Cummins [1990], Feldblum [1992], and Kahley and Halliwell [1992].

[23] Ferrari [1967] concluded that the optimal capital structure for insurance companies was as little capital as possible.

[24] For further discussion of capital structure, see especially Modigliani and Miller [1958; 1963], Miller [1963; 1977; 1999], Stiglitz [1969]; DeAngelo and Masulis [1980], Myers [1984; 1999], Titman and Wessels [1988], Harris and Raviv [1991], and Meggison [1997], chapter 7, "Capital Structure Theory," pages 305-352. For an analysis of capital structure for property-casualty insurers, see Meyers [1989].

[25] See Daykin *et al.* [1987], Daykin and Hey [1991, Pentikäinen, *et al.*, [1989], Pentikäinen and Rantala [1982], and Daykin, Pentikäinen, and Pesonen [1994])

[26] See also Kreps [1990] and Butsic [2000 on reinsurance] for additional actuarial approaches.

[27] See Feldblum and Thandi [2003J], who use an expected policyholder deficit analyses to allocate company capital to lines of business, not to determine the required capital for the company itself.

[28] See Feldblum and Thandi [2002J] for a complete treatment of this subject.

[29] Financial analysts sometimes differentiate between the cost of internal equity (retained earnings) and the cost of external equity (new stock floatation). The difference is the floatation costs of a new stock issue. For simplicity, we consider only the cost of internal equity.

[30] The effect of double taxation is mitigated if the implied equity flow from the insurance company is in the form of capital gains instead of stockholder dividends from the insurance company. This topic is treated in the general finance literature.

[31] In addition to the cost discussed in the text, some analysts argue that corporate managers with discretionary control over excess capital may not invest it solely in the interests of equityholders. They cite examples from various industries, such as the oil industry in the 1970's, to show that managers often use excess capital to increase market share at the expense of profitability. Investors may be reluctant to provide more capital than is essential for the company's operations. This is particularly relevant to those analysts who believe the insurance industry is over-capitalized.

[32] Some financial analysts retort that the lower risk of the Treasury securities increases the present value of their returns to a level approximately equal to the return on other securities. This assumes that equityholders consider the risk of placing their capital in an insurance enterprise that invests in risk-free securities similar to the risk of investing in those risk-free securities themselves. The alternative perspective is that the risk of placing capital in an insurance enterprise, regardless of its investment policy, is an equity risk. There are sound arguments on both sides, and we do not judge the issue here; see Miccolis [1987].

[33] There are additional taxes paid for the double discounting of loss reserves if held reserves are less than full value reserves and for personal income taxes paid by the equityholders; see Feldblum and Thandi [2003D].

[34] The deferral of the tax on capital gains until the gains are realized lowers the effective tax rate; see Jeffrey [1995] or Feldblum and Thandi [2003G]

[35] Debt financing has one layer of federal income taxes (only personal); equity financing has two layers (corporate and personal).

Credibility Theory for Dummies

Gary G. Venter, FCAS, MAAA

CREDIBILITY THEORY FOR DUMMIES

Gary G Venter

Guy Carpenter Instrat

Least squares credibility is usually derived from some fairly complicated looking assumptions about risk across a collective. It turns out, however, that the basic results can be developed from some standard statistical operations with weighted regression. This is outlined, and some more advanced models are tied to the same approach, in this note.

Credibility Theory for Dummies

Credibility theory is usually presented as a mathematically dense body of formulas. Here is something a little different: a short, simple approach. "Dummies" is of course a relative term. Algebra, differential calculus, and some background in statistics are all assumed.

What is credibility?

Credibility theory is all about weighted averages. Different estimates of a quantity are to be weighted together. The more credible estimates get more weight.

In the context of estimating expected losses for a member of a class, there are two natural estimates: the experience of the member itself, and the average of the entire class. The former is more relevant but also more volatile than the latter. Two general approaches have been taken to calculating weights in this case. The limited fluctuation approach is willing to accept the member experience at face value if it meets a pre-defined standard of stability (full credibility) and if not reduces the weight enough for the weighted average to meet the stability requirement. The greatest accuracy approach measures relevance as well as stability and looks for the weights that will minimize an error measure. The average of the entire class could be a very stable quantity, but if the members of the class tend to be quite different from each other, it could be of less relevance for any particular class. So the relevance of a wider class average to any member's mean is inversely related to the variability among the members of the class.

The error measure used in the greatest accuracy approach is almost always expected squared error, so this method is often called "least squares credibility." In Europe it is sometimes called "classical credibility." The limited fluctuation approach is called classical in North America. Thus "classical" is a term worth avoiding, not only because of its geographic ambiguity, but also because it is a historical rather than a methodological description.

Least squares credibility

Suppose you have two independent estimates x and y of a quantity, with respective expected squared errors u and v. Take a weighted average $a = zx + (1-z)y$. The expected squared error of a is $w = z^2u + (1-z)^2v$. What z minimizes w? Here is where the calculus comes in. The derivative dw/dz is $2zu + 2(z-1)v$. If you set that to zero you get: $zu+zv = v$, or $z = v/(u+v)$. Then $1-z$ is $u/(u+v)$. This makes it look like each estimate gets a weight proportional to the expected squared error of the other. To express the weights as properties of the estimates themselves, note that $(1/u)/[1/u + 1/v] = 1/[1+u/v] = v/(u+v) = z$. This shows that each estimate gets a weight proportional to the reciprocal of its expected squared error[1]. Least squares credibility is an application of this principle.

As an example, consider a class of risks. Suppose the losses L_{ij} in year j for the ith member of the class are randomly distributed as follows:

$$L_{ij} = C + M_i + \varepsilon_{ij} \qquad (1)$$

where C is the class mean loss, $C+ M_i$ is the mean loss for the ith member, and ε_{ij} is the random component for the jth period for this member. It is not much of a restriction to assume that the M_i's average to zero as do the ε_{ij}'s. Suppose the variance of the M_i's is t^2 and the variance of the random components ε_{ij} all are s_i^2. Denote their average $E(s_i^2)$ by s^2.

Sometimes t^2 is called the variance of the hypothetical means and s^2 the expected process variance. "Hypothetical" refers to the fact that the means $C + M_i$ are not observed.

With this setup, consider two estimates of member i mean losses: x, the average losses of the member for n periods, and y, the class mean loss C, which for now we will assume to know or at least be able to estimate well enough to ignore the error. To apply the inverse variance weightings, we needed to know the expected squared errors of x and y from the true value of $C + M_i$. By the definitions, y's expected squared error is just t^2. The expected squared error of x is the expected value of its variance s_i^2/n, i.e., s^2/n. Then applying the inverse expected squared error principle gives a weight to x of $z = (n/s^2)/[n/ s^2 + 1/t^2] = n/[n + s^2/t^2]$. This is the original Bühlmann credibility formula.

---

[1] This assumes the expected squared error is minimized rather than maximized at this z. The second derivative of w is $2u + 2v$ which is positive, so this assumption is valid.

The above would be an appropriate set of assumptions for a class where all members had roughly the same exposure, such as single cars. If the exposure varies much across members, like in territory ratemaking or commercial experience rating, the variances of the random components could not reasonably be assumed to be constant over time. To address this case, introduce an exposure measure P_{ij} for the ith member in period j, and assume that the variance of its random loss component is $P_{ij}s_i^2$, so each unit of exposure has a variance of s_i^2. In this case it would not be right to assume that M_i has mean zero, in that different members of the class would depart from the class mean loss in differing amounts depending on exposure. However, if in equation (1) L is reinterpreted as losses per unit of exposure, i.e., pure premium, this assumption could be reasonable. In that case, the variance of ε_{ij} would be s_i^2/P_{ij}. So here, x is the average loss per exposure for the ith member for n periods, and y is the mean pure premium for the class.

Thus the expected squared error of y from $C + M_i$ would still be t^2. Assume further that x is calculated as the sum of the n period losses divided by the sum of the exposures. Use a "~" in a subscript to denote summation, so the total exposures for the ith class over the n periods is $P_{i\sim}$. Then the variance of x is just $P_{i\sim}s_i^2/P_{i\sim}^2 = s_i^2/P_{i\sim}$, with expected value $s^2/P_{i\sim}$. So what is the credibility of the pure premium? The inverse expected squared error weighting gives $z_i = (P_{i\sim}/s^2)/[P_{i\sim}/s^2 + t^{-2}] = P_{i\sim}/[P_{i\sim} + s^2/t^2]$. This is often expressed more simply as $z = P/[P+K]$, which is the Bühlmann-Straub credibility formula.

C can be estimated by a weighted average of the x's, the member means. The expected squared error of x from C is $t^2 + s^2/P_{i\sim}$, so x should get a weight inversely proportional to that, so proportional to $t^{-2}P_{i\sim}/[P_{i\sim} + s^2/t^2]$, which is proportional to z_i. Thus C can be estimated as a weighted average of the x's where the weights for each member are proportional to the member's credibility.

What has been lost by the simplified approach? First, instead of (1), L_{ij} is often considered to be a conditional process with a parameter, say q_i .and a conditional mean and variance given the parameter. The conditional means are assumed to average to the class mean C with a variance t^2 and the conditional variances average to s^2. Then defining M_i as the conditional mean less C is equivalent to the additive formulation (1). However the full usual derivation gets an additional result: a weighted average of member means is the best linear combination of any sort of the individual member observations by time period. However, this is a fairly general statement itself, and it might be true in general separate from the credibility formulation.

Both this formulation and the usual credibility derivation ignore the estimation error for C in the credibility formula. Empirical Bayes theory addresses this issue, which does make a difference in small samples. It might be possible to get the empirical Bayes results from the inverse squared error principle as well.

Beyond Bühlmann-Straub: Large vs. small risk differences

The assumption that each unit of exposure generates the same amount of loss variance is sometimes described as assuming that a large risk behaves like an independent combination of small risks. Hewitt in his 1967 paper presented some data showing this was not the case[2]. Actually large risks have more variance than would be expected from treating them as independent combinations of smaller risks. One thing that contributes to this is that risk conditions change over time. Size of exposure does not provide much stability against changing economic and business sector changes. A way to model this would be to assume that the variance of the observed loss for each risk for each period has the usual component that increases with risk size plus another component that increases with risk size squared, i.e., assume that the loss variance is $P_{ij}^2u^2 + P_{ij}s^2$. Then the variance of the pure premium would be $u^2 + s^2/P_{ij}$.

The credibility formula now gets more complicated, but is not too bad in the special case where there is just one time period. With the inverse expected squared error formula, $z = [P_{i\sim}/(P_{i\sim}u^2+s^2)]/[P_{i\sim}/(P_{i\sim}u^2+s^2)+t^{-2}] = P_{i\sim}/[P_{i\sim} + P_{i\sim}u^2/t^2 + s^2/t^2]$. This could be written as $z = P/[P + AP + K]$. For larger values of P this makes the denominator larger, so decreases the credibility compared to $P/[P+K]$.

In this case risk stability is a more complicated function of exposure than in the original model. In experience rating workers compensation another phenomenon has sometimes been observed:

---

[2] *Loss Ratio Distributions – A Model*, PCAS LIV.

the large risks' mean loss exposures are less different from the overall mean than are the small risks'. This could be a matter of regulation, where large risks must follow more safety precautions, but other reasons are possible. Whatever causes this phenomenon, the result is that the variance of M_i (i.e., the variance among risk means) also becomes a function of exposure. Since it is the smaller risks that have more potential for large departures from the overall average pure premium, this average becomes less relevant for the small risks, which increases the credibility of their own experience. A reasonable formula for the variance among risk means in this situation might be $t^2+v^2/P_{i\sim}$ in the single time period case for member i. Suppressing the subscripts on P, z becomes $z = [P/(Pu^2+s^2)]/[P/(Pu^2+s^2)+P/(Pt^2+v^2)] = (Pt^2+v^2)/[Pu^2+s^2+Pt^2+v^2]$. This can be simplified to $z = (P + B)/(P + AP + K + B)$. The extra B in the numerator and denominator increases z, especially for smaller risks where P is smaller, which is what was anticipated.

When linear estimates don't work

So far this discussion has been non-parametric. That is, the forms of the distributions have not entered in. That is the advantage of linear estimates with squared error penalties. If you have some information about the type of distribution available, you can give up the restriction to linear functions. In a Bayesian framework the class experience becomes the prior distribution for the member experience, and then the Bayesian conditional expected value of the member mean given the data is the least squares estimator of the member mean of any sort, linear or not. In some cases the conditional mean is a linear function of the data (e.g., normal and gamma distributions) so the linear restriction of credibility theory does not reduce the accuracy. However in highly skewed distributions, like some lognormal cases, the Bayes estimate is highly non-linear, and credibility weighting can give large errors for classes with small means.

If the distribution type is fairly well understood, Bayesian methods would be preferable in such cases. However, an alternative when the member means can be very different from each other is to do the usual credibility estimation in the logs of the data, then exponentiate the results. This introduces a downward bias, however, which has to be adjusted multiplicatively to balance to the overall data.

Effects of Parameters of Tranformed Beta Distributions

Gary G. Venter, FCAS, MAAA

Effects of Parameters of Transformed Beta Distributions

Gary G Venter

Guy Carpenter Instrat

The transformed beta distribution was introduced to the insurance literature in Venter(1983) and independently to the economics literature in McDonald (1984). The parameterization discussed here was introduced by Rodney Kreps in order to make the parameters more independent of each other in the estimation process. The resulting parameters have somewhat separate roles in determining the shape of the distribution, and this note examines those effects.

Effects of Parameters of Transformed Beta Distributions

The transformed beta is considered parameterized so that $f(x) \propto (x/d)^{b-1}(1+(x/d)^c)^{-(a+b)/c}$. Each of the parameters will be considered in alphabetical order. In general terms, a determines the heaviness of the tail, b the shape of the distribution and the behavior near zero, c moves the middle around, and d is a scale parameter.

a

All positive moments $E(X^k)$ exist for $k < a$, but not otherwise. Thus a determines the heaviness of the tail. One way to measure tail heaviness is to look at the ratio of a high percentile to the median. For a large company, say with 50,000 expected claims, a pretty large claim would be one of the five largest – say the 1/10,000 probability claim.

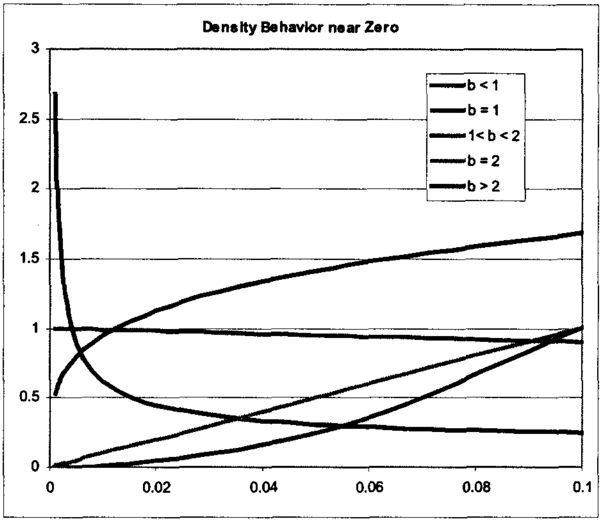
The ballasted Pareto $F(x) = 1 - (1+ x/d)^{-a}$ will be used to illustrate tail heaviness, as it is easy to deal with and has essentially the same tail heaviness as the general case. If B is a (big) number, the 1 - 1/Bth percentile is $d(B^{1/a} - 1)$. For the example with B=10,000, this is $d(10^{4/a} - 1)$. The ratio of this to the median (B=2) is thus $(10^{4/a} - 1)/(2^{1/a} - 1)$. This is very sensitive to a, especially for a between 1 and 2, where it often is. The ratio of pretty large claim to median in this case is 9254 at a=1.01 down to 788 for a=1.5. Thus the estimate of a could have a big impact on excess losses.

b

Negative moments $E((1/X)^k)$ exist for $k<b$ and not otherwise. This parameter governs the behavior of the distribution near 0. In that region, the density is close to constant*$(x/d)^{b-1}$so the derivative of the density is proportional to $(b-1)x^{b-2}$. This can be used to ascertain the shape of the density for smaller claims, which really determines the overall shape of the distribution.

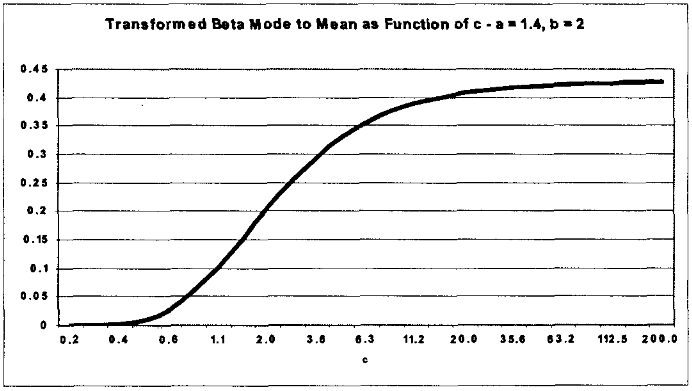
If $b < 1$, the slope of the density at zero is negative infinity, so the density is asymptotic to the vertical axis. For b=1, the other factor in the density becomes significant, and the slope is a negative number. The mode of the distribution is at zero in both of these cases. For $1 < b < 2$, the slope at zero is positive infinity, so the density is rising and tangent to the vertical axis. For b=2, the slope is a positive number, and for $b > 2$, the slope is zero, so the density is tangent to the horizontal axis. For $b > 1$, then, there is a positive mode.

This graph show the behavior near zero and how that depends on b. Since the right tail is an inverse power curve, the behavior near zero determines the overall look of the distribution. The case b > 2 gives the usual shape of a density function people think of, which rises gradually then more steeply before falling off with the inverse power relationship. The transformed beta in this case looks like a heavy tailed lognormal. The case b = 1 is also seen a lot, for instance in the exponential and ballasted Pareto distributions.

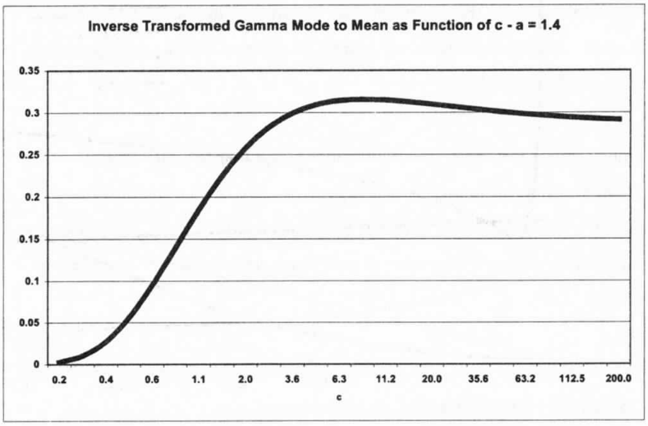


c

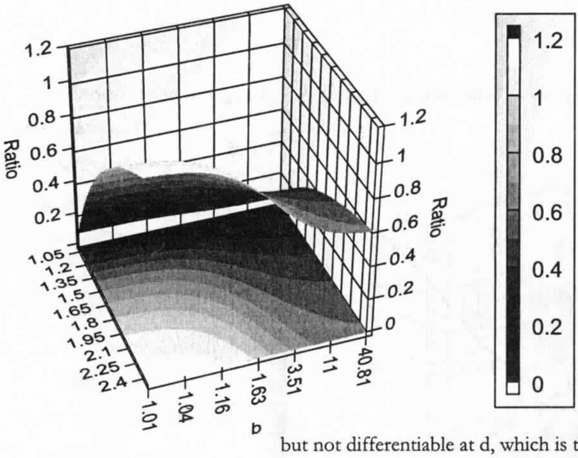
The c parameter introduces a power transform x –> x^c into the transformed beta. This tends to move the middle of the distribution around. A useful measure of where the middle is is the mode, as related to the mean. The ratio of mode to mean, when the mean exists (a > 1) and the mode is positive (b > 1), is for the most part an



increasing function of c for fixed a and b. The graph above shows the case a=1.4, b=2, and the graph below shows the same a for the inverse transformed gamma, which is the limit of the transformed beta as b goes to infinity. This is a high enough a to have some cases with a



reasonably high mode, say 40% of the mean, but a is still small enough to be of potential use in US liability insurance. For b=2, the ratio is a strictly increasing function of c. For the limiting case, the ratio reaches a peak and then declines slightly after that. This will also be the behavior for large values of b.



Thus something near the highest value of the mode-to-mean ratio for a given a and b is provided by the limit of the transformed beta when c goes to infinity, which is in fact the split simple Pareto distribution. Its density f(x) is proportional to $(x/d)^{b-1}$ for x<d and to $(x/d)^{-a-1}$ for x>d. The density is continuous but not differentiable at d, which is the mode when b > 1. The mean is dab/[(b+1)(a-1)]. Thus the ratio of mode to mean is (a-1)(b+1)/ab = (1–1/a)(1+1/b). This is increasing in a and decreasing in b. The graph above shows this ratio for a from 1.05 to

2.5 and b-1 from 0.01 to 40 on a geometric scale.

For low values of a, the ratio cannot get very high, as the mean is increased by the heavy tail. The ratio for this distribution is close to the upper limit for the transformed beta with the same a and b, so for low values of a, the c parameter is not going to be able to have much effect on the mode for any transformed beta distribution.

The ratio declines for increasing b, but rather slowly. It is interesting that for this limiting distribution, the maximum mode-to-mean ratio is as b approaches 1, while for the transformed beta the mode is zero at b=1. The split simple Pareto at b=1 is the uniform Pareto, which is uniform up to d and Pareto after that. Thus its mode is undefined, or it could be considered to be the whole interval [0,b].

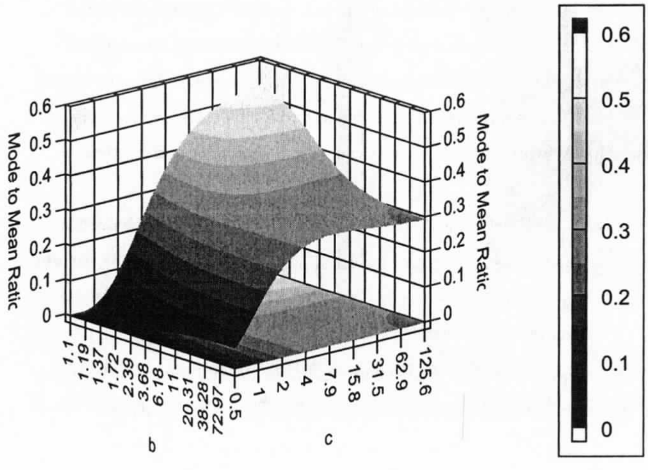
The split simple Pareto shows the maximum, and thus the range of mode-to-mean ratios for any a and b. How is this ratio affected by c? The transformed beta mean is:

$d\Gamma(b/c+1/c)\Gamma(a/c-1/c)/[\Gamma(a/c)\Gamma(b/c)]$, for $a > 1$, and the mode is:

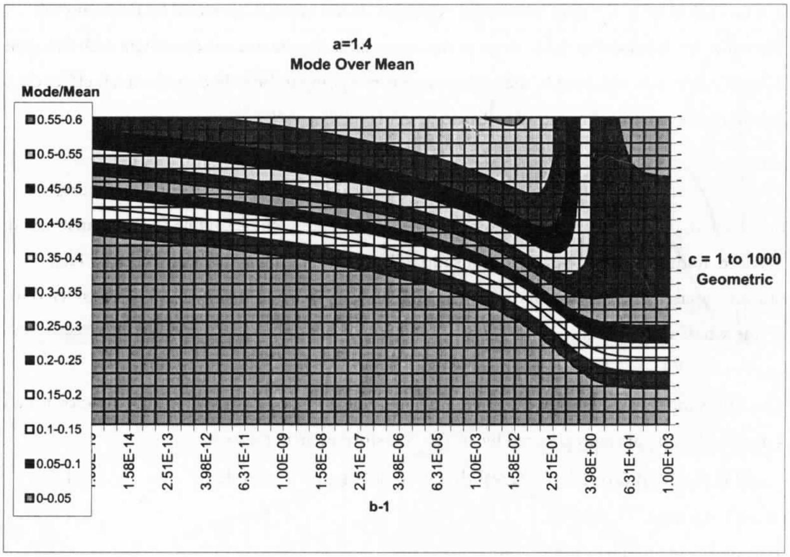
$d[(b-1)/(a+1)]^{1/c}$, for $b>1$

This makes the mode to mean ratio:

$[(b-1)/(a+1)]^{1/c}\Gamma(a/c)\Gamma(b/c) / [\Gamma(b/c+1/c)\Gamma(a/c-1/c)]$ for $a, b >1$. The graph below shows the ratio for a=1.4 and a range of b's. The high values of b can be seen to have a relationship to c similar to that for the inverse transformed gamma, with a decrease in the ratio for higher values of c. The



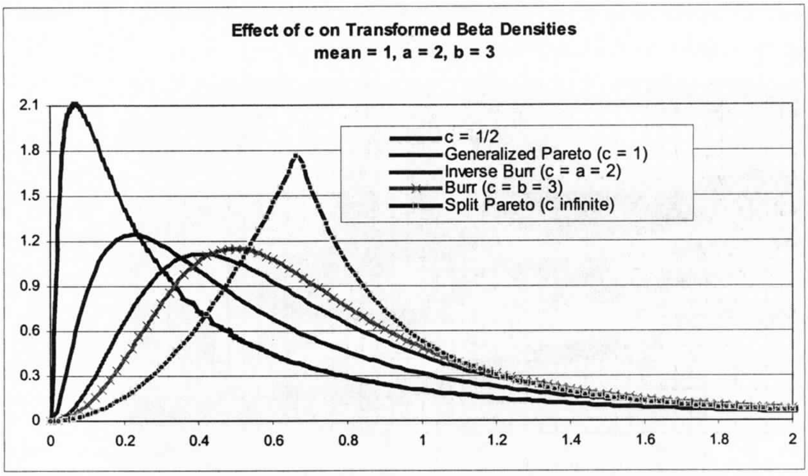
lower values of b have a strictly increasing function of c, like the case b=2 above. The contours of this surface are shown below for a wider range of values.



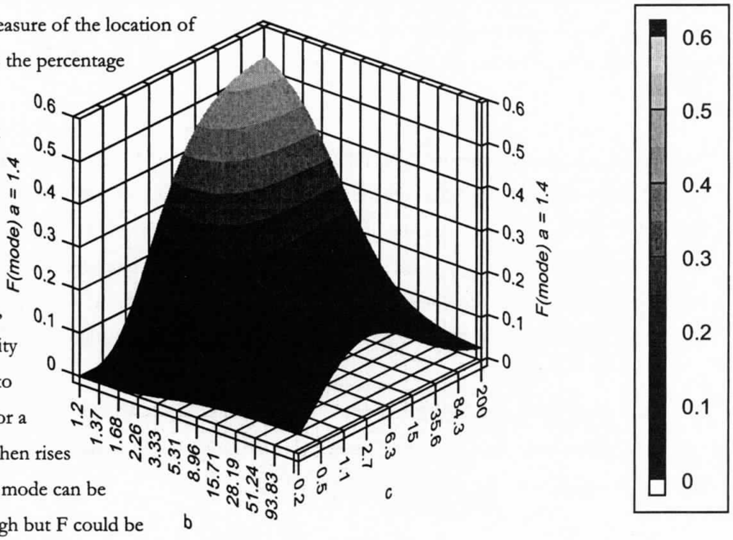
The vertical lines are the contours of the function for fixed values of b, like what was graphed for b=2 above. The horizontal lines are the contours of the function for fixed values of c. For the smaller c's the ratio is an increasing function of b, but very slowly increasing, so b has little impact. For larger c's the function increases then decreases. The maximum seems to hit fairly early, like around 1.01 to 1.25. This is somewhat surprising, in that for b=1, the mode is zero. Thus the mode increases rapidly for b just above 1, especially for higher values of c.

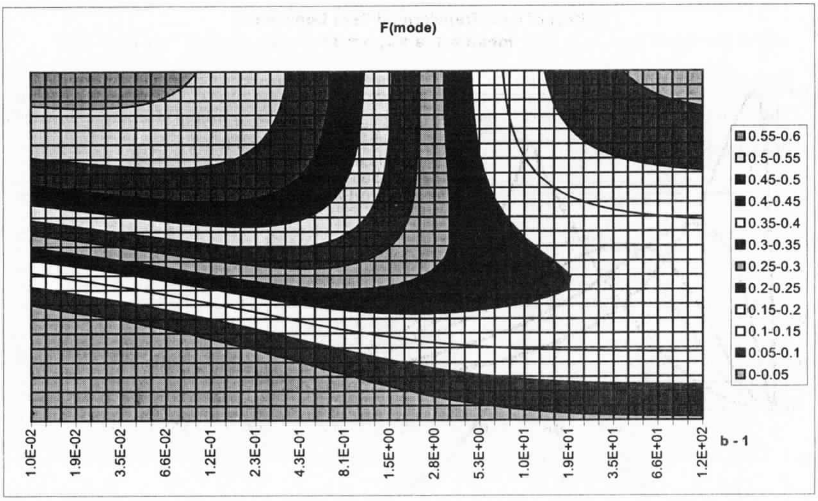
The graph above starts at b–1= 10^{-15}, which is the smallest value for which Excel can do this calculation. Even at this level, higher values of c give modes substantially above zero. The mode is above 1% of the mean for c as low as 9 for b–1= 10^{-15}.

To illustrate the effect of the c parameter, and thus the mode, on the density function, several cases are illustrated on the graph below. All the distributions have a=2, b=3, and mean=1.

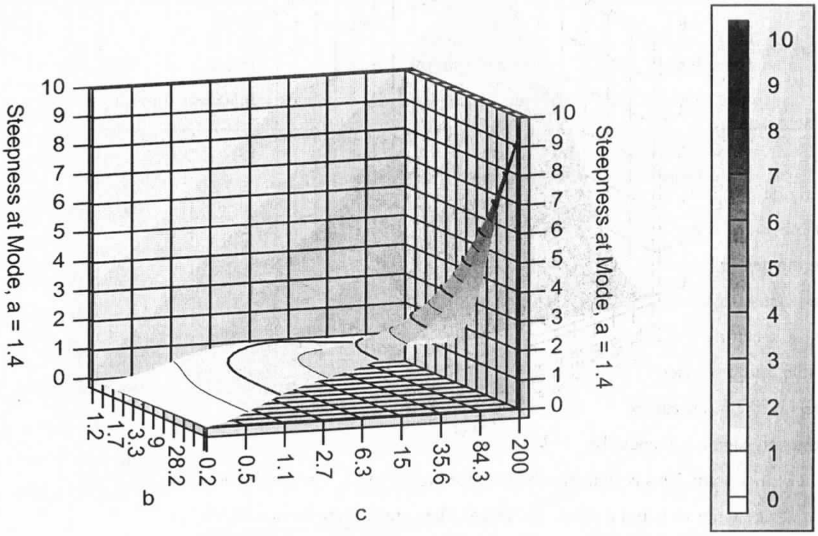


Another measure of the location of the mode is the percentage of the distribution that is below it, i.e., F(mode). If b is high, so the density stays close to the x-axis for a while, and then rises steeply, the mode can be relatively high but F could be low at that point. This is generally the case for high b and c. The graph above shows F(mode) for a wide range of b and c values for a=1.4. The contours are shown below.

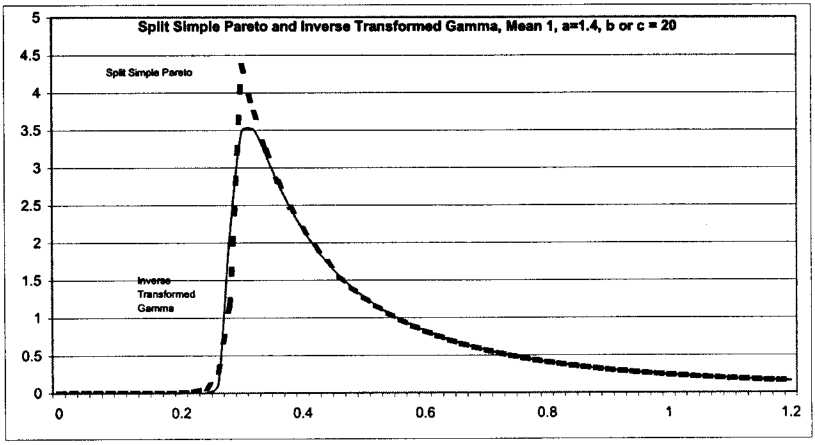




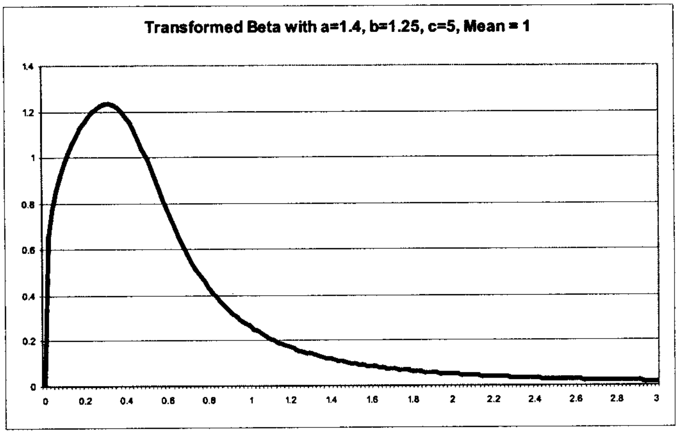
This has a very similar shape to the mode-to-mean ratio. Dividing that by this gives an indication of how steep is the distribution just below the mode. A graph of the steepness is below.



The steepness measure is between zero and two for most of the range of b and c values. It is only in the upper right, with high values of b and c, that the steepness gets very high. The limiting distribution where both b and c go to infinity is the simple Pareto: $F(x) = 1 - (d/x)^a$. This is both the limit of the inverse transformed gamma as c goes to infinity and the split simple Pareto as b goes to infinity.



Examples of both that goes towards the simple Pareto limit are graphed above. Both show a steep rise to the mode. For the split simple Pareto, the mode is at 0.30, and F(mode) = 0.065. This is closer to the limiting case of 0.3 and 0 than is the inverse transformed gamma, with mode of 0.31 and F(mode) = 0.11.

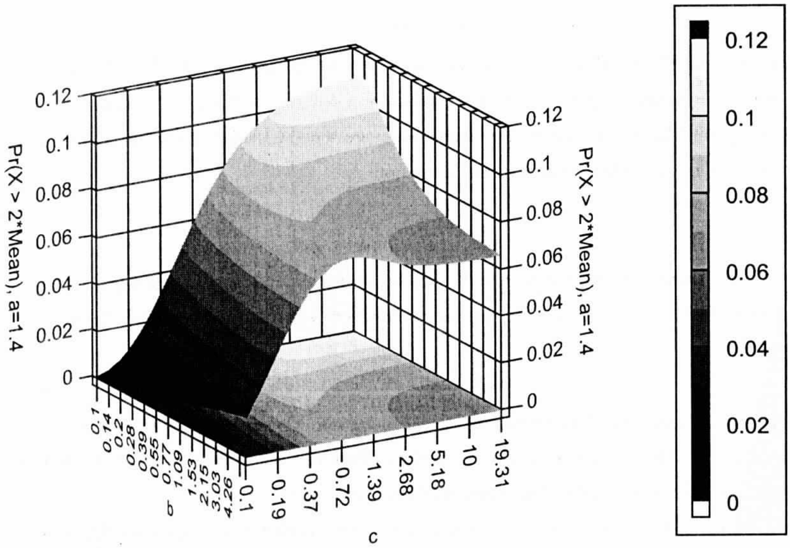


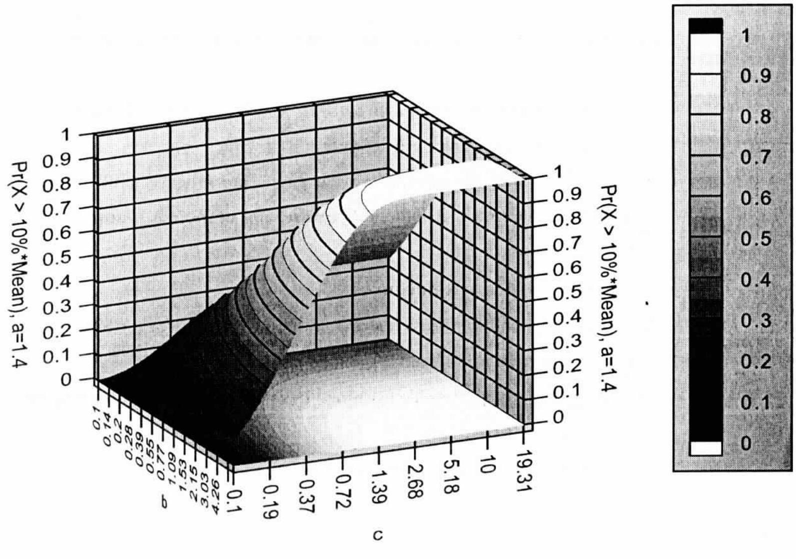
In contrast, a low b with a high c puts a lot more of the probability below the mode. The graph here shows the case b=1.25, c=5, which also has a mode of 0.31 but a higher F(mode) of 0.32. It is clear from the graph that a lot more of the distribution is below the mode in this case.

Another effect of low values of c can be to increase the tails, even though this might not show up in moments. An interesting example is the Weibull distribution, for which a is infinite so all positive moments exist, and b=c. Taking b=c=0.2 gives a fairly heavy-tailed distribution for which all positive moments exist. This has been traditionally used in the US workers compensation line. As an example, take d=100, with a mean of 12,000. In this case, the pretty large loss – 1-in-10,000 claim – is 6.6M, or 550 times the mean. This is heavier-tailed by this measure than most Pareto distributions. For instance, with a=1.4, this ratio is 287. The cv^2 for this Weibull distribution is 251, so the cv itself is almost 16. For contrast, a lognormal with the same mean and cv would have the 1-in-10,000 loss about 4,750,000. The Weibull has another strange feature, however. As b is so small, negative moments do not exist except for powers closer to zero than –0.2. This means that a lot of the distribution is packed in towards zero. In fact, about 33% of the claims are less than 1, and the median claim is 16. The comparable lognormal, which can be given in the limit of a and b both going to infinity, has only 0.2% of its claims less than 1, even though the mode is 3. The median claim is 756 for this distribution.

A small c can pump up the tail of the transformed beta as well. For instance, taking a=1.4, and b=c=0.2 gives a Burr distribution where the 1-in-10,000 claim is 910 times the mean, and over 50% of the claims are below $1/12{,}000^{th}$ of the mean. Keeping this value of c, but letting b get larger, can allow the pretty large claim to be a high multiple of the mean without so many small claims. For instance, taking b=1 (which gives the Pareto T), the pretty large loss is 795 times the mean, and only 7% of claims are below $1 when the mean is $12,000. Taking b up to 5, keeping a=1.4 and c=0.2, these numbers come down to 673 times and 0.1%, which is still very heavy tailed without pushing so many claims to unrealistically small sizes. Although this distribution has a positive mode, it is at 0.065% of the mean, or 7.8 for a mean of 12,000, so is close to zero.

To get an idea of how b and c influence the tail heaviness, the probability that a loss is greater than twice the mean is shown by b and c for a=1.4 below. The graph after that shows the probability of being greater than 10% of the mean.


Pr(X > 2*Mean), a=1.4
b
c
0.1
0.14
0.2
0.28
0.39
0.55
0.77
1.09
1.53
2.15
3.03
4.26
0.19
0.37
0.72
1.39
2.68
5.18
10
19.31
0
0.02
0.04
0.06
0.08
0.1
0.12


Pr(X > 10%*Mean), a=1.4
b
c
0.1
0.14
0.2
0.28
0.39
0.55
0.77
1.09
1.53
2.15
3.03
4.26
0.19
0.37
0.72
1.39
2.68
5.18
10
19.31
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

d

The parameter d is a scaling factor. It's effect is just like re-scaling the x-axis. For instance, to convert a distribution expressed in pounds to Canadian dollars, just multiply the scale parameter by 3 (typically). Then a probability for an amount expressed in Canadian dollars would be the same as for the equivalent amount expressed in pounds.

Where did b and c go?

Several two-parameter cases of the transformed beta have just the a and d parameters. To understand what they are doing, it is helpful to know how b and c were disposed of. Some examples:

Ballasted Pareto: b=c=1, so moments in (-1,a), mode zero. Closed form and invertible.
Loglogistic: a=b=c, so moments in (-a,a), and thus the mode is positive if the mean exists, but is probably pretty small with a low steepness. Closed form and invertible.
Inverse Weibull: b infinite, c=a, also closed form and invertible for simulation. Mode is always positive .
Inverse Gamma: b infinite, c=1, so mode positive but usually less than for inverse Weibull. Not closed form.
Simple Pareto: b=c=infinity, so positive mode, infinite steepness, F(mode) = 0. The opposite extreme from the ballasted Pareto for b and c. Invertible.
Uniform Pareto: c infinite, b=1. Mode ambiguous – whole range from 0 to d is uniform. Intermediate between ballasted and simple Paretos and mirror image of inverse gamma in parameters. Invertible.

References

James McDonald (james_mcdonald@byu.edu) Some Generalized Functions for the Size Distribution of Income. Econometrica, 1984, vol. 52, issue 3, pages 647-63
Gary G. Venter (gary.g.venter@guycarp.com) Transformed Beta and Gamma Distributions and Aggregate Losses. PCAS LXX, 1983 pages 156-93
Rodney Kreps (rodney.e.kreps@guycarp.com) Continuous Distributions.doc

MLE for Claims with Several Retentions

Gary G. Venter, FCAS, MAAA

MLE for Claims with Several Retentions

Gary G Venter, Guy Carpenter Instrat

In the *Loss Models* readings, CAS students learn how to fit severity distributions by MLE, including the case of fitting a ground-up distribution where only losses above a deductible are available. In that case the MLE looks for the ground-up distribution parameters that provide the best fit to the known excess losses. This procedure falls apart, however, when different deductibles are used and there are different degrees of exposure to each. This note derives the likelihood function for that situation.

Acknowledgment

This problem was posed by Claude Yoder, who recognized the shortcomings of the usual method. The likelihood function was worked out one day after lunch by Rodney Kreps, Paul Silberbush, John Pagliaccio and the author. The usual caveat applies.

MLE for Claims with Several Retentions

A not atypical fitting problem for reinsurance losses is trying to find the severity distribution that is generating claims, where the data is provided for groups of excess policies, each with its own retention and limit. The case considered here is where information is also presented about how much exposure is included in each group. Losses are of course truncated from below at their retentions and censored from above at their limits.

What is the likelihood function for severity in this situation? It turns out that that question is intrinsically linked with the likelihood function for frequency, as the exposure information comes in through frequency. First some notation. To simplify the typing and also to increase the visibility of the sub-variables, subscripts will not be used.

So suppose you have k groups of claims, and the jth group has retention Rj, upper limit or plafond (i.e., retention plus policy limit) Uj, and Ej exposures. The data for the group consists of Mj claims at the policy limit, (some of which are probably censored by the limit, so would have been larger without the limit) and Nj claims less than the limit. All the ground-up claim sizes are assumed to come from the same distribution, with severity distribution function F and density f. The exposures for all the groups are assumed to have the same ground-up Poisson loss frequency h per unit of exposure, so the observed frequency for the jth group is hEj(1-F(Rj)), which will be denoted by hj, and is still Poisson. Estimating h is part of the problem to be addressed.

With this setup, what is the likelihood function for the set of losses observed in the jth group? This is the product of the frequency and severity probabilities of observing that many claims of those sizes. Let **a** denote the severity parameters, considered to be a vector, and Xji the ground-up amount of the ith loss in the jth group. Then the likelihood function at h and **a** for the jth group is:

$$Lj(h,\mathbf{a}) = hj^{Nj+Mj}\exp(-hj)[\Pi_{i=1}{}^{Nj}f(Xji\,|\,\mathbf{a})][1-F(Uj\,|\,\mathbf{a})]^{Mj}$$

The log-likelihood for all the groups combined is the log of the product of these:

(1) $$LL(h,\mathbf{a}) = \Sigma_{j=1}{}^{k}\{\ln(hj)(Nj+Mj) - hj + \Sigma_{i=1}{}^{Nj}\ln[f(Xji\,|\,\mathbf{a})] + Mj\ln[1-F(Uj\,|\,\mathbf{a})]\}$$

Since hj is a function of F, this cannot be separated into frequency and severity sections.

Formula (1) is the answer, in the sense that this is the function that has to be maximized to estimate h and **a**. Some insight can be gained by considering its partials, however. First, wrt h:

$\partial LL/\partial h = \Sigma_{j=1}^{k}\{(Nj+Mj)/h - Ej(1-F(Rj\,|\,\mathbf{a}))\}$, which setting to zero gives:

$$h = \Sigma_{j=1}^{k}(Nj+Mj) \;/\; \Sigma_{j=1}^{k}(Ej(1-F(Rj\,|\,\mathbf{a}))$$

This gives the ground up frequency in terms of the severity parameters.

The partial of the LL wrt **a** can be seen to have two components – the partial of the first two terms is a frequency component, and the partial of the second two is the usual severity component that does not consider exposures. If these separately become zero at the maximum, then the exposure information is not affecting the parameters. But it can be shown that for this to happen, a different value of h would result. Thus the exposure information does make a difference.

A large number of exposures with high retentions would be expected to produce several large ground-up claims, but no small ones. But if retentions are small, the same sample would suggest that the severity distribution tends to produce larger claims. Thus including the exposure information should make a difference of this kind. Some practical testing of MLE with this LL could discern if this is the case.

Testing Stochastic Interest Rate Generators for Insurer Risk and Capital Models

Gary G. Venter, FCAS, MAAA

Testing Stochastic Interest Rate Generators for Insurer Risk and Capital Models

Gary G Venter, Guy Carpenter Instrat

Stochastic models for interest rates are reviewed and fitting methods are discussed. Tests for the dynamics of short-term rates are based on model fits. A method of testing yield curve distributions for use in insurer asset scenario generators is introduced. This uses historical relationships in the conditional distributions of yield spreads given the short-term rate. As an illustration, this method is used to test a few selected models.

Acknowledgement

I would like to acknowledge the invaluable assistance of Andrei Salomatov in both preparing the exhibits and helping to develop the methodology for this paper.

Testing Stochastic Interest Rate Generators for Insurer Risk and Capital Models

P&C insurers are looking to financial modeling to address how risk is diversified among assets, liabilities and current underwriting results. Before fast computer models, actuaries measured asset risk by a few simple constants, like duration and convexity. Asset managers have their own collection of risk measurement constants for hedging issues, identified by Greek letters, and so often referred to as "the Greeks." Appendix 1 summarizes these measures.

These asset risk scalars typically measure the sensitivity of the asset portfolio to changes in some particular risk event, such as a change in the average interest rate, or a change in the volatility of interest rates. With stochastic generators, however, two degrees of specificity are added. First the dimension of probability of risk events is incorporated. Risk scalars show the sensitivity to a change but not the probability of that change. Second is the response to a much broader range of possible risks. Complex combinations of risk situations can occur, and stochastic modeling can quantify the combined risk picture.

These added dimensions come from representing the distribution of possible outcomes for an asset portfolio. Models can then combine asset outcomes with liability development and underwriting return outcomes to give a more comprehensive risk profile. Asset models generate a large variety of asset scenarios, ideally each showing up by the probability of its occurrence, and apply them to the asset portfolio to measure the distribution of asset risk.

Although useful and general in theory, the possible weakness of this approach is that in practice the model might not capture the full range of economic outcomes, or it could over-weight the chances of some occurrences that are in fact not all that likely to happen. Thus a significant risk to this methodology is generating the wrong distribution of financial events.

This paper looks at evaluating interest rate generators by testing the distribution of yield curves. Empirical research on the dynamics of the short-term rate is reviewed, then tests of the generated distribution of yield curves are introduced and applied to a few models. Interest rate models in other areas of finance tend to be used to price options, so they are evaluated on how well they can match option prices. Insurer models are more focused on the risks inherent in holding various asset mixtures for a period of time, and sometimes on trading strategies, and so realistic distributions of ending yield curves and probabilities for movements in yields are of more direct concerns than are option prices.

I. Models of Interest Rates

The primary focus here is on arbitrage-free models of interest rates. There is still some debate among actuaries on whether this is the best approach, and some of this debate is summarized in Appendix 2, but it is such models which will be emphasized here. The tests on the yield curve distributions introduced below, however, can be used on any model that generates yield curve scenarios. Interest rates are further assumed to be default free. Modeling default probabilities adds a degree of complexity that is not addressed here.

There are a few ways to generate arbitrage-free interest rate scenarios. The method illustrated below models the short-term interest rate, denoted by r, directly, and uses the implied behavior of r, along with market considerations, to infer the behavior of longer-term rates. For these models, r is usually treated as a continuously fluctuating process. This is somewhat of an approximation as actual trades occur at discrete times, but at scales longer than a few minutes it seems appropriate, at least during trading hours.

The most common financial models for continuous processes are based on Brownian motion. A Brownian motion has a simple definition in terms of the probabilities of outcomes over time: the change in r from the current position between time zero and time t is normally distributed with mean zero and variance $\sigma^2 t$ for some σ. In differential notation , the instantaneous change in r is expressed as $dr = \sigma dz$. Here z represents a Brownian motion with $\sigma=1$, and so its variance after a time period of length t is just t. If r also has a drift (i.e., a trend) of bt during time t, the process could be expressed as $dr = bdt + \sigma dz$.

Cox, Ingersoll and Ross (*A Theory of the Term Structure of Interest Rates* Econometrica 53 March 1985) provide a model of the motion of the short-term rate that has been widely studied. In the CIR model, r follows the following process:

$$dr = a(b - r)dt + sr^{1/2}dz.$$

Here b is the level of mean reversion. If r is above b, then the trend component is negative, and if r is below b it is positive. Thus the trend is always towards b. The speed of mean reversion is expressed by a. Note that the volatility depends on r itself, so higher short-term rates would be associated with higher volatility. Also, if r=0 there is no volatility, so the trend takes over. With r=0 the trend would be positive, so r would move to a positive value. The mean reversion combined with rate-dependent volatility thus puts a reflective barrier at r=0.

If this model were discretized it could be written:

$r_t - r_{t-1} = a(b - r_{t-1}) + sr_{t-1}^{1/2}\varepsilon$, where ε is a standard normal residual.

This is a fairly standard autoregressive model, so the CIR model can be considered a continuous analogue of an autoregressive model.

Some other models of the short rate differ from CIR only in the power of r in the dz term. The Vasicek model takes the power to be zero. Another choice is taking a power of unity.

Most of the models incorporate mean reversion, but constant mean reversion is problematic. The rates sometimes seem to gravitate towards a temporary mean for a while, then shift and revert towards some other. One way to account for this is to let the reversion mean b itself be stochastic. This can be done by adding a second stochastic equation to the model:

$$db = j(q - b)dt + wb^{1/2}dz_1$$

Here dz_1 is a second, independent standard normal variate, and so b follows a mean reverting process gravitating towards q. Again different powers can be taken for b in the stochastic term. Such two factor models are popular in actuarial literature. For instance, Hibbert, Mowbray and Turnbull in , "A Stochastic Asset Model & Calibration for Long-Term Financial Planning Purposes," Technical Report, Barrie & Hibbert Limited, use a two factor model which generalizes the Vasicek model by taking b and r both to the zero powers, so they both drop out of the stochastic terms.

The volatility can also be stochastic. For instance, Hull, J. and A. White, 1987, "The Pricing of Options on Assets with Stochastic Volatilities," The Journal of Finance, XLII, 2, pp. 281-300 consider such a model.

Combining stochastic volatility and stochastic mean reversion, Andersen and Lund (Working Paper No. 214, Northwestern University Department of Finance) use the model:

$$dr = a(b - r)dt + sr^k dz_1 \quad k>0$$

$$d\ln s^2 = c(p - \ln s^2)dt + v dz_2$$

$$db = j(q - b)dt + wb^{1/2} dz_3$$

This model uses three standard Brownian motion processes, z_1, z_2, and z_3. The volatility parameter s^2 now also varies over time, but via a mean reverting geometric Brownian motion process (i.e., Brownian motion on the log). In total there are eight parameters: a, c, j, k, p, q, v, and w and three varying factors r, b, and s. It is thus labeled a three-factor model. The power k on r in the stochastic term is a parameter that can be estimated.

2. Dynamics of Short-Term Rates – Empirical Findings

Estimation of model parameters should be distinguished from calibration to current states. The permanent parameters of the models are estimated from historical data, whereas the variable factors are re-calibrated to current yield curves to capture the latest market conditions. Different techniques might be used for estimation vs. calibration.

Multi-factor Brownian motion models can be difficult to estimate. Some single-factor models, such as CIR, can be can be integrated out to form a time series, which can be estimated by maximum likelihood. In the case of CIR, the conditional distribution of the short rate at time t+T given the rate at time t follows a non-central chi-squared distribution:

$f(r_{t+T} \mid r_t) = ce^{-u-v}(v/u)^{q/2} I_q(2(uv)^{1/2})$, where

$c = 2as^{-2}/(1-e^{-aT})$, $q=-1+2abs^{-2}$, $u=cr_t e^{-aT}$, $v=cr_{t+T}$ and I_q is the modified Bessel function of the first kind, order q, $I_q(2z)= \sum_{k=0}^{\infty} z^{2k+q}/[k!(q+k)!]$, where factorial off integers is defined by the gamma function

This is not usually possible for multi-factor models, where the volatility and other factors can change stochastically. Further, the short-term rate is observed, or is closely related to observed rates for very short terms, but the other factors, like the reverting mean and the volatility scalar, are not typically observed. Thus fitting techniques that match models to data will not be applicable for these factors.

A few fitting techniques have been developed for stochastic processes. The general topic of what these techniques are and how they work is beyond the scope of this paper, but one method which has been used successfully – the efficient method of moments (EMM) – is briefly discussed below. This method was introduced by Gallant, A. and Tauchen, G.: 1996, Which moments to match?, Econometric Theory 12, 657-681, and they provide further analysis in 1999, The relative efficiency of method of moments estimators, Journal of Econometrics 92 (1999) 149-172. However the optimal methodology for estimating models of this type is far from settled.

In any case, EMM is a special case of GMM, the generalized method of moments. A generalized moment is any quantity that can be averaged over a data set, such as $(3/x)\ln x$. GMM fits a model by matching the modeled and empirical generalized moments for some selection of generalized moments. EMM is a particular choice of generalized moments that has some favorable statistical properties when used to fit stochastic models.

EMM for a particular data set starts by finding the best time series model, called the auxiliary model, that can be fit to that data. If the auxiliary model is fit by maximum likelihood, then the scores of that model (i.e., the first partial derivatives of the log-likelihood function with respect to each model parameter)will be zero at the MLE estimates. These score functions can be viewed as generalized moments, which are all zero when averaged over the data. The fitted value of the scores of the auxiliary model might be hard to calculate for the stochastic model, but they can be approximated numerically by simulating a large sample from the stochastic model, and computing the scores of the auxiliary model for that sample. The parameters of the stochastic model can then be adjusted to match these moments, i.e., until all the scores approximate zero for the generated data.

The result of this technique is a parameterized stochastic model whose simulated values have all the same dynamics as the data, as far as anyone can tell by fitting time-series models to both. With this fitting done, the modeled factors then can be calibrated to current economic conditions to provide a basis for simulating future possible outcomes.

Andersen and Lund (AL) did an empirical study of short-term rate dynamics by EMM-fitting their above model to four decades of US Treasury notes, incorporating data from the 1950's through the 1990's. Their results provide empirical background to evaluate other models as well.

AL estimate k as about 0.55, which supports the power of ½ in the CIR model. In fact the AL model with this parameter is close to the CIR model at any instant of time, but the CIR parameters are subject to change over time. Other models with k=0 or k=1 appear to be disindicated for US data by this result.

The period 1979-81 had high rates and high volatility, and studies that emphasize this period have found the power of ½ on r too low. There has been some debate about whether or not to exclude this period in fitting models. These results happened, so they can happen, but it was an unusual confluence of conditions not likely to be repeated. By taking a longer period which incorporates this interval AL do not exclude it but reduce its influence.

All parameters in the AL model were statistically significant. This implies that dependence of the volatility on r is not enough to capture the changes in volatility of interest rates. There have been periods of high volatility with low interest rates, for example. Thus the one and two-factor models without stochastic volatility appear to be insufficient to capture US interest rate dynamics.

3. Generating Yield Curves

The modeled dynamics of the short-term rate can produce implied yield curves. This is done by modeling the prices of zero-coupon bonds with different maturities, from which the implied interest rates can be backed out. P(T), the current price of a bond paying $1 at maturity T, can be calculated as the risk adjusted discounted expected value of $1 using the continuously evolving interest rate r from the short-term model. Here "expected value" indicates that the discounted mean is calculated over all possible paths for r. This can be expressed as:

$$P(T) = E^*[\exp(-\int r_t dt)],$$

where r_t is the interest rate at time t, the integral is over the time period 0 to T, and E^* is the risk-adjusted expected value of the discounted value over all paths r can take.

If E were not risk adjusted, the expectation that gives P(T) could be approximated by simulating many instances of the r process to time T over small increments and then discounting back over each increment. The risk-adjusted expected value is obtained instead by using a risk-adjusted process to simulate the r's. This process is like the original process except that it tends to generate higher r's over time. These higher rates usually produce an upward-sloping yield curve.

What is the risk adjusted process for r that with this procedure will generate the yield curves? If you write the price at time t for a bond maturing at time T as a Brownian motion with drift u and volatility v, i.e.,

$$dP(t,T) = u(t,T)dt + v(t,T)dz$$

then it can be shown (Vasicek 1977) that the drift u can be expressed as a function of the risk-free rate $r^{\#}$, the volatility v and a quantity λ called the market price of risk, by:

$$u(t,T) = r^{\#}P(t,T) + \lambda(t,r)v(t,T)$$

Thus the value of the bond grows by the risk-free rate plus the product of the bond's volatility with the market price of risk, plus the stochastic term $v(t,T)dz$. The market price of risk $\lambda(t,r)$ does not depend on the maturity date T, but it could depend on the interest rate r and the current time t.

The market price of risk in the bond price process is the link that specifies the risk-adjustment to the interest rate process that will generate the bond prices as the discounted expected value. As for the bond price process, only the drift of the interest rate process needs to be risk-adjusted, and the adjustment is to add the market price of risk times a function of the volatility of the interest rate process. For instance, AL suggest using the following adjusted process to simulate the interest rates in the bond price calculation:

$$dr = a(b - r + \lambda_1 rs)dt + sr^k dz_1 \qquad k>0$$

$$d\ln s^2 = c(p - \ln s^2)dt + v dz_2$$

$$db = j(q - b + \lambda_3 b)dt + wb^{1/2}dz_3$$

This adds terms to the drift of the first and third equations but not the second, as AL feel there is little price effect of stochastic volatility. The risk-price factors λ_1 and λ_3 can be calibrated to the current yield curve along with r, s, and b. These factors do not depend on T, so are held constant throughout any simulated yield curve calculation, but they can change stochastically when a new yield curve is calculated from a new time 0.

In the AL model you have to actually simulate the dynamics of the risk-adjusted process to get the yield curves. However, in the case of the CIR model, a closed form solution exists which simplifies the calculation. The yield rate for a zero coupon bond of maturity T is given by:

Y(T) = A(T) + rB(T) where:

A(T) = -2(ab/s2T)lnC(T) – 2aby/s2

B(T) = [1 – C(T)]/yT

C(T) = (1 + xyeT/x – xy)-1

$x = [(a - \lambda)^2 + 2s^2]^{-1/2}$

$y = (a - \lambda + 1/x)/2.$

Note that the only occurrence of r is in the Y equation, so Y is a linear function of r – but not of course of T. The linearity will come into play when we look at the distribution of Y across the generated scenarios. Since all the yield rates for different maturities are linear functions of r, they will also be linear functions of each other.

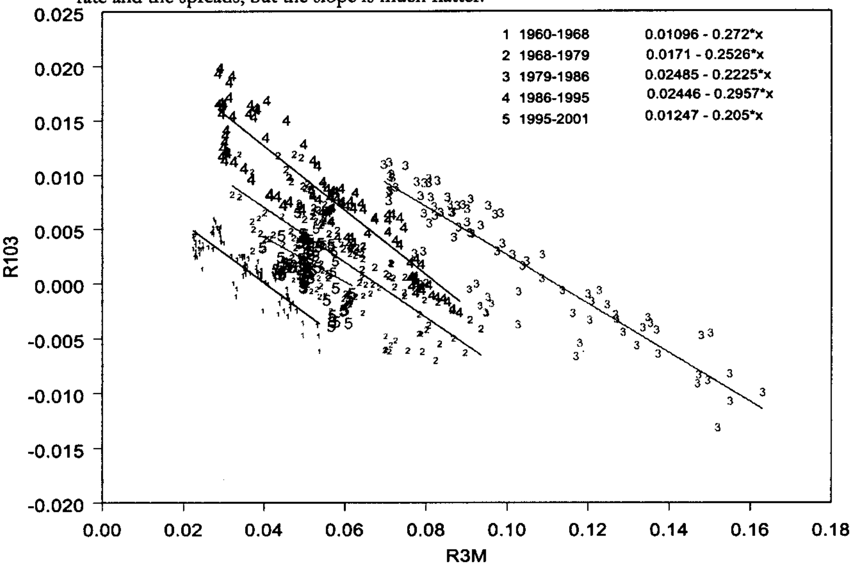
4. Historical Distributions of Yield Curves

To develop tests of distributions of yield curves, it is necessary to find some properties of these distributions which remain fairly constant over time. As it is difficult to describe properties of the distribution of the entire curve, the focus will be on the distribution of yield spreads, i.e., the differences between yields.

For a property to test the models against, however, the historical distribution of a given yield spread is not necessarily all that germane. When short-term rates are high, the yield curve tends to get compressed or even inverted, so spreads get low or even negative. This is related to the mean reversion of the short-term rate. Over time it tends to move back towards its long-term average, though with a large random deviation. Thus when it is high, a downward movement is anticipated, which produces lower long-term rates and thus negative yield spreads. If the period being projected by the model is not likely to have such high short-term rates, the yield spreads will be higher in the model than in the history.

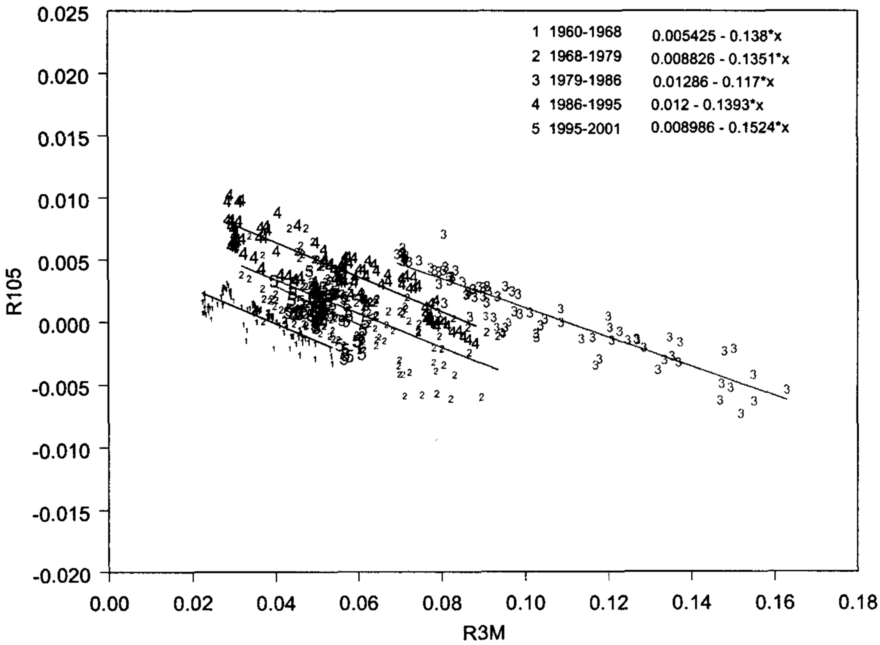
An alternative is looking at the conditional distribution of the yield spreads given the short-term rate. Over time, these conditional yield-spread distributions are more consistent than the unconditional distributions of yield spreads. The conditional distributions themselves do change in certain ways over time, however, but there are some consistencies remaining.

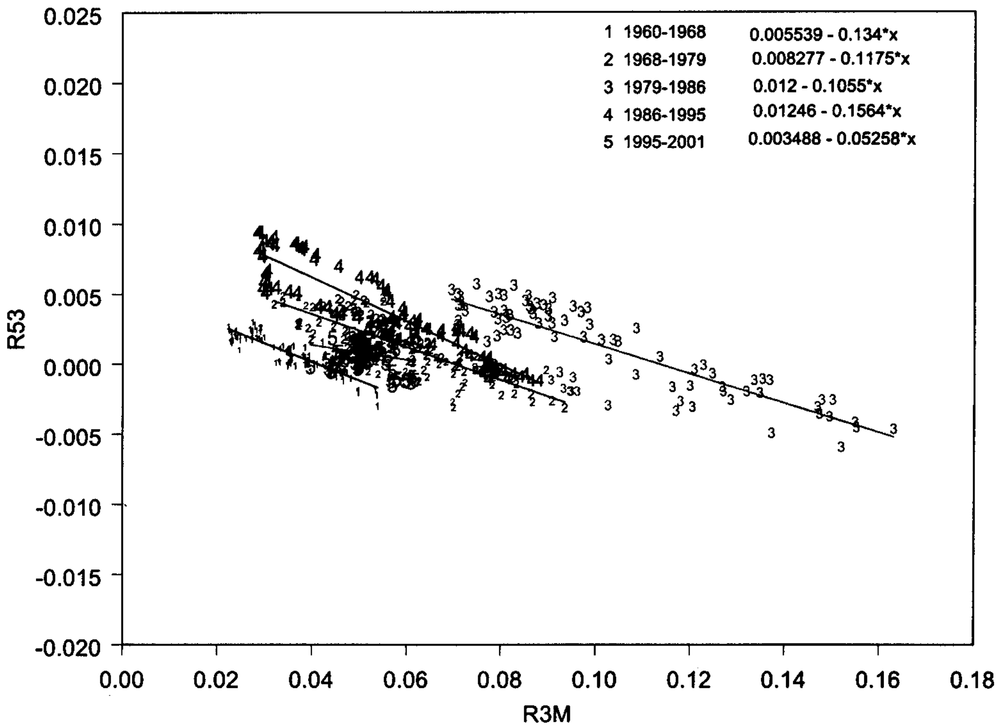
The graph below shows the US treasury three-year to ten-year yield spread as a function of the three-month rate for a 40+ year period. This period is divided up into five sub-periods, which were selected to maintain somewhat consistent relationships between the spread and the short-term rate. From the 60's to the early 80's, the short-term rates increased (sub-periods 1 – 3), then came back down after that (4 and 5). Each sub-period shows a negative slope for the spread as a function of the short-term rate, with the slopes in the range of –0.2 to –0.3. For the entire forty year period, there still seems to be a negative relationship between the short-term rate and the spreads, but the slope is much flatter.



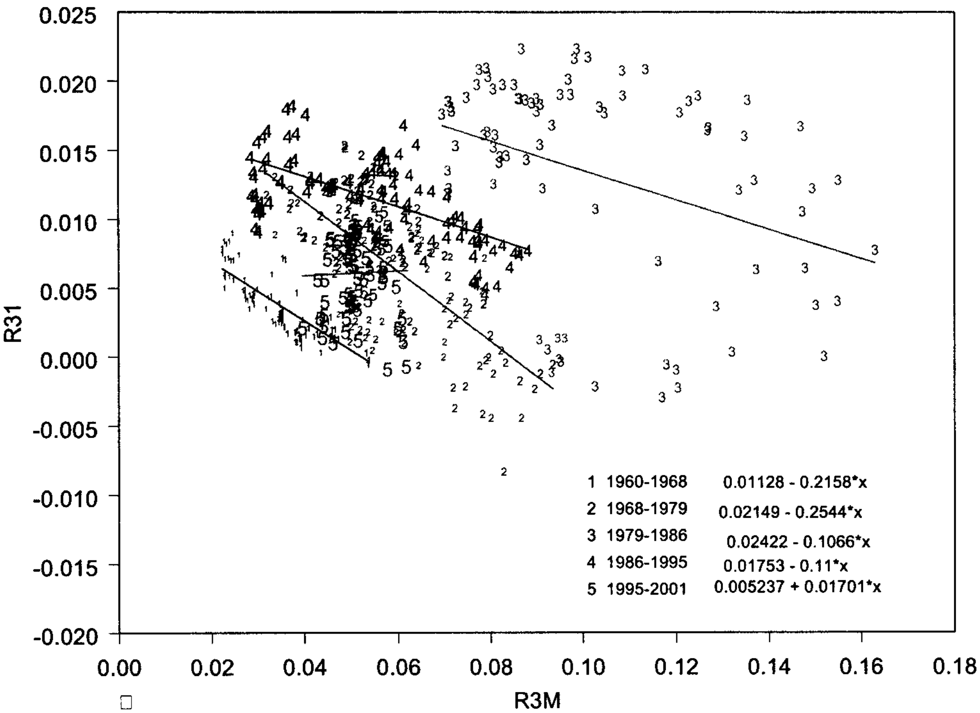
This behavior suggests that it would not be appropriate to use the conditional distribution from the entire period as a test of a scenario generator, especially if it is generating scenarios for a horizon of a few years. Over a several-year period the steeper slopes as in the historical sub-periods would be more likely to prevail. For a model projecting a few years into the future, the yield spreads would be expected to vary across scenarios, with generally lower spreads expected in those scenarios with higher short-term rates. From the historical record, it would be reasonable to expect a basically linear relationship, with a fair amount of spread around a slope in the range of –0.2 to –0.3. This could be tested by graphing the scenarios generated by the model to see if they were generally consistent with this pattern.

The graph below shows the same thing for the five-year to ten-year spreads as a function of the three-month rate. The main difference is that the relationship of the spread to the short-term rate is less dramatic, with sub-period slopes about half what they are for the 10 – 3 spread.

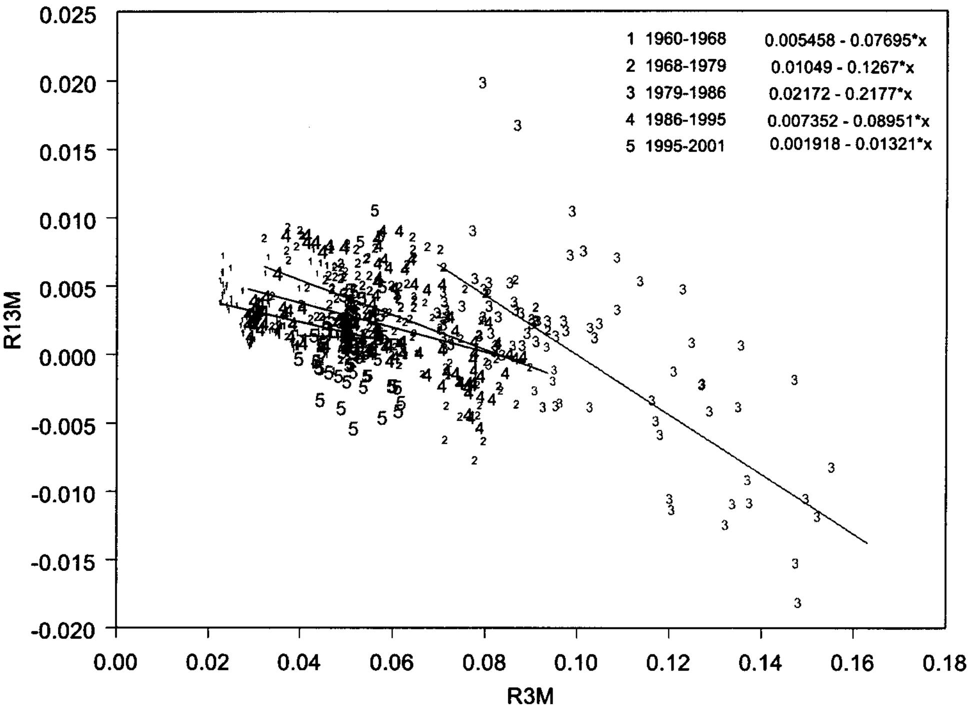




The three-year to five-year spreads show similar slopes to the 10 – 5 case, except for the latest period, which has a much flatter slope. The short-term rates in the last period have stayed in a fairly narrow range, however, making it harder to estimate the slope. In any case, relying more on the latest observations, it would seem that models producing a somewhat flatter slope in the near future should be reasonable.



The one-year to three-year spreads above show something different. Here the trend was below –0.2 in the 60's and 70's, around –0.11 in the 80's to mid-90's, and actually insignificant in the last period. Thus a flat relationship might be most appropriate in a short-horizon model.



□

The three-month to one-year spread shows even more of a break from the pattern of the longer spreads. Here the slope appears to be steeper when the short-term rates are higher, and the spreads can easily be negative. The slope is less in sub-period 5 than 1, and less in 4 than 2, suggesting that for a given short-term rate the slopes are less than they used to be. Thus a significant negative trend would not be expected for the near future, although a fair amount of randomness would still be anticipated.

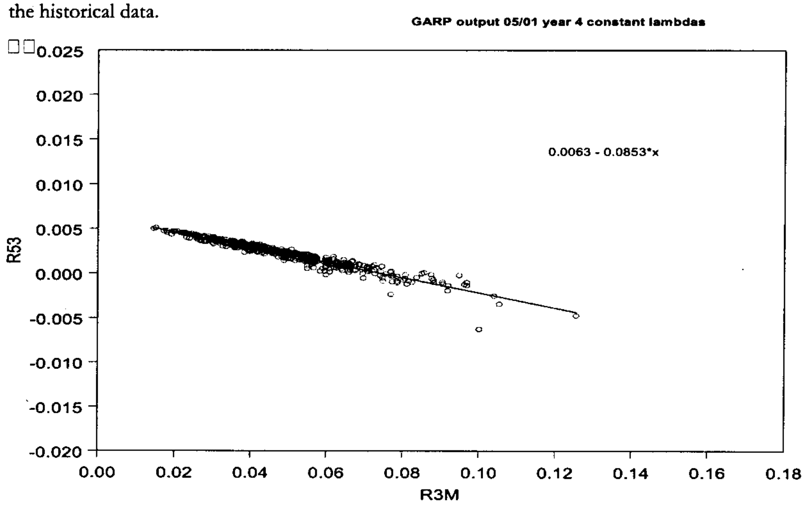
5. Testing Models Against Historical Distributions

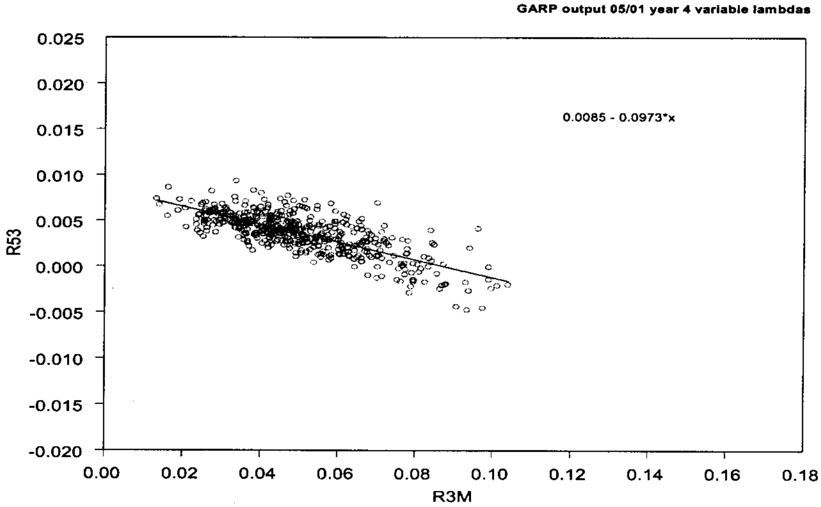
Models can be tested against the historical patterns by comparing the conditional distribution of yield spreads across the scenarios to the historical patterns to see if the patterns that have been produced historically are produced by the models. Initially two models will be compared. Both are produced by Guy Carpenter's proprietary scenario generator Global Asset Realization Processor, or GARP. They are both based on the AL specification for the short-term rate generator, but they differ in the treatment of the market price of risk. The CIR model will be included also.

The market price of risk has to be a deterministic function across all maturities to guarantee arbitrage-free yield curves at a given time. But it can change stochastically when generating scenarios for the yield curves at another time period. Allowing the market price of risk to change stochastically produces somewhat more variability among the yield curves generated. In one model, the constant lambda model, the two AL market price of risk parameters are held constant across all simulations. In the variable lambda model, on the other hand, stochastic changes are generated from one period to the next. How best to do that is a subject of ongoing research. The variable lambda model tested here is one of many possible models of this type and has not been optimized for this test. It probably introduces a bit too much variability into the market-price of risk.

The market price of risk parameters, as well as the current values of the three factors r, b, and s are calibrated to the current yield curve to get starting values for the simulations. For this example, a yield curve from May 2001 was used for calibration. The parameters are selected that generate a current yield curve that most closely matches the selected target curve. Then yield curves are simulated at various projected periods. For periods in the near future, the curves would not be expected to be too much different from the current curves. But going out a few years produces a wider variety of yield curve scenarios. In this case the sets of curves generated for year end 2004 are used in the distributional tests. This seems like a long enough projection period to expect to see the kind of variability that exists in the sub-periods historically.

Models can be tested for the conditional distributions of all of the yield spreads. First examined is the three-year to five-year spread. Recall that the slope for this was about –0.05 in the latest sub-period, but ranged from –0.11 to –0.16 in earlier segments. The graphs below show the relationship for the simulated spreads under the two models. The constant lambda model shows a slope of about –0.09, vs. –0.1 for the variable lambda, which are both reasonable. There is a difference apparent in the spread around the trend line, with the constant lambda model showing little spread, and the variable lambda showing a good deal more, which is more compatible with the historical data.





For the CIR model it was shown above that any yield spread is a linear function of the three-month rate. Although this model does have a fair amount of flexibility in determining the slope of that relationship, there will be no variability possible around the trend line. Graphically this would look narrow like the constant lambda case, only more so. This suggests that the CIR model will necessarily produce a restricted set of yield curve scenarios, and these will not have all the variability present in historical yield curves. Thus yield curve scenarios will not be present in proportion to their probability of occurring, contrary to the criteria established above for DFA asset generators.

The table below summarizes the historical and modeled slopes and the residual standard errors from the trend lines for the sub-periods and models considered.

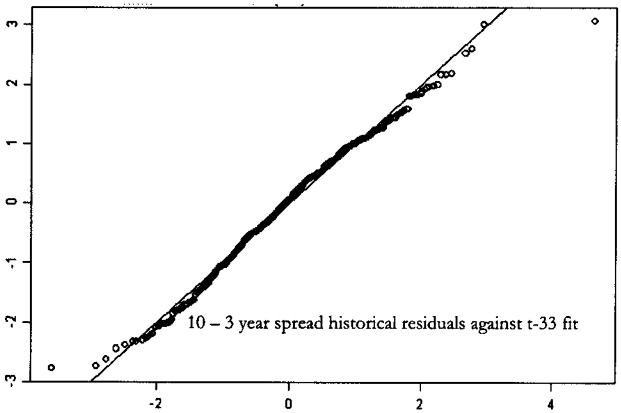
	R10 3	**R10 5**	**R5 3**	**R3 1**	**R1 3MO**
Period 1	(0.2720)	(0.1380)	(0.1340)	(0.2158)	(0.0769)
Period 2	(0.2526)	(0.1351)	(0.1175)	(0.2544)	(0.1267)
Period 3	(0.2225)	(0.1170)	(0.1055)	(0.1066)	(0.2177)
Period 4	(0.2957)	(0.1393)	(0.1564)	(0.1100)	(0.0895)
Period 5	(0.2050)	(0.1524)	(0.0526)	0.0170*	(0.0132)*
Constant λ	(0.2489)	(0.1635)	(0.0853)	(0.0721)	0.0299
Variable λ	(0.2960)	(0.1987)	(0.0973)	(0.0615)	0.0475
Period 1 se	0.0013	0.0009	0.0006	0.0013	0.0019
Period 2 se	0.0031	0.0022	0.0013	0.0037	0.0030
Period 3 se	0.0026	0.0013	0.0017	0.0070	0.0051
Period 4 se	0.0022	0.0012	0.0012	0.0024	0.0029
Period 5 se	0.0020	0.0013	0.0009	0.0028	0.0028
Constant λ se	0.0008	0.0005	0.0004	0.0013	0.0023
Variable λ se	0.0042	0.0028	0.0015	0.0021	0.0028

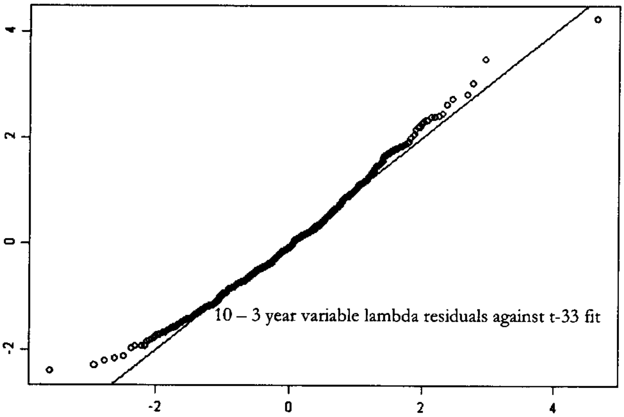
* Not significantly different from zero

These results indicate that the constant lambda model tends to produce too little variability around the trend, whereas this formulation of the variable lambda model produces perhaps too much in the longer spreads. This suggests that allowing somewhat less variability in the stochastic processes that generates the market prices of risk could lead to still more realistic models.

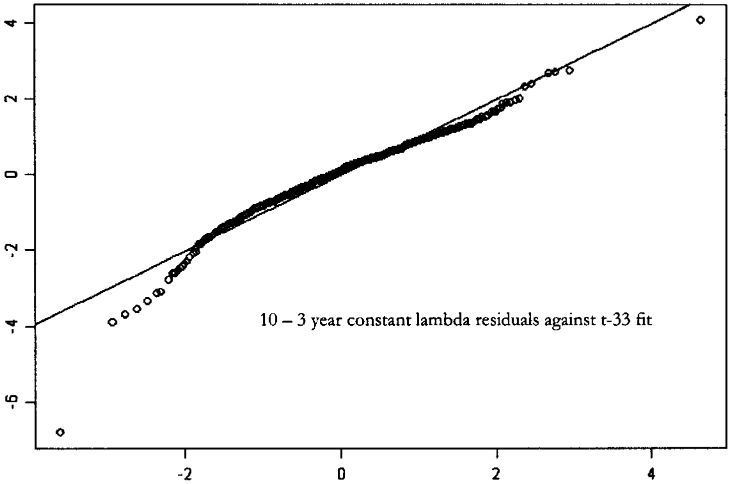
6. Testing Residual Distributions

The conditional distributions of the generated yield spreads given the short-term rate have been tested against the slopes and standard errors of historical data. What about the actual distributions of the residuals around the trend lines? Are these the same historically and for the generated scenarios? This was tested by fitting t-distributions to the residuals from the model and the combined set of residuals from the historical periods. The graphs below show QQ-plots, which graph the percentiles of the residuals against the same percentiles of the fitted t's.

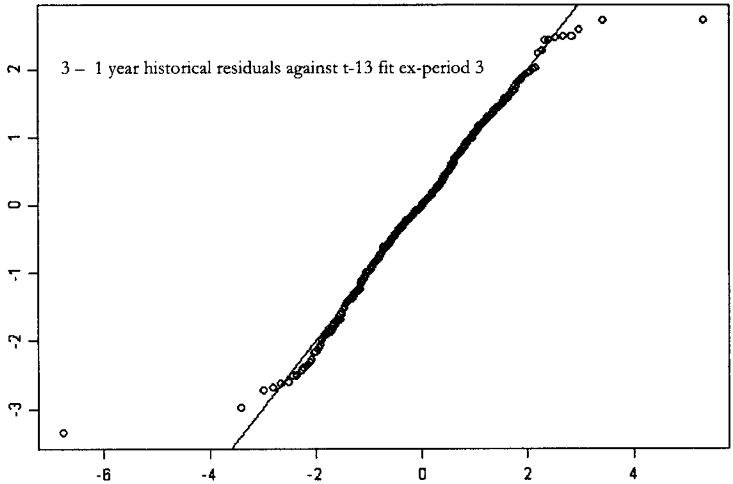


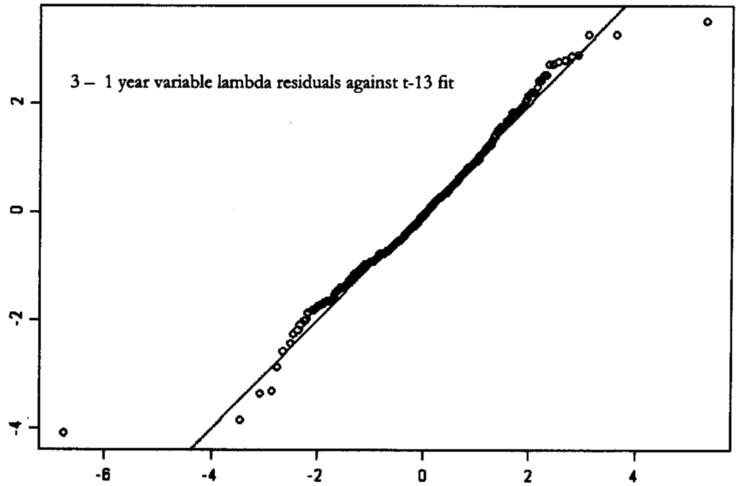


The 10 – 3 constant and variable lambda residuals look a lot like the data except in the left tail, where the constant lambda diverges. The t-distribution with 33 degrees of freedom was fit here.

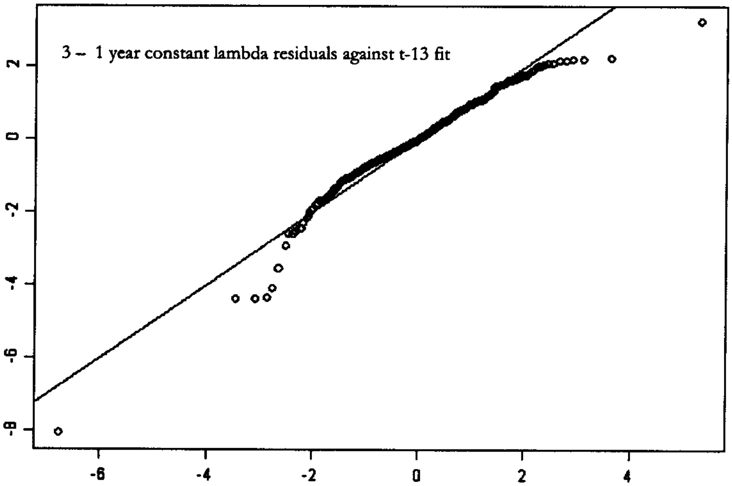


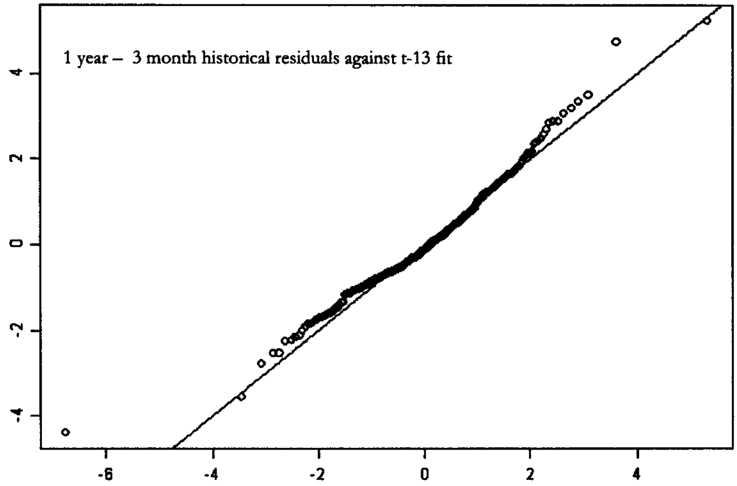
The 3 – 1 year residuals were done excluding period 3, which was unusual.

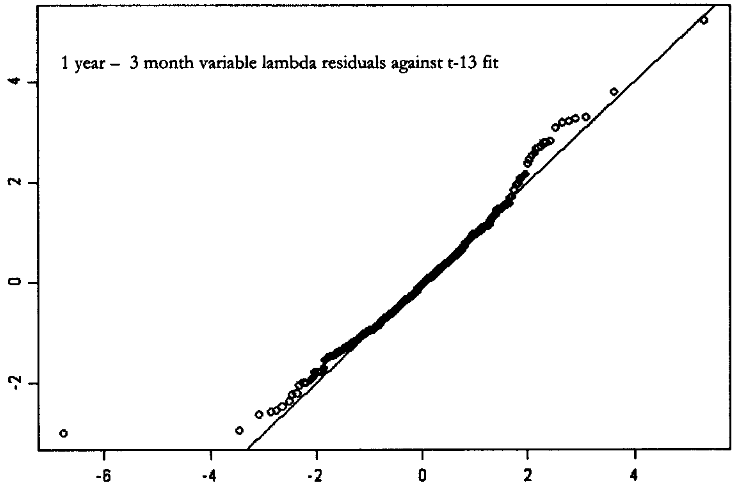


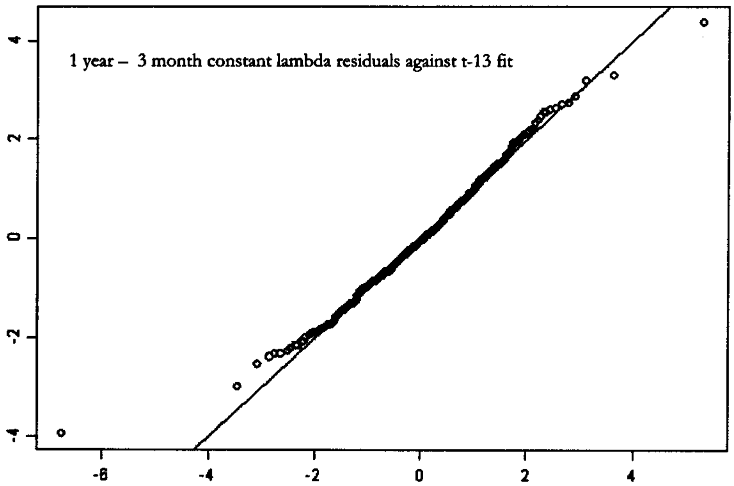


For the 3 – 1 spreads the variable lambda model provided residuals distributed similarly to those from the data, when compared to the t with 13 degrees of freedom.




1 year – 3 month historical residuals against t-13 fit
4
2
0
-2
-4
-6
-4
-2
0
2
4


1 year – 3 month variable lambda residuals against t-13 fit
4
2
0
-2
-6
-4
-2
0
2
4



Again for the 1 year – 3 month spread residuals, the data and the variable lambda model compare similarly to the t-13 fit, where the constant lambda is a little different.

Conclusion

Many models of interest rates have been proposed. For one survey, see R. Rebonato (1997) "Interest Rate Option Models," John Wiley NY. Models of the dynamics of the short-term rate apparently need to incorporate mean reversion, stochastic changes in mean reversion over time, mean sensitive volatility, proportional approximately to the square-root of the mean, and stochastic volatility as well.

Testing the conditional distribution of various yield spreads given the short-term rate appears to be a reasonable way to see if a model is generating a realistic distribution of yield curves. The unconditional distribution of generated yield spreads would not necessarily be comparable to the historical distribution, because different spreads are associated with different short-term rates, and the simulation might not be generating a distribution of short-term rates that matches the historical record, due to the particular economic conditions that prevail at the time of the simulation. The slopes of the conditional fitted lines are fairly consistent over different historical periods.

As with most tests of distributional issues, this one is not a formulaic system that gives a strict "yes/no" answer to a model's output. But it does provide a realm of reasonable results so you can give an opinion of the "probably ok/ probably not" type. For example, having no variability around the conditional trend line would seem to be too limiting. Slopes that are much steeper than historical would also seem disindicated, as would distributions of residuals around the slopes that differ substantially from the t-distributions fit. Even though these tests are not strict, better results could be sought than those of any of the models tested.

An application issue is how much variability you should have for projection periods of different lengths. When projecting out four or five years, a conditional distribution similar to those of the historical sub-periods might be appropriate. However there is some chance of entering a new realm – i.e., changing sub-periods – over that much time. In all the sub-periods graphed, changing to an adjacent sub-period would tend to flatten the conditional trend.

Appendix I – Scalar Measures of Response

A number of risk measures have been devised to look at the effect on an investment holding or portfolio of a small change in some quantity. For example, Macaulay duration measures the change in the value of a portfolio due to a change in the annualized average yield to maturity. It can be expressed as the weighted average of the times to each cash flow of the portfolio, where the weights are the cash flow amounts discounted at the average yield. Thus duration is expressed in units of time. (Duration measures value per interest rate, but as interest rate is value per time, duration is time.) One way to produce a given change in the average yield to maturity is to shift the entire yield curve by the same amount, so duration is often described as the sensitivity of the portfolio to a parallel shift in the curve.

Macaulay convexity is the weighted average of the squares of the times to the cash flows, using the same weights as for duration. It can be shown to be the square of duration less the derivative of duration with respect to the instantaneous average yield.

The analysis of derivative instruments has produced several similar measures, denoted by Greek letters, and so called "the Greeks." These measure the change in the value of a position brought on by the change in something else that affects value. For instance, the change in the value of an option due to the change in the value of the underlying security is called delta.

For bond portfolios, each bond could be thought of as a holding of a combination of future positions in the short-term rate, which could thus be considered to be the underlying security. With the short-term rate as the underlying security, the delta risk is the change in the value of the portfolio with respect to a small change in the short-term rate. This is different than duration, as even though all the rates will change in response to a change in the short-term rate, they will not necessarily change by the same amount. This is clear in the CIR model where a change in r makes all the rates change, but each by its own B(T). If the underlying security is taken to be the average yield to maturity, then delta is duration.

Gamma risk is the change in delta due to a small change in the value of the underlying security. With the short-term rate as the underlying security, in CIR gamma is zero, but for a typical asset or liability portfolio it will not be. Gamma is somewhat analogous to convexity, but as defined here focuses on the actual short-term rate, not the average yield.

Vega measures the change in value due to a change in the volatility of the underlying instrument. The volatility of the short-term rate Brownian motion is an element in bond pricing, so vega risk is present in bond portfolios. CMO's probably have a fair degree of this risk as well, as greater interest rate volatility can increase the probability of pre-payment.

Theta is just the sensitivity of the position to a small change in the valuation date.

Rho for any portfolio measures its change in value due to a small change in the interest rate. In most asset pricing models the yield curve is assumed to be constant, so rho could be considered to be the effect of a shift in the average yield, i.e., duration.

Appendix 2 – The Arbitrage Debate

Most finance theory takes the impossibility of arbitrage as a given, but some actuaries use interest rate models that are not arbitrage-free. This may be just a matter of convenience, but two arguments are sometimes advanced for using such models:

1. Actual published yield curves are not always arbitrage-free
2. It is more important to get the statistical properties of the set of scenarios right than to avoid arbitrage.

One problem from having arbitrage possibilities in generated scenarios is that searching for optimal investment strategies would find the arbitrage strategy, and that will appear the best. It seems pretty unlikely, however, that a DFA model could identify truly risk-free high-profit investment strategies that insurers could work in practice. Even if the search disallowed the arbitrage strategies, their presence in the scenario set could have a distorting effect. However, a model that allows arbitrage only in unrealistic cases, like being able to borrow huge amounts at the risk-free rate, could be considered arbitrage-free in practice.

With this in mind, the two arguments can be reviewed separately. First, there may occasionally be some arbitrage possibilities in published yield curves. But this does not mean that these can be taken advantage of in practice. For one thing, the published curves look at trades that took place at slightly different times, so are not snapshots of one moment in time. Looking at a combination of positions in different deals that have happened recently could yield a hypothetical arbitrage, but that possibility could be gone before it could be realized. A related issue is that some of the deals might have to be scaled up significantly to get the arbitrage to work, and doing this could change the prices. In short, finding some historical published yield curves with hypothetical arbitrage possibilities in them is not reason enough to use a modeled set of scenarios that have specific arbitrage strategies built in.

The second argument is more interesting. This paper argues for the importance of getting the statistical issues right, focusing on the distribution of yield spreads across scenarios. This does not appear to be in any way inconsistent with no arbitrage. Using models like AL also emphasizes that the movement of interest rates across time should be statistically correct.. Thus both the statistics of changes in rates over time and the distribution of yield spreads at each time are compatible with arbitrage-free scenarios. It would be interesting to see what other statistical issues there are that would require using scenarios with arbitrage built in.