

# 2008 Winter E-Forum

## Table of Contents

### Data Management Call Papers

#### **Dirty Data on Both Sides of the Pond**

Findings of the GIRO Data Quality Working Party—Robert Campbell, FCAS, FCIA,  
*Louise Francis, FCAS, MAAA, Virginia Prevosto, FCAS, MAAA, Mark Rothwell, FIA,*  
*Simon Sheaf, FIA* .....1  
*Triangle data for Dirty Data paper.XLS*

#### **Data Organization and Analysis in Mortgage Insurance: The Implications of Dynamic Risk Characteristics**

*Tanya Havlicek, and Kyle Mrotek, FCAS, MAAA* .....71

#### **ROOT: A Data Analysis and Data Mining Tool from CERN**

*Ravi Kumar, ACAS, MAAA, and Arun Tripathi, Ph.D.* .....90

#### **Staying Ahead of the Analytical Competitive Curve: Integrating the Broad Range Applications of Predictive Modeling in a Competitive Market Environment**

*Jun Yan, Ph.D., Mo Masud, and Cheng-sheng Peter Wu, FCAS, ASA, MAAA* .....121

### Research Working Party Report

#### **Actuarial I.Q. (Information Quality)**

*A Report of the CAS Data Management Educational Materials Working Party* .....136  
*Box Plot and Histogram for Actuarial IQ.xls*

### Additional Papers

#### **Capital Allocation by Percentile Layer**

*Neil M. Bodoff, FCAS, MAAA*.....196

#### **Uncertainty-based Credibility and its Application to Excess-of-loss Reinsurance**

*Pietro Parodi, PhD, and Stephane Bonche, IA*.....224

#### **Consideration of Bias in Chain Ladder Estimates**

*Rajesh Sahasrabudhe, FCAS, MAAA*.....252

## ***E-Forum* Committee**

Glenn M. Walker, *Chairperson*

Mark A. Florenz  
Karl Goring  
Dennis L. Lange  
Darci Z. Noonan  
Zongli Sun  
Windrie Wong  
Yingjie Zhang

For information on submitting a paper to the *E-Forum*, visit  
<http://www.casact.org/pubs/forum/>.

# Dirty Data on Both Sides of the Pond

Findings of the GIRO Data Quality Working Party  
by Robert Campbell, FCAS, FCIA, Louise Francis, FCAS, MAAA, Virginia  
Prevosto, FCAS, MAAA, Mark Rothwell, FIA, Simon Sheaf, FIA

---

**Motivation.** This paper takes a multi-faceted approach to quantifying the significance of data quality issues for property/casualty actuaries, addressing both the prevalence of data quality issues across areas of practice and the significance of those issues. The conclusion gives some guidance to improve data quality.

**Method.** This paper

- describes some actual data quality disasters in non-insurance and insurance businesses;
- presents the results of a data quality survey of practicing actuaries in the United States, Canada, Great Britain and Bermuda;
- presents the results of a data quality experiment where data was altered to change its quality and the effect on analyses using the data was quantified; and
- provides advice on what can be done to improve the state of data quality, including introducing some freeware that can be used to screen data.

**Results.** Both the survey results and the data quality experiment suggest that data quality issues affect the accuracy and increases the uncertainty associated with actuarial estimates

**Conclusions.** Data quality issues significantly impact the work of property/casualty insurance actuaries; and such issues could have a material impact on the results of property/casualty insurance companies.

**Availability.** Excel spreadsheets containing the data used in the data quality experiment as well as the spreadsheet containing the bootstrap procedure will be available on the CAS web site.

**Keywords.** Data, data quality, reserve variability, exploratory data analysis, data diagnostics

---

## 1. INTRODUCTION

*“Poor data quality can be insidious.*

*Insidious a. 1. Characterized by craftiness or shyness... 2. Operating in a slow, not easily apparent manner; more dangerous than seems evident.”*

—Redman, *Data Quality: The Field Guide*

While the quality of data used in many insurance ratemaking analyses may be regarded as poor, little has been done to quantify the prevalence of poor data or its impact on analyses. In 2006, a paper was produced by the GIRO (General Insurance Research Organization) Data Quality Working Party and presented at the 2006 GIRO conference (Campbell et al., 2006). The Working Party was formed because of the perception that data quality is an important issue that is given insufficient attention by the managements of insurance industry companies. The Working Party’s report presented several arguments to support applying increased resources to data quality including recounting of data quality “horror stories,”

## *Dirty Data on Both Sides of the Pond*

presenting the results of a survey of actuaries and insurance professionals and an examination of the impact of data quality issues on an actuarial database. The authors of the 2006 paper decided to continue their research. In particular, the data quality survey that attempts to quantify the extent of data quality problems has been distributed to a considerably wider audience and the number of respondents has more than doubled. In addition, significant changes have been made to a data quality experiment that attempts to quantify the extent of data quality problems in property/casualty insurance, by simulating data quality problems in data used in an actuarial analysis. The authors also wished to present their results to North American as well as U.K. actuaries.

Data quality is an important issue affecting all actuaries. Whether one is engaged in reserving, pricing, claims or premium fraud detection, or other actuarial applications, or whether one is using conventional actuarial techniques or more advanced data intensive techniques (e.g., predictive modeling), virtually all actuaries encounter data that is either incomplete or inaccurate. Recently enacted laws in both Europe (Basel II) and the United States (Sarbanes-Oxley) addressing record keeping issues would seem to justify more attention to data quality, but a general increase in concern about data quality is not obvious.

### **1.1 Research Context**

In this section we review some of the literature addressing data quality issues in insurance.

The U.K. General Insurance Reserving Task Force (GRIT) working party report recommended more focus on data quality (Copeman et al., 2006) and suggested that U.K. professional guidance notes incorporate standards from Actuarial Standards of Practice 23, Data Quality (ASOP 23). ASOP 23 provides a number of guidelines to actuaries when selecting data, relying on data supplied by others, reviewing and using data, and making disclosures about data quality. The Casualty Actuarial Society Committee on Management Data and Information and the Insurance Data Management Association (IDMA) also produced a white paper on data quality (CAS Committee on Management Data and Information, 1997). The white paper states that evaluating the quality of data consists of examining the data for validity, accuracy, reasonableness, and completeness. This CAS committee also promotes periodic calls for papers on data management and data quality, which are published by the CAS.

More recently, the CAS Data Management and Information Educational Materials

## *Dirty Data on Both Sides of the Pond*

Working Party (CAS DMIWP) has completed two papers relevant to data quality: The first (CAS DMIWP, 2007) is a survey of data quality texts. The survey is intended to provide guidance to actuaries who seek a more detailed and comprehensive exposure to data quality literature. The texts reviewed in the paper are rated on a number of qualities, such as actuarial relevance and introductory versus advanced focus, which are intended to assist actuaries in selecting appropriate texts for their particular needs.

The CAS DMIWP also completed the paper “Actuarial IQ” (CAS DMIWP, to be published in 2008) which distills and summarizes much of the current literature on data quality and data management as it relates to the assurance of the quality of information used by actuaries.

In general, the literature on data quality and its effect on the insurance business is limited. In Section 2, we provide some background on the effect of poor data quality on businesses, but many of the studies cited only address the issue for non-insurance businesses.

### **1.2 Objective**

The GIRO Data Quality Working Party was constituted to act as a catalyst to the profession and the industry to improve data quality practices.

In this paper we will

- Recount some anecdotes illustrating the real cost of poor data both in insurance and other ventures.
- Present the results of a data quality survey of practicing actuaries in the United States, Canada, Great Britain, and Bermuda.
- Present the results of a data quality experiment where data was intentionally altered to change its quality and the effect on analyses using the data was quantified.
- Provide advice on what can be done to improve the state of data quality research.

### **1.3 Disclaimer**

While this paper is the product of a GIRO working party, its findings do not represent

the official view of the General Insurance Research Organization. It also does not represent the views of the authors' employers. Moreover, while we believe the approaches we describe are good examples of how to address the issue of data quality, we do not claim they are the only acceptable ones.

## **1.4 Outline**

The remainder of the paper proceeds as follows. Section 2 will review literature on the cost to business of poor data. It will then provide a number of data quality "horror stories" in both non-insurance and insurance contexts. Section 3 will present the result of a data quality survey that was distributed to actuaries on both sides of the "pond." Section 4 presents the results of our data quality experiments that measure the effect of data quality on an actuarial analysis. First, a deterministic experiment is performed that introduces data quality problems into a dataset used to estimate loss reserves. For this dataset the "true" ultimate losses are known and can be used to evaluate the quality of the deterministic estimates. Next, a stochastic data quality experiment using a bootstrap procedure is used to evaluate the effect of data quality problems. In Section 5 we suggest a number of actions actuaries can take, including data quality advocacy, data quality measurement and routine screening of data before performing an analysis. Software for screening data is also discussed. In section 6 we summarize our findings from the data quality survey and data quality experiment. Appendix A describes the open source software ViSta and presents data screening graphs obtained from the software data used in our analysis is presented in Appendices B through D.

## **2. BACKGROUND AND METHODS**

### **2.1 The Cost of Poor Data Quality**

In the literature on data quality there is a virtually universal agreement that poor data quality imposes a significant cost on companies and on the economy. For instance, Moore predicts that there is a significant likelihood that a data quality error will cause the downfall of at least one large corporation (Moore, 2006). In this section we summarize some of the published findings with respect to the magnitude and cost of data quality problems.

There are various rules of thumb found in the literature concerning the cost of poor data

*Dirty Data on Both Sides of the Pond*

quality. Both the IDMA and Olson cite an estimate that data quality problems cost companies 15% - 20% of operating profits.<sup>1</sup> The IDMA value proposition<sup>2</sup> also cites an estimate that poor data costs the U.S. economy \$600 billion a year<sup>3</sup>. The IDMA believes that the true cost is higher than these figures reflect, as they do not depict “opportunity costs of wasteful use of corporate assets.” (IDMA Value Proposition – General Information).

According to Eckerson, in many customer databases 2% of records per month become obsolete because of deaths and address changes (Eckerson). This is in addition to data entry, merging data from different systems and other sources of errors. Eckerson mentions that most organizations overestimate the quality of their data stating, “On one hand, almost half of the companies who responded to our survey believe the quality of their data is excellent or good.” Yet more than one-third of the respondent companies think the quality of their data is “worse than the organization thinks.” Eckerson also cites a study done by The Data Warehouse Institute that indicates that data quality is a leading cause of problems when implementing CRM (Customer Relationship Management) systems (46% of survey respondents to a 2000 survey selected it as a challenge). According to Wand and Wang (1996), 60% of executives from 500 medium-sized surveyed firms reported data quality problems.

Poor data quality can also have credibility consequences and motivate regulatory intervention to curb the use of some information deemed important by corporations. In property and casualty insurance in the United States, the use of credit information in underwriting and pricing insurance is a very controversial practice. A key argument of consumer groups opposed to the use of credit is the poor quality of credit data. Among actuaries who price and reserve small (self-insured or alternative market) accounts, there is a general belief that the quality of data from third-party administrators (TPA) is perhaps worse than that of insurance companies. Popelyukhin (1999) reviewed the loss runs of 40 TPAs and concluded that no TPA provided data that satisfied his data quality definition (similar to that in the CAS-IDMA White Paper above).

In 2004, PricewaterhouseCoopers LLP (PricewaterhouseCoopers LLP, 2004) distributed

---

<sup>1</sup> Olson, p9.

<sup>2</sup> This citation is apparently from a study done by The Data Warehouse Institute

<sup>3</sup> Based on information at econostats.com the 2006 gross domestic product of the U.S. was about \$13,000 billion.

## *Dirty Data on Both Sides of the Pond*

a data management survey to executives at 450 companies in the U.S., U.K., and Australia. The following results were cited by PricewaterhouseCoopers:

- Almost half of all respondents do not believe that senior management places enough importance on data quality.
- Only 18% of respondents whose organizations share data with third parties are very confident in the quality of that data.
- On average respondents thought data represented 37% of the value of their company (but only 15% actually measured the value of data to their company).
- The survey indicated that when data improvement initiatives were undertaken and when their value was measured, significant returns on investment were realized.

Note that while a number of surveys have been conducted to evaluate the extent of the data quality problem, there appears to be very little literature where an attempt has been made to quantify the impact of data quality problems on the accuracy and variability of financial quantities being computed. In this paper we add to the results of prior surveys on data quality by conducting a survey of actuaries. We also perform several experiments where the effect of data quality problems is measured on an actuarial database used for reserving.

## **2.2 Data Quality Anecdotes**

### **2.2.1. Non-Insurance Industry Stories**

As the anecdotes below illustrate, data errors can result in very serious consequences. In some cases the result is serious embarrassment. In other cases, the result is a large financial loss. In yet other cases, loss of life results, demonstrating that data quality can be a matter of life and death. Many of the most highly publicized data quality horror stories are from non-insurance industries. It should be noted that non-insurance industry errors sometimes have implications for insurance as they may result in errors and omissions or medical malpractice claims as in the first example below.

- A 17-year old Mexican girl received a heart-lung transplant at Duke University Hospital in South Carolina. She soon fell into a coma as it was discovered that the organs she received were of the wrong blood type (Archibald, 2003). Apparently none of the medical personnel at the hospital performing the transplant requested or

### *Dirty Data on Both Sides of the Pond*

verified that proper documentation of a match in blood types was provided. A subsequent transplant with organs of the correct blood type failed and the girl died.

- The Web site [www.iqtrainwrecks.com](http://www.iqtrainwrecks.com) reports that surgery on the wrong site, i.e., removing the wrong kidney, occurs too frequently and is in large part preventable. It is noted, that many wrong site surgeries occur as a result of reading x-rays from the wrong side. They note that since most x-rays are produced digitally, it would be trivial to label the x-ray as to which side is which.
- During the conflict in Bosnia, American pilots accidentally bombed the Chinese embassy in Belgrade as a result of faulty information. “It was the result of neither pilot nor mechanical error,” Cohen and Tenet stated. “Clearly, faulty information led to a mistake in the initial targeting of this facility. In addition, the extensive process in place used to select and validate targets did not correct this original error.” (CNN, 1999a)
- In Porter County, Illinois, a house worth a little over \$100,000 was accidentally valued at \$400 million. This caused the county to bill the owner \$8 million for what should have been a \$1,500 real estate tax bill. Due to the glitch, the county significantly overestimated its tax revenue and experienced significant budget shortfalls.
- Statscan, the Canadian statistical agency, reported that it had understated the inflation rate for five years due to a software glitch. The effect was estimated to be one tenth of a point on average. (Infoimpact, 2006). In addition, Statistics South Africa reported that, due to an error, it had greatly overstated inflation for five months, causing interest rates to be significantly higher than they would otherwise have been. (Data Quality Solutions, 2007)

#### **2.2.2 Insurance Industry Stories**

Although we contacted a number of insurance regulators, we are not at this time aware of any insolvency that resulted primarily from data quality errors. On the other hand, there is a lot of sentiment that data quality often deteriorates badly after insolvency occurs and that it significantly impairs the quality of post-insolvency estimates of liabilities. It is possible that the role of data quality issues in insolvencies is obscured by other management issues.

### **2.2.2.a Reserving stories**

- In June 2001, The Independent went into liquidation and became the U.K.'s largest general insurance (i.e., property/casualty) failure. A year earlier, its market valuation had reached £1B. Independent's collapse came after an attempt to raise £180M in fresh cash by issuing new shares failed because of revelations that the company faced unquantifiable losses. The insurer had received claims from its customers that had not been entered into its accounting system, which contributed to the difficulty in estimating the company's liabilities.
- The National Association of Insurance Commissioners<sup>4</sup> stated that it often cannot rely on typical domiciliary country data when reviewing the condition of alien (non-U.S.) insurers. However, they indicated that when they request data from the companies themselves, it is usually supplied. (Otis, 1977)
- The Canadian federal regulator (the Office of the Superintendent of Financial Institutions, or OSFI for short) has uncovered instances of:
  - Inaccurate accident year allocation of losses and double-counted IBNR loss estimates (i.e., the actuary calculated IBNR from triangles that already included IBNR).
  - Claims reported after a company is insolvent and it is discovered that the original notices (sometimes from years before) were not properly recorded in the company's systems.
- In the U.S., actuaries providing statements of actuarial opinion to insurance regulators concerning the adequacy of reserves for an insurance company are required to supply an exhibit balancing totals from data used in their actuarial analysis to totals in the statutory financial statement. A former regulator indicated this requirement is motivated by disclaimers in opinions letters (i.e., the data was supplied by the company and responsibility for its accuracy was deemed to be theirs) and concerns that invalid data would be used in the actuary's reserve analyses.
- It is widely believed by U.S. actuaries that the quality of an insolvent insurance company's data declines after the company is declared insolvent. A report by the

---

<sup>4</sup> An association of state insurance regulators in the United States

### *Dirty Data on Both Sides of the Pond*

California Auditors Office on the California Conservation and Liquidation office found numerous data quality problems (Sonnett, 2005). For instance, due to manual processing of many bills, one employee retired without billing a reinsurer for \$900,000. The error was discovered months later only after the reinsurer inquired about the bill. A finding of the report (California Auditor's Office, 2004) was that "the information technology controls were not sufficient to ensure the overall reliability and integrity of data."<sup>5</sup>

#### **2.2.2.b. Ratemaking Stories**

Advisory organizations in the United States such as the National Council on Compensation Insurance (NCCI) for workers compensation and the Insurance Services Office, Inc. (ISO) for most of the remaining property/casualty lines of insurance devote significant resources to finding and correcting errors in data.

The stories below are a just a few examples of data anomalies that have been faced by ISO over the years in its role as an advisory organization, along with other examples drawn from the consulting community. These are cases where the anomaly was found during the rate-level experience review and caused extra expense to either correct the error or remove the data in error from the rate-level experience review. It is not a complete list but rather gives a flavor of the data quality glitches that typically occur.

- A company reported its homeowners exposure (the amount of insurance on the dwelling) in units of \$10,000 instead of units of \$1,000. Since the exposure was understated by a factor of 10, applying current manual base loss costs (or manual rates) and rating factors to the exposure would have resulted in greatly understated aggregate loss costs at current manual level (or aggregate premium at present rates). Therefore the experience loss ratio (= incurred losses/aggregate loss costs at manual level) and the statewide rate-level indication would have been overstated.
- One of the ten largest insurers in a state reported all of its personal auto data under a miscellaneous coverage code. Since miscellaneous coverage code data are excluded from the rate-level review for the core coverages, this would have had a significant effect on ratemaking results if it had not been detected.
- A company reported all its homeowners losses as fire in the state of Florida. It is

---

<sup>5</sup> This finding is stated in the Executive Summary of the report.

### *Dirty Data on Both Sides of the Pond*

evident what this error can do for any homeowners rate-level review especially when the experience period included the hurricane-heavy accident years of 2004 and 2005.

- Another common error occurs when the premium and loss records for the same policy are not coded identically for the common fields. For example, a company may record all their liability premium records as composite rated, but the corresponding liability loss records are recorded otherwise. This is commonly known as a premium-loss mismatch error. A recent occurrence of this type of anomaly in homeowners affected about 25% of a company's book of business.

### **3. DATA QUALITY SURVEY**

We conducted a brief survey of actuaries<sup>6</sup> to verify that data quality issues have a significant impact on the work undertaken by general insurance actuaries. The precise wording of the survey questions was as follows:

- Based on the time spent by both you and your actuarial staff, what percentage of this effort is spent investigating and rectifying data quality issues?
- What percentage of the project results are adversely affected by data quality issues? Adversely affected includes re-working calculations after data is corrected; or stating results/opinions/conclusions but allowing for greater uncertainty in results; or finding adverse runoff over time due to initial work based on faulty data; etc.

In order to improve our response rate, we decided to adopt a targeted and personal approach. Copies of the survey were sent to the following groups:

- All original members of the GIRO Data Quality Working Party, including those who had subsequently chosen not to take part in our work
- Members of the CAS Committee on Management Data and Information
- Members of the CAS Data Management and Information Educational Materials Working Party

---

<sup>6</sup> In some cases, other quantitative analysts and systems people who work with and support actuaries were included in the survey.

*Dirty Data on Both Sides of the Pond*

- A sample of attendees at a WRG<sup>7</sup> Predictive Modeling Conference
- A sample of attendees at the 2007 CAS Ratemaking Seminar
- A sample of attendees at the 2007 CAS Reinsurance Seminar

In addition, each member of the GIRO Data Quality Working Party contacted a handful of people to ask them to answer the survey questions. This survey was carried out by phone.

As a result of these efforts, we received 76 responses to the survey.

The tables below summarize the results of the survey. We have split the results between those actuaries who work for insurers or reinsurers, those who work as consultants, and the remainder. The last category includes insurance and reinsurance brokers, rating agencies, and statistical agents, as well as those respondents who we were unable to categorize. We show the highest and lowest responses to give an indication of the range of the responses.

---

<sup>7</sup> World Research Group, March 2007

*Dirty Data on Both Sides of the Pond*

**Question 1: Percentage of Time Spent on Data Quality Issues**

<b>Employer</b>	<b>Number of Responses</b>	<b>Mean</b>	<b>Median</b>	<b>Minimum</b>	<b>Maximum</b>
Insurer/Reinsurer	40	25.0%	20.0%	2.0%	75.0%
Consultancy	17	26.9%	25.0%	5.0%	75.0%
Other	17	29.6%	25.0%	1.0%	80.0%
All	74	26.5%	25.0%	1.0%	80.0%

**Question 2: Percentage of Projects Adversely Affected by Data Quality Issues**

<b>Employer</b>	<b>Number of Responses</b>	<b>Mean</b>	<b>Median</b>	<b>Minimum</b>	<b>Maximum</b>
Insurer/Reinsurer	40	32.5%	20.0%	3.5%	100.0%
Consultancy	17	37.6%	30.0%	5.0%	100.0%
Other	17	35.4%	25.0%	1.0%	100.0%
All	74	34.3%	25.0%	1.0%	100.0%

The discrepancy between the total numbers of 76 responses received and the numbers of responses to the two questions arises because some respondents only provided quantitative answers to one of the two questions.

The first point to make about these results is that they support the hypothesis that data issues have a significant impact on the work undertaken by general insurance actuaries. The mean response to question 1 implies that actuarial staff spends about a quarter of their time on issues of data quality. There was relatively little variation among employer groupings here with all three means covered by a span of less than five percentage points.

The responses to the second question also indicate that data quality is a major issue for general insurance actuaries since about a third of projects are adversely affected by data issues among responders. Again, there is relatively little variation among the means for the employer groupings with all three covered by a span of just over five percentage points.

### *Dirty Data on Both Sides of the Pond*

For both questions, the mean and median for insurers and reinsurers are lower than the mean and median for other actuaries. This may reflect that actuaries working for insurers and reinsurers will be more familiar with the data they are using than actuaries working for consultants, brokers or rating agencies.

It is clear from the above tables that we received a wide range of responses, with answers to question 1 varying between 1% and 80%, and those to question 2 varying between 1% and 100%. The range of responses was wide everywhere—of the two questions and three employer groupings, the narrowest range of responses was 70 percentage points. The wide range of responses on the significance of data quality issues within each employer grouping suggests that there may be something driving differences in data quality within each employer category. It could be that certain employers (or their designates) have been able to materially improve data quality over that of their peers. It should be noted that two responders attributed their low answers (<5% of projects adversely affected) to their companies' data scrubbing efforts.

Despite the wide variation in responses, data quality issues appear to be significant for most general insurance actuaries. Only 14% of the responses to question 1 were below 10%, and only 38% were below 20%. Similarly, on question 2, only 12% of the responses were below 10% and only 39% were below 20%. Only three respondents (4%) provided answers that were below 10% to both questions, and only 26% answered both questions with figures that were below 20%.

These survey results support our initial hypothesis that data quality problems impose a significant cost on industry.

## **4. DATA QUALITY EXPERIMENT**

While some of the anecdotal information communicated in the data quality stories in Section 2 support the claim that data quality issues can have a significant effect on businesses, the working party also wanted to provide quantitative information based on research about data quality issues. The data quality survey presents information on how actuaries and insurance professionals assess the severity of the problem, but is based on a limited sample. As a result, it is of only limited assistance in assessing the magnitude of the effect data quality problems have on the accuracy of estimates. In order to examine the

### *Dirty Data on Both Sides of the Pond*

effect of data quality problems on critical financial quantities, the working party conducted a data quality experiment with actual data used for an actuarial application. This experiment was designed to examine the effect of incomplete and/or erroneous data on loss reserve estimates. Real loss triangle data was felt to be more persuasive than conducting the experiment on a simulated dataset<sup>8</sup>. Data of sufficient maturity were obtained—all years are fully developed and the true ultimate losses are known—and various methods were employed to estimate ultimate losses using the data as of past valuation dates.

One of the data challenges that practicing actuaries frequently encounter relates to datasets that are severely limited with respect to the completeness of information provided. That is, the data may be limited with respect to the numbers of years of history (e.g., only five years of history for a long tail line where claims take 20 years to fully settle) or the types of data provided (e.g., only paid and incurred losses, but no reported claim count, closed claim count or exposure data). To simulate these situations, various projection methods were used on subsets of the original data to estimate the ultimate losses on the subsets.

Another data quality challenge that we investigated is data accuracy. Modifications were intentionally introduced into the data to simulate data errors and data quality problems commonly encountered. The various estimates of ultimate losses, based both on error-modified and unmodified datasets, were compared to the true ultimate losses to measure the accuracy of the estimates. In addition, the bootstrapping technique was used to compute measures of uncertainty for the reserve estimates for complete, incomplete, and error-modified data.

We begin with a brief discussion of the methods used to project ultimate losses in subsection 4.1. Subsection 4.2 summarizes the data. In subsection 4.3, we examine the impact of varying the size of the dataset by methodology. Subsection 4.4 discusses the modifications and errors introduced into the datasets and examines their impact on the estimates. Subsection 4.5 discusses a simple bootstrap analysis of the unmodified and error modified data. Finally, in subsection 4.6, we compare the results from the different estimates of the ultimate losses and we provide our observations and conclusions.

---

<sup>8</sup> Note that a working party of the Casualty Actuarial Society is developing a database incorporating known underlying trends and patterns and ultimate claim amounts to be used in reserving and other actuarial research, but their simulated data base is not yet available

## 4.1 Projection Methods

We restricted methodologies to mechanical approaches in order to filter out the effect of different actuaries making different subjective judgments. However we attempted to address material violations of the underlying assumptions of the methods. For example, a typical assumption of actuarial methods such as the chain ladder method is that the patterns and trends in the historic data do not change over time. As often happens in actual practice, our quick review of the loss-triangle data indicated that this assumption was not appropriate. It is clear that closing rates (see Closing Rate Triangle, Appendix B) on the most recent diagonals of the triangle are significantly higher than those of earlier years. Thus, looking at the 12-month development age, the closing rate for the most recent year, 1991, exceeds that of the earliest year, 1974, by a significant margin. A similar change in settlement rates over time can be observed through at least age 84 months. To adjust for the effect on loss development patterns, we applied a Berquist-Sherman (B-S) settlement rate adjustment (Berquist and Sherman, 1977) to one of the methods, the paid chain ladder. Note that the adjustment can only be applied if reported and closed claim counts are included in the data provided to the actuary for the reserve analysis. In addition, because the age-to-ultimate factors are very high (greater than 4.00) for the two most recent years, a Bornhuetter-Ferguson (B-F) method was used in addition to the chain ladder method for the paid data. Note that exposure data was used in estimating the B-F a priori estimate<sup>9</sup>. We believe the quality of the B-F estimate would be adversely affected if exposure data were unavailable.

The selected approaches for estimating ultimate losses are: (1) incurred chain ladder, (2) paid chain ladder, and (3) paid B-F and (4) paid chain ladder adjusted for accelerated closing rates using a B-S adjustment. We also provide some results for incurred chain ladder adjusted for closing rates. Note that we also tested a claim count times severity method (where each component's estimate is based on incurred data and the chain ladder method). Since the results were very similar to those of the incurred chain ladder, we chose not to report them.

## 4.2 The Data

A database with 18 accident years of data from accident years 1974 to 1991 was obtained.

---

<sup>9</sup> Losses were trended at a rate of 7% per year and divided by exposures (earned vehicle years). The trend rate was selected based on 1) our knowledge of the line of business during the 1980s and 2) testing of several trends to determine which seemed to perform best. An all-year average loss cost was selected as the B-F prior.

### *Dirty Data on Both Sides of the Pond*

The triangles contain an accident year in each row with annual evaluations of the statistic in each column (e.g., the second column is the cumulative value of the statistic at two years or 24 months of development). The data are from primary, private passenger automobile bodily injury liability business from a single no-fault American state. The data are direct with respect to reinsurance and limited to policy limits written. Policy limits distributions remained somewhat constant during the experience period. Although the data have been slightly adjusted to guard against identification, they are reflective of an actual situation. The data include paid losses, outstanding losses, number of reported claims, number of claims closed with payment, number of open claims, and exposures.

The “ultimate losses” were supplied by the provider of the triangles. However, because the original data were altered to hide the identity of the source, the “actual” ultimate losses do not exactly track the true actual numbers. The data are shown in Appendix B.

#### **4.3 Experiment 1: Impact of Reduced Completeness of Data**

It is not uncommon for actuaries to perform analyses on sparse data sets containing only a few years of data and only a few types of information. An example would be the actuary who is sent five accident years of incurred and paid loss data, including history for triangles, and is asked to estimate loss reserves. How much better would the estimate be if the actuary had 10 or 20 years of data, and had claim count and exposure data, as well as paid and incurred loss data?

In order to evaluate the effect of lack of completeness, subsets of the data were analyzed. Subsets were created with (1) all years, (2) only accident years 1986 to 1991, and (3) the latest three diagonals of information. The loss development pattern selected for each dataset is the volume-weighted average of all years. Note that the inverse power curve (Sherman, 1984) is used to estimate the tail factor for the 1986 to 1991 dataset. The ultimates estimated for each of the datasets is shown in Appendix C.

Two overall measures of accuracy were used in the analysis: 1) bias, that is, whether the overall estimate is near the “true” estimate, and 2) variability, as measured by the standard error, is used to assess the dispersion of estimates around the “true” value.

The projections based on paid loss triangle data are summarized in Figures 4.1 (unadjusted data) and 4.2 (B-S adjusted data). In each graph, the solid line with no markers represents the actual answer known with the benefit of hindsight, whilst the lines with

*Dirty Data on Both Sides of the Pond*

markers show the results based on the three datasets.

Figure 4.1: Estimated Ultimate Losses by Year Based on Unadjusted Paid Data

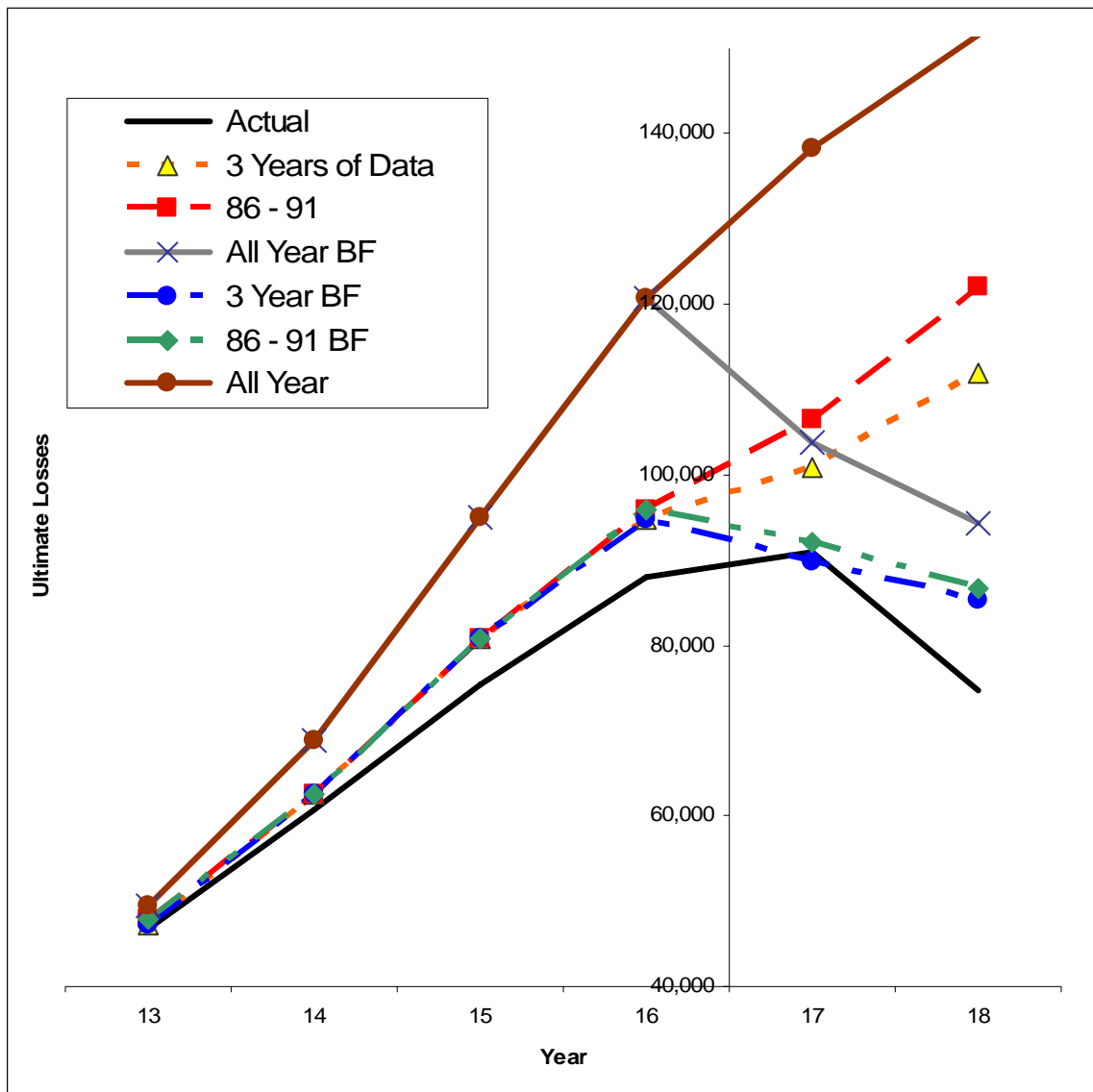
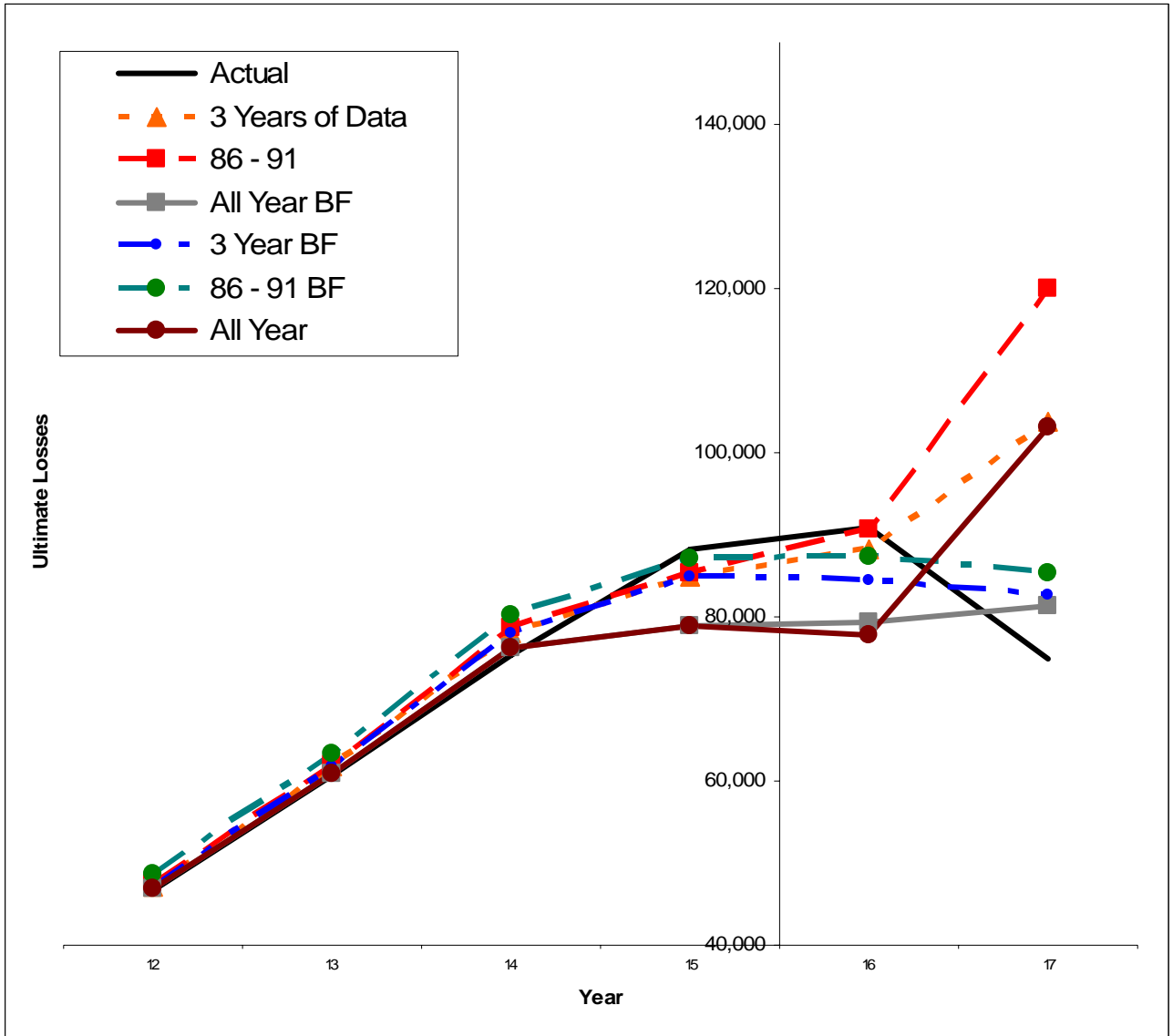


Figure 4.2: Estimated Ultimate Losses by Year Based on Adjusted Paid Data



A brief inspection of the estimated ultimate losses arising from paid (Figure 4.1) chain ladder method indicates that the paid chain ladder estimated ultimate losses tend to be higher than the actual ultimate losses. This is largely due to the impact of the 12-to-ultimate factor and to a lesser extent to the factors from other immature years. A more stable approach such as a B-F model is appropriate in this situation, but our implementation of the Bornhuetter–Ferguson required additional data, namely exposures. Thus to improve on the paid chain ladder estimate, additional data beyond just paid and incurred loss aggregates was

### *Dirty Data on Both Sides of the Pond*

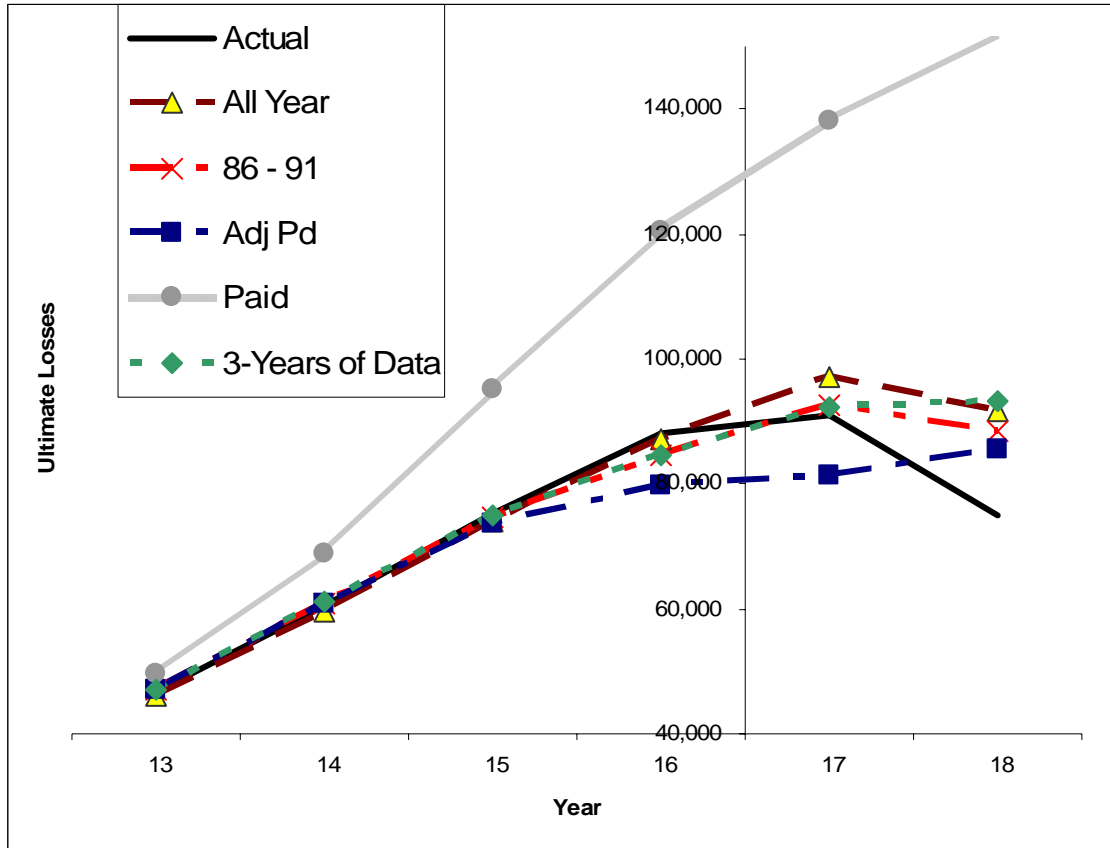
required. As the exposures varied considerably over the historic period, the absence of this data would likely have significantly affected the quality of the estimates. We note that the smaller datasets (3-diagonals and 1986-1991) performed better on the chain ladder paid ultimates than the all-year dataset. This reflects that these data were more responsive to recent changing patterns in the data.

Figure 4.2 indicates that there is a significant improvement in the quality of the estimates when the B-S adjustment is used. The B-S adjustment adjusts the historic paid loss diagonals to match the claim closing rates of those diagonals to that of the latest diagonal.<sup>10</sup> Such an adjustment requires data that is often not present in small datasets supplied to actuaries for reserving and pricing analyses. For the adjusted data, the ultimate losses based on 1986-1991 only are the least accurate, while the all-year and 3-year datasets perform about the same. As with the unadjusted data, the B-F method performs better than the chain ladder method.

---

<sup>10</sup> More advanced methods using regression modeling (Zehnwirth, 1994) and generalized linear models (Taylor, 2004) might be applied by actuaries encountering dynamic patterns in their data. For this analysis, the working party restricted itself to approaches that could be applied mechanically.

Figure 4.3: Estimated Ultimate Losses by Year Based on Unadjusted Incurred Data



*Dirty Data on Both Sides of the Pond*

Figure 4.3 presents the results for estimated losses based on incurred loss data. For comparison, the graph also displays the ultimate losses from the all years paid and adjusted paid techniques. It is clear from this graph that the estimated ultimate losses based on incurred loss data are considerably more accurate than the unadjusted paid chain ladder ultimate losses. All the incurred loss datasets appear to provide reasonable estimates of ultimate losses.

Some statistics from the data quality experiment are presented in Tables 4.1 and 4.2. The statistics presented are 1) the overall bias of the method, defined as the sum of the actual ultimate losses minus the sum of the estimates of the ultimate losses for the methods/datasets, and 2) the standard error of the estimate, which is the average of the squared deviations of actual ultimate losses from estimated ultimate losses:

$$(4.1) \quad se = \sqrt{\frac{\sum_{i=1}^N (Y_i - \hat{Y}_i)^2}{N-1}}, \quad Y_i \text{ is actual ultimate, } \hat{Y}_i \text{ is estimate}$$

**Table 4.1: Bias of Estimation Methods and Datasets**

	<b>All Years</b>	<b>3- Years</b>	<b>86 - 91</b>	<b>All Year BF</b>	<b>3-Year BF</b>	<b>86 - 91 BF</b>
<b>Paid</b>	188,759	62,011	98,353	97,019	24,140	44,377
<b>Adjusted Paid</b>	6,599	26,502	59,234	-13,401	1,552	16,571
<b>Incurred</b>	17,803	16,100	-9,490			
<b>Adjusted Incurred</b>	-8,435	18,753	12,344			

**Table 4.2: Standard Error of Estimation Methods and Datasets**

	<b>All Years</b>	<b>3-Years</b>	<b>86 - 91</b>	<b>All Year BF</b>	<b>3-Year BF</b>	<b>86 - 91 BF</b>
<b>Paid</b>	5,460	2,197	3,137	2,525	765	1,098
<b>Adjusted Paid</b>	1,806	1,633	2,679	896	618	705
<b>Incurred</b>	1,003	1,048	566			
<b>Adjusted Incurred</b>	933	1,276	1,053			

Table 4.1 indicates that the unadjusted paid loss estimates have a significant bias that is somewhat mitigated by applying the B-F technique. The adjusted paid methods perform

significantly better, although the 3-year and 1986-1991 adjusted paid chain ladder methods still have significant bias. While the incurred chain ladder method has less bias than the paid chain ladder method, the size of the dataset does not appear to improve the overall bias of the estimates—indeed, the smallest bias for the incurred data (based on absolute values) is for the 1986-1991 dataset. For informational purposes we also show the results for the incurred method when the B-S adjustment is applied. The all-year incurred chain ladder method bias is improved by using data with the settlement rate adjustment.

For the paid datasets, the standard error of the estimate (Table 4.2) is highest for the chain ladder method applied to the all-year unadjusted paid loss data. It is least for the B-S adjusted data using the B-F method. All the incurred loss estimates have relatively modest standard errors. It is not clear that the size of the dataset significantly impacts the incurred ultimates.

Observations:

- The adjusted paid and the incurred methods produce reasonable estimates for all but the most immature points (however, these points contribute the most dollars to the reserve estimate).
- The paid chain ladder method, which is based on less information (no case reserves, claim data or exposure information), produces worse estimates than the methods based on the incurred data or the adjusted paid data.
- It is not clear from this analysis that datasets with more historical years of experience produce better estimates than datasets with fewer years of experience.

## **4.4 Experiment 2: Impact of Reduced Data Accuracy**

### **4.4.1 Data Modifications to Simulate Data Quality Problems**

Based on actual experiences of members of the working party, we postulated various events that cause data glitches such as systemic misclassification of claims to the wrong accident year and erroneous entries escaping systems edits. The datasets were then modified to reflect the effects of such issues. The working party decided to introduce more than one error at a time to improve the realism of the scenario and to explore how the interaction of errors can affect estimates.

### *Dirty Data on Both Sides of the Pond*

The error-modified triangles simulate the following data quality issues:

1. Losses from accident years 1983 and 1984 have been misclassified as 1982 and 1983 respectively.
2. Approximately half of the financial movements from 1987 were processed late in 1988.
3. The incremental paid losses for accident year 1988 development period 12-24 has been overstated by a multiple of 10. This was corrected in the following development period. Similarly, an outstanding reserve for a claim in accident year 1985 at the end of development month 60 was overstated by a multiple of 100 and was corrected in the following period.
4. Data prior to the 1982 calendar year is not available.
5. The paid losses in the latest diagonal are crude estimates rather than actual losses.
6. From 1988 onwards, the definition of “reported claims” was changed to exclude claims closed without payment.

The projections based on the modified data appear in Appendix D.

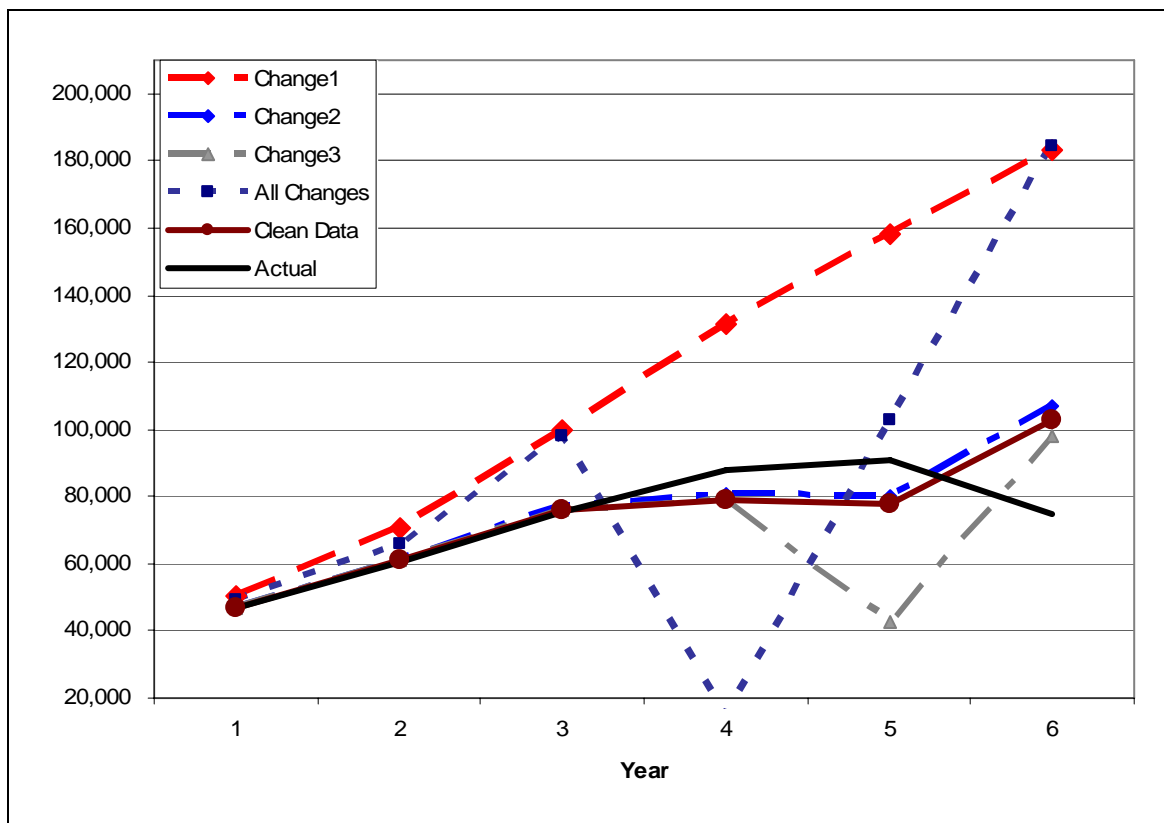
For simplicity of presentation, results are presented only for the “all year” datasets. Again, all of the methods used to project the claims are mechanical: there is no judgment involved. This means, for example, that in places where there is missing data, the development factors based on volume-weighted averages will be wrong because there is a mismatch between the numbers of years containing claims figures in the numerator and the denominator. In practice, an actuary may well spot this and correct the data glitches, but we wanted to use a mechanical approach and demonstrate the more extreme distortion caused by a failure to do so.

Since analyzing data containing all the errors seems somewhat extreme, we also selected some “errors” to be applied to the data individually. In order to keep the number of permutations of scenarios to a manageable level, only the first three “errors” were applied separately to the data. Results are presented for each of error modifications 1 through 3 and for data reflecting all 6 modifications.

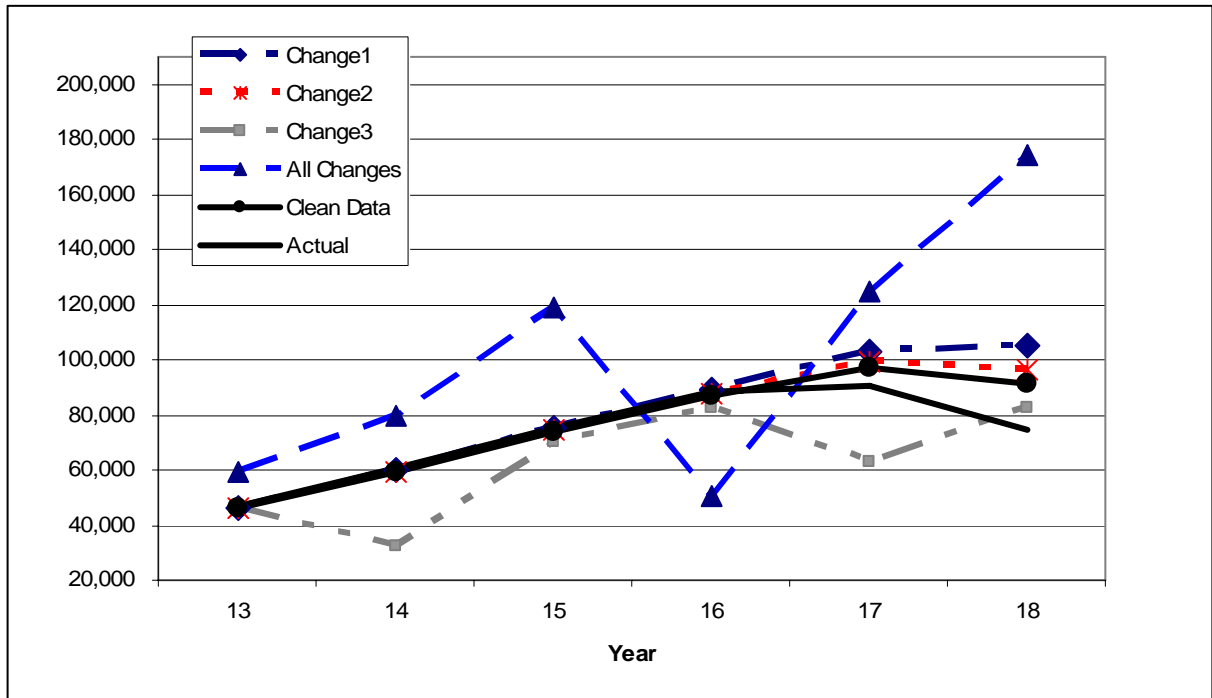
#### 4.4.2 Results

Figures 4.4 and 4.5 show the comparison of the actual ultimate losses to estimates of ultimate losses based on “clean” (unmodified) data and on data modified to introduce errors. The results shown are for chain ladder method applied to adjusted paid loss data (Figure 4.4) and to unadjusted incurred loss data (Figure 4.5).

**Figure 4.4: Comparison of Actual and Estimated Ultimate Losses Based on Error-Modified Paid Loss Data**



**Figure 4.5: Comparison of Actual and Estimated Ultimate Losses Based on Error-Modified Incurred Loss Data**



The graphs indicate that some of the projections based on error-modified data are extremely volatile, particularly for reserve values based on paid losses. When compared to the unmodified or clean data, the results for the error-modified data show a large amount of both additional volatility and bias. In practice an actuary will likely spot many of the errors and try to correct for them. Nevertheless the actuary will often be unable to get back to the correct data and will be forced to compensate for the problem with a data adjustment. Thus some of the additional volatility and error will almost certainly remain. Indeed, in some cases, an attempt to correct the data may introduce additional volatility and bias.

Table 4.3 presents the bias (i.e., the overall error between actual and estimated ultimates) for each of the error-modified datasets for four different methods of estimating ultimate losses. In general, the error-modified data results in estimates that have a higher bias than the clean data, but there are a couple of exceptions. The exceptions occur in the use of two of the paid methods on the data reflecting change 3 (an error in the 1988 paid losses at 24 months and 1986 outstanding losses at 60 months).

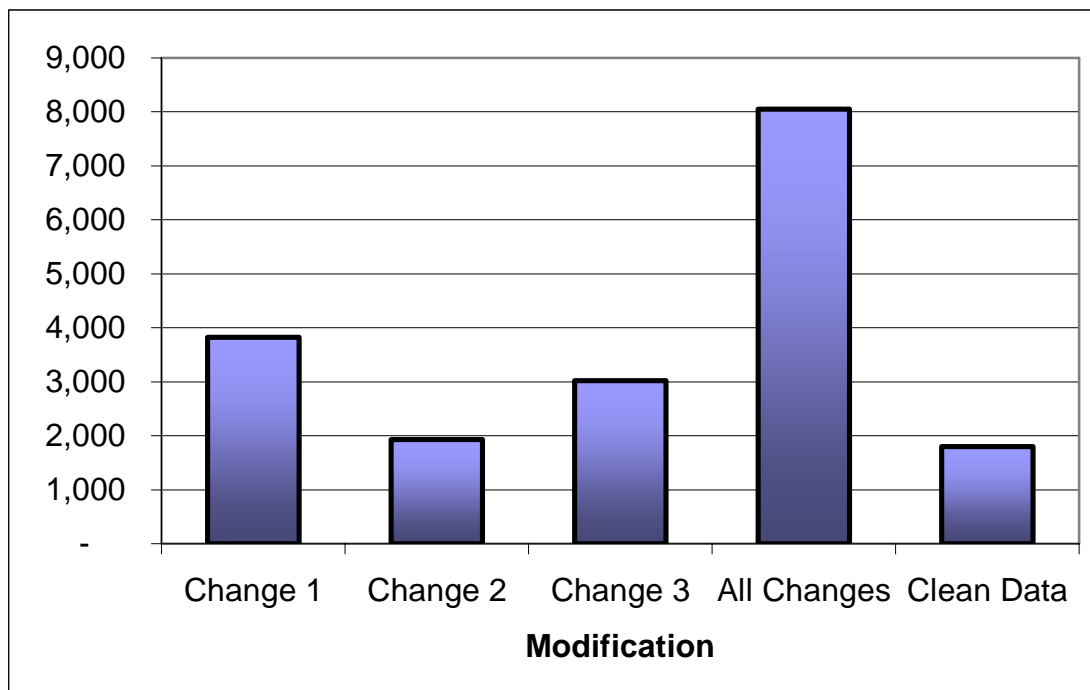
*Dirty Data on Both Sides of the Pond*

Since positive and negative errors that offset each other could produce results that exhibit low bias overall, we also present the standard error of the estimates. These are displayed graphically in Figure 4.6 for the adjusted paid estimation methods and Figure 4.7 for incurred data.

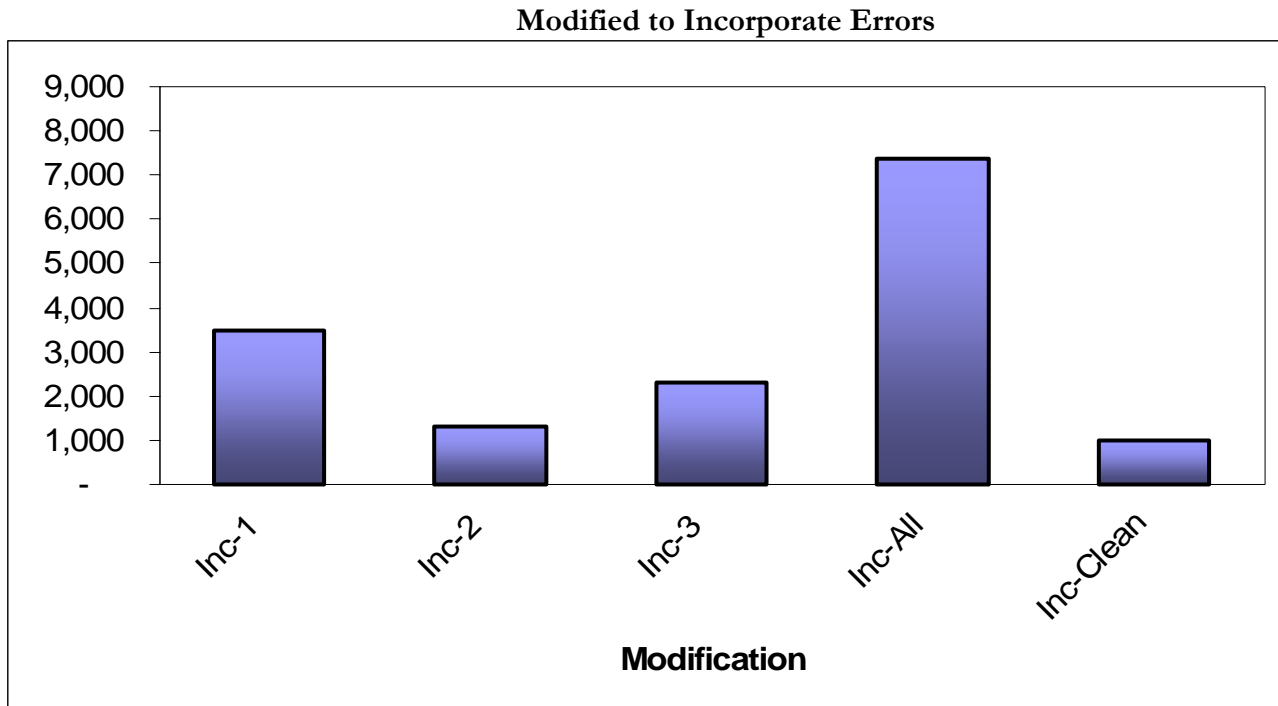
**Table 4.3 Bias of Estimation Methods and Datasets**

	<b>Change 1</b>	<b>Change 2</b>	<b>Change 3</b>	<b>All Changes</b>	<b>Clean Data</b>
<b>Paid</b>	257,669	206,735	103,081	231,168	188,759
<b>Adjusted Paid</b>	38,862	15,994	-33,454	98,673	6,599
<b>B-F Paid</b>	126,833	104,716	63,857	108,220	97,019
<b>Incurred</b>	41,392	25,948	-61,542	173,703	17,803

**Figure 4.6: Standard Errors for Adjusted Paid Loss Data Modified to Incorporate Errors**



**Figure 4.7: Standard Errors for Incurred Loss Data**



The error-modified data reflecting all changes results in estimates having a higher standard error than that for the clean data. From Table 4.3 and Figure 4.7 it is also clear that for incurred ultimate losses, the clean data has the lowest bias and lowest standard error.

We suspect that some of the results for the paid loss data, especially results obtained for our analysis of reduced-size datasets, are a result of happenstance and the unique features of the dataset used in this analysis. The accuracy of estimates is particularly sensitive to the instability of paid ultimate losses for a few recent accident years. Thus the results may reflect the quirks of one particular dataset, which is itself a single realization of many possible loss scenarios. That is, process variance may be the source of unexpected results when comparing the accuracy of different datasets due to the happenstance of how particular random realizations affect ultimate estimates for a few key years. A more representative assessment of the impact of data quality issues might be provided by a stochastic analysis, where many possible realizations are considered.

## **4.5 Bootstrapping**

When measuring the quality of different estimation procedures, actuaries often quantify their uncertainty by estimating a probability distribution for ultimate losses (or reserves). In this section, the bootstrap approach is used to derive a probability distribution for estimated reserves for 1) clean data, 2) incomplete data, and 3) modified data containing errors.

### **4.5.1 Description of Bootstrapping**

A limitation of the deterministic analyses we have performed is that they are based on single realizations of reported claim counts, closed counts, paid losses, and incurred losses from a distribution of potential outcomes. Other realizations would have resulted in different development factors and different ultimate loss estimates using the same estimation methods and based on the same underlying stochastic processes generating the data. In order to augment our analysis with information about a distribution of realizations for the development factors, the technique of bootstrapping was used. Bootstrapping is a computationally simple way of obtaining prediction errors and probability distributions of the predictions. In its simplest form, bootstrapping assumes that the empirical data supply a probability distribution that can be sampled to derive uncertainty measures of functions (such as means, sums, and projected ultimates and reserves) based on the data. For instance, one could randomly sample loss development factors from each column of a triangle of loss development factors and use these to randomly compute new estimates of ultimates. However, because the size of the sample for each factor is limited, particularly for more mature development periods, a bootstrap procedure that uses all the observations on the triangle for each sampling has become popular with actuaries. The procedure is based on sampling from deviations of observations from their means. A description of the procedure is provided by England and Verrall (1999, 2002). The procedure is widely used in quantifying the uncertainty of loss reserve estimates.

We refer to the implementation of the bootstrapping technique used here as the chain ladder bootstrap method. The approach is based on recreating many realizations of the incurred and paid triangles by sampling from a distribution of standardized deviations of incremental triangle values. The method uses link ratios to estimate the “expected” amounts in each cell of the loss development triangle. It then computes the deviation of the actual incremental loss value for an accident year and development age from its expected value. The paid and incurred link ratio methods were used in the bootstrapping. Based on the

*Dirty Data on Both Sides of the Pond*

outperformance of the adjusted paid ultimates above, our paid bootstrap analysis was performed on adjusted data only.

The original and error-modified data for both the paid and incurred losses were passed through a mechanical bootstrapping process. The process used a freeware Microsoft Excel bootstrapping spreadsheet that is currently being distributed at a Limited Attendance Seminar on Reserve Variability<sup>11</sup>. The following broad steps were followed in the calculation:

- A link ratio model was fitted to derive the best estimate of the development pattern underlying the data. Link ratio selections were based on a weighted average of all years of data.
- An “expected triangle” of data was derived by applying the development factors backwards from the latest values on the diagonal of the triangle. Thus, the current latest point of each origin year can be arrived at by following the derived fitted loss development pattern precisely.
- A triangle of raw incremental residuals was calculated by subtracting the actual data from the expected incremental data triangle.
- Pearson residuals were derived from the raw residuals. The Pearson residual is a generalization of the well known z-score or standardized residual. For the Pearson residual, the raw residual is divided by the square root of the variance of the expected value, which is dependent on the distribution assumed. Under the assumption of normality, the Pearson residual and the z-score are the same. The Pearson residual is a concept commonly used in the generalized linear models context<sup>12</sup>:

$$(4.2) \quad r = \frac{x - \mu}{\sqrt{\text{Var}(\mu)}}, \text{ x=actual value, } \mu \text{ its expected value}$$

- 5,000 simulations were run on each set of data. During each simulation, the adjusted residuals were sampled and added to the expected triangle to generate a

---

<sup>11</sup> The seminar was sponsored in 2006 and 2007 by the Casualty Actuarial Society. In November of 2007, the U.K. Actuarial Profession will sponsor the seminar for its members. Significant modification of the formulas in the spreadsheet was required to tailor it to the datasets and methods used in the data quality experiment.

<sup>12</sup> Following the procedures described by England and Verrall (England and Verrall, 1999), the variance is assumed to be proportional to the expected value.

### *Dirty Data on Both Sides of the Pond*

new data triangle. The link ratio projection method was then applied to each of the generated data triangles to produce an estimate of the ultimate losses. The estimated ultimate losses resulting from this process reflect parameter (i.e., estimation) variance, but not process variance.

- During each simulation, a parametric distribution assumption (the gamma distribution) was applied to add process variance to the future realizations of incurred and paid losses (to “square” the triangle).

It should be noted that underlying the chain ladder bootstrap method is the assumption that the chain ladder is an appropriate model for the data. Venter (1997) describes a number of statistical and graphical tests that can be performed to test the assumptions of the chain ladder. For the purposes of this “experiment,” we assumed that the chain ladder model was appropriate and used the bootstrap to create random samples of possible triangles and “true” ultimate losses and then tested the impact of various data quality impairments on the accuracy of estimated reserves.

Bootstrap results for the total reserves were generated based on each of the complete unmodified, reduced unmodified, and error-modified data. In addition, the “true” ultimates and reserves were computed for each simulation. The deviations of estimated from “true” reserves was then computed. Percentiles were calculated from the bootstrapped results.

#### **4.5.2 Results**

Table 4.4 presents some summary statistics from the bootstrap analysis using the incurred method for selected datasets. The datasets displayed are 1) the complete (i.e., all 18 years of data) clean dataset, 2) the 1986-1991 dataset, no errors, and 3) the complete 18-year dataset containing all six errors. The table also presents the distribution of “true” reserves. Descriptive statistics from the bootstrap are presented at the top of the table followed by a display of the results at various percentiles of reserves from the selected datasets.

The table indicates that reserve distributions based on small datasets and on error-modified datasets have a lot more variation than those reserve distributions based on clean data that includes the entire sample. Note that the distribution of “actual” reserves includes process variance, while the distribution of reserve estimates from the various samples includes only parameter variance, i.e., variability from the estimates in reserves, while not reflecting how far the reserves are from the “true” simulated ultimate and its “true” reserve.

*Dirty Data on Both Sides of the Pond*

Also, note that the bootstrap sample that generated the 1st percentile of the actual reserve distribution may be different from the sample that generated the 1st percentile of the modified data sample. While Table 4.4 provides information regarding the variability of estimates from different datasets, our focus is actually on the deviation of actual needed reserve from estimated reserves (or alternatively, of estimated ultimate loss from true ultimate loss).

**Table 4.4: Bootstrap Results Based on Incurred Chain Ladder Method**

	<b>Incurred Actual Reserve</b>	<b>Incurred Clean Data</b>	<b>Incurred 1986-1991 Data</b>	<b>Incurred Modified Data</b>
Mean	178,677	181,257	159,743	341,943
standard dev	27,034	25,927	47,282	41,760

**Percentile**

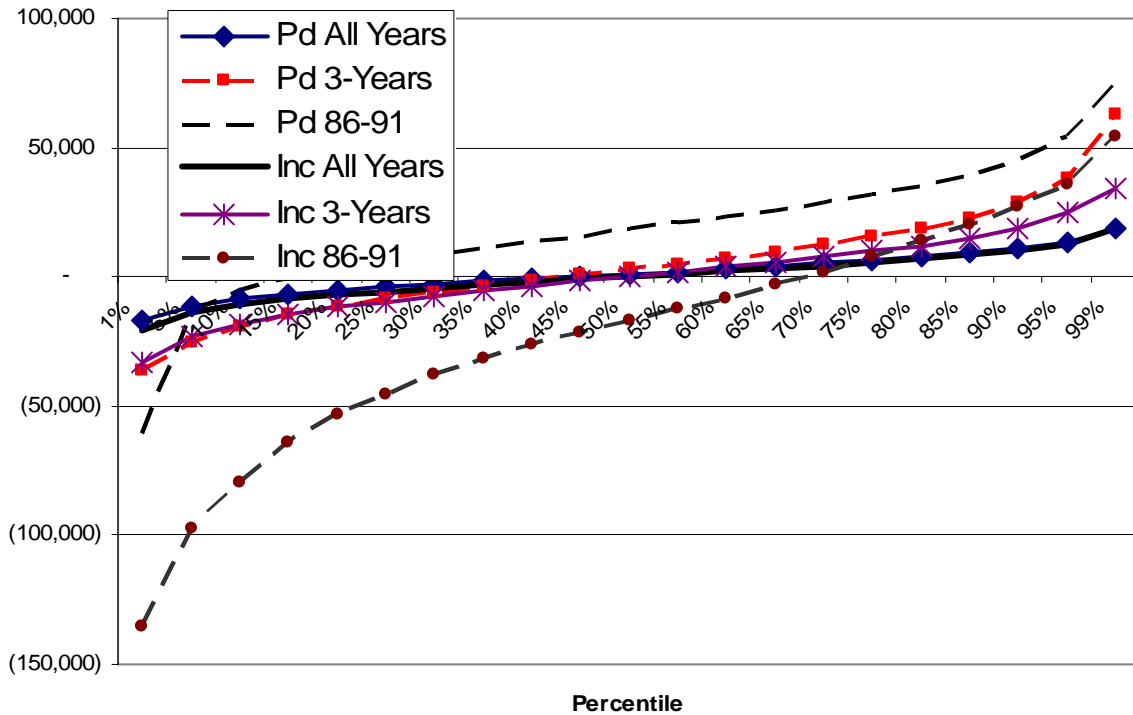
1%	118,025	123,185	28,726	104,951
5%	134,744	140,407	74,652	127,661
10%	144,637	148,772	98,521	140,483
20%	156,301	159,570	122,800	155,756
30%	164,503	167,227	139,616	166,702
40%	171,097	174,247	152,335	175,493
50%	177,926	180,991	163,383	184,596
60%	185,082	187,202	174,401	193,748
70%	192,162	194,103	185,269	203,228
80%	200,151	202,399	198,520	213,917
90%	214,139	214,794	215,468	231,400
95%	224,207	225,016	230,045	244,574
99%	243,599	244,943	259,289	268,835

From each bootstrap simulation the difference or “error” between the reserve estimate and the “true” reserve<sup>13</sup> was tabulated. Figure 4.8 displays the cumulative distribution of errors from the bootstrap experiment for the complete and incomplete data sets.

---

<sup>13</sup> Note that the “true” reserve was also a stochastic variable that varied for each bootstrap simulation.

**Figure 4.8: Distribution of Reserve Estimation Errors for Datasets of Different Sizes**



This graph indicates that the “errors” are much larger for both paid and incurred reserve estimates for the incomplete data, and are largest for the 1986-1991 datasets. It can also be observed that the incomplete data is more variable than the complete data and that, at the extreme low and high percentiles, the incomplete 1986-1991 paid and incurred datasets show very large deviations from the “true” values.

*Dirty Data on Both Sides of the Pond*

For each dataset and method the average error was computed and is displayed for all bootstraps in Table 4.5. From Table 4.5, it is also apparent that the overall bias of the estimated reserves is greater for the incomplete and error-modified data. The table indicates that the incurred reserves using all years of clean data have minimal bias while the incurred estimates computed with data containing all six errors have a mean error approximately equal to 100% of the “true” reserve value of \$170,000.

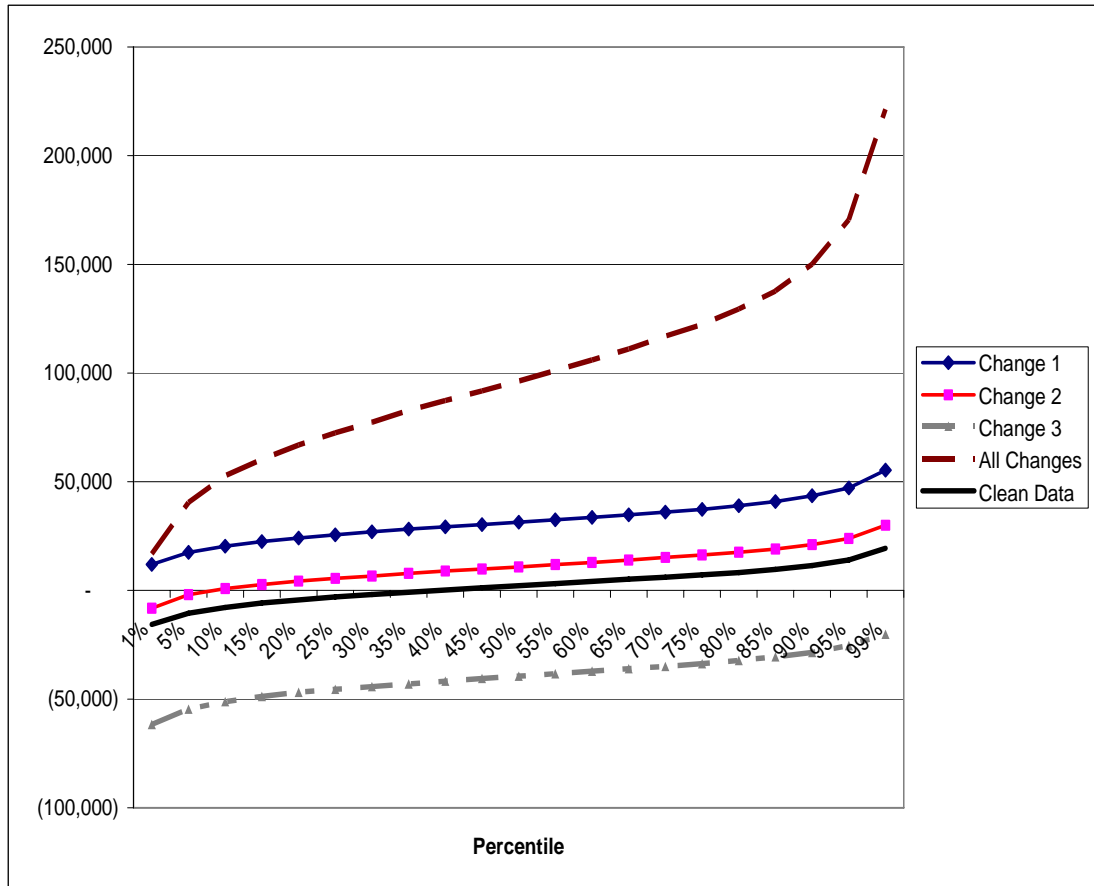
**Table 4.5: Bias in Estimated Reserve by Method and Dataset**

	<b>Paid Estimates</b>	<b>Incurred Estimates</b>
<b>Unmodified Data</b>		
All Year Clean	1,196	(94)
3-Year	4,238	214
86-91	11,774	(21,605)
<b>Modified Data</b>		
Change 1	31,716	21,861
Change 2	10,915	6,556
Change 3	(39,678)	(36,779)
<b>All Changes</b>	<b>99,552</b>	<b>163,266</b>

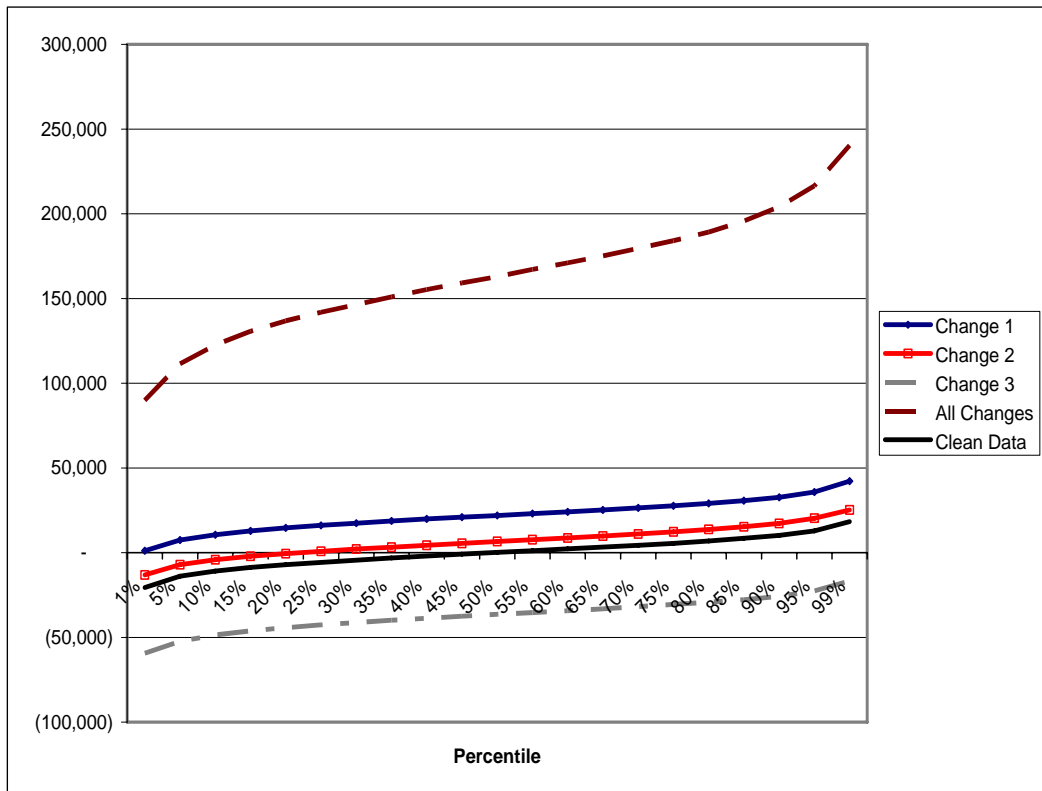
Figures 4.9 and 4.10 display the average error of the cumulative distribution of errors for the data modified to contain inaccuracies. For both adjusted paid estimates (Figure 4.9) and incurred estimates (Figure 4.10) the reserves based on the clean data are clearly more accurate and less uncertain than the modified data.

The distributions and statistics from the bootstrap analysis confirm our original hypothesis—the accuracy of ultimate loss estimates based on poor quality data is significantly worse than the accuracy of ultimate loss estimates based on accurate data, and that the variability is significantly higher. While actuarial estimates usually contain uncertainty, when estimating loss reserves using data not processed through a rigorous quality review process, the uncertainty is likely to be much greater, and therefore the magnitude of any under- or over-estimation is likely much higher than for data that have been screened.

Figure 4.9: Distribution of Reserve Estimation Errors for Paid Ultimates Based on Modified Data



**Figure 4.10: Distribution of Reserve Estimation Errors for Incurred Ultimates Based on Modified Data**



#### 4.6 Summary of Results from the Data Quality Experiment

Two different experimental approaches were applied to various sets of a reserving database that were 1) clean and complete, 2) clean but incomplete, and 3) contained intentionally introduced errors. The first approach, a deterministic approach, applied several traditional actuarial techniques to actuarial loss development data where actual ultimates were known. The second approach applied a stochastic bootstrapping method to the datasets to evaluate both the bias and uncertainty of the various data impairments.

The results of the deterministic approach did not support the claim that datasets with fewer historic observations result in less accurate estimates than data sets with many historic observations. But, in general, the results did support the claim that inaccurate data may 1) on average produce biased estimates and 2) provides more uncertain estimates. In addition, the deterministic analysis indicated that dramatic improvements were observed when paid

### *Dirty Data on Both Sides of the Pond*

estimates included certain modifications to the methodology. These modifications included a Berquist-Sherman adjustment for accelerated settlement rates and a Bornhuetter-Ferguson method applied to the two most recent accident years. Both adjustments require additional types of data not contained in the incurred and paid triangles.

The stochastic approach, based on applying the chain ladder bootstrap procedure to the incurred and paid data, produced more consistent results. The bias of the reserve estimates increased and their precision decreased for both the reduced datasets and the datasets containing inaccuracies. For the datasets with inaccuracies, the dataset containing all six errors produced estimates with a large bias and extreme volatility. As this represents an extreme scenario, we selected some of the errors to model individually. Each of these individual errors had a significant impact on the quality of the estimated reserve.

Our research is only a beginning in examining the consequences to insurance companies of data quality problems. It was limited to one relatively small dataset. A variety of datasets from a variety of lines of business would provide a more complete picture of the impact of data quality problems on loss reserve estimates. In addition, we examined the effect of data quality on only one kind of insurance application. We did not address the effect of data quality problems on other common actuarial analyses such as pricing and classification reviews. Also as insurance companies continue to expand their use of predictive models, a very data-intensive activity, actuaries and predictive modelers must be aware of the impact on their work of errors in large corporate databases and in the other external datasets relied on in building the models.

The data quality experiment supports the conclusion that more accurate and complete, error-free data yields more accurate results. Consequently, we believe our research indicates that the most efficient way to mitigate the consequences is to minimize errors in the data by ensuring that quality data enters systems, that errors are corrected promptly, and that the systems and processes handling the data are error-free.

## **5. DATA QUALITY ADVOCACY**

Because actuaries are typically heavy users of data and must frequently contend with poor quality data, we believe actuaries should become data quality advocates. In the next section, we describe some actions that can be taken by actuaries and insurance company

managements to improve data quality.

## **5.1 Data Quality Advocacy**

Currently, two organizations in the United States are working to increase the profile of data quality issues in the property/casualty insurance industry:

- The CAS is sponsoring the Committee on Data Management and Information and the Data Management and Information Education Materials Working Party. The Working Party sponsors a number of activities, including presentations at seminars, and has authored two papers on data quality. The Committee sponsors a Call Paper Program jointly with the IDMA on data management every other year.
- The Insurance Data Management Association (IDMA) is an excellent source of information on insurance data quality.
  - The IDMA Web Site contains “value propositions” that describe the value of data quality from the perspective of various insurance stakeholders, e.g., senior management, claims, marketing, and actuaries.
  - The IDMA also sponsors an annual conference where data quality is typically a topic on the schedule and its Web site contains suggested readings on data quality.

These are examples of data quality advocacy that can be undertaken by professional actuarial and industry organizations. More specific actions that can be taken to improve data quality within organizations are discussed next.

### **5.1.1 Data Quality Measurement**

As a tool for promoting data quality improvement, a number of authors recommend regular measurement of an organization’s data quality (Dasu and Johnson, 2003; Redman, 2001). Among the advantages of measurement noted by Redman<sup>14</sup> are that measurement replaces anecdotal information with factual data, quantifies the severity of the problem, and identifies where the problems are so they can be acted upon.

---

<sup>14</sup> Redman, p. 107

### *Dirty Data on Both Sides of the Pond*

Some of the measures recommended by Dasu and Johnson quantify traditional aspects of quality data such as accuracy, consistency, uniqueness, timeliness, and completeness. Some capture systems-related aspects of data quality such as the extent of automation (sample some transactions, follow them through the database creation processes, and tabulate the number of manual interventions) and successful completion of end-to-end processes (count the number of instances in a sample that, when followed through the entire process, have the desired outcome). Yet others are intended to measure the consequences of data quality problems (measure the number of times in a sample that data quality errors cause errors in analyses, and the severity of those errors). Dasu and Johnson recommend that the different metrics be weighted together into an overall data quality index using business considerations and the analysts' goals to develop weights.

Redman points out that the most appropriate measure depends on the organization. An organization that is just beginning its data quality initiative probably only needs simple measures, while a more advanced organization might employ more sophisticated measures. Redman offers the following algorithm for implementing a simple data quality measure<sup>15</sup>:

- determine who will take the action
- select a business operation
- select needed data fields
- draw a small sample
- inspect sampled records
- estimate impact on business operation
- summarize and present results
- follow up

#### **5.1.2 Advocating Data Quality—Management Issues**

In this section we briefly summarize some of the recommendations in the data quality literature for implementing data quality programs.

For data originating within one's company, Redman suggests managing the information

---

<sup>15</sup> Redman, p108

### *Dirty Data on Both Sides of the Pond*

chain. Redman notes that most information is distributed horizontally. For instance, an information technology department programs and maintains a claims system that collects and stores claims data, and performs edits on data as they are entered. Claim adjusters record information into the claims system. Actuaries use the claims data, perhaps after aggregation by yet another department. The flow of this data is from department to department, not hierarchically. Redman notes that departments often do not communicate effectively with each other and this exacerbates data quality problems. He suggests that once departments understand the needs of the users of the data, they will be more motivated to satisfy those needs. Redman describes a formal program for information chain management including<sup>16</sup>

- establish management responsibilities
- describe information chain (information flow)
- understand customer needs
- establish measurement system
- establish control and check performance
- identify improvement opportunities
- make improvements

Redman suggests that some middle managers will resist data quality initiatives, thinking their jobs may be eliminated (because as data processes become more efficient fewer people are needed) and that managers should be assured that this will not occur.

Redman advocates supplier management for data originating outside the company, stating, “The most difficult aspect of supplier management for most organizations is coming to the realization that they have contributed to the inadequate data quality they currently receive. They believe that these suppliers are simply incompetent, don’t care, don’t have enough good people or use old technology.”<sup>17</sup> On the contrary, Redman suggests that organizations do not provide adequate communication and feedback to their data suppliers. Thus Redman suggests<sup>18</sup>

- customers define for the supplier the quality of the data they need

---

<sup>16</sup> Redman, p.162

<sup>17</sup> Redman, p. 154

<sup>18</sup> Redman, p.155

### *Dirty Data on Both Sides of the Pond*

- the supplier measures baseline performance as to how well the requirements are met
- the supplier and user agree on improvements
- the supplier regularly remeasures performance

## **5.2 Screening Data**

Even when data quality initiatives have been undertaken, actuaries and other analysts will need to screen their data. Moreover, a point made in the data quality literature (Redman and CAS DMIWP) is that everyone who uses data has a role in assuring its quality. A fairly extensive literature relevant to data quality exists in statistical journals and publications. This includes the tools of exploratory data analysis (EDA), pioneered by Tukey (Hartwig and Dearing, 1979 discuss Tukey's contribution). Exploratory data analysis techniques are particularly useful for detecting outliers. While outliers, or extreme values, may represent legitimate data, they are often the result of data processing glitches and/or coding errors. The CAS DMIWP and Francis (CAS DMIWP, 2008; and Francis, 2005) describe a number of exploratory techniques useful for screening data and illustrate their application insurance data. Some of the EDA methods recommended include:

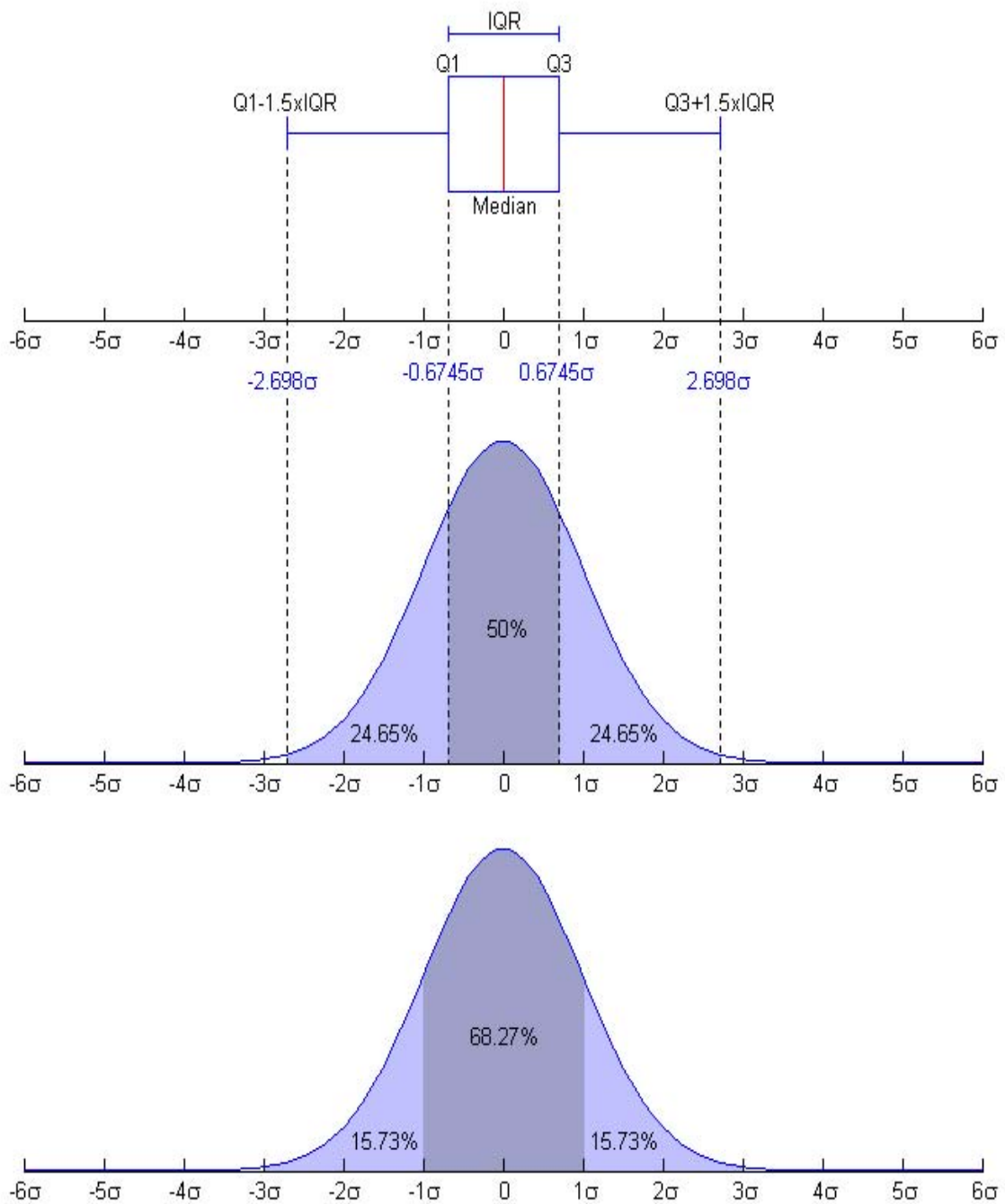
- produce and examine descriptive statistics such as mean, median, minimum, maximum, and standard deviation of each numeric field
- for categorical variables, tabulate the frequency of records in the database for each value of the categorical variable
- tabulate the percentage of records with missing values for each variable
- produce histograms of numeric fields (possibly on a log scale for loss amounts) and categorical variables
- produce box-and-whisker plots of numeric fields (possibly on a log scale for loss amounts)
- examine databases for records with duplicate values in fields which should be unique (such as claimant identifier)
- apply multivariate techniques that screen multiple variables for outliers simultaneously or that screen for invalid combinations, e.g., state and ZIP Code.

### **5.2.1 Primer on Box-and-whisker Plots**

Simple summaries or descriptive statistics can be used to describe the basic characteristics of a database. These statistics usually include the mean (either the arithmetic or geometric), median, mode, minimum, maximum, variance, and standard deviation.

John Tukey introduced the box plot concept in 1977 as a visual tool for summarizing these descriptive statistics in a one dimensional chart. A box plot (also known as a box-and-whisker diagram or plot) is an easy-to-view, graphical way of depicting the five-number summary, which consists of the smallest observation, lower quartile (Q1), median, upper quartile (Q3), and largest observation. The box plot also indicates which observations, if any, are considered unusual, or outliers. Figure 5.1 compares the box plot against a probability density function for a normal  $N(0,1\sigma^2)$  distribution and provides a pictorial for understanding the box plot. The commonly used box-and-whisker plot incorporates a refinement of separately displaying outliers beyond the range of the “whiskers.” The box-and-whisker plot is a very useful graphical tool for EDA. Appendix A shows an example of using it to screen data.

Figure 5.1: Schematic of the Box Plot from Wikipedia (www.wikipedia.org)  
Based on Normally Distributed Data



### **5.2.3. Software for Screening Data**

The CAS DMIWP (CAS DMIWP, 2008) paper describes how to obtain many descriptive statistics and EDA graphs using Microsoft Excel. In addition, a number of free open-source products, including the popular statistical language R, are available to analysts wishing to augment the capabilities of Microsoft Excel. In this section we introduce a lesser known shareware software package, ViSta. The ViSta software is an open source product with an exclusive focus on techniques for visualizing data. Appendix A of this paper provides a brief introduction to ViSta and describes a procedure for importing data into the software. The ViSta product is based on the XLisp language and the free statistical package XLisp-Stat. After data have been read by ViSta, it is relatively simple to create graphs using the software's GUI menus.

The book *Visual Statistics* (Young et al., 2006), which makes heavy use of the ViSta software, provides an excellent introduction to many graphs that are useful in EDA and in detecting data quality problems. Other shareware software for visualizing data is also described by Young et al. (2006). Appendix A also provides a number of examples of graphs useful for data screening that were created with ViSta.

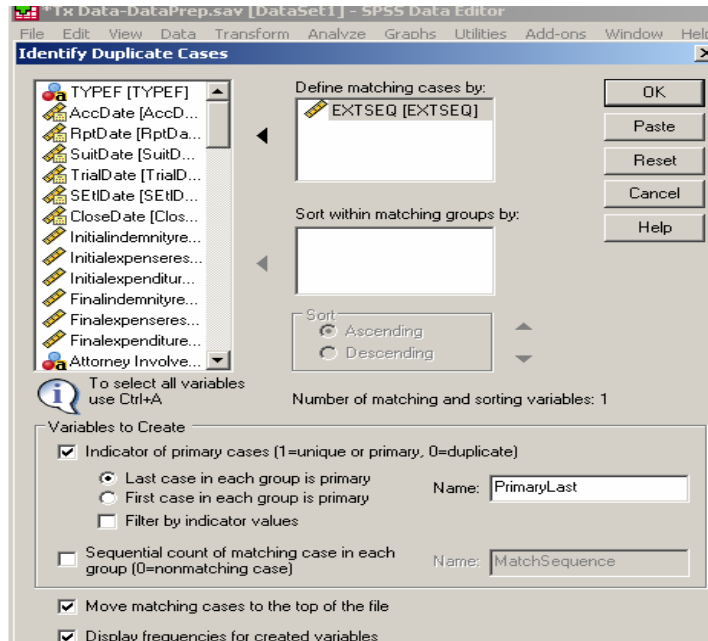
#### **5.2.3.a Screening for duplicates**

The problem of redundant records (two records with identical values for a variable that should only have unique values in the database) is so widespread that at least one major statistical software vendor, SPSS, includes the capability of screening for duplicates in its base statistical package. An example of screening the claim sequence unique identifier variable from a database<sup>19</sup> is presented below:

---

<sup>19</sup> The data used in this example is also used in Appendix D and is described there.

**Figure 5.7: Menu for Duplicate Screen/ Duplicate Report from SPSS**



**Table 5.1: Indicator of Duplicate Case**

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid Duplicate Case	1	.1	.1	.1
Primary Case	1817	99.9	99.9	100.0
Total	1818	100.0	100.0	

### 5.2.3.c Commercial software for automatically screening data

SPSS recently began selling a Data Preparation add-on to its basic statistical software that performs many key data cleaning functions. These include screening data for invalid values, identifying missing values and patterns of missing values, and identifying records with outlying (and possibly erroneous) values.

While we are not aware of a similar package for the popular statistical software, SAS, Cody provides detailed recipes for programming data cleaning capabilities into SAS. Some of his recipes make use of SQL, while others use some of SAS's built-in procedures such as Proc Freq and Proc Univariate. Here is a list of some common ones that can be used:

*Dirty Data on Both Sides of the Pond*

- Proc Compare is used for comparing the contents of two SAS datasets.
- Proc Univariate is used to look for outliers in the output under the “Extreme Observations” section.
- Proc Freq is used for finding duplicate records—essentially forming a “key” or concatenation of one or more fields on a record and then counting the number of observations in the dataset for each unique key.<sup>20</sup>
- Proc Freq is also a descriptive as well as a statistical procedure that produces one-way to n-way frequency and crosstabulation tables. Frequency tables concisely describe your data by reporting the distribution of variable values. Crosstabulation tables, also known as contingency tables, summarize data for two or more classification variables by showing the number of observations for each combination of variable values. See Table 5.2 below for an Example of using Proc Freq to screen data.

The availability of tools such SPSS Data Preparation and Cody’s data cleaning recipes can make the implementation of data screening procedures more efficient. In addition commercial availability of data screening tools likely raises general user awareness of the importance of data screening prior to an analysis.

---

<sup>20</sup> See Cody, pg. 113

**Table 5.2 Example of SAS's Proc Freq for Age with Error**

The FREQUENCY Procedure

---

Age	Frequency	Cumulative Percent	Frequency	Cumulative Percent
0	1	0.06	1	0.06
14	1	0.06	2	0.11
16	2	0.11	4	0.23
17	7	0.4	11	0.62
18	4	0.23	15	0.85
19	12	0.68	27	1.53
20	20	1.13	47	2.66
75	1	0.06	1757	99.6
77	3	0.17	1760	99.77
81	3	0.17	1763	99.94
83	1	0.06	1764	100.00

## 6. RESULTS AND CONCLUSIONS

As discussed in Section 1, the GIRO Data Quality Working Party was formed because of the view that

- data quality issues significantly impacted the work of general insurance actuaries
- such issues could have a material impact on the results of general insurance companies

The Working Party wants to encourage the insurance industry and the actuarial profession to improve practices for collecting and handling data and, in order to do so, much of our work was designed to test the accuracy of the statements in the two bullet points above.

In Section 2, we highlighted a number of anecdotal incidents in which data errors had very serious repercussions.

In Section 3, we discussed the results of a survey of general insurance actuaries that

### *Dirty Data on Both Sides of the Pond*

demonstrated that data quality issues have a significant impact on the work they undertake. The survey indicated that, on average, about a quarter of the effort expended by actuarial teams is spent on data quality issues, and about a third of the projects they undertake are adversely affected by data quality issues. A wide range of responses was noted. One possible explanation for the wide range of responses within each area of practice is that, rather than being a sad fact of business, clients or actuaries or both can take action to improve the quality of the data actuaries use.

In Section 4, we described an experiment we conducted in order to examine the impact of data issues on an insurer's required claims reserves. In order to test the effect of only having access to restricted information, we then created various subsets of the data that varied in their level of completeness. In addition, in order to test the effect of errors in the data, the dataset was modified to reflect the effect of various hypothetical data errors and various projections were repeated using the modified data. From the results of this analysis, we drew the following conclusions:

- There was some positive correlation between the number of historic evaluations in the dataset and the accuracy of the estimates although the strength of this relationship varied with the method used to project losses and the analytical approach (i.e., deterministic versus bootstrap).
- Estimates based on unadjusted paid claims produced worse estimates than those based on incurred claims, presumably because they utilize less data (that is, the case reserve information is not used which particularly impacts immature years).
- When data errors were introduced, the accuracy of the estimates deteriorated significantly.
- When data errors were introduced, the volatility of the estimates increased.

The outcome of the data experiment indicated that there is a significant increase in the uncertainty of results and a significant decrease in the accuracy of results when data quality problems are present. The errors resulting from poor data can significantly reduce the reliability of actuarial analyses, and this could have a direct effect on an insurer's financial statements.

Sections 2, 3, and 4 support the working party's initial hypotheses that were stated at the start of this section, namely that

### *Dirty Data on Both Sides of the Pond*

- data quality issues significantly impacted the work of general insurance actuaries.
- data quality issues could have a material impact on the results of general insurance companies.

It follows that, if insurers improved the quality of their data, it could have a number of highly beneficial effects:

- profitability could increase
- the accuracy and reliability of financial statements could increase
- actuarial resources could be freed up (as well as resources in other areas such as finance and IT) to concentrate on other assignments that could add more value to the organization

The GIRO Working Party believes that insurers should devote more time and resources to increasing the accuracy and completeness of their data by improving their practices for collecting and handling data. In particular, insurers would benefit from the investment of increased senior management time in this area. By taking such action, they could improve their efficiency and hence their profitability.

The Working Party also believes that actuaries are well suited to be data quality advocates. In order to fulfill such a role, actuaries will need to familiarize themselves with the data quality literature, perhaps by reading one of the books recommended by the CAS Data Management Educational Materials Working Party or the IDMA. They will need to participate in data quality initiatives that manage data quality both from within their company and from external suppliers. Finally, even in the best of scenarios where both their internal and external suppliers initiate data quality programs, they will need to screen data for problems. Vigilance is never ending!

### **Acknowledgment**

The authors acknowledge:

- Don Mango for suggesting a GIRO Data Quality working party
- The General Insurance Research Organization of the United Kingdom for forming the working party
- Jane Taylor for assistance with graphs
- The Casualty Actuarial Society for supporting the working party

*Dirty Data on Both Sides of the Pond*

**Supplementary Material**

Excel spreadsheets containing the data used in the data quality experiment as well as the spreadsheet containing the bootstrap procedure will be available on the CAS Web Site

**Appendix B: Data for Experiment**

**Appendix C: Experiment Projections based on Unmodified Data**

**Appendix D: Experiment Projections based on “Erroneous” Data**

## Appendix A: Exploratory Analysis Using The ViSta Visual Statistics System

In this appendix we explain how to download and install the ViSta data visualization software. We also alert potential users to some of ViSta's limitations and unusual (and sometimes annoying) features. We then illustrate some graphs that are useful in data screening that can be obtained with ViSta.

To download ViSta, go to <http://forrest.psych.unc.edu/research/index.html>. You will see an image like one below:

The screenshot shows the ViSta website homepage. At the top left is a small icon with the text "Show All Links". Below it is a navigation menu with the following items: **Hot News**, **About ViSta** (with sub-links: Overview, Look and Feel, New Features, Audience, **!!Free & Open!!**, About The Author), **Download**, **Be A ViSta Developer**, and **User's Info**.

The main content area features the **ViSta** logo and the text "THE VISUAL STATISTICS SYSTEM". Below the logo is the copyright notice: "Copyright © 1990-9 by Forrest W. Young. All rights reserved." The main heading is "ViSta's Dynamic Visualizations help you see what your data seem to say." There are three main sections:

- WorkMaps**: structure your data analysis. It shows a flowchart of analysis steps: CarRating, Car-Prefs, Norm, Prn Cmp, Norm-CarR, PCA-Car-P, rats-scor, Scores-PC, MulReg, MRG-rats, and two BarScore plots.
- Interactive Graphs**: show you your data's structure. It displays a plot of Normalized Data vs PC0, with points for "volvo dl" and "ford pinto" and lines connecting them across PC1, PC2, and PC3.
- GuideMaps**: guide your data analysis. It shows a flowchart with boxes for "Link: Explore", "Link: Transform", and "Link: Analyze".

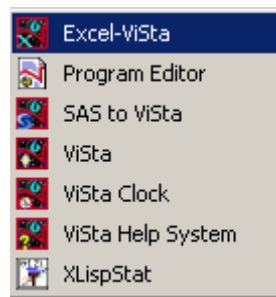
Going down the left side of the screen you will see several options offered including

*Dirty Data on Both Sides of the Pond*

“About ViSta,” “Download,” etc. Choose the download link. On the next screen, choose language (English, French, and Spanish) and an operating system (i.e., Windows, Macintosh, and Unix). On the next screen, click download (for windows users, WinVista6.4). Then download the installation file to your hard drive.

After downloading, run the ViSta installation file by clicking on it. Once it is installed, visit the Users help screen on the ViSta Web Site and download the Users Guide which documents how to use the software. ViSta also comes with a help menu that documents some of the system’s features. As the documentation is somewhat sparse, a few key items are covered below.

The first challenge to overcome is bringing data into ViSta. Because ViSta is programmed in the XLisp language, it reads Lisp files. However, it also has the capability of reading Excel files, text files and SAS files. Since a lot of actuarial analyses are done in Excel, it is relatively easy to read data from Excel files once one becomes familiar with the actual procedure for performing this task. Under the program menus for ViSta, there is an “Excel-ViSta” option similar to the drop-down shown below. To get the drop down, click on your computer’s Start/All Programs menu items; then go to the ViSta6 option. When you place the mouse over “ViSta6”, you see the drop down.

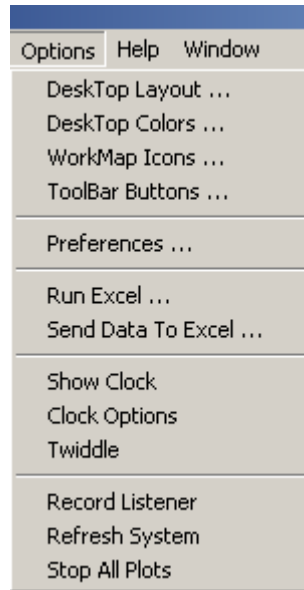


The first time the Excel-ViSta option is chosen, the user is asked to supply the location of the “Excel.exe” file. This typically resides in the Program/Office directory, but its location should be identified by using the search option of Windows explorer before attempting to use Excel and ViSta together.

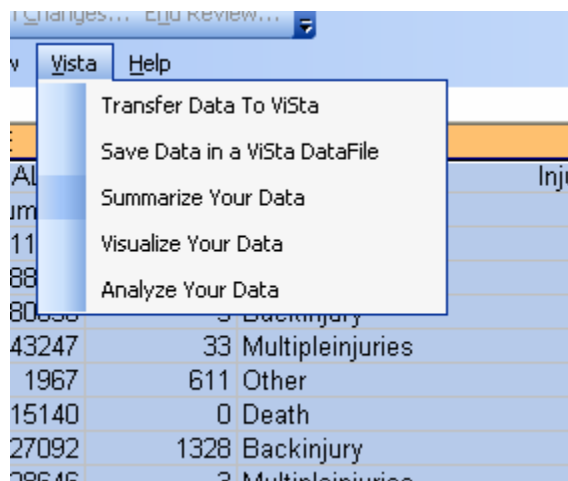
Once the Excel-ViSta macros have been installed, the following procedure can be used to read data from Excel to ViSta:

*Dirty Data on Both Sides of the Pond*

- Launch ViSta.
- In the ViSta Options tab at the top of the ViSta screen, select “Run Excel.”



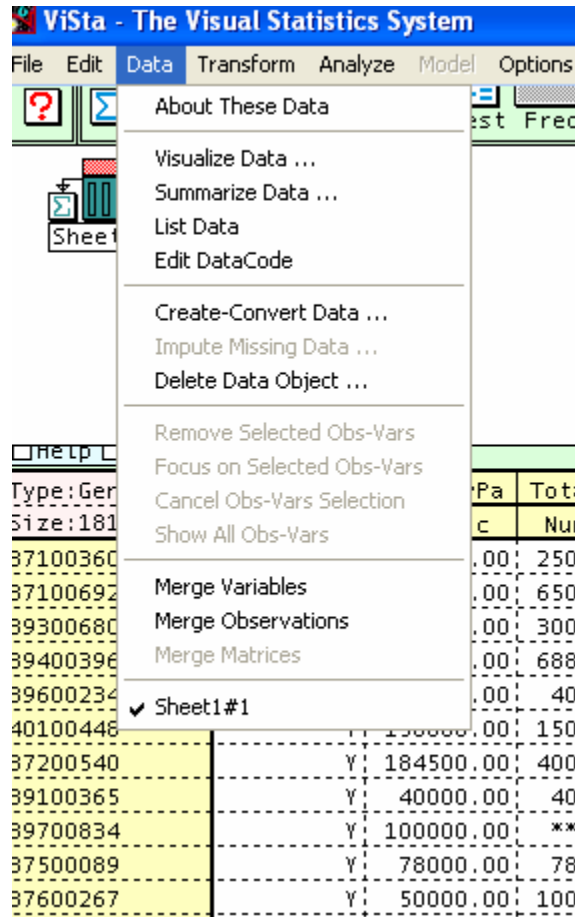
- When Excel is launched, make sure to enable macros.
- From within Excel open the database you want to analyze.
- Highlight all the data you want read into ViSta, while in Excel.
- In Excel, click on the ViSta tab on the tool bars at the top of the worksheet, then click “Transfer Data to ViSta.”



*Dirty Data on Both Sides of the Pond*

- Wait a little while, until the ViSta screen appears.
- Proceed with your analysis.

Graphs are typically created via the Data Menu in ViSta and then selecting Visualize Data.



Data that is read into ViSta, whether in a text file or an Excel file, must be in a very specific format. Any deviation from the format causes an error in attempting to read the data. The first line must contain the word “Cells” in the left column and the variable names in subsequent columns. The second line contains the word “Labels” in the left column and the variable type, either Category or Numeric, below the variable names. The left column is a record label or identifier. The actual data begins in the third row.

*Dirty Data on Both Sides of the Pond*

Cells	Attorney Involvement insurer	Attorney involvement insured	Primary Paid
Labels	Category	Category	Numeric
37100360	Y	N	250,000
37100692	Y	N	250,000
39300680	Y	N	300,000
39400396	Y	N	33,000
39600234	Y	N	40,000
40100448	Y	N	150,000

A few other limitations of ViSta are:

- Only up to 4 categorical variables are allowed in any one database, though the number of numeric variables does not seem to be limited
- The categorical variables can have no more than 12 categories
- After finishing the analysis on one Excel database, it is easiest to close ViSta and launch it again if you wish to use another Excel database. However, multiple Lisp databases can be used without closing ViSta. Once Excel data has been transferred to ViSta, it can be saved as a Lisp file.
- We believe that ViSta will not perform well on very large databases. We have used it on databases with up to 6,000 records.
- To print a ViSta graph, it is necessary to first copy it (by clicking on it and typing control-C) to other software such as Microsoft Word.

In addition, it can be helpful to join the ViSta users group (from the ViSta Web Site), as answers to user's questions can be answered by another user.

In summary, once initial challenges of using ViSta, especially those associated with transferring data to it, are overcome, ViSta provides some very useful visualization tools. We have provided only a cursory introduction to its graphical capabilities, which include dynamic graphs. A more thorough introduction to its capabilities is provided by Young et al. (2006). ViSta also provides some statistical functionality, including ANOVA, regression and principal components analysis. It has a number of limitations and does not appear to be suited for use on large databases.

### **Graphical examples**

Below we present a sample of graphs that are useful in data quality screening. The graphs are based on publicly available closed claim data on work-related injuries from the Texas Department of Insurance Web Site. Although some of the claims were closed without payment, most exceed a trigger of \$10,000 that is used for collecting detailed information on a claim. The fields available in the data include accident date, report date, settlement date, primary paid losses, total paid losses (all parties), claimant age, and injury type.

To illustrate how these tools can be used to uncover potential data quality problems, errors were intentionally introduced into the data for some of the graphs. A bold arrow points to the outliers or intentional errors. We show illustrations for:

- Box-and-whisker Plots
  - Simple dot plots (Figure A.1)
  - Box Plots (Figure A.2)
- Histogram-type Plots
  - Frequency Polygons (Figure A.3)
  - Histogram with smooth curve (Figure A.4)
    - Normal curve
    - Kernel smoothing<sup>21</sup>
  - Bar Plots (Figure 5.6)

---

<sup>21</sup> Kernel smoothing uses a non-parametric technique to fit a smooth curve to histogram data. See Young et al. (2007) for a discussion of smoothing.

**Figure A.1: Dot Plot for Claimant Age with Error**

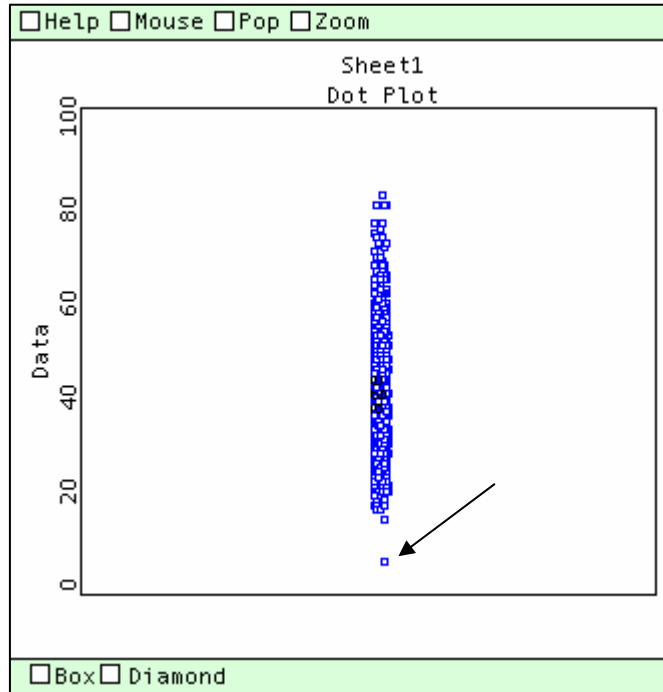
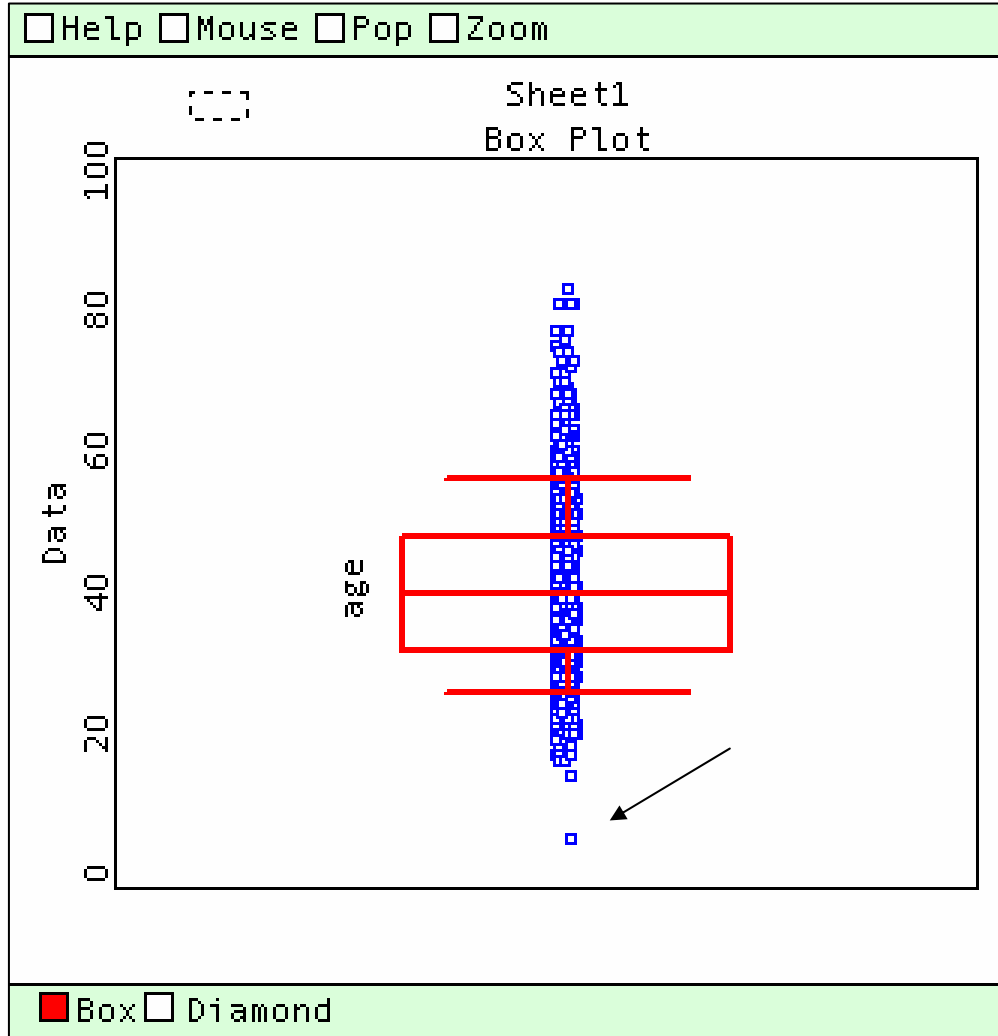
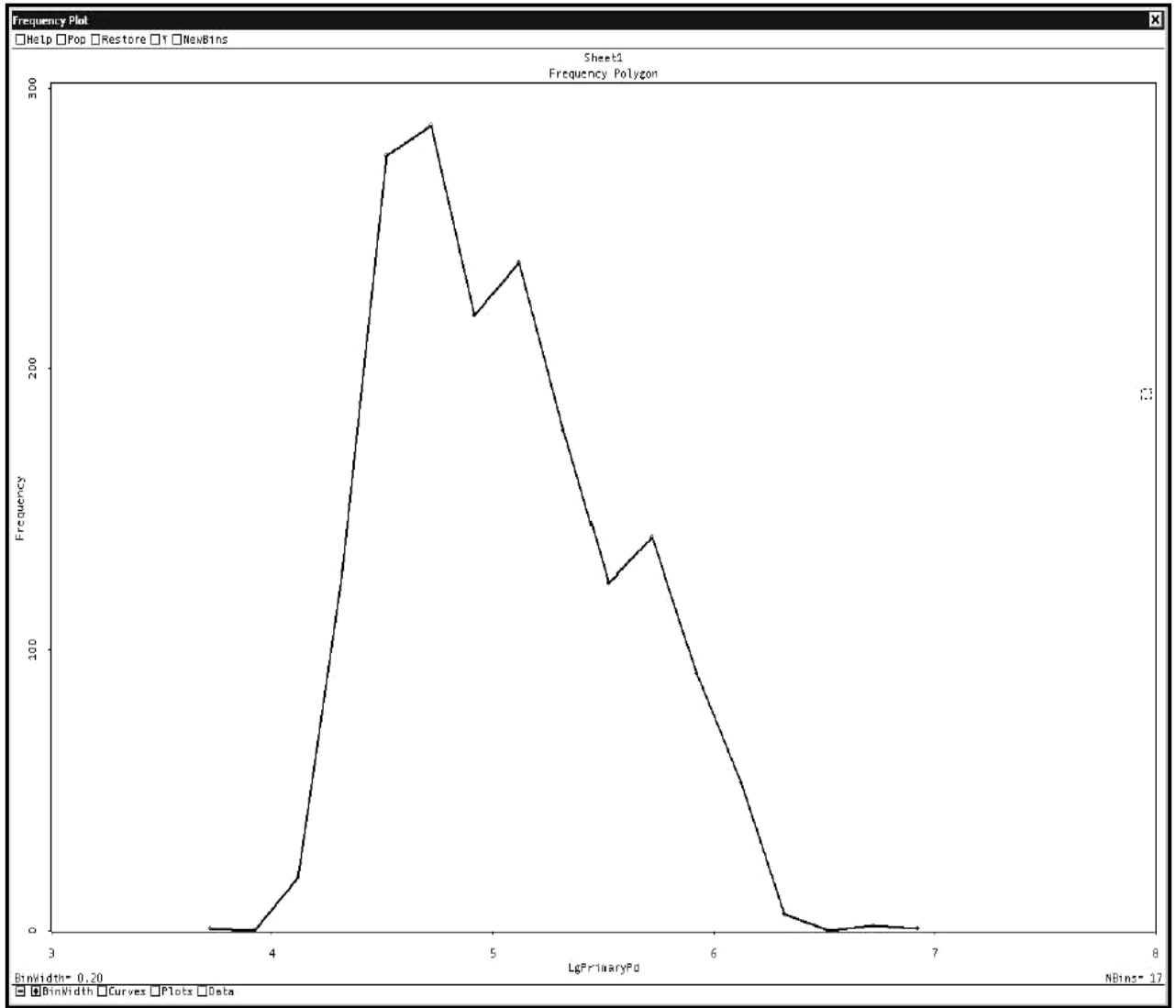


Figure A.2: Box-and-whisker Plot for Claimant Age with Error



**Figure A.3: Frequency Polygon of Log of Primary Paid Losses – No Errors in Data**



**Figure A.4: Histogram of Log of Primary Paid Losses – Errors in Data**

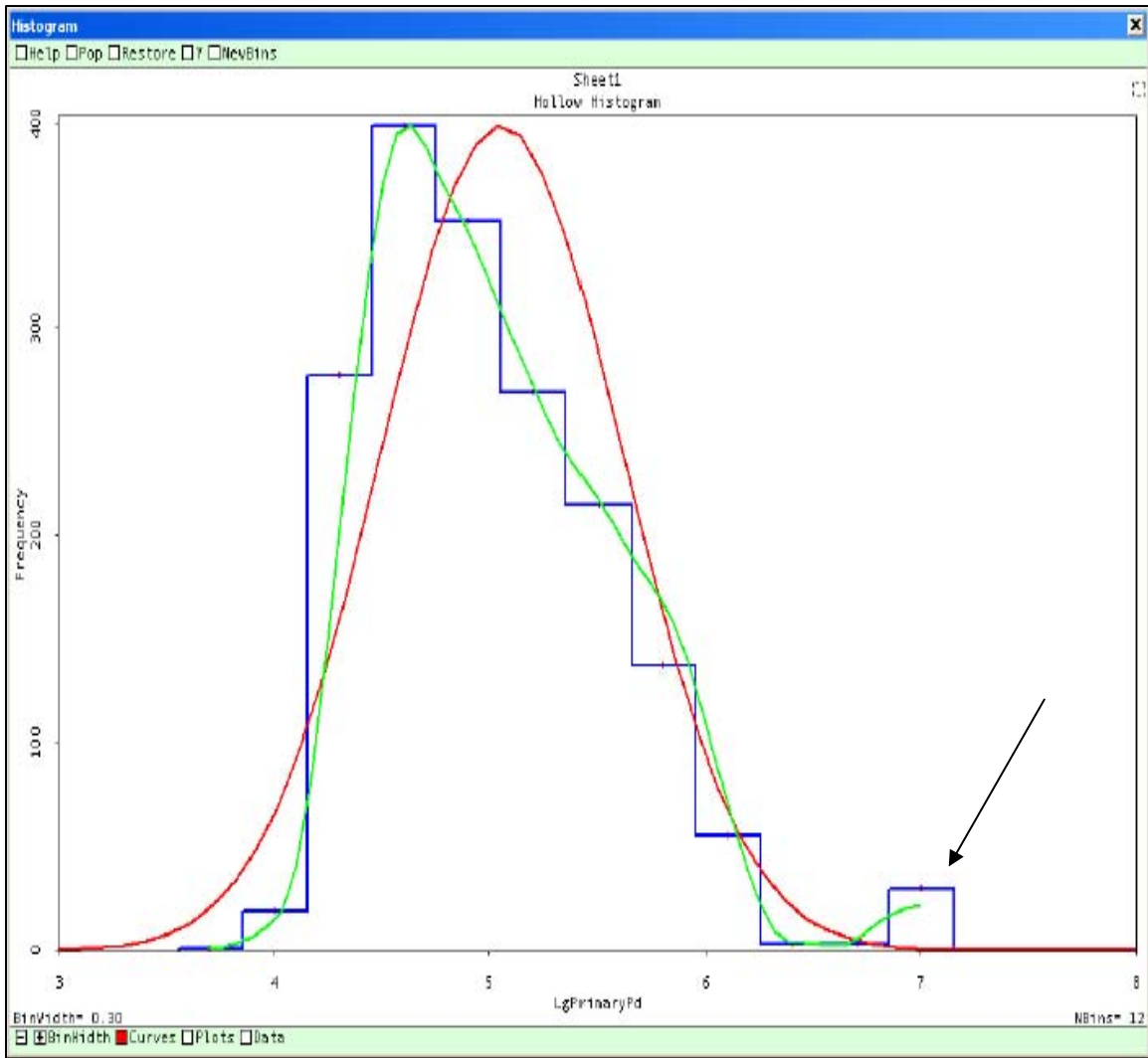
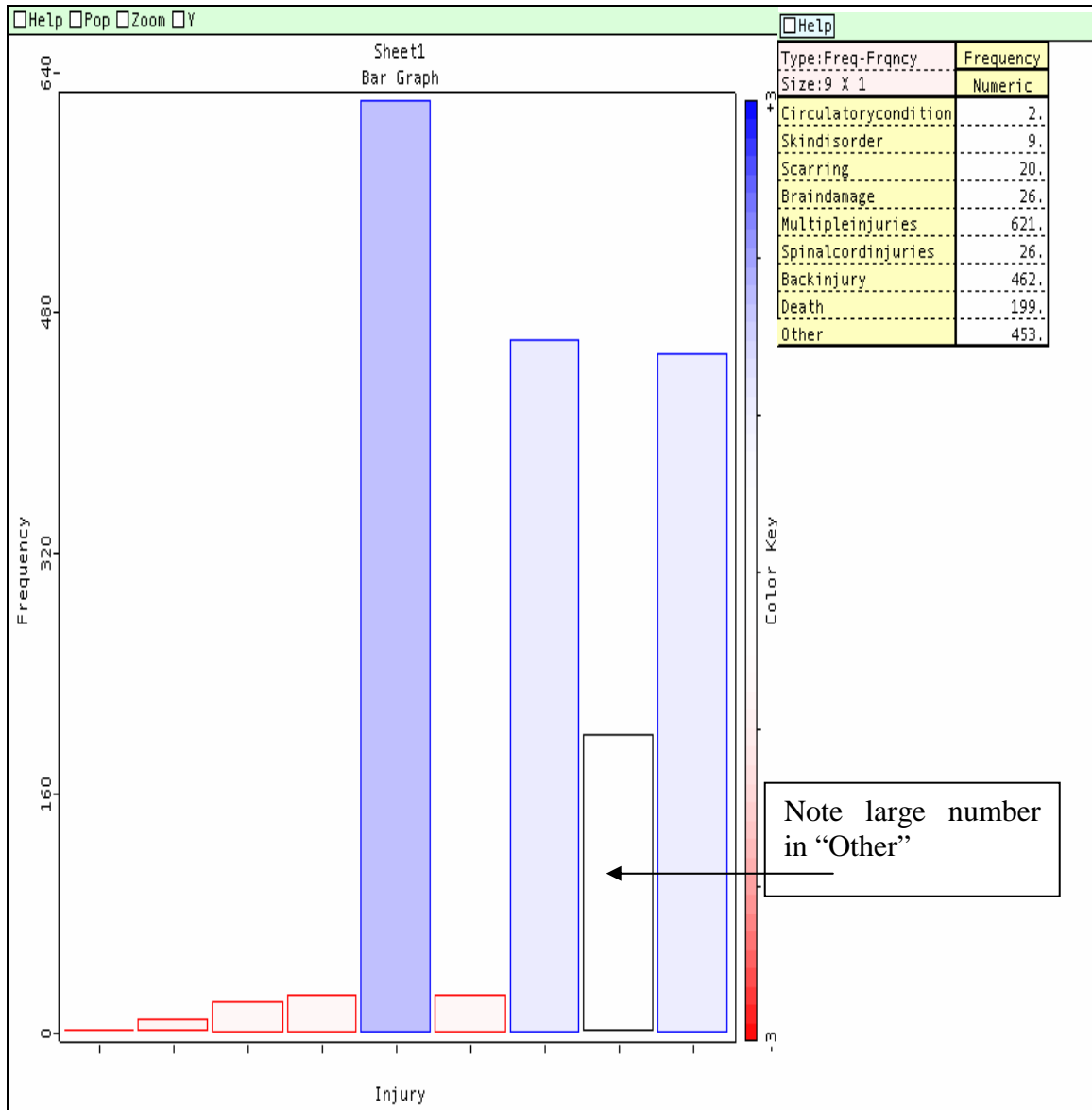


Figure A.5. Bar Plot for Categorical Data



## 7. REFERENCES

- [1] Actuarial Standards Board, Actuarial Standard of Practice 23, Data Quality, Revised Edition, December 2004, [www.actuaries.org](http://www.actuaries.org).
- [2] Anderson, Bolton, Callen, Cross, Howard, Mitchell, Murphy, Rakow, Sterling, Welsch, "General Insurance Premium Rating-The Way Forward," presented to Institute of Actuaries, May 2007.
- [3] Archibald, R., "Girl in Heart Transplant Dies After Two Weeks," *New York Times*, Feb 23, 2003.
- [4] Arenson, K., "Technical Problems Cause Problems in SAT Scores," *New York Times*, March 8, 2006.
- [5] Berquist J and Sherman, R, "Loss Reserve Adequacy Testing," *Proceedings of the Casualty Actuarial Society*, 1977, 123-184.
- [6] California Auditor's Office, "Auditors Examine Conservation and Liquidation Office," <http://www.wcexec.com/Resources.aspx>.
- [7] Campbell, R., Francis, L., Prevosto, V., Rothwell, M., and Sheaf, S., "Report of the Data Quality Working Party," completed 2006, [http://www.actuaries.org.uk/Display\\_Page.cgi?url=/research/wp\\_reports.html](http://www.actuaries.org.uk/Display_Page.cgi?url=/research/wp_reports.html).
- [8] CAS Committee on Management Data and Information, "White Paper on Data Quality," *Casualty Actuarial Society Winter Forum*, 1997.
- [9] CAS Data Management Educational Materials Working Party, "Survey of Data Management and Data Quality Texts," *Casualty Actuarial Society Winter Forum*, 2007:273-306.
- [10] CAS Data Management Educational Materials Working Party, "Actuarial IQ (Information Quality)" to be published in the CAS Winter 2008 *E-Forum*.
- [11] Chambers, J., Cleveland, W., Kleiner, B., and Tukey, P., *Graphical Methods for Data Analysis*, Wadsworth International Group, 1983.
- [12] Cleveland, W., *Visualizing Data*, Hobart Press, 1993.
- [13] Cody, R. Cody's Data Cleaning Techniques Using the SAS Software, SAS Institute, 1999.
- [14] COPLFR, Property & Casualty Practice Note, December 2004.
- [15] CNN, "Amid Protest, U.S. says Faulty Data Led to Chinese Embassy Bombing," [cnn.com](http://cnn.com), May 9, 1999a.
- [16] CNN, "NASA's Metric Confusion Causes Mars Orbiter Loss," September 30, 1999b
- [17] Consumer Federation of America, "Millions of Americans Jeopardized by Inaccurate Credit Scores," 2002.
- [18] Cornejo, R., "Searching for a Cause", *Bests Review*, February, 2006.
- [19] Copeman, P., Gibson, L., Jones, T., Line, N., Lowe, J., Martin, P., Mathews, P., Powell, D., "A Change Agenda for Reserving: A Report of the General Insurance Reserving Issues Task Force," 2006, [www.actuaries.org.uk](http://www.actuaries.org.uk).
- [20] Data Quality Solutions, [www.dataqualitysolutions.com](http://www.dataqualitysolutions.com), September 6, 2007.
- [21] Dasu, T. and Johnson, T., *Exploratory Data Mining and Data Cleaning*, Wiley 2003.
- [22] Eckerson, W., "Data Warehousing Special Report: Data Quality and the Bottom Line," <http://www.adtmag.com/article.aspx?id=6321&page=>.
- [23] England, P. and Verrall, R. "Analytic and bootstrap estimates of prediction errors in claims reserving," *Insurance Mathematics and Economics* 25 (1999) pp. 281-293.
- [24] England, P. and Verrall, R., "Stochastic Claims Reserving in General Insurance," Presented to the Institute of Actuaries, January, 2002, pp. 443-518.
- [25] Francis, L., "Dancing With Dirty Data", *Casualty Actuarial Society Winter Forum*, 2005.
- [26] Hartwig, F. and B. Dearing, *Exploratory Data Analysis*, Sage Publications, 1979.
- [27] Infoimpat, "Statscan admits 5-year Inflation Mistake," [www.infoimpact.com/newspdf/Statscan%20admits%20five-year%20inflation%20mistake.pdf](http://www.infoimpact.com/newspdf/Statscan%20admits%20five-year%20inflation%20mistake.pdf) , August 16, 2006.
- [28] IQ Trainwrecks, [www.iqtrainwrecks.com](http://www.iqtrainwrecks.com).
- [29] Insurance Data Management Association, "Value Proposition," [www.idma.org](http://www.idma.org).
- [30] Moore, L., "Data Business—No data quality control? Expect to count the cost," 2006, *Business Media Europe*.

## *Dirty Data on Both Sides of the Pond*

- [31] Olson, J., *Data Quality: The Accuracy Dimension*, Morgan Kauffman Publishers, 2003.
- [32] *New York Times*, January 4, 1995, “Magellan Error is Explained.”
- [33] Popelyukhin, A., “Watch Your TPA”, *Casualty Actuarial Society Winter Forum*, 1999, pp. 239 - 254.
- [34] PriceWaterhouseCoopers LLP, “Global Data Management Survey 2004,” 2004
- [35] Redman, T., *Data Quality: The Field Guide*, Digital Press, 2001.
- [36] RMS, “Hurricane Katrina: Profiles of a Supercat: Lessons and Implications for Catastrophe Risk Management, October 2005.
- [37] Ruethling, G., *New York Times*, February 15, 2006.
- [38] Schneier, Bruce, [http://www.schneier.com/blog/archives/2006/02/database\\_error.html](http://www.schneier.com/blog/archives/2006/02/database_error.html), “Database Error Causes Unbalanced Budget.”
- [39] Schnacker, B. and Chute, E., “Colleges Adjust to SAT Errors,” *Pittsburgh Post Gazette*.
- [40] Sherman, R., “Extrapolating, Smoothing and Interpolating Development Factors,” *Proceedings of the Casualty Actuarial Society*, 1984, pp. 122 – 155.
- [41] Sonnett, Sharon, “Resolving Insolvencies,” *Best’s Review*, February, 2005.
- [42] Venter, G., “Testing the Assumptions of Age to Age Factors,” *Proceedings of the Casualty Actuarial Society*, 1998.
- [43] Wand Y. and Wang R., “Anchoring Data Quality Dimensions in Ontological Foundations,” *Communications of the ACM*, November 1996.
- [44] Westfall, C., “How Did Catastrophe Models Weather Katrina?,” *Insurance Journal*, October 3, 2005.
- [45] Williams D., “NATO Missiles Hit Chinese Embassy,” *Washington Post*, May 8, 1999.
- [46] Wright, T., “Stochastic Reserving When the Past Claims Numbers are Known,” *Proceedings of the Casualty Actuarial Society*, 1992, pp. 255 – 361.
- [47] Young, F., Valerio-Mora, P., Friendly, M., *Visual Statistics: Seeing Data with Dynamic Interactive Graphics*, 2006, Wiley.

### Abbreviations and notations

CL, chain ladder	GLM, generalized linear models
DFA, dynamic financial analysis	GUI, Graphical user interface
General Insurance, non-life insurance	GLM, generalized linear models
GIRO: General Insurance Research Organization	OLS, ordinary least squares
	ERM, enterprise risk management

### Biographies of the Authors

**Robert Campbell, FCAS, FCIA**, is Director, Commercial Lines Actuarial at Lombard Canada in Toronto, Canada. He has a Bachelor of Mathematics in Business Administration from the University of Waterloo. He is a Fellow of the CAS and a Fellow of the Canadian Institute of Actuaries. He is chair of the Data Management Educational Materials working party, participates on the CAS Committee on Data Management and Information, and was a participant on the 2006 GIRO Data Quality working party.

**Louise Francis (chair), FCAS, MAAA**, is a Consulting Principal at Francis Analytics and Actuarial Data Mining, Inc. She is involved in data mining projects as well as conventional actuarial analyses. She has a BA degree from William Smith College and an MS in Health Sciences from SUNY at Stony Brook. She is a Fellow of the CAS and a Member of the American Academy of Actuaries. She serves on several CAS committees/working parties and is a frequent presenter at actuarial and industry symposia. She is a four-time winner of the Data Quality, Management and Technology call paper prize including one for “Dancing with Dirty Data: Methods for Exploring and Cleaning Data (2005).”

**Virginia R. Prevosto, FCAS, MAAA**, is a Vice President at Insurance Services Office, Inc. Ms. Prevosto is a Phi Beta Kappa graduate of the State University at Albany with a Bachelor of Science degree in Mathematics, summa cum laude. She is a Fellow of the CAS and a Member of the American Academy of

*Dirty Data on Both Sides of the Pond*

Actuaries. She serves as General Officer of the CAS Examination Committee and as liaison to various other CAS admission committees. She also serves on the CAS Committee on Management Data and Information. In the past Ms. Prevosto also served on the Data Quality Task Force of the Specialty Committee of the Actuarial Standards Board that wrote the first data quality standard of practice. Virginia has been a speaker at the Casualty Loss Reserve Seminar on the data quality standard and to various insurance departments on data management and data quality issues. Ms. Prevosto authored the papers “Statistical Plans for Property/Casualty Insurer” and “Study Note: ISO Statistical Plans” and co-authored “For Want of a Nail the Kingdom was Lost—Mother Goose was Right: Profit by Best (Data Quality) Practices” for the IAIDQ.

**Mark Rothwell, FIA**, is the U.K. Actuary from Brit Insurance. He is responsible for the reserving and pricing of a wide range of personal and commercial lines products sold through the U.K. Company Market. Mr. Rothwell received an MA in Mathematics from Cambridge University, is a Fellow of the Institute of Actuaries and a Chartered Mathematician. He has presented at several GIRO workshops and been a member of various GIRO working parties, including most recently, the ROC working party on the effects of the reserving cycle. Mr. Rothwell is also a member of the London Market Actuaries Group.

**Simon Sheaf, FIA**, is Actuarial Director and Head of General Insurance at Grant Thornton U.K. LLP. He advises clients on such areas as reserving, rating, financial and capital modeling, and management information systems. Mr. Sheaf received an MA in Mathematics from Oxford University and is a Fellow of the Institute of Actuaries. He is the deputy chairman of the U.K. Actuarial Profession’s General Insurance Education and Continuing Professional Development Committee, and a member of the profession’s Education Committee. He is also a member of the London Market Actuaries Group. Mr. Sheaf has co-authored papers on topics as diverse as reinsurance pricing, data quality, Lloyd’s reinsurance-to-close, claims inflation, and reinsurance bad debts.

*Dirty Data on Both Sides of the Pond*

**Appendix B**  
Cumulative Paid Losses

Accident Year	Months of Development																	
	12	24	36	48	60	72	84	96	108	120	132	144	156	168	180	192	204	216
1974	\$267	\$1,975	\$4,587	\$7,375	\$10,661	\$15,232	\$17,888	\$18,541	\$18,937	\$19,130	\$19,189	\$19,209	\$19,234	\$19,234	\$19,246	\$19,246	\$19,246	\$19,246
1975	310	2,809	5,686	9,386	14,884	20,654	22,017	22,529	22,772	22,821	23,042	23,060	23,127	23,127	23,127	23,127	23,159	
1976	370	2,744	7,281	13,287	19,773	23,888	25,174	25,819	26,049	26,180	26,268	26,364	26,371	26,379	26,397	26,397		
1977	577	3,877	9,612	16,962	23,764	26,712	28,393	29,656	29,839	29,944	29,997	29,999	29,999	30,049	30,049			
1978	509	4,518	12,067	21,218	27,194	29,617	30,854	31,240	31,598	31,889	32,002	31,947	31,965	31,986				
1979	630	5,763	16,372	24,105	29,091	32,531	33,878	34,185	34,290	34,420	34,479	34,498	34,524					
1980	1,078	8,066	17,518	26,091	31,807	33,883	34,820	35,482	35,607	35,937	35,957	35,962						
1981	1,646	9,378	18,034	26,652	31,253	33,376	34,287	34,985	35,122	35,161	35,172							
1982	1,754	11,256	20,624	27,857	31,360	33,331	34,061	34,227	34,317	34,378								
1983	1,997	10,628	21,015	29,014	33,788	36,329	37,446	37,571	37,681									
1984	2,164	11,538	21,549	29,167	34,440	36,528	36,950	37,099										
1985	1,922	10,939	21,357	28,488	32,982	35,330	36,059											
1986	1,962	13,053	27,869	38,560	44,461	45,988												
1987	2,329	18,086	38,099	51,953	58,029													
1988	3,343	24,806	52,054	66,203														
1989	3,847	34,171	59,232															
1990	6,090	33,392																
1991	5,451																	

Claims Closed with Payment

Accident Year	Months of Development																	
	12	24	36	48	60	72	84	96	108	120	132	144	156	168	180	192	204	216
1974	268	607	858	1,090	1,333	1,743	2,000	2,076	2,113	2,129	2,137	2,141	2,143	2,143	2,145	2,145	2,145	2,145
1975	294	691	913	1,195	1,620	2,076	2,234	2,293	2,320	2,331	2,339	2,341	2,343	2,343	2,343	2,343	2,344	
1976	283	642	961	1,407	1,994	2,375	2,504	2,549	2,580	2,590	2,596	2,600	2,602	2,603	2,603	2,603		
1977	274	707	1,176	1,688	2,295	2,545	2,689	2,777	2,809	2,817	2,824	2,825	2,825	2,826	2,826			
1978	269	658	1,228	1,819	2,217	2,475	2,613	2,671	2,691	2,706	2,710	2,711	2,714	2,717				
1979	249	771	1,581	2,101	2,528	2,816	2,930	2,961	2,973	2,979	2,986	2,988	2,992					
1980	305	1,107	1,713	2,316	2,748	2,942	3,025	3,049	3,063	3,077	3,079	3,080						
1981	343	1,042	1,608	2,260	2,596	2,734	2,801	2,835	2,854	2,859	2,860							
1982	350	1,242	1,922	2,407	2,661	2,834	2,887	2,892	2,911	2,915								
1983	428	1,257	1,841	2,345	2,683	2,853	2,908	2,920	2,925									
1984	291	1,004	1,577	2,054	2,406	2,583	2,622	2,636										
1985	303	1,001	1,575	2,080	2,444	2,586	2,617											
1986	318	1,055	1,906	2,524	2,874	2,958												
1987	343	1,438	2,384	3,172	3,559													
1988	391	1,671	3,082	3,771														
1989	433	1,941	3,241															
1990	533	1,923																
1991	339																	

*Dirty Data on Both Sides of the Pond*

Cumulative Reported Claims

Accident Year	Months of Development																	
	12	24	36	48	60	72	84	96	108	120	132	144	156	168	180	192	204	216
1974	1,912	2,854	3,350	3,945	4,057	4,104	4,149	4,155	4,164	4,167	4,169	4,169	4,169	4,170	4,170	4,170	4,170	4,170
1975	2,219	3,302	3,915	4,462	4,618	4,673	4,696	4,704	4,708	4,711	4,712	4,716	4,716	4,716	4,716	4,716	4,717	
1976	2,347	3,702	4,278	4,768	4,915	4,983	5,003	5,007	5,012	5,012	5,013	5,014	5,015	5,015	5,015	5,015		
1977	2,983	4,346	5,055	5,696	5,818	5,861	5,884	5,892	5,896	5,897	5,900	5,900	5,900	5,900	5,900			
1978	2,538	3,906	4,633	5,123	5,242	5,275	5,286	5,292	5,298	5,302	5,304	5,304	5,306	5,306				
1979	3,548	5,190	5,779	6,206	6,313	6,329	6,339	6,343	6,347	6,347	6,347	6,348	6,348					
1980	4,583	6,106	6,656	7,032	7,128	7,139	7,147	7,150	7,151	7,153	7,154	7,154						
1981	4,430	5,967	6,510	6,775	6,854	6,873	6,883	6,889	6,892	6,894	6,895							
1982	4,408	5,849	6,264	6,526	6,571	6,589	6,594	6,596	6,600	6,602								
1983	4,861	6,437	6,869	7,134	7,196	7,205	7,211	7,212	7,214									
1984	4,229	5,645	6,053	6,419	6,506	6,523	6,529	6,531										
1985	3,727	4,830	5,321	5,717	5,777	5,798	5,802											
1986	3,561	5,045	5,656	6,040	6,096	6,111												
1987	4,259	6,049	6,767	7,206	7,282													
1988	4,424	6,700	7,548	8,105														
1989	5,005	7,407	8,287															
1990	4,889	7,314																
1991	4,044																	

Outstanding Claims

Accident Year	Months of Development																	
	12	24	36	48	60	72	84	96	108	120	132	144	156	168	180	192	204	216
1974	1,381	1,336	1,462	1,660	1,406	772	406	191	98	57	23	13	3	4	0	0	0	0
1975	1,289	1,727	1,730	1,913	1,310	649	358	167	73	30	9	6	4	2	2	1	1	
1976	1,605	1,977	1,947	1,709	1,006	540	268	166	79	48	32	18	14	10	10	7		
1977	2,101	2,159	2,050	1,735	988	582	332	139	66	38	27	21	21	8	3			
1978	1,955	1,943	1,817	1,384	830	460	193	93	56	31	15	9	7	2				
1979	2,259	2,025	1,548	1,273	752	340	150	68	36	24	18	13	4					
1980	2,815	1,991	1,558	1,107	540	228	88	55	28	14	8	6						
1981	2,408	1,973	1,605	954	480	228	115	52	27	15	11							
1982	2,388	1,835	1,280	819	354	163	67	44	21	10								
1983	2,641	1,765	1,082	663	335	134	62	34	18									
1984	2,417	1,654	896	677	284	90	42	15										
1985	1,924	1,202	941	610	268	98	55											
1986	1,810	1,591	956	648	202	94												
1987	2,273	1,792	1,059	626	242													
1988	2,403	1,966	1,166	693														
1989	2,471	2,009	1,142															
1990	2,642	2,007																
1991	2,366																	

*Dirty Data on Both Sides of the Pond*

Outstanding Losses

Accident Year	Months of Development																	
	12	24	36	48	60	72	84	96	108	120	132	144	156	168	180	192	204	216
1974	\$5,275	\$8,867	\$12,476	\$11,919	\$8,966	\$5,367	\$3,281	\$1,524	\$667	\$348	\$123	\$82	\$18	\$40	\$0	\$0	\$0	\$0
1975	6,617	11,306	13,773	14,386	10,593	4,234	2,110	1,051	436	353	93	101	10	5	5	3	3	
1976	7,658	11,064	13,655	13,352	7,592	4,064	1,895	1,003	683	384	216	102	93	57	50	33		
1977	8,735	14,318	14,897	12,978	7,741	4,355	2,132	910	498	323	176	99	101	32	14			
1978	8,722	15,070	15,257	11,189	5,959	3,473	1,531	942	547	286	177	61	67	7				
1979	9,349	16,470	14,320	10,574	6,561	2,864	1,328	784	424	212	146	113	38					
1980	11,145	16,351	14,636	11,273	5,159	2,588	1,290	573	405	134	81	54						
1981	10,933	15,012	14,728	9,067	5,107	2,456	1,400	584	269	120	93							
1982	13,323	16,218	12,676	6,290	3,355	1,407	613	398	192	111								
1983	13,899	16,958	12,414	7,700	4,112	1,637	576	426	331									
1984	14,272	15,806	10,156	8,005	3,604	791	379	159										
1985	13,901	15,384	12,539	7,911	3,809	1,404	827											
1986	15,952	22,799	16,016	8,964	2,929	1,321												
1987	22,772	24,146	18,397	8,376	3,373													
1988	25,216	26,947	17,950	8,610														
1989	24,981	30,574	19,621															
1990	30,389	34,128																
1991	28,194																	

*Dirty Data on Both Sides of the Pond*

Accident Year	Earned Exposures	TRUE Ultimates
1974	11,000	19,256
1975	11,000	23,161
1976	11,000	26,400
1977	12,000	30,049
1978	12,000	31,991
1979	12,000	34,529
1980	12,000	35,984
1981	12,000	35,207
1982	11,000	34,418
1983	11,000	38,354
1984	11,000	37,175
1985	11,000	36,446
1986	12,000	46,777
1987	13,000	60,676
1988	14,000	75,418
1989	14,000	88,115
1990	14,000	90,938
1991	13,000	74,807

Closing Rates

Accident Year	Months of Development																	
	12	24	36	48	60	72	84	96	108	120	132	144	156	168	180	192	204	216
1974	0.278	0.532	0.564	0.579	0.653	0.812	0.902	0.954	0.976	0.986	0.994	0.997	0.999	0.999	1.000	1.000	1.000	1.000
1975	0.419	0.477	0.558	0.571	0.716	0.861	0.924	0.964	0.984	0.994	0.998	0.999	0.999	1.000	1.000	1.000	1.000	
1976	0.316	0.466	0.545	0.642	0.795	0.892	0.946	0.967	0.984	0.990	0.994	0.996	0.997	0.998	0.998	0.999		
1977	0.296	0.503	0.594	0.695	0.830	0.901	0.944	0.976	0.989	0.994	0.995	0.996	0.996	0.999	0.999			
1978	0.230	0.503	0.608	0.730	0.842	0.913	0.963	0.982	0.989	0.994	0.997	0.998	0.999	1.000				
1979	0.363	0.610	0.732	0.795	0.881	0.946	0.976	0.989	0.994	0.996	0.997	0.998	0.999					
1980	0.386	0.674	0.766	0.843	0.924	0.968	0.988	0.992	0.996	0.998	0.999	0.999						
1981	0.456	0.669	0.753	0.859	0.930	0.967	0.983	0.992	0.996	0.998	0.998							
1982	0.458	0.686	0.796	0.875	0.946	0.975	0.990	0.993	0.997	0.998								
1983	0.457	0.726	0.842	0.907	0.953	0.981	0.991	0.995	0.998									
1984	0.428	0.707	0.852	0.895	0.956	0.986	0.994	0.998										
1985	0.484	0.751	0.823	0.893	0.954	0.983	0.991											
1986	0.492	0.685	0.831	0.893	0.967	0.985												
1987	0.466	0.704	0.844	0.913	0.967													
1988	0.457	0.707	0.846	0.914														
1989	0.506	0.729	0.862															
1990	0.460	0.726																
1991	0.415																	

*Dirty Data on Both Sides of the Pond*

**Appendix C**

**Ultimate Losses - Incomplete Data**

**Ultimate Paid Losses**

Accident Year	Paid Ultimate All Years	Paid Ultimate 3 Years	Paid Ultimate 86 - 91	BF Paid Ultimate All Years	BF Paid Ultimate 3 Years	BF Paid Ultimate 86 - 91
1974	19,246	19,246	19,246	19,246	19,246	19,246
1975	23,159	23,159	23,159	23,159	23,159	23,159
1976	26,417	26,417	26,405	26,417	26,417	26,405
1977	30,072	30,072	30,075	30,072	30,072	30,075
1978	32,020	32,020	32,043	32,020	32,020	32,043
1979	34,581	34,601	34,632	34,581	34,601	34,632
1980	36,053	36,066	36,144	36,053	36,066	36,144
1981	35,279	35,285	35,448	35,279	35,285	35,448
1982	34,574	34,504	34,782	34,574	34,504	34,782
1983	38,084	37,874	38,179	38,084	37,874	38,179
1984	37,739	37,392	38,036	37,739	37,392	38,036
1985	37,289	36,478	37,647	37,289	36,478	37,647
1986	49,475	47,268	49,448	49,475	47,268	49,448
1987	68,911	62,628	64,537	68,911	62,628	64,537
1988	95,093	80,904	83,371	95,093	80,904	83,371
1989	120,591	94,869	99,048	120,591	94,869	99,048
1990	138,214	100,918	109,831	103,782	89,851	93,851
1991	151,661	112,010	126,025	94,353	85,207	88,029

**Ultimate Adjusted Paid Losses**

Accident Year	Adj Paid Ultimate All Years	Adj Paid Ultimate 3 Years	Adj Paid Ultimate 86 - 91	BF Paid Ultimate All Years	BF Paid Ultimate 3 Years	BF Paid Ultimate 86 - 91
1974	19,246	19,246	19,246	19,246	19,246	19,246
1975	23,159	23,159	23,159	23,159	23,159	23,159
1976	26,417	26,417	26,393	26,417	26,417	26,393
1977	30,072	30,072	30,046	30,072	30,072	30,046
1978	32,012	32,010	31,991	32,012	32,010	31,991
1979	34,554	34,553	34,550	34,554	34,553	34,550
1980	35,996	35,997	36,026	35,996	35,997	36,026
1981	35,231	35,221	35,291	35,231	35,221	35,291
1982	34,433	34,425	34,579	34,433	34,425	34,579
1983	37,775	37,762	38,041	37,775	37,762	38,041
1984	37,185	37,175	37,774	37,185	37,175	37,774
1985	36,470	36,453	37,219	36,470	36,453	37,219
1986	46,967	47,097	48,564	46,967	47,097	48,564
1987	60,881	61,689	63,395	60,881	61,689	63,395
1988	76,147	78,056	80,216	76,147	78,056	80,216
1989	78,998	84,925	87,181	78,998	84,925	87,181
1990	77,709	88,184	92,645	79,397	84,345	87,230
1991	103,048	103,760	122,619	81,361	82,649	85,370

**Ultimate Incurred Losses**

Accident Year	Incurred Ultimate All Years	Incurred Ultimate 3 Years	Incurred Ultimate 86 - 91	Adj Incurred Ultimate All Years	Adj Incurred Ultimate 3 Years	Adj Incurred Ultimate 86 - 91
1974	19,246	19,246	19,246	19,246	19,246	19,246
1975	23,162	23,162	23,162	23,162	23,162	23,162
1976	26,450	26,450	26,364	26,450	26,450	26,450
1977	30,077	30,074	29,910	30,077	30,074	30,074
1978	31,997	32,001	31,747	32,020	32,031	32,031
1979	34,548	34,538	34,211	34,583	34,596	34,596
1980	35,982	35,978	35,548	36,030	36,043	36,043
1981	35,181	35,210	34,665	35,260	35,258	35,258
1982	34,344	34,411	33,805	34,484	34,483	34,483
1983	37,780	37,856	37,206	37,976	37,980	37,980
1984	36,821	37,053	36,301	37,229	37,227	37,227
1985	36,183	36,637	35,778	36,709	36,821	36,821
1986	46,069	47,092	45,959	47,005	47,281	47,163
1987	59,577	61,020	59,731	60,692	61,307	61,108
1988	74,101	74,995	73,507	73,655	75,356	75,112
1989	87,227	84,445	82,575	79,835	83,423	82,907
1990	97,147	92,393	88,169	81,257	90,445	89,432
1991	91,612	93,242	82,327	85,596	97,272	92,953

*Dirty Data on Both Sides of the Pond*

**Appendix D**

**Ultimate Losses - Modified Data**

**Ultimate Paid Losses**

Accident Year	Paid Ultimate Change 1	Paid Ultimate Change 2	Paid Ultimate Change 3	Paid Ultimate All Changes	BF Paid Ultimate Change 1	BF Paid Ultimate Change 2	BF Paid Ultimate Change 3	BF Paid Ultimate All Changes
1974	19,246	19,246	19,246	19,246	19,246	19,246	19,246	19,246
1975	23,159	23,159	23,159	23,127	23,159	23,159	23,159	23,159
1976	26,417	26,417	26,417	26,397	26,417	26,417	26,417	26,417
1977	30,072	30,072	30,072	30,070	30,072	30,072	30,072	30,072
1978	32,020	32,020	32,020	32,035	32,020	32,020	32,020	32,020
1979	34,581	34,590	34,581	34,576	34,581	34,590	34,581	34,590
1980	36,053	36,065	36,053	36,065	36,053	36,065	36,053	36,065
1981	35,279	35,288	35,279	35,287	35,279	35,288	35,279	35,288
1982	72,471	34,600	34,574	72,411	72,471	34,600	34,574	34,600
1983	37,486	38,099	38,084	37,446	37,486	38,099	38,084	38,099
1984	-	37,789	37,739	-	-	37,789	37,739	37,789
1985	37,414	37,353	37,289	41,679	37,414	37,353	37,289	37,353
1986	50,083	49,636	49,475	63,723	50,083	49,636	49,475	49,636
1987	70,906	69,419	68,911	95,852	70,906	69,419	68,911	69,419
1988	99,986	96,302	95,093	146,063	99,986	96,302	95,093	96,302
1989	131,146	123,191	120,591	26,622	131,146	123,191	120,591	123,191
1990	158,013	143,701	65,422	124,043	110,507	105,534	76,153	105,534
1991	183,037	159,489	138,776	206,227	99,707	95,636	88,821	95,636

**Ultimate Adjusted Paid Losses**

Accident Year	Adj Paid Ultimate Change 1	Adj Paid Ultimate Change 2	Adj Paid Ultimate Change 3	Adj Paid Ultimate All Changes	BF Paid Ultimate Change 1	BF Paid Ultimate Change 2	BF Paid Ultimate Change 3	BF Paid Ultimate All Changes
1974	19,246	19,246	19,246	19,246	19,246	19,246	19,246	19,246
1975	23,159	23,159	23,159	23,127	23,159	23,159	23,159	23,127
1976	26,417	26,417	26,417	26,397	26,417	26,417	26,417	26,397
1977	30,072	30,072	30,072	30,070	30,072	30,072	30,072	30,070
1978	32,012	32,012	32,012	32,020	32,012	32,012	32,012	32,020
1979	34,554	34,554	34,554	34,530	34,554	34,554	34,554	34,530
1980	35,996	35,996	35,996	35,999	35,996	35,996	35,996	35,999
1981	35,231	35,231	35,231	35,228	35,231	35,231	35,231	35,228
1982	72,175	34,431	34,433	72,061	72,175	34,431	34,433	72,061
1983	37,188	37,775	37,775	37,140	37,188	37,775	37,775	37,137
1984	-	37,224	37,185	-	-	37,224	37,185	-
1985	36,575	36,498	36,470	35,872	36,575	36,498	36,470	40,758
1986	47,348	47,040	46,967	49,070	47,348	47,040	46,967	60,212
1987	62,396	61,291	60,881	65,938	62,396	61,291	60,881	86,230
1988	79,013	76,908	76,147	98,040	79,013	76,908	76,147	125,490
1989	84,315	80,628	78,998	14,364	84,315	80,628	78,998	18,093
1990	84,865	80,290	42,651	102,665	83,912	80,869	50,348	121,278
1991	118,001	106,922	98,053	184,627	84,387	82,093	78,740	66,243

**Ultimate Incurred Losses**

Accident Year	Incurred Ultimate Change 1	Incurred Ultimate Change 2	Incurred Ultimate Change 3	Incurred Ultimate All Changes
1974	19,246	19,246	19,246	19,246
1975	23,162	23,162	23,162	23,130
1976	26,450	26,450	26,450	26,430
1977	30,077	30,077	30,077	30,075
1978	31,997	32,001	31,997	32,015
1979	34,548	34,546	34,548	34,528
1980	35,982	35,981	35,982	35,982
1981	35,181	35,172	35,181	35,172
1982	72,196	34,345	34,344	72,087
1983	37,041	37,761	37,780	36,966
1984	-	36,815	36,821	-
1985	36,220	36,172	36,183	40,695
1986	46,031	46,065	46,069	59,299
1987	59,897	59,651	32,221	79,368
1988	75,538	74,519	69,942	118,947
1989	89,341	87,952	82,331	50,573
1990	102,929	99,297	62,892	124,885
1991	105,256	96,438	82,932	174,007

*Dirty Data on Both Sides of the Pond*

**Appendix B**  
Cumulative Paid Losses

Accident Year	Months of Development																	
	12	24	36	48	60	72	84	96	108	120	132	144	156	168	180	192	204	216
1974	\$267	\$1,975	\$4,587	\$7,375	\$10,661	\$15,232	\$17,888	\$18,541	\$18,937	\$19,130	\$19,189	\$19,209	\$19,234	\$19,234	\$19,246	\$19,246	\$19,246	\$19,246
1975	310	2,809	5,686	9,386	14,884	20,654	22,017	22,529	22,772	22,821	23,042	23,060	23,127	23,127	23,127	23,127	23,159	
1976	370	2,744	7,281	13,287	19,773	23,888	25,174	25,819	26,049	26,180	26,268	26,364	26,371	26,379	26,397	26,397		
1977	577	3,877	9,612	16,962	23,764	26,712	28,393	29,656	29,839	29,944	29,997	29,999	29,999	30,049	30,049			
1978	509	4,518	12,067	21,218	27,194	29,617	30,854	31,240	31,598	31,889	32,002	31,947	31,965	31,986				
1979	630	5,763	16,372	24,105	29,091	32,531	33,878	34,185	34,290	34,420	34,479	34,498	34,524					
1980	1,078	8,066	17,518	26,091	31,807	33,883	34,820	35,482	35,607	35,937	35,957	35,962						
1981	1,646	9,378	18,034	26,652	31,253	33,376	34,287	34,985	35,122	35,161	35,172							
1982	1,754	11,256	20,624	27,857	31,360	33,331	34,061	34,227	34,317	34,378								
1983	1,997	10,628	21,015	29,014	33,788	36,329	37,446	37,571	37,681									
1984	2,164	11,538	21,549	29,167	34,440	36,528	36,950	37,099										
1985	1,922	10,939	21,357	28,488	32,982	35,330	36,059											
1986	1,962	13,053	27,869	38,560	44,461	45,988												
1987	2,329	18,086	38,099	51,953	58,029													
1988	3,343	24,806	52,054	66,203														
1989	3,847	34,171	59,232															
1990	6,090	33,392																
1991	5,451																	

Claims Closed with Payment

Accident Year	Months of Development																	
	12	24	36	48	60	72	84	96	108	120	132	144	156	168	180	192	204	216
1974	268	607	858	1,090	1,333	1,743	2,000	2,076	2,113	2,129	2,137	2,141	2,143	2,143	2,145	2,145	2,145	2,145
1975	294	691	913	1,195	1,620	2,076	2,234	2,293	2,320	2,331	2,339	2,341	2,343	2,343	2,343	2,343	2,344	
1976	283	642	961	1,407	1,994	2,375	2,504	2,549	2,580	2,590	2,596	2,600	2,602	2,603	2,603	2,603		
1977	274	707	1,176	1,688	2,295	2,545	2,689	2,777	2,809	2,817	2,824	2,825	2,825	2,826	2,826			
1978	269	658	1,228	1,819	2,217	2,475	2,613	2,671	2,691	2,706	2,710	2,711	2,714	2,717				
1979	249	771	1,581	2,101	2,528	2,816	2,930	2,961	2,973	2,979	2,986	2,988	2,992					
1980	305	1,107	1,713	2,316	2,748	2,942	3,025	3,049	3,063	3,077	3,079	3,080						
1981	343	1,042	1,608	2,260	2,596	2,734	2,801	2,835	2,854	2,859	2,860							
1982	350	1,242	1,922	2,407	2,661	2,834	2,887	2,892	2,911	2,915								
1983	428	1,257	1,841	2,345	2,683	2,853	2,908	2,920	2,925									
1984	291	1,004	1,577	2,054	2,406	2,583	2,622	2,636										
1985	303	1,001	1,575	2,080	2,444	2,586	2,617											
1986	318	1,055	1,906	2,524	2,874	2,958												
1987	343	1,438	2,384	3,172	3,559													
1988	391	1,671	3,082	3,771														
1989	433	1,941	3,241															
1990	533	1,923																
1991	339																	

*Dirty Data on Both Sides of the Pond*

Cumulative Reported Claims

Accident Year	Months of Development																	
	12	24	36	48	60	72	84	96	108	120	132	144	156	168	180	192	204	216
1974	1,912	2,854	3,350	3,945	4,057	4,104	4,149	4,155	4,164	4,167	4,169	4,169	4,169	4,170	4,170	4,170	4,170	4,170
1975	2,219	3,302	3,915	4,462	4,618	4,673	4,696	4,704	4,708	4,711	4,712	4,716	4,716	4,716	4,716	4,716	4,717	
1976	2,347	3,702	4,278	4,768	4,915	4,983	5,003	5,007	5,012	5,012	5,013	5,014	5,015	5,015	5,015	5,015		
1977	2,983	4,346	5,055	5,696	5,818	5,861	5,884	5,892	5,896	5,897	5,900	5,900	5,900	5,900	5,900			
1978	2,538	3,906	4,633	5,123	5,242	5,275	5,286	5,292	5,298	5,302	5,304	5,304	5,306	5,306				
1979	3,548	5,190	5,779	6,206	6,313	6,329	6,339	6,343	6,347	6,347	6,347	6,348	6,348					
1980	4,583	6,106	6,656	7,032	7,128	7,139	7,147	7,150	7,151	7,153	7,154	7,154						
1981	4,430	5,967	6,510	6,775	6,854	6,873	6,883	6,889	6,892	6,894	6,895							
1982	4,408	5,849	6,264	6,526	6,571	6,589	6,594	6,596	6,600	6,602								
1983	4,861	6,437	6,869	7,134	7,196	7,205	7,211	7,212	7,214									
1984	4,229	5,645	6,053	6,419	6,506	6,523	6,529	6,531										
1985	3,727	4,830	5,321	5,717	5,777	5,798	5,802											
1986	3,561	5,045	5,656	6,040	6,096	6,111												
1987	4,259	6,049	6,767	7,206	7,282													
1988	4,424	6,700	7,548	8,105														
1989	5,005	7,407	8,287															
1990	4,889	7,314																
1991	4,044																	

Outstanding Claims

Accident Year	Months of Development																	
	12	24	36	48	60	72	84	96	108	120	132	144	156	168	180	192	204	216
1974	1,381	1,336	1,462	1,660	1,406	772	406	191	98	57	23	13	3	4	0	0	0	0
1975	1,289	1,727	1,730	1,913	1,310	649	358	167	73	30	9	6	4	2	2	1	1	
1976	1,605	1,977	1,947	1,709	1,006	540	268	166	79	48	32	18	14	10	10	7		
1977	2,101	2,159	2,050	1,735	988	582	332	139	66	38	27	21	21	8	3			
1978	1,955	1,943	1,817	1,384	830	460	193	93	56	31	15	9	7	2				
1979	2,259	2,025	1,548	1,273	752	340	150	68	36	24	18	13	4					
1980	2,815	1,991	1,558	1,107	540	228	88	55	28	14	8	6						
1981	2,408	1,973	1,605	954	480	228	115	52	27	15	11							
1982	2,388	1,835	1,280	819	354	163	67	44	21	10								
1983	2,641	1,765	1,082	663	335	134	62	34	18									
1984	2,417	1,654	896	677	284	90	42	15										
1985	1,924	1,202	941	610	268	98	55											
1986	1,810	1,591	956	648	202	94												
1987	2,273	1,792	1,059	626	242													
1988	2,403	1,966	1,166	693														
1989	2,471	2,009	1,142															
1990	2,642	2,007																
1991	2,366																	

*Dirty Data on Both Sides of the Pond*

Outstanding Losses

Accident Year	Months of Development																	
	12	24	36	48	60	72	84	96	108	120	132	144	156	168	180	192	204	216
1974	\$5,275	\$8,867	\$12,476	\$11,919	\$8,966	\$5,367	\$3,281	\$1,524	\$667	\$348	\$123	\$82	\$18	\$40	\$0	\$0	\$0	\$0
1975	6,617	11,306	13,773	14,386	10,593	4,234	2,110	1,051	436	353	93	101	10	5	5	3	3	
1976	7,658	11,064	13,655	13,352	7,592	4,064	1,895	1,003	683	384	216	102	93	57	50	33		
1977	8,735	14,318	14,897	12,978	7,741	4,355	2,132	910	498	323	176	99	101	32	14			
1978	8,722	15,070	15,257	11,189	5,959	3,473	1,531	942	547	286	177	61	67	7				
1979	9,349	16,470	14,320	10,574	6,561	2,864	1,328	784	424	212	146	113	38					
1980	11,145	16,351	14,636	11,273	5,159	2,588	1,290	573	405	134	81	54						
1981	10,933	15,012	14,728	9,067	5,107	2,456	1,400	584	269	120	93							
1982	13,323	16,218	12,676	6,290	3,355	1,407	613	398	192	111								
1983	13,899	16,958	12,414	7,700	4,112	1,637	576	426	331									
1984	14,272	15,806	10,156	8,005	3,604	791	379	159										
1985	13,901	15,384	12,539	7,911	3,809	1,404	827											
1986	15,952	22,799	16,016	8,964	2,929	1,321												
1987	22,772	24,146	18,397	8,376	3,373													
1988	25,216	26,947	17,950	8,610														
1989	24,981	30,574	19,621															
1990	30,389	34,128																
1991	28,194																	

*Dirty Data on Both Sides of the Pond*

Accident Year	Earned Exposures	TRUE Ultimates
1974	11,000	19,256
1975	11,000	23,161
1976	11,000	26,400
1977	12,000	30,049
1978	12,000	31,991
1979	12,000	34,529
1980	12,000	35,984
1981	12,000	35,207
1982	11,000	34,418
1983	11,000	38,354
1984	11,000	37,175
1985	11,000	36,446
1986	12,000	46,777
1987	13,000	60,676
1988	14,000	75,418
1989	14,000	88,115
1990	14,000	90,938
1991	13,000	74,807

Closing Rates

Accident Year	Months of Development																	
	12	24	36	48	60	72	84	96	108	120	132	144	156	168	180	192	204	216
1974	0.278	0.532	0.564	0.579	0.653	0.812	0.902	0.954	0.976	0.986	0.994	0.997	0.999	0.999	1.000	1.000	1.000	1.000
1975	0.419	0.477	0.558	0.571	0.716	0.861	0.924	0.964	0.984	0.994	0.998	0.999	0.999	1.000	1.000	1.000	1.000	
1976	0.316	0.466	0.545	0.642	0.795	0.892	0.946	0.967	0.984	0.990	0.994	0.996	0.997	0.998	0.998	0.999		
1977	0.296	0.503	0.594	0.695	0.830	0.901	0.944	0.976	0.989	0.994	0.995	0.996	0.996	0.999	0.999			
1978	0.230	0.503	0.608	0.730	0.842	0.913	0.963	0.982	0.989	0.994	0.997	0.998	0.999	1.000				
1979	0.363	0.610	0.732	0.795	0.881	0.946	0.976	0.989	0.994	0.996	0.997	0.998	0.999					
1980	0.386	0.674	0.766	0.843	0.924	0.968	0.988	0.992	0.996	0.998	0.999	0.999						
1981	0.456	0.669	0.753	0.859	0.930	0.967	0.983	0.992	0.996	0.998	0.998							
1982	0.458	0.686	0.796	0.875	0.946	0.975	0.990	0.993	0.997	0.998								
1983	0.457	0.726	0.842	0.907	0.953	0.981	0.991	0.995	0.998									
1984	0.428	0.707	0.852	0.895	0.956	0.986	0.994	0.998										
1985	0.484	0.751	0.823	0.893	0.954	0.983	0.991											
1986	0.492	0.685	0.831	0.893	0.967	0.985												
1987	0.466	0.704	0.844	0.913	0.967													
1988	0.457	0.707	0.846	0.914														
1989	0.506	0.729	0.862															
1990	0.460	0.726																
1991	0.415																	

*Dirty Data on Both Sides of the Pond*

**Appendix C**

**Ultimate Losses - Incomplete Data**

**Ultimate Paid Losses**

Accident Year	Paid Ultimate All Years	Paid Ultimate 3 Years	Paid Ultimate 86 - 91	BF Paid Ultimate All Years	BF Paid Ultimate 3 Years	BF Paid Ultimate 86 - 91
1974	19,246	19,246	19,246	19,246	19,246	19,246
1975	23,159	23,159	23,159	23,159	23,159	23,159
1976	26,417	26,417	26,405	26,417	26,417	26,405
1977	30,072	30,072	30,075	30,072	30,072	30,075
1978	32,020	32,020	32,043	32,020	32,020	32,043
1979	34,581	34,601	34,632	34,581	34,601	34,632
1980	36,053	36,066	36,144	36,053	36,066	36,144
1981	35,279	35,285	35,448	35,279	35,285	35,448
1982	34,574	34,504	34,782	34,574	34,504	34,782
1983	38,084	37,874	38,179	38,084	37,874	38,179
1984	37,739	37,392	38,036	37,739	37,392	38,036
1985	37,289	36,478	37,647	37,289	36,478	37,647
1986	49,475	47,268	49,448	49,475	47,268	49,448
1987	68,911	62,628	64,537	68,911	62,628	64,537
1988	95,093	80,904	83,371	95,093	80,904	83,371
1989	120,591	94,869	99,048	120,591	94,869	99,048
1990	138,214	100,918	109,831	103,782	89,851	93,851
1991	151,661	112,010	126,025	94,353	85,207	88,029

**Ultimate Adjusted Paid Losses**

Accident Year	Adj Paid Ultimate All Years	Adj Paid Ultimate 3 Years	Adj Paid Ultimate 86 - 91	BF Paid Ultimate All Years	BF Paid Ultimate 3 Years	BF Paid Ultimate 86 - 91
1974	19,246	19,246	19,246	19,246	19,246	19,246
1975	23,159	23,159	23,159	23,159	23,159	23,159
1976	26,417	26,417	26,393	26,417	26,417	26,393
1977	30,072	30,072	30,046	30,072	30,072	30,046
1978	32,012	32,010	31,991	32,012	32,010	31,991
1979	34,554	34,553	34,550	34,554	34,553	34,550
1980	35,996	35,997	36,026	35,996	35,997	36,026
1981	35,231	35,221	35,291	35,231	35,221	35,291
1982	34,433	34,425	34,579	34,433	34,425	34,579
1983	37,775	37,762	38,041	37,775	37,762	38,041
1984	37,185	37,175	37,774	37,185	37,175	37,774
1985	36,470	36,453	37,219	36,470	36,453	37,219
1986	46,967	47,097	48,564	46,967	47,097	48,564
1987	60,881	61,689	63,395	60,881	61,689	63,395
1988	76,147	78,056	80,216	76,147	78,056	80,216
1989	78,998	84,925	87,181	78,998	84,925	87,181
1990	77,709	88,184	92,645	79,397	84,345	87,230
1991	103,048	103,760	122,619	81,361	82,649	85,370

**Ultimate Incurred Losses**

Accident Year	Incurred Ultimate All Years	Incurred Ultimate 3 Years	Incurred Ultimate 86 - 91	Adj Incurred Ultimate All Years	Adj Incurred Ultimate 3 Years	Adj Incurred Ultimate 86 - 91
1974	19,246	19,246	19,246	19,246	19,246	19,246
1975	23,162	23,162	23,162	23,162	23,162	23,162
1976	26,450	26,450	26,364	26,450	26,450	26,450
1977	30,077	30,074	29,910	30,077	30,074	30,074
1978	31,997	32,001	31,747	32,020	32,031	32,031
1979	34,548	34,538	34,211	34,583	34,596	34,596
1980	35,982	35,978	35,548	36,030	36,043	36,043
1981	35,181	35,210	34,665	35,260	35,258	35,258
1982	34,344	34,411	33,805	34,484	34,483	34,483
1983	37,780	37,856	37,206	37,976	37,980	37,980
1984	36,821	37,053	36,301	37,229	37,227	37,227
1985	36,183	36,637	35,778	36,709	36,821	36,821
1986	46,069	47,092	45,959	47,005	47,281	47,163
1987	59,577	61,020	59,731	60,692	61,307	61,108
1988	74,101	74,995	73,507	73,655	75,356	75,112
1989	87,227	84,445	82,575	79,835	83,423	82,907
1990	97,147	92,393	88,169	81,257	90,445	89,432
1991	91,612	93,242	82,327	85,596	97,272	92,953

*Dirty Data on Both Sides of the Pond*

**Appendix D**

**Ultimate Losses - Modified Data**

**Ultimate Paid Losses**

Accident Year	Paid Ultimate Change 1	Paid Ultimate Change 2	Paid Ultimate Change 3	Paid Ultimate All Changes	BF Paid Ultimate Change 1	BF Paid Ultimate Change 2	BF Paid Ultimate Change 3	BF Paid Ultimate All Changes
1974	19,246	19,246	19,246	19,246	19,246	19,246	19,246	19,246
1975	23,159	23,159	23,159	23,127	23,159	23,159	23,159	23,159
1976	26,417	26,417	26,417	26,397	26,417	26,417	26,417	26,417
1977	30,072	30,072	30,072	30,070	30,072	30,072	30,072	30,072
1978	32,020	32,020	32,020	32,035	32,020	32,020	32,020	32,020
1979	34,581	34,590	34,581	34,576	34,581	34,590	34,581	34,590
1980	36,053	36,065	36,053	36,065	36,053	36,065	36,053	36,065
1981	35,279	35,288	35,279	35,287	35,279	35,288	35,279	35,288
1982	72,471	34,600	34,574	72,411	72,471	34,600	34,574	34,600
1983	37,486	38,099	38,084	37,446	37,486	38,099	38,084	38,099
1984	-	37,789	37,739	-	-	37,789	37,739	37,789
1985	37,414	37,353	37,289	41,679	37,414	37,353	37,289	37,353
1986	50,083	49,636	49,475	63,723	50,083	49,636	49,475	49,636
1987	70,906	69,419	68,911	95,852	70,906	69,419	68,911	69,419
1988	99,986	96,302	95,093	146,063	99,986	96,302	95,093	96,302
1989	131,146	123,191	120,591	26,622	131,146	123,191	120,591	123,191
1990	158,013	143,701	65,422	124,043	110,507	105,534	76,153	105,534
1991	183,037	159,489	138,776	206,227	99,707	95,636	88,821	95,636

**Ultimate Adjusted Paid Losses**

Accident Year	Adj Paid Ultimate Change 1	Adj Paid Ultimate Change 2	Adj Paid Ultimate Change 3	Adj Paid Ultimate All Changes	BF Paid Ultimate Change 1	BF Paid Ultimate Change 2	BF Paid Ultimate Change 3	BF Paid Ultimate All Changes
1974	19,246	19,246	19,246	19,246	19,246	19,246	19,246	19,246
1975	23,159	23,159	23,159	23,127	23,159	23,159	23,159	23,127
1976	26,417	26,417	26,417	26,397	26,417	26,417	26,417	26,397
1977	30,072	30,072	30,072	30,070	30,072	30,072	30,072	30,070
1978	32,012	32,012	32,012	32,020	32,012	32,012	32,012	32,020
1979	34,554	34,554	34,554	34,530	34,554	34,554	34,554	34,530
1980	35,996	35,996	35,996	35,999	35,996	35,996	35,996	35,999
1981	35,231	35,231	35,231	35,228	35,231	35,231	35,231	35,228
1982	72,175	34,431	34,433	72,061	72,175	34,431	34,433	72,061
1983	37,188	37,775	37,775	37,140	37,188	37,775	37,775	37,137
1984	-	37,224	37,185	-	-	37,224	37,185	-
1985	36,575	36,498	36,470	35,872	36,575	36,498	36,470	40,758
1986	47,348	47,040	46,967	49,070	47,348	47,040	46,967	60,212
1987	62,396	61,291	60,881	65,938	62,396	61,291	60,881	86,230
1988	79,013	76,908	76,147	98,040	79,013	76,908	76,147	125,490
1989	84,315	80,628	78,998	14,364	84,315	80,628	78,998	18,093
1990	84,865	80,290	42,651	102,665	83,912	80,869	50,348	121,278
1991	118,001	106,922	98,053	184,627	84,387	82,093	78,740	66,243

**Ultimate Incurred Losses**

Accident Year	Incurred Ultimate Change 1	Incurred Ultimate Change 2	Incurred Ultimate Change 3	Incurred Ultimate All Changes
1974	19,246	19,246	19,246	19,246
1975	23,162	23,162	23,162	23,130
1976	26,450	26,450	26,450	26,430
1977	30,077	30,077	30,077	30,075
1978	31,997	32,001	31,997	32,015
1979	34,548	34,546	34,548	34,528
1980	35,982	35,981	35,982	35,982
1981	35,181	35,172	35,181	35,172
1982	72,196	34,345	34,344	72,087
1983	37,041	37,761	37,780	36,966
1984	-	36,815	36,821	-
1985	36,220	36,172	36,183	40,695
1986	46,031	46,065	46,069	59,299
1987	59,897	59,651	32,221	79,368
1988	75,538	74,519	69,942	118,947
1989	89,341	87,952	82,331	50,573
1990	102,929	99,297	62,892	124,885
1991	105,256	96,438	82,932	174,007

# Data Organization and Analysis in Mortgage Insurance: The Implications of Dynamic Risk Characteristics

Tanya Havlicek and Kyle Mrotek, FCAS, MAAA

---

## Abstract

The capability for mortgage guaranty insurance companies to establish loss reserves conditioned on a dynamic risk characteristic, delinquency status, presents particular data issues. There is a need to collect, organize, warehouse, and analyze large data sets that contain loan-level detail over consecutive evaluation dates in order to measure the probability of claim, conditioned on delinquency status. The generally accepted methodology of reserving for mortgage guaranty insurance claim liabilities requires evaluation of dynamic risk characteristics because mortgage guaranty insurance companies need only reserve for loans currently delinquent, both known and IBNR. Because each loan's delinquency status is usually revised monthly by the mortgage servicing company, the cohort of insured loans currently delinquent changes each month and therefore is dynamic with respect to time. Coincidentally, delinquency status has been found to be a strong predictor of future losses, so it is imperative for mortgage guaranty insurance companies to estimate reserves as a function of delinquency status, a dynamic risk characteristic. Maintaining historical economic factors in step with the historical delinquency and claim data can also enhance the reserving approach.

---

## 1. INTRODUCTION

The generally accepted methodology of reserving for mortgage guaranty insurance claim liabilities is to reserve for loans currently delinquent, both known and IBNR. Mortgage guaranty insurance companies do not reserve for loans insured but not delinquent [1]. Estimating reserves requires the evaluation of dynamic risk characteristics because each loan's delinquency status is, typically, revised monthly by the mortgage servicing company. Therefore, the *cohort* of insured loans currently delinquent in a given month for which the mortgage guaranty insurance company needs to reserve changes each month and is dynamic with respect to time. Many delinquent loans do not result in a loss. However, the delinquency status of a loan is an established strong predictor of future losses [2], so it is imperative for mortgage guaranty insurance companies to estimate reserves as a function of delinquency status, a dynamic risk characteristic.

The capability for mortgage guaranty insurers to establish loss reserves conditioned on delinquency status presents particular data issues. There is a need to collect, organize, warehouse, and analyze large data sets that contain loan-level detail over consecutive monthly evaluation dates in order to measure the probability of claim conditioned on delinquency status. When a loan becomes delinquent, it can maintain the same delinquency status, become progressively more delinquent, or move back and forth between delinquency stages before eventually resolving into one of two fates: it may become current in payments

and be considered cured, or it may remain in default and result in a claim. There is then a need to track the eventual fate of each delinquency over consecutive monthly evaluations to its ultimate cure or claim.

The ability to distinguish and quantify delinquency trips and subsequent fates for all delinquent loans and then aggregate that data along risk-characteristic dimensions to develop reserving factors requires data availability and storage over consecutive monthly evaluation dates. Otherwise, the capacity to track fates and calculate empirical conditional claim probabilities is lost in data uncertainties. The dynamic nature of loan delinquency status manifests itself in mortgage guaranty insurance reserving in two aspects: calculating conditional claim frequencies from historical delinquencies to create reserving factors and identifying the current cohort of delinquent loans that need reserves, both reported and unreported.

## **2. BACKGROUND**

### **2.1 Mortgage Guaranty**

The Mortgage Guaranty Model Act of the NAIC defines mortgage guaranty insurance as insurance against financial loss by reason of nonpayment of principal, interest, or other sums agreed to be paid under the terms of any note or bond or other evidence of indebtedness secured by a mortgage, deed of trust, or other instrument constituting a lien or charge on real estate, providing the improvement on such real estate is designed for residential occupancy or industrial or commercial purposes [3].

The rationale for the existence of mortgage guaranty insurance is to disperse the credit risk of borrowers defaulting on their mortgages [4]. Lenders can offer borrowers mortgages more cheaply when the cost of mortgage insurance is factored in [4]. This is true because, like any insurable risk, the law of large numbers makes the variance around the mean smaller for more insured risks. Investors providing funds to the mortgage lenders (through several channels and ultimately the purchase of mortgage-backed securities) require lower returns when the credit enhancement of mortgage guaranty insurance is applicable and when this cost savings is more than the additional cost to the borrower [4]. In the end, the coupling of these phenomena allows more people to buy homes than would be able to otherwise.

Mortgage guaranty insurance is considered a property and casualty line of business, but it has notable differences from more traditional property and casualty lines of business.

The NAIC requires mortgage guaranty insurers to be monoline insurers. That is, in general, mortgage guaranty insurers are only allowed to underwrite mortgage guaranty

insurance and not other lines of business [5]. As a result, mortgage guaranty insurers generally diversify through geographic and temporal underwriting initiatives. Mortgage guaranty insurance losses are strongly correlated with macroeconomic events such as home price appreciation, unemployment, and interest rates [6]. Economic recessions tend to be concentrated regionally, so, to the extent a mortgage guaranty insurer is diversified geographically, the mortgage guaranty insurer's performance should vary less. Further, because mortgage guaranty insurance losses are strongly correlated with macroeconomic events, loans insured over extended underwriting periods are affected. As a result, mortgage guaranty insurers benefit from having a portfolio of insured loans underwritten over an extended period of time because the houses of loans insured years ago tend to have appreciated more in home price than houses of loans insured recently.

In contrast, traditional property and casualty insurers in general are allowed to underwrite multiple lines of business and often do so. A review of the 2006 annual statements for nearly 3,100 U.S. property casualty insurance companies indicates that only 2% of the filed direct and assumed earned premium for calendar year 2006 was from monoline insurance companies as measured by Schedule P lines of business.

Further, mortgage guaranty insurance policy terms are generally several years long and, depending on the amortization period of the mortgage, can be as long as 20 years. This is in contrast to traditional property and casualty policies with terms of one year or even six months. Additionally, the policies are generally noncancellable by the insurer except for nonpayment of premium. In other words, a policy can not be re-underwritten periodically as is common with traditional property and casualty lines of business. As a result, the premium rate schedule is stipulated at policy issuance.

In general, mortgage guaranty insurers offer three types of premium payment: monthly, annual, and up-front. The majority of mortgage loan borrowers engage in policies requiring premium payment on a monthly basis. The mortgage loan borrower submits a payment to the mortgage loan servicer each month. Depending on the mortgage product, the monthly payment includes amounts for principal, interest, hazard insurance, property taxes, and mortgage guaranty insurance premium. The mortgage loan servicer then submits the mortgage guaranty insurance premium to the mortgage guaranty insurer on a monthly basis. The premium is earned immediately by the mortgage guaranty insurer, as there is not an unearned premium reserve affiliated with monthly policies. Much less frequently, the policy may call for annual premium payments instead of monthly and, depending on regulations and specifics, there may or may not be an unearned premium reserve. Finally, single up-front premium policies generally include a provision for an unearned premium reserve.

Another feature of mortgage guaranty insurance that differs from traditional property and casualty lines of business is the relationship between the beneficiary and the premium payer. In traditional property and casualty lines of business, the premium payer is also the beneficiary. For example, auto liability insureds pay premiums and are covered against liabilities against them. By contrast, in mortgage guaranty insurance, borrowers pay the premium, and in the event of default, the mortgagee is the beneficiary and is reimbursed by the mortgage guaranty insurance company.

## **2.2 Reserving**

Property and casualty insurance companies are generally required to maintain loss and loss-expense reserves. Mortgage guaranty insurance companies are generally classified as property casualty insurance companies, so it follows that mortgage guaranty insurers must also maintain loss and loss-expense reserves (mortgage guaranty insurance is classified as line “S” Financial Guaranty/Mortgage Guaranty in Schedule P of annual statements for NAIC property and casualty insurance companies).

The Mortgage Guaranty Model Act of the NAIC reads “A mortgage guaranty insurance company shall compute and maintain adequate case basis and other loss reserves which accurately reflect loss frequency and loss severity and shall include components for claims reported and for claims incurred but not reported, including estimated losses on:

1. Insured loans which have resulted in the conveyance of property which remains unsold;
2. Insured loans in the process of foreclosure;
3. Insured loans in default for four months or for any lesser period which is defined as default for such purposes in the policy provisions; and
4. Insured leases in default for four months or for any lesser period which is defined as default for such purposes in policy provisions.”

As a note, mortgage guaranty insurance policy provisions generally stipulate that a loan is in default (a.k.a., delinquent) the moment one monthly payment is not made and until the time at which that payment and accrued interest have been repaid.

The list of four items above presents particular data and projection issues for the actuary in estimating loss and loss-expense reserves. As mentioned previously, the delinquency status of the mortgage is a strong predictor of the likelihood of claim. Conveyance, foreclosure, and length of default indicate various delinquency statuses. Further, within default, the

duration in which the loan has been in default also provides information on the likelihood of a claim. In general, the longer a loan has been in default, the more likely the loan will not cure and potentially lead to a claim. As such, loans in default for three months tend to be more likely to cure than loans in default for nine months. To better estimate the reserve, it serves the actuary well to be able to differentiate the probability of claim between loans in default with various statuses (i.e., one month, two months, three months, etc.). In order to estimate the probability of claim conditioned on delinquency statuses, the actuary may want data on the resolution of historical delinquencies, given delinquency status.

A reserving method commonly employed to estimate reserves consistent with the Mortgage Guaranty Model Act of the NAIC is a frequency-severity methodology. The frequency component of the method is incorporated by applying a probability of claim given that a loan is delinquent. In choosing the frequency factor (i.e., probability of claim conditioned on being delinquent), the reserving actuary will want to consider delinquency status, underwriting risk characteristics, and macroeconomic variables. Delinquency status can be based on the number of monthly payments missed or how long the loan has been consecutively delinquent. Potential underwriting risk characteristics include loan-to-value, borrower credit rating (e.g., FICO<sup>®</sup> Score), and property geography. Consideration for economic variables is addressed later in the article.

The severity component of the frequency-severity method can be viewed either as one factor net of salvage/subrogation, or as two components: loss given default (before salvage/subrogation) and recovery (salvage/subrogation). As is typically assumed with the frequency-severity method, the severity factor is the estimate of loss given that a claim occurred. Because the severity factor is often conditioned on a claim having occurred, the actuary may not want the quantification of the severity factor to be a function of delinquency status, in which case the premise of challenges posed by a dynamic input variable is moot. However, to the extent the actuary wants to reflect the dynamic delinquency status as an input into the severity estimate, it would pose further challenges not addressed in this article. The challenge in particular is the need to track not only a binary result (i.e., cure or claim) but also a loss (potentially relative to a coverage amount) along a continuum with respect to the dynamic delinquency status from month to month.

### **3. DATA ORGANIZATION**

There are two types of loan characteristics that must be stored: dynamic and static. Static characteristics are those that do not change over the lifetime of the loan. The static

characteristics database will contain loan information that the actuary may want to use as dimensions in developing reserving factors. Examples of static characteristics include original loan-to-value and borrower original FICO score. The static characteristics can be stored in a single policy-record database that is updated as new policies are insured.

Dynamic characteristics are those that can change monthly, such as delinquency status, and should be stored in a database that contains every monthly evaluation date. The database of dynamic characteristics contains a record for every month of the loan's lifetime and is compiled by appending the revised values each successive month. The dynamic status needs to be stored for every monthly evaluation so that historical delinquency cohorts and their fates can be accurately reconstructed and analyzed.

Size requirements and processing time may make it infeasible or impractical to store all attributes of all loans at every month, thus the segregation between the static and dynamic databases. Further, it is not necessary to store the static characteristics along with the dynamic fields. Consider the database size necessary to store monthly records of 100 fields for 100,000 loans for 156 months (a single book year of business over its lifetime) when only five of the 100 fields can change during the loan's lifetime. Then repeat this to add five book years of business. Clearly the database size will grow rapidly. It suffices to have a unique primary key ID field in the two databases such that information from the databases can be merged one-to-many correctly. The field typically used for this purpose is the policy's certificate number and is generally assigned by the mortgage guaranty insurer.

Before data storage space became relatively abundant and inexpensive in recent years, historical performance data was frequently purged or overwritten. One result of this practice is a "data vacuum," where information on prior delinquencies of a loan that cured is lost when the loan again becomes delinquent later on. In this instance, "prior delinquencies" for a particular mortgage guaranty insurance policy refer to all delinquencies except the most recent one. The absence of exhaustive historical performance data will almost always occur if only a single record with key status dates is kept for each insured loan policy, as opposed to storing delinquency information from every monthly evaluation date in the dynamic characteristics database. Overwriting key historical status dates, or purging, makes it impossible to reconstruct or analyze the delinquency behavior and exposure over time.

Other data organizational challenges may occur if the mortgage guaranty insurance company does not keep delinquency and claims information in the same database environment because the two types of data are handled by different departments. This

makes it challenging to match delinquency cohorts and resolution fates, particularly if the separate departments store data with incompatible database systems or key ID fields.

The preferred time window for collection and storage of loan-level data is, simply, as long as is possible. Data availability will determine, in part, the constraints of the analysis. Obviously, newer companies will have less performance data accumulated than older companies. At a minimum, to develop frequency and severity factors, there need to be enough consecutive months of complete data to observe a credible amount of delinquency resolutions. The more granular the reserving methodology, the more voluminous the data must be.

The time window of delinquency performance data should be long enough to allow many historical cohorts of delinquent loans to fully resolve into cure or claim. Although data maintained only quarterly can be used for such analyses, it reduces both the amount of data available to build credibility for analysis and the resolution of the analysis. Quarterly interval data also requires the use of additional assumptions because a loan could cure and become delinquent again during the three-month interval comprising a quarter, resulting in decay of the resolution accuracy of delinquency performance. Also, the mortgage guaranty insurance company need only carry reserves when the loan is delinquent.

Ideally, the historical time window the actuary considers for analysis should be long enough to capture an entire economic cycle, because delinquency, foreclosure, and claim rates are influenced by economic factors (e.g., unemployment rate) [6]. If the insurer's volume or product mix is volatile or heterogeneous, the ideal time window would capture behavior changes that could result from these changes and shifts. Also, mortgage insurance policies have extended policy terms relative to other property and casualty insurance policies, so the time window also should be long enough to observe the claims development for several policy years of business from inception to ultimate resolution.

Not only is the historical time period of data collection important, it is also imperative that the periods for which data are collected be contiguous. If there are "holes" where some months of delinquency activity are absent, it is impossible in many cases to determine the ultimate fate of any loan actively delinquent prior to and leading into the missing evaluation date. Further, the delinquency cohorts of the missing months cannot be used for data analysis. Data holes arise for various reasons. A loan-servicing company may not report to the mortgage guaranty insurer at every monthly evaluation date. Consultants have client relationships that may not be engaged in perpetuity, in which case the client would not provide data to the consultant when there is not an active engagement. If the client-

consultant relationship re-engages, the performance data from the period of relationship inactivity may no longer be available in its entirety because it was purged by the client for storage reasons or because the client contractually cannot provide the consultant with data from that period.

Despite the hole of historical performance data and its negative consequences, it may be possible to salvage information from the data. Although each data hole can lead to a significant loss in full information on many loans, depending on the number of policies in force, there may still be complete information on enough loans for credible analysis. The actuary will need to determine on a case-by-case basis whether enough information remains. Unfortunately, in addition to there being less information available to harvest from an incomplete historical data set, the amount of effort required to compensate for the inadequacies can also be far more than with a full data set. The actuary's programming code may require more "do loops," "if-then-else" conditions, and likely run-time. When there are data holes, there is less information on what has happened to a delinquency and, thus, there are more fate conditions or resolution possibilities to be considered. In some cases, it may be impossible to extract any information from the vicinity of the data hole.

The desired data organization is comprised of two databases that have been maintained for a period long enough to contain a credible amount of resolutions on delinquent loans. Both databases have a unique or key ID field specific to each loan, typically loan number or certificate number. Further, the unique ID per loan should be the same across databases. One database contains a single record for each loan, sometimes called the master policy file, of static loan and borrower characteristics (e.g., underwriting characteristics) that do not change over the life of the loan or that identify when the loan became inactive (termination or claim). Records for newly issued insurance policies can be appended to this database as new policies are written.

The second database is made up of dynamic loan characteristics that can change monthly and, therefore, it is updated monthly. However, just because it is updated monthly does not mean old information should be purged to make way for the new information. For each insured policy, the database of dynamic loan characteristics contains each evaluation date applicable to that policy and the dynamic status of each loan as of each of those dates. Specifically, the database of dynamic loan characteristics contains information on whether the loan is in force (an active policy) or not in force and the loan's status (current, delinquent, claim) for each evaluation date. As a note, a current, or nondelinquent, loan that is no longer in force is a terminated loan that is no longer insured by the mortgage guaranty company.

### **3.1 Storage Considerations**

By including dynamic date and status fields in the master policy database, which would need to be updated monthly, accurate delinquency histories can be constructed from dynamic databases that contain only the delinquent loans from each evaluation date. Some additional fields needed for reconstruction are current delinquency status and the date the loan achieved current delinquency status.

Warehousing only delinquent loans in the dynamic database requires more assumptions, merging, and date logic to program and process than using a dynamic database that tracks all loans ever written at every evaluation date; however, the decrease in data storage and program-processing-time requirements may make this organization more desirable than the “desired” organization described above. The design decision on how to organize the database will be based on the business requirements of the user and the hardware and software platforms that will support the data. If all loans ever written are included in the dynamic database, there is no ambiguity associated with an omitted loan. Examples of ambiguity include delinquency status as of evaluation dates and whether the loan is in force or terminated. By storing only delinquent loans as of each evaluation date, a loan may not appear at a given evaluation date for at least two reasons: the loan is no longer delinquent or there has been a data error. If all loans ever written are in the dynamic database, an omitted loan indicates a data error, either because the loan was accidentally omitted or because the loan did not belong to the mortgage guaranty insurance company and should be removed. However, depending on the size and age of the business, these files can rapidly become quite large. Clearly, warehousing only delinquent loans will use less storage than keeping a monthly status on all loans ever written.

## **4. DATA PROCESSING**

A delinquency cohort is the group of all loans delinquent as of the reserving evaluation date. Consider the following table, which presents a simplistic example of five loans over six months. The group of all delinquent loans at each evaluation date comprises six delinquency cohorts. Note that a particular loan delinquent for, say, three consecutive months will be part of those three delinquency cohorts. Table 1 is an example of the record layout from a dynamic characteristics database with four fields added for processing the data. Columns A-D come from the dynamic characteristics database (column A is implicit and shown for explanatory purposes), whereas columns E-H are added on during the program processing.

Table 1.

A Record #	B Evaluation Date	C Loan ID	D Status*	E Delq	F Delq Trip	G Cure	H Claim
1	Jan-06	1	0				
2	Jan-06	2	30				
3	Jan-06	3	90				
4	Jan-06	4	60				
5	Jan-06	5	0				
6	Feb-06	1	0				
7	Feb-06	2	30				
8	Feb-06	3	90				
9	Feb-06	4	0				
10	Feb-06	5	30				
11	Mar-06	1	30				
12	Mar-06	2	30				
13	Mar-06	3	120				
14	Mar-06	4	30				
15	Mar-06	5	60				
16	Apr-06	1	0				
17	Apr-06	2	30				
18	Apr-06	3	FCL				
19	Apr-06	4	60				
20	Apr-06	5	30				
21	May-06	1	30				
22	May-06	2	30				
23	May-06	3	FCL				
24	May-06	4	FCL				
25	May-06	5	0				
26	Jun-06	1	0				
27	Jun-06	2	30				
28	Jun-06	3	CLM				
29	Jun-06	4	CLM				
30	Jun-06	5	0				

\* 0 = Current; 30, 60, 90, 120 = days past missed mortgage payment; FCL = foreclosure; CLM = claim

(Note: Loans need not progress through delinquency categories consecutively or unidirectionally. For example, a loan can go from 90 days delinquent to 120 days delinquent to 30 days delinquent over three consecutive months. The jump backward from 120 days delinquent to 30 days delinquent in just one month can occur when the borrower makes up for several missed monthly mortgage payments at once).

The goal in processing the data is to determine the fate of each loan for every month it is delinquent, while distinguishing delinquency trips, so that claim ratios can be calculated for each cohort of loans. Delinquency trips are important because if a loan cures, it no longer needs a reserve. If a loan cures on a given delinquency trip but then becomes delinquent and

results in a claim at a later date on a subsequent delinquency trip, the former delinquency trip should not result or tally as a claim. The earlier delinquency trip should get full credit for the cure because, as discussed earlier, mortgage guaranty insurance companies do not reserve on ultimate claims for all insured loans, but only for losses related to loans that are currently delinquent and will not cure before leading to the insurance loss. Once delinquency fates are determined, the empirical conditional probability of claim for each monthly delinquency cohort and each delinquency status can be calculated via aggregation. Tallies are summed by delinquency cohort and risk characteristics and then claim probability is calculated as number of claims divided by number of delinquencies. This process is illustrated later.

Delinquency fates are determined by looking forward in time from each evaluation month to determine the resolution of each delinquency. Table 2 shows Table 1 condensed and tallied for Loan ID 3.

Table 2.

A Record #	B Evaluation Date	C Loan ID	D Status*	E Delq	F Delq Trip	G Cure	H Claim
3	Jan-06	3	90	1	1	0	1
8	Feb-06	3	90	1	1	0	1
13	Mar-06	3	120	1	1	0	1
18	Apr-06	3	FCL	1	1	0	1
23	May-06	3	FCL	1	1	0	1
28	Jun-06	3	CLM	0	1	0	1

\* 0 = Current; 30, 60, 90, 120 = days past missed payment; FCL = foreclosure; CLM = claim; columns E-H quantified via binary 0/1

Considering the delinquency cohort as of January 2006 (from record #3), Loan ID 3 is 90 days past due. Loan ID 3 becomes progressively more delinquent until Loan ID 3 results in a claim in June 2006. Loan ID 3 has a single delinquency trip that results in a claim in June 2006 (record #28). Therefore, for the delinquency cohort January 2006, delinquency category 90, Loan ID 3 results in a claim and is tallied as such in column H. Similarly, for delinquency cohort March 2006 (from record #13), delinquency category 120, Loan ID 3 results in a claim and is tallied as such in column H. This does not mean there are multiple claims on Loan ID 3, but rather, it is affiliated with multiple delinquency cohorts.

Alternatively, consider Table 3, condensed from Table 1, which highlights Loan ID 4.

Table 3.

A Record #	B Evaluation Date	C Loan ID	D Status*	E Delq	F Delq Trip	G Cure	H Claim
4	Jan-06	4	60	1	1	1	0
9	Feb-06	4	0	0	NA	NA	NA
14	Mar-06	4	30	1	2	0	1
19	Apr-06	4	60	1	2	0	1
24	May-06	4	FCL	1	2	0	1
29	Jun-06	4	CLM	0	2	0	1

\* 0 = Current; 30, 60, 90, 120 = days past missed mortgage payment; FCL = foreclosure; CLM = claim

In delinquency cohort January 2006 (from record #4), Loan ID 4 is 60 days past due on its first delinquency trip and results in a cure. This is because Loan ID 4 becomes current on payments during February 2006 (from record #9). However, for the evaluation months and delinquency cohorts that follow, Loan ID 4 tallies fate as a claim because its resolution from delinquency trip 2 results in a claim. Note that the hindsight delinquency segregation, categorization, and tallying can only occur because there is a contiguous history of delinquency status and evaluation dates. As previously mentioned, in practice, the mortgage guaranty insurance company only needs to reserve for a loan whenever it is delinquent or during any of the monthly cohorts in the tables where Delq = 1 (column E).

For completeness, Table 4 presents all the fate tallies from the dynamic characteristics database presented in Table 1.

Table 4.

A Record #	B Evaluation Date	C Loan ID	D Status*	E Delq	F Delq Trip	G Cure	H Claim
1	Jan-06	1	0	0	NA	NA	NA
2	Jan-06	2	30	1	1	0	0
3	Jan-06	3	90	1	1	0	1
4	Jan-06	4	60	1	1	1	0
5	Jan-06	5	0	0	NA	NA	NA
6	Feb-06	1	0	0	NA	NA	NA
7	Feb-06	2	30	1	1	0	0
8	Feb-06	3	90	1	1	0	1
9	Feb-06	4	0	0	NA	NA	NA
10	Feb-06	5	30	1	1	1	0
11	Mar-06	1	30	1	1	1	0
12	Mar-06	2	30	1	1	0	0
13	Mar-06	3	120	1	1	0	1
14	Mar-06	4	30	1	2	0	1
15	Mar-06	5	60	1	1	1	0
16	Apr-06	1	0	0	NA	NA	NA
17	Apr-06	2	30	1	1	0	0
18	Apr-06	3	FCL	1	1	0	1
19	Apr-06	4	60	1	2	0	1
20	Apr-06	5	30	1	1	1	0
21	May-06	1	30	1	2	1	0
22	May-06	2	30	1	1	0	0
23	May-06	3	FCL	1	1	0	1
24	May-06	4	FCL	1	2	0	1
25	May-06	5	0	0	NA	NA	NA
26	Jun-06	1	0	0	NA	NA	NA
27	Jun-06	2	30	1	1	0	0
28	Jun-06	3	CLM	0	1	0	1
29	Jun-06	4	CLM	0	2	0	1
30	Jun-06	5	0	0	NA	NA	NA

\* 0 = Current; 30, 60, 90, 120 = payment days past due; FCL = foreclosure; CLM = claim

In practice, there are not 30 records for five loans to analyze, but potentially millions of records for hundreds of thousands of loans. At the end of 2006, the private mortgage insurance industry had nearly \$800 billion of primary insurance in force [7]. This tallying procedure is executed with a programming language that can handle the logic of do loops and consecutive record comparison, so that key ID fields, delinquency statuses, and evaluation dates can be compared and processed. Two examples of programming languages that can accomplish these tasks are C++ and Visual Basic. For each record, tallies depend on

what happens in later records for the same certificate number and delinquency trip. Delinquency trip is determined by delinquency status and evaluation date, as was illustrated with Loan ID 4 in Table 3.

Table 5 illustrates the aggregation of tallies and calculation of empirical claim rate for one delinquency cohort, March 2006. Column 3 shows three delinquent loans of status 30. These are Loan IDs 1, 2, and 4 from record numbers 11, 12, and 14. Loan ID 1 results in a cure, for a sum of 1 for cure, status 30, in column D. Loan ID 2 is still delinquent at the end of the time window under consideration. The empirical claim rate can only be calculated based on those loans whose fate, or resolution, is known. Therefore, unresolved loans should be excluded from the calculation. Loan ID 4 results in a claim, for a sum of 1 for claim, status 30, in column E. Column G is calculated as the number of claims for the status divided by the number of resolved delinquencies, or the sum of cures and claims.

Table 5.

A Delinquency Cohort	B Status*	C Delqs	D Cures	E Claims	F = D+E Resolved Delqs	G = E/F Claim Rate on Resolved Delinquencies
Mar-06	30	3	1	1	2	50%
Mar-06	60	1	1	0	1	0%
Mar-06	90	0	0	0	0	NA
Mar-06	120	1	0	1	1	100%
Mar-06	FCL	0	0	0	0	NA

\*30, 60, 90, 120=payment days past due; FCL= foreclosure

As a note, it may also be of interest to the reserving actuary to calculate the maximum possible claim rate for a delinquency category. In the previous example, the max claim rate would be 67% (two-thirds). The ratio is calculated by summing every claim plus unresolved delinquencies (assumes all unresolved loans with claim) divided by number of loans in the delinquency cohort (3).

When fates are comprehensively tallied, the loan risk characteristics from the static database can be merged onto each record, such that resolution ratios (i.e., probability of claim versus probability of cure) for each cohort can be calculated along various risk dimensions. The fewer fields within each record to be processed, the more program run performance is optimized; therefore, record-by-record tallying is best done prior to merging the static characteristics. The risk dimensions that can or should be used depend on the robustness of the data and the judgment of the actuary (and are beyond the scope of this

discussion). Table 5 shows the most basic risk-dimension calculation based only on delinquency status and not including other characteristics.

Table 6 presents an example of what summarized tallies might look like for a single delinquency cohort aggregated along the risk dimension loan-to-value (the ratio of loan amount to purchase price). In general, the higher the percentage of loan relative to the home's value, the larger the likelihood of default and, similarly, claim. In general, higher loan-to-value ratios result in borrowers with less equity in the property and therefore less to lose in the case of default, versus borrowers with loans that have low loan-to-value. As mentioned previously, in general, the more severely a loan's delinquency status has progressed along the spectrum of delinquency status (i.e., 30, 60, 90+, FCL), the higher likelihood of claim. The authors have observed exceptions to this, but even then, the anti-intuitive empirical result is not significant. Table 6 is similar to Table 5 but with the addition of a second, albeit static, risk characteristic that allows the actuary to analyze the interaction of these two risk characteristics, delinquency status and loan-to-value.

Table 6.

A Status*	B Loan-To-Value	C Delqs	D Cures	E Claims	F = D+E Resolved Delqs	G = E/F Claim Rate
30	90	1000	930	70	1000	7%
	95	1200	1092	108	1200	9%
	100	1400	1232	168	1400	12%
60	90	800	720	80	800	10%
	95	900	792	108	900	12%
	100	1000	860	140	1000	14%
90	90	600	528	72	600	12%
	95	700	595	105	700	15%
	100	800	664	136	800	17%
120	90	300	240	60	300	20%
	95	350	266	84	350	24%
	100	400	288	112	400	28%
FCL	90	100	65	35	100	35%
	95	120	72	48	120	40%
	100	140	77	63	140	45%

\*30, 60, 90, 120 = payment days past due; FCL = foreclosure

Claim-rate frequency indications can be calculated using summary statistics of the actuary's choice by using different groupings of delinquency cohorts. From these indications, along with other sources for consideration, the actuary can select frequency factors to be applied to the current, and potentially future, cohort of delinquent loans for loss-reserving purposes.

## **5. ECONOMIC VARIABLES**

As mentioned previously, mortgage guaranty insurance performance is strongly dependent on macroeconomic factors. Macroeconomic factors found to be predictive of mortgage default include home price appreciation, unemployment, and interest rates (this list is not exhaustive). As such, the actuary may choose to include a loss-reserving methodology dependent on forecasted macroeconomic factors such as these.

Depending on the granularity of the modeling approach, the actuary may want to have available selected macroeconomic factors associated with historical mortgage loan defaults, loss given default, and recoveries. Collection of the corresponding macroeconomic variables is relatively easy. Generally, a high-speed Internet connection and time to gather and download the information is all that is required. The first pass at collecting all the historical information may require a fair amount of time up front, but updating the series periodically should be less onerous.

For example, assume the actuary wishes to estimate loss reserves each month and incorporate interest rates, home price appreciation, and unemployment into the loss-reserving process as leading factors.

The actuary may want to estimate loss reserves as a function of forecasted market mortgage interest rates, in addition to the dynamic delinquency status and other static underwriting risk characteristics. One possibility is to collect Freddie Mac's Primary Mortgage Market Survey<sup>®</sup> (PMMS) as a historical information source for mortgage interest rates. It provides a proxy for market mortgage rates for four mortgage products and also reports for the nation and five geographic regions. According to Freddie Mac, "Freddie Mac's Primary Mortgage Market Survey surveys lenders each week on the rates and points for their most popular 30-year fixed-rate, 15-year fixed-rate, 5/1 hybrid amortizing adjustable-rate, and 1-year amortizing adjustable rate mortgage products." Additionally, "Average rates and points (and margin for ARMs) for each product are reported for the nation and the five Freddie Mac regions."

The actuary can evaluate PMMS historical interest rates as predictors of claim probability, loss given default, and recovery rates. Possible models include logit models for default where the input variables include economic variables such as interest rate, as well as underwriting characteristics and delinquency status. Once a model relating interest rates as a leading indicator to mortgage loan default and mortgage insurance loss is developed, interest rates can be incorporated into the reserving process. Interest rates can be forecast using various interest rate models, or the actuary can rely on readily available deterministic estimates of

future mortgage interest rates. Freddie Mac offers mortgage rate forecasts in its weekly “Economic and Housing Market Outlook.” The Mortgage Bankers Association (MBA) offers on its website an economic forecast of Treasury interest rates and unemployment in its “MBA Long-term Economic Forecast.”

As mentioned earlier, home price appreciation and unemployment are other economic variables that can be collected and tested for significance of estimating loss reserves. Sources for historical home price appreciation data include the Office of Federal Housing Enterprise Oversight (OFHEO) House Price Index, Freddie Mac’s Conventional Mortgage Home Price Index (CMHPI) and the S&P/Case-Shiller<sup>®</sup> Home Price Indices. OFHEO’s House Price Index is published quarterly and geographically for the U.S. as a whole, nine U.S. Census divisions, state, and metropolitan statistical area (MSA). Freddie Mac’s CMHPI is also provided for the same geographic regions, while the S&P/Case-Shiller<sup>®</sup> Home Price Indices are only available for 20 large MSAs (and two composites), but broken out monthly. Finally, historical unemployment data can be obtained from the U.S. Department of Labor’s Bureau of Labor Statistics monthly and at the state level.

Depending on the granularity of the historical economic data along dimensions of frequency and geography (i.e., monthly versus quarterly or state versus Census division), preparing it for mapping to the preferred reserving methodology may require additional consideration. Conceptually, this tends to be straightforward. For example, using loan-level performance data where each loan record contains a field for property state but historical Freddie Mac mortgage rates provide only geographic regions (where these geographic regions contain multiple states) would require mapping the states to Freddie Mac’s geographic regions. In practice, this requires another step in the approach and generally leads to fewer field categories (i.e., 50 states, Washington, D.C., and territories get aggregated into five geographic regions).

Next, merging the collected historical economic data to test its predictive significance on default, loss given default, and recovery will require further effort. The actuary may want to test the historical economic variables with respect to the mortgage-loan performance at various time leads (e.g., one month, one quarter, or one year), and this adds another dimension to the considerations for historical economic data manipulation.

## **6. CONCLUSIONS**

Mortgage guaranty insurance loss reserves are provisions for losses due to insured loans currently delinquent, both reported and unreported. Specifically, there need not be a

provision for losses due to loans insured but not delinquent. As a result, the status of whether a loan is delinquent or not is integral to the reserve estimate. The extent of a loan's delinquency has been found to have significance as a predictor of loan default and therefore insured loss. Because of the dynamic nature of each loan's delinquency status over time, the reserving actuary will want a contiguous historical performance data set with enough information to reconstruct the month-by-month status of each insured loan so as to quantify the relationship between delinquency status (dynamic) and other characteristics (generally static but potentially dynamic, such as borrower's current FICO<sup>®</sup> Score) to ultimate fate and claim loss. The ability to reconstruct this history requires monthly database updating, relational database fields with integrity (i.e., unique ID keys that can be referenced across different data sets) and maintenance without purging.

## 7. REFERENCES

- [1] “Mortgage Guaranty Insurance Model Act” Model #630-1, National Association of Insurance Commissioners, Section 16.
- [2] DeFranco, Ralph, “Modeling Residential Mortgage Termination and Severity Using Loan Level Data”, Spring 2002, page 74.
- [3] “Mortgage Guaranty Insurance Model Act” Model #630-1, National Association of Insurance Commissioners, Section 2.
- [4] Dennis, Marshall W. and Robertson, Michael J., Residential Mortgage Lending, Fourth Edition, 1995, page 131.
- [5] “Mortgage Guaranty Insurance Model Act” Model #630-1, National Association of Insurance Commissioners, Section 9.
- [6] Siegel, Jay, “Moody’s Mortgage Metrics: A Model Analysis of Residential Mortgage Pools”, April 1, 2003, page 10.
- [7] *Inside Mortgage Finance*, Feb. 16, 2007.

### Abbreviations and notations

ARM, adjustable rate mortgage  
CMHPI, Freddie Mac’s Conventional Mortgage Home Price Index  
IBNR, incurred but not reported  
MBA, Mortgage Banker’s Association  
MSA, Metropolitan Statistical Area  
NAIC, National Association of Insurance Commissioners  
OFHEO, Office of Federal Housing Enterprise Oversight  
PMMS, Freddie Mac’s Primary Mortgage Market Survey

### Glossary of Terms

1-year amortizing adjustable rate mortgage, a mortgage with an interest rate that changes annually  
5/1 hybrid amortizing adjustable-rate mortgage, a mortgage with an initial five-year fixed-interest rate; thereafter the interest rate begins to adjust on an annual basis  
Conveyance, the transfer of property from one person to another  
Delinquent, mortgage overdue in payment  
Delinquency cohort, group of loans with the same accident month  
Delinquency status, categorical classification of a mortgage’s overdue payment  
Delinquency trip, series of monthly delinquency statuses beginning on a loan’s accident month and only ending with a status of cure or claim  
Fate, ultimate resolution of delinquent loan  
Foreclosure, proceeding in which the financier of a mortgage seeks to regain property  
Length of default, time elapsed between evaluation date and accident month

### Biographies of the Authors

**Tanya Havlicek** is an Actuarial Assistant and Statistician at Milliman in Milwaukee, WI. She has a B.S. in mathematics from The Ohio State University and an M.S. in Land Resources from the University of Wisconsin - Madison. She is an expert in SAS and has 15+ years of programming experience from coding in a variety of languages. She works on projects involving loss reserve analysis, reinsurance, government and international markets. Prior to joining Milliman, Tanya was a research assistant at the University of Wisconsin - Madison and developed models to analyze complex interactions in a natural resource management context.

**Kyle Mrotek** is an Actuary at Milliman in Milwaukee, WI. He has B.B.A. degrees in both Actuarial Science and Finance from the University of Wisconsin - Madison. Kyle is a Fellow of the CAS and a Member of the American Academy of Actuaries. He has performed work for mortgage insurers, mortgage lenders, state housing finance agencies and other government guaranty insurers. Recently, Kyle worked out of Milliman’s London office for one year.

# ROOT: A Data Analysis and Data Mining Tool from CERN

Ravi Kumar ACAS, MAAA, and Arun Tripathi, Ph.D.

---

## Abstract

This note briefly describes ROOT, which is a free and open-source data mining tool developed by CERN, the same lab where the World Wide Web (WWW) was invented. Development of ROOT was motivated by the necessity to address the challenges posed by the new generation High Energy Physics experiments, which are expected to produce and analyze thousands of terabytes of very complex data every year.

ROOT is an object-oriented data analysis framework, written in C++. It contains several tools designed for statistical data exploration, fitting, and reporting. In addition, ROOT comes with powerful high-quality graphics capabilities and interfaces, including an extensive and self-contained GUI development kit that can be used to develop easy to use customized interfaces for the end users. This note provides some simple examples of how ROOT can be used in an insurance environment.

---

## INTRODUCTION

In this paper, we provide an introduction to some features of ROOT [1] by using it to simulate and analyze the simulated data. We also show some very basic, but necessary, first steps needed for one to become familiar with ROOT. Going through this process will hopefully give the reader a flavor of some of the analysis tasks that can be accomplished within ROOT. Also, hopefully this will provide the reader enough of a familiarity and hands-on experience with ROOT so that they can start using its more advanced features, customized to their own needs.

We want to emphasize that this is just a preview, intended for readers who might not be familiar with ROOT at all. The scope of various tasks that can be accomplished using ROOT is much more comprehensive. We will provide Web links and references at the end of this paper for the curious reader who wants to learn more about this tool.

ROOT is a free, open-source, object-oriented data analysis framework based on C++. This tool was developed at CERN [2], which is a particle physics lab located near Geneva, Switzerland. It is interesting to note that CERN is the same lab where the World Wide Web was born [3, 4].

Development of ROOT was motivated by the need to address the challenges posed by the experimental high-energy physics community, where scientists produce and analyze vast amounts of very complex data. For example, the ATLAS [5, 6] experiment at the Large

Hadron Collider (LHC) [7] at CERN will be generating over 1,000 terabytes of data per year. And this is just one of the experiments running at LHC.

ROOT is being used widely by several experiments in high-energy physics, astrophysics, etc. [8]. In terms of the cost of these research projects, and the people involved, the ROOT user community comprises a multibillion dollar “industry,” with the labs and the users located pretty much across the whole planet.

## **WHY ROOT?**

ROOT is a very appropriate tool for use by actuaries and other insurance analysts who do ad hoc data analysis and predictive modeling type work.

ROOT is a framework that is specifically designed for large scale data analysis. ROOT stores data in a very efficient way in a hierarchical object-oriented database. This database is machine independent and highly compressed. If one loads a 1 GB text file into a ROOT file, it will take up much less disk space than the original text file. ROOT also has tools to interact with data in a very efficient way. It has built in tools to do multi-dimensional histograms, curve fitting, modeling and simulation. All these tools are designed to handle large volumes of data.

Conversely, relational databases (databases where the data is organized as tables and rows) were originally designed for transactional systems and not for data analysis. Thus a relational database is very good for use in a policy administration system, which looks at one policy at a time, or claim administration system, which looks at one claim at a time. But, when one is interested in segmenting the data across all the policies or across all the claims, a relational solution falls apart. In order to make the relational solution work for large scale data analysis, we use the brute force method. A typical brute force method will involve adding considerable computing power, adding sophisticated I/O capabilities such as cache, etc., adding numerous indices to tables, creating additional summaries of the data (like OLAP cubes), and other similar techniques. If one loads a 1 GB text file into a relational database, it will take up multiple gigabytes to just store the data. When one further tweaks the database for performance with additional indices, pre-summaries and such, the original 1 GB data would have exploded to something very large. Most (if not all) of the commercial software for data analysis is built for accessing data from relational databases. These commercial tools cannot overcome the fundamental flaw in the way data is stored (tables and rows) except by using brute force.

Some data analysis tools are very memory intensive. Some data analysis tools are very I/O intensive. Some data analysis tools are both memory intensive and I/O intensive (like most commercial business intelligence tools operating on relational databases). In these systems, even if the data grows on a linear scale, the performance of the system degenerates on an exponential scale. Thus, these systems are not easily scalable, whereas ROOT stores and retrieves data in an optimal way that is conducive for data analysis. It avoids most memory issues and I/O performance issues by seamlessly buffering the data between memory and storage. One can thus get a very reasonable throughput from ROOT even from a small PC (all the analysis reported in this paper was done on a PC). A laptop with ROOT as a data analysis tool may be able to give a better performance than a powerful mainframe using one of the commercially available data analysis tools. ROOT can thus be a solution adopted by one person in an insurance company. Once proven, it can be easily extended to an entire team of data analysts or as a corporate wide solution. A ROOT solution is very highly scalable.

ROOT might be an appropriate solution even for smaller data sets. Typically, predictive modeling and ad hoc data analysis involve presenting the data in different graphical/tabular forms. These presentations are best done in a notebook device. This is one of the reasons why Excel is very popular among the actuaries. Using Excel, one can play with the data and once a story emerges from the data, it becomes easy to share the story with the rest of the team. This concept can be loosely termed as interactive computing. When one wants to do analysis on one column in an Excel spreadsheet, the entire spreadsheet must be read into memory. Like Excel, other technologies also suffer similar inefficiencies. When data is stored as tables and rows as in a relational database, subsets of the data cannot be accessed or modified in an efficient way without touching other parts of the data. The design of ROOT allows access to subsets of data without the need to touch the rest of the data. An entire ROOT file can be read sequentially if all the information must be processed. With no data explosion issues, a ROOT file can also be read randomly to process just a few attributes if that is what the analysis requires. ROOT is thus able to give us interactive computing capabilities where other solutions fail.

There are many other reasons why ROOT is an appropriate tool for predictive modeling. But efficiency in storing and accessing the data is where ROOT stands out from any other tool that is in the market today.

## HOW TO GET ROOT

ROOT can be downloaded under GNU Lesser General Public License [9] from the ROOT download page [10]. Installation instructions are also provided there. A ROOT user's guide [11], complete class reference [12], tutorials [13], and useful how-to's [14] are also available online. A searchable ROOT user forum, called Root Talk [15], is a useful resource to find answers to several of the questions an average user might come up with.

Any references in this paper to the ROOT user's guide correspond to version 5.16, which is the current production version of ROOT as of the writing of this paper.

Throughout this paper, we will sometimes provide the CPU time taken for a given analysis. These times were measured on a PC running Windows XP, with a 1.73 GHz Intel Pentium M processor, and 1 GB of memory. Also, all these analyses were performed using CINT, the C interpreter provided by ROOT. See chapter 7 of the ROOT users guide to learn about CINT.

## STARTING ROOT

If ROOT was installed correctly, a tree-shaped icon, shown in Figure 1, should automatically appear on your Windows desktop.



*Figure 1: The ROOT shortcut icon on Windows desktop.*

In order to start ROOT in Windows, just double-click on this icon. This will start the ROOT console, which is shown Figure 2.

In Unix/Linux environment, ROOT can be started by issuing the following command from the command line:

```
$ROOTSYS/bin/root
```

ROOTSYS is the environment variable pointing to the directory where ROOT was installed. If the directory containing the ROOT executable is already in the system path, then

one just needs to type “root” from command line to start ROOT. Regardless of the operating system, the resulting ROOT console will appear the same (as shown in Figure 2).

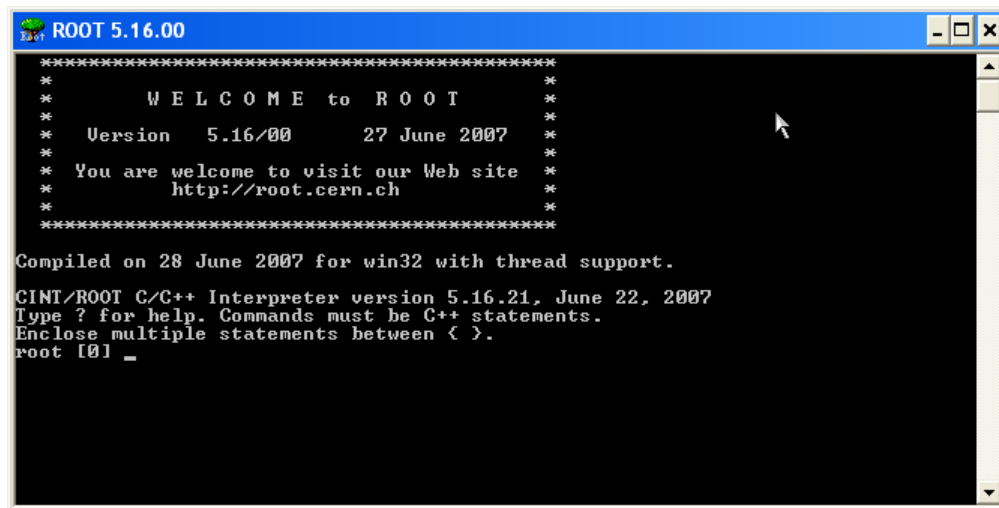


Figure 2: The ROOT console.

## LOADING DATA INTO ROOT

ROOT provides TTree and TNtuple classes to store and access data efficiently. Chapter 12 of the ROOT user’s guide provides a detailed discussion of ROOT Trees, why one should use them for storing data, and how to read data into a ROOT tree.

ROOT trees are designed specifically to store large volumes of data very efficiently, resulting in much smaller files on disk. Also, since a tree stores data in hierarchical branches, each branch can be read independently from any other branch. This can make for a very fast access to the data, since only the necessary information is read from disk, and not necessarily the whole file.

A very simple example of how to read data into a ROOT tree is given in appendix A. This example converts a space delimited file into a ROOT file, which can then be explored/manipulated further using ROOT.

ROOT also provides interfaces using ODBC to relational databases such as ORACLE, MYSQL, etc.

## EXPLORING A ROOT FILE

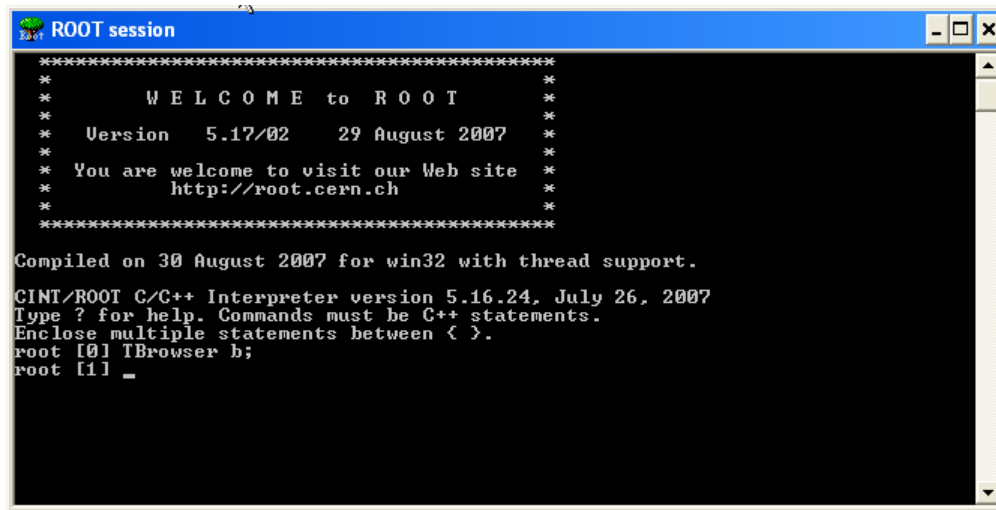
The TTreeView class [16] of ROOT provides a GUI for convenient exploration of data, once it has been converted into a ROOT tree. In order to illustrate some of its

functionality, we will use the ROOT file generated by the sample data load program mentioned in the previous section. Chapter 12 of the ROOT user's guide describes how to start a Tree Viewer.

First, one has to start a ROOT object browser (TBrowser class [17]) from the ROOT console:

```
root [] TBrowser b;
```

Figure 3 shows a screen shot of the ROOT console, with this command.



*Figure 3: A screen shot of the ROOT console, with the command to start the ROOT browser.*

This will start the ROOT object browser, which looks like figure 4.

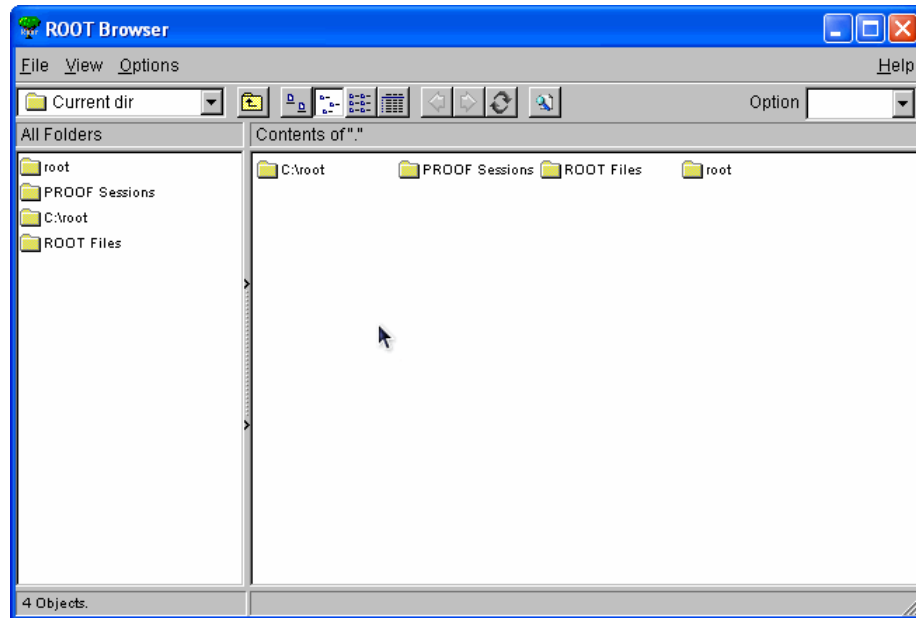


Figure 4: A screen shot of ROOT object browser.

Now, one can use this object browser to open a ROOT file by using **File->Open** menu. In this case, we will navigate to the ROOT file generated by the sample data load program of Appendix A. Using the File menu, open the root file called "SampleData.root". Figure 5 shows a screen shot of the file selection dialog, used to open the file.

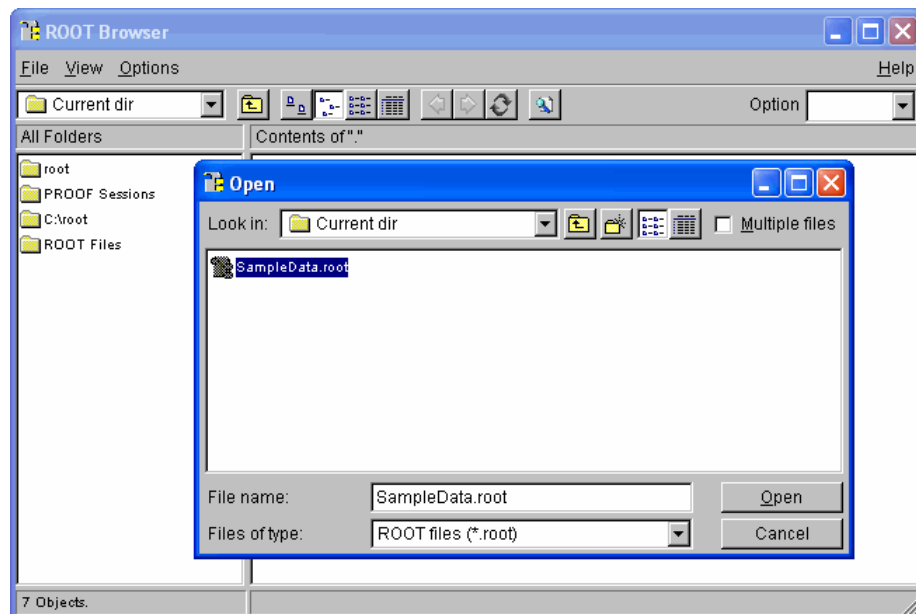
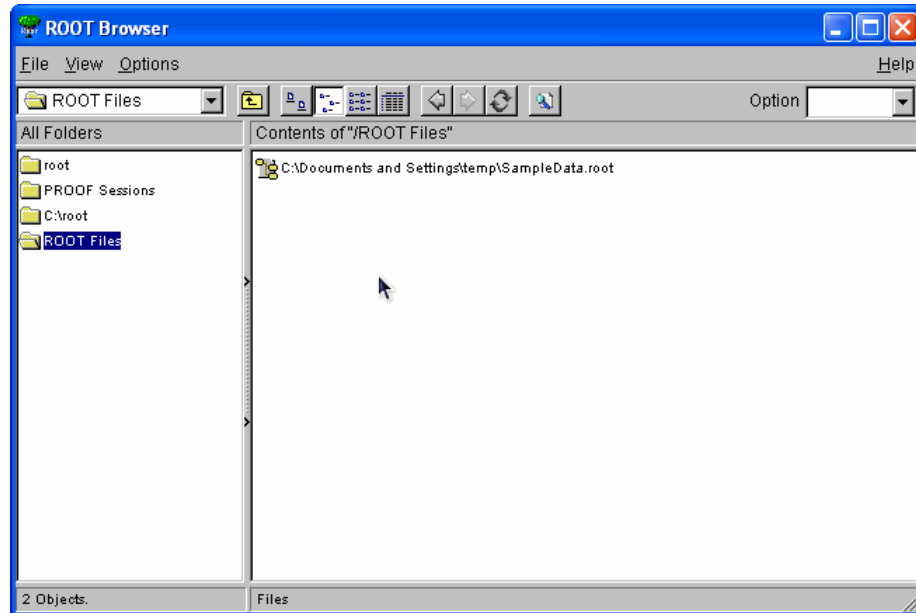


Figure 5: Screen shot of the ROOT object browser and file selection window, after navigating to the ROOT file generated by the sample data load program.

After selecting the appropriate ROOT file, click the “open” button in the file selection dialog. This will close the file-selection window, and the object browser will again appear like Figure 4. At this point, double-click the icon labeled “ROOT Files” in the right-hand panel of the object browser. After this action, the browser looks like Figure 6.



*Figure 6: Appearance of the ROOT browser after double-clicking on the “ROOT Files” icon.*

Notice that an icon representing the selected ROOT file appears in the right panel of the browser. The absolute path indicating the file name and location is also shown. Now, double-click on this ROOT file icon. The browser will now look like Figure 7.

Notice that a tree icon appears in the right panel of the browser. This is the tree that we created using the sample data load program. One can create several ROOT trees in a single ROOT file, but in this case, we have just one.

Now, right-click on the tree icon and a menu appears. From this menu, select “StartViewer”. A new Tree Viewer window will appear. A screen shot of this window is shown in Figure 8.

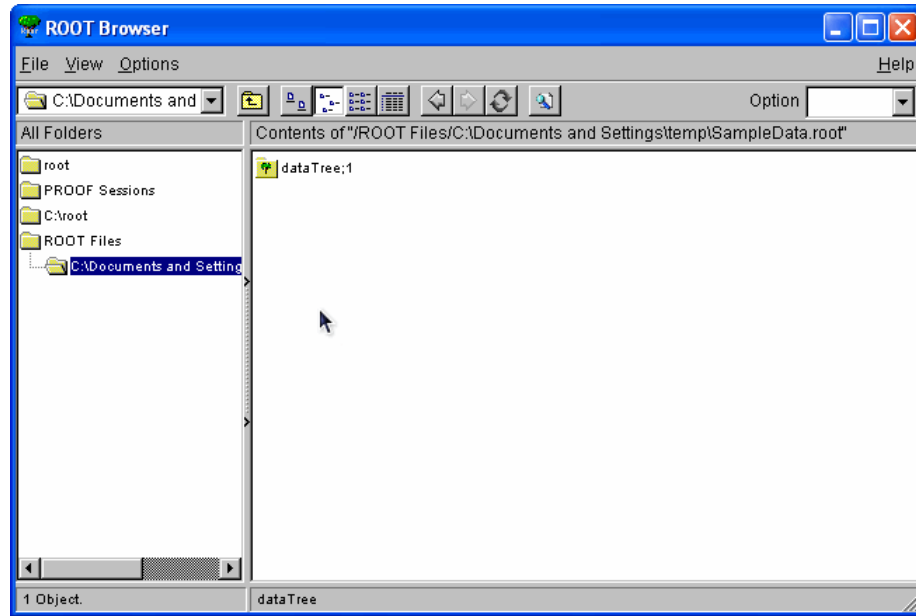


Figure 7: Appearance of the ROOT object browser, after double-clicking on ROOT file icon (shown in Figure 3).

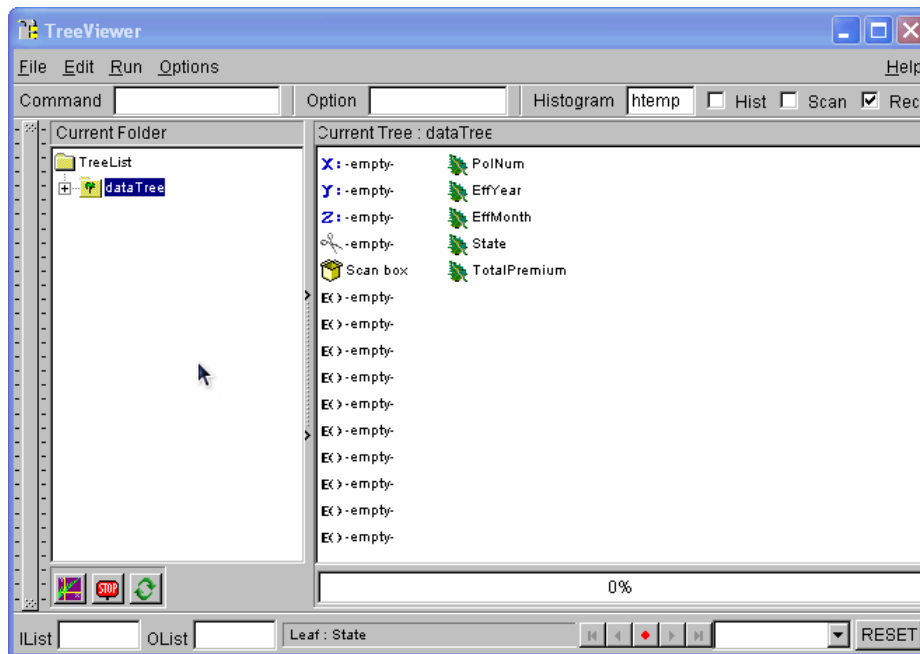


Figure 8: The ROOT tree viewer window, displaying the information contained in the sample data tree.

Notice in Figure 8 the leaf shaped icons in the right panel of the tree viewer. These are the leaves of the ROOT tree we created in the sample data load program. Next to each leaf is the name of the corresponding variable.

The ROOT tree viewer is a powerful data exploration tool, which can be used for one-, two-, and three-dimensional data exploration, fitting, etc. In this section, we will just see how to quickly generate one- and two-dimensional histograms from this simple data.

Suppose we want to see a histogram of all the states in our data set, which will show us the number of policies written in each state. In order to get this histogram, simply double-click the leaf labeled “State” in the tree viewer. A separate window (object of type TCanvas) appears with this histogram. A screen shot of this window is shown in Figure 9. We see that there are four policies in California, and two each in New York, Kansas, and Arizona; as expected from our input data.

The TCanvas object itself is a very complex object, which allows the user to interactively explore the data, and customize the visual appearance of the graphics that appears on the canvas. See chapter 9 of the user’s guide for more details.

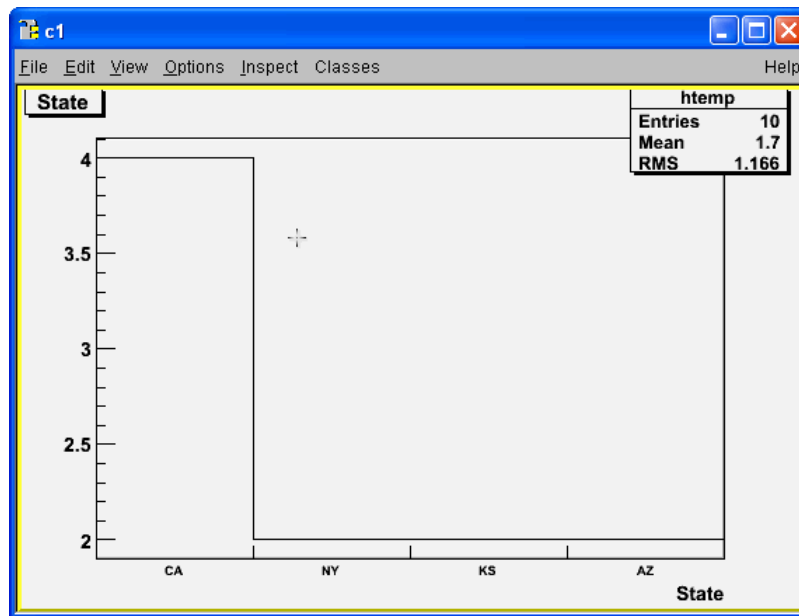


Figure 9: A histogram of all the states in the data set used by the sample data load program.

The histogram in Figure 9 can easily be saved to disk in various formats, e.g., gif, pdf, png, etc., by using **File->Save** (or Save As) menu in the menu bar at the top of the canvas.

The reader is encouraged to explore this interactive canvas, all the objects on it, and the associated context menus. Right-click on different parts of the canvas, e.g., the title box, statistics box, the histogram area, the lines of the histogram, the axes, etc., and explore the large amount of interactive functionality available in the context menus.

Now, suppose we are interested in finding out the average premium collected in each state. We can do that in the following manner. In the tree viewer (Figure 8), drag the leaf labeled “TotalPremium” to the icon (in the same panel) labeled Y (which is empty by default). Then, drag the leaf labeled “State” to the icon labeled X. This tells the tree viewer to plot TotalPremium (Y-axis) vs. State (X-axis). Next, using the Options menu at the top menu bar in the tree viewer, set the drawing option to Profile (**Options->2D Options->Profile**). This tells the tree viewer to plot a profile histogram, which plots the average Y value for each bin on the X-axis. After these steps, the tree viewer window should look like Figure 10.

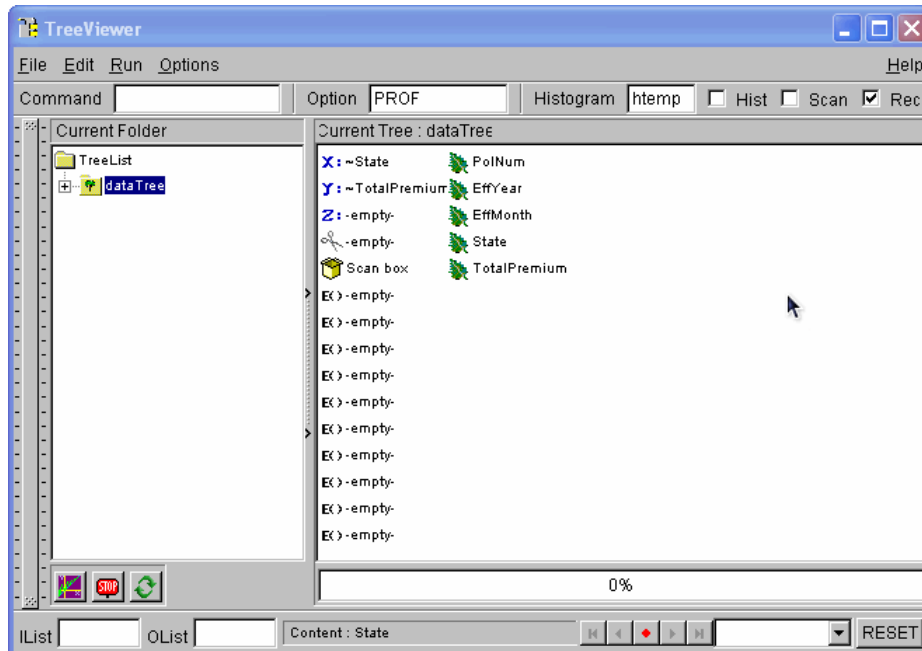


Figure 10: The appearance of tree viewer window, after preparing it to plot average total premium in each state.

In Figure 10, notice that the X icon says State next to it and the Y icon says TotalPremium next to it. Also notice that the Option is set to PROF in the small Option window, below the top menu bar. Now we are ready to produce the graph we want. Just click on the graphing icon near the bottom left corner of the panel, on the left of the icon labeled STOP. Again, a new canvas pops up, with the desired graph. A screen shot of this canvas is shown in Figure 11. Notice how the average premium, as well as RMS (root mean square) is plotted for each state. This allows for a visual exploration of the relationships between any two variables quickly. One can also perform fits to this relationship. The reader is referred to the ROOT user’s guide to learn how to perform interactive fits to the data points on a canvas. For example, see chapter 5 of the ROOT user manual.

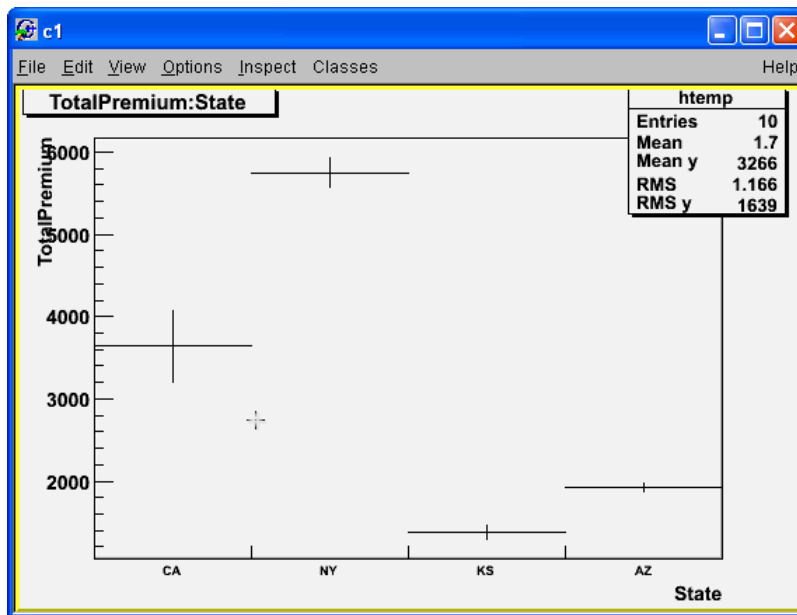


Figure 11: Graph of average premium vs. state.

Now, suppose we want to apply filters to our data. The tree viewer allows us to do that easily. In order to illustrate how to do this, we will create the profile histogram of average total premium vs. state, but this time, we only want to look at the states of California and New York.

There is a scissors shaped icon in the tree viewer, just below the “Z” icon (see Figure 10). This is used to apply “cuts” or selection criteria to the data. The cut can be any valid C++ expression, involving one or more of the variables stored in the tree. This expression must first be defined, using the “E()” icons in the tree viewer. All these expressions are empty by default.

In order to achieve our goal of looking at the average premium only in California and New York, double-click on the first “E()” icon in the tree viewer of Figure 10 (we will assume that the tree viewer is in the state shown in Figure 10; if not, follow the instructions above to bring it this state). A dialog appears that allows us to type in the selection criterion and also give it a name. Figure 12 shows the screen shot of the tree viewer after double-clicking on the “E()” icon.

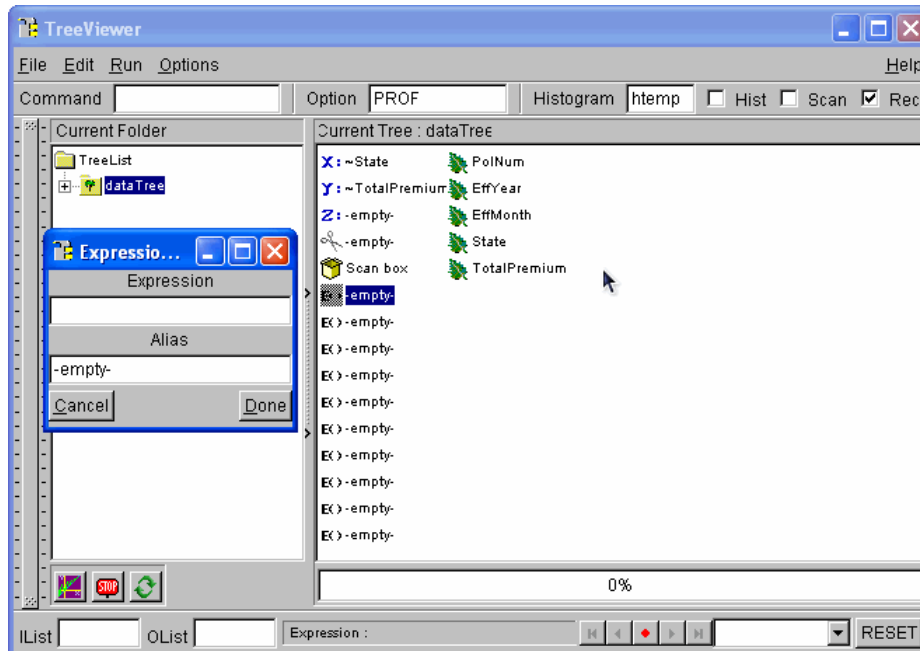


Figure 12: The tree viewer after double-clicking on the “E()” icon. Notice the dialog box at left-center of the tree viewer.

Now, we can type in our selection criterion (a valid C++ statement, involving variables in the tree) in this box, and also assign it a name (alias). We just want to keep California and New York. So in the Expression dialog, we will type `(State == \"CA\") || (State == \"NY\")`. Note that the double quotes are preceded by backslash, since the expression itself is a string, and within this expression we are comparing with a string. Also keep in mind that both C++ and ROOT are case-sensitive.

In the alias box, we will give it the name “Cut1”. After these steps, the screen shot of the tree viewer is shown in Figure 13.

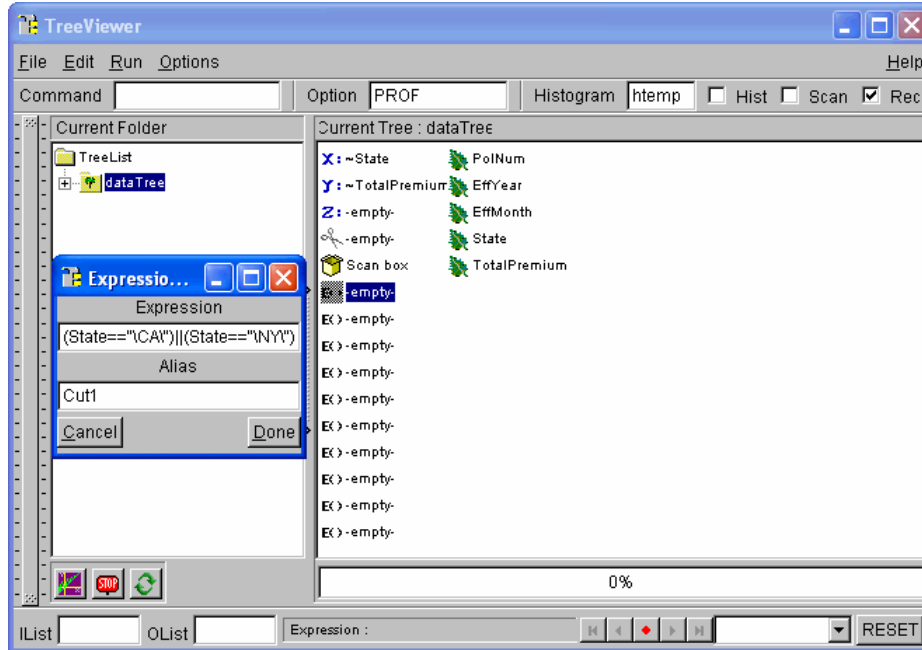


Figure 13: The tree viewer after typing in the selection criteria in the expression dialog.

Now, click the “done” button in the selection expression dialog. This will close the selection expression dialog. The screen shot of the tree viewer after this step is shown in Figure 14. Notice how the appearance of the selection expression icon we just edited has now changed. It no longer shows “E()”, and it is no longer empty. Now it contains a valid selection criterion, which we can use to filter data. Notice how the scissors icon is still “empty”; meaning our selection criterion has not been activated yet.

In order to activate this filter, simply drag this selection expression icon (Cut1) to the scissors icon. After this step, the tree viewer should look like Figure 15.

Now we are ready to make the plot we need. Simply click on the drawing icon (near bottom left, just left to the “STOP” icon). The canvas will now show the updated plot, which is shown in Figure 16. Notice that only the states of California and New York appear on the X-axis now, according to our selection. One can apply any arbitrarily complex filter to the data in this manner.

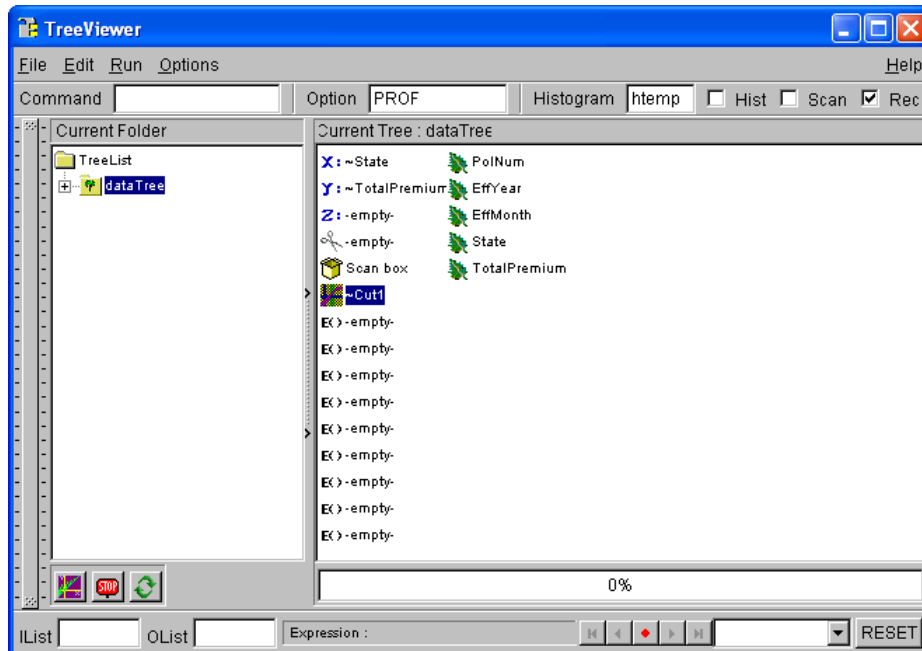


Figure 14: Appearance of the tree viewer after editing the first selection (cut) expression.

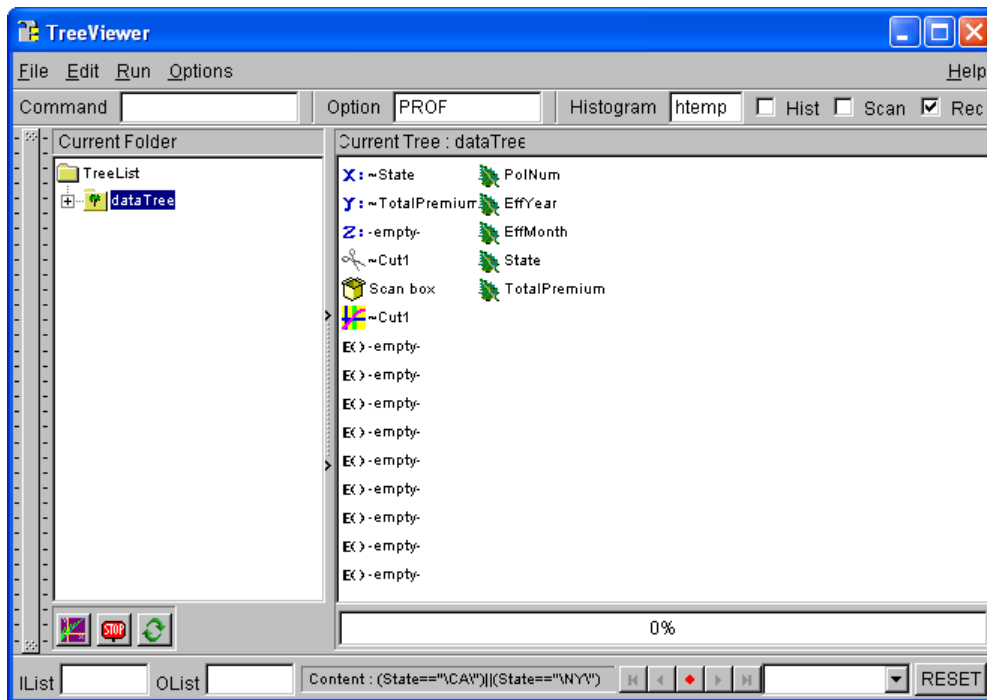


Figure 15: Appearance of the tree viewer after activating the selection criterion (Cut1).

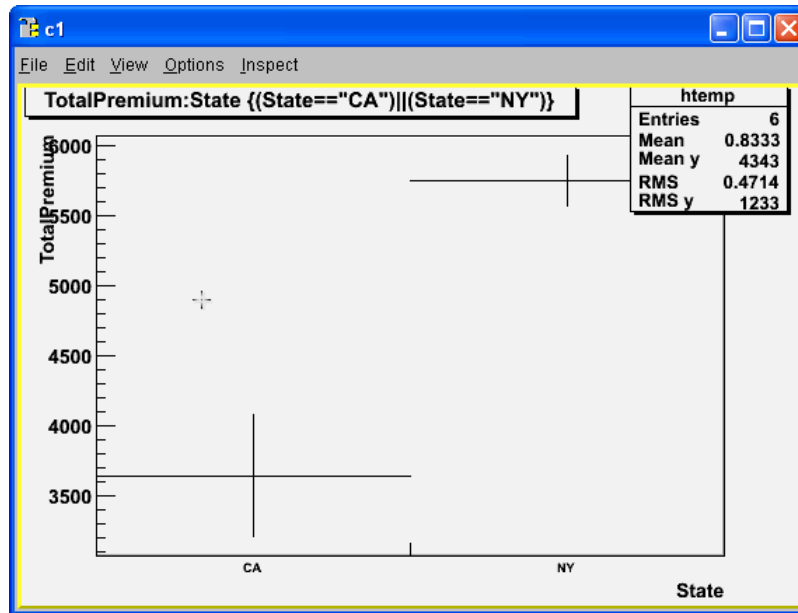


Figure 16: Plot of average total premium vs. State, after applying the state selection criterion.

Finally, all steps we have performed so far in the tree viewer can easily be saved as a ROOT macro (which is basically a C++ function, using ROOT classes), which can be run from the ROOT console to quickly reproduce these steps at a later time. The reader is referred to the ROOT user's guide to learn how to accomplish that.

## DATA SIMULATION

Complex simulations are a significant part of research in a typical high-energy physics experiment, and ROOT provides tools to facilitate this process.

For the purpose of this paper, we prepared some simulated “insurance-like” data. The code used for this simulation is provided in appendix B. For this study, we chose to simulate 5 Million BOP policies. The loss ratio for each of these policies was generated based on five predictive variables:

- Age of business: Number of years the business has been around.
- Building Count: Number of buildings insured under the policy.
- Credit score: The commercial credit score of the business.
- Policy Age: Number of years the policy has been in force.
- Total Building Insurance: The total amount of building insurance on the policy.

First, for each observation, each of these variables was generated independently. Age of Business, Building Count, and Total Building Insurance were generated using Landau distribution. Credit Score was simulated using a flat distribution (on a scale of 1-100), and

Policy Age was made to follow a linear distribution, subject to the condition that policy age must be less than or equal to the building age.

In addition to these predictive variables, we also generated three random variables, called ran1, ran2 and ran3. These variables are uniformly distributed between 0-10. These variables don't have any predictive power, by design. But we will include them in our list of predictive variables while searching for and ranking the most predictive variable. The tools being used to search for the predictive variable should be able to identify these random variables as non-predictive. This works as a sanity check of the tool.

The loss ratio from the five predictive variables was generated using the following equation:

$$\text{LossRatio} = 0.5 - 0.00053 * \text{AgeOfBusiness} + 0.0025 * \text{BuildingCount} - 0.00057 * \text{CreditScore} - 0.0227 * \text{PolicyAge} + 0.0437 * \text{TotalBuildingInsurance} \quad (1)$$

In addition, a random Gaussian noise is added to the LossRatio, with mean 0 and standard deviation of 0.04. Roughly speaking, this corresponds to smearing the true loss ratio by about 10%.

Figure 17 shows histograms of all the five predictive variables, and the loss ratio. We have chosen, on purpose, to simulate only the "lossy" policies.

We would like to point out one thing here. Since we simulated our data such that the relationship between the target variable (Loss Ratio) and the five predictive variables follows equation (1) above, we know exactly what results to expect from a correctly done regression analysis of this simulated data. The regression should give us back, within statistical uncertainties, the relationship defined in equation (1) above; otherwise something is wrong with the analysis.

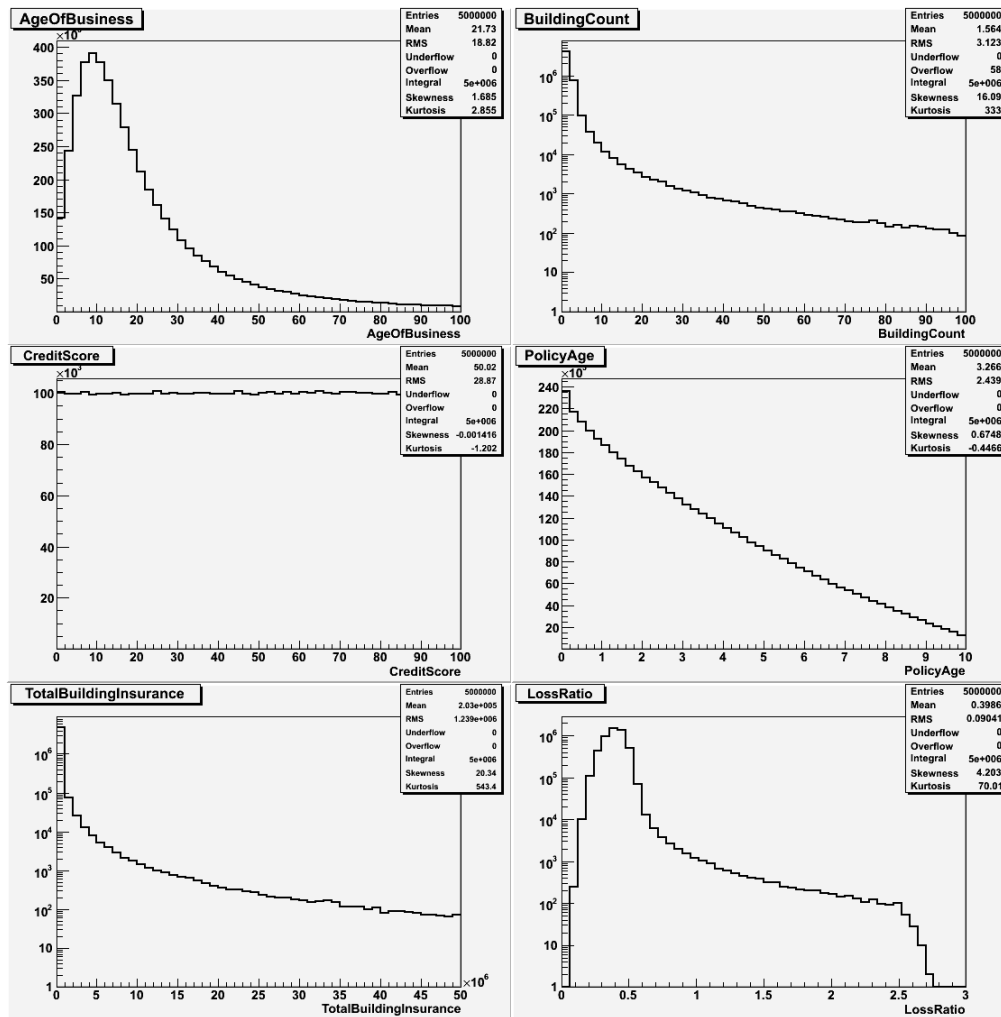


Figure 17: Raw histograms of the five simulated predictive variables and the loss ratio. The figure titles show which figure corresponds to which variable.

## EXPLORATION OF SIMULATED DATA

Figure 18 shows several profile histograms, plotting LossRatio vs. the five predictive variables. These profile histograms show the mean value of the loss ratio, for a given value of the X-axis. This is a useful way to visualize the relationship between the target and predictive variables. In our simulated data set, each of the five predictive variables is correlated, by design, with the target variable. Consequently, we should see a non-flat pattern in all these profile histograms, which we do.

It may be useful to point out here the time taken to generate the 11 figures shown in Figures 17 and 18. Each of these figures was generated, as mentioned earlier, from the

simulated data set of 5 Million policies. It took a total of 50 CPU seconds to generate these 11 figures.

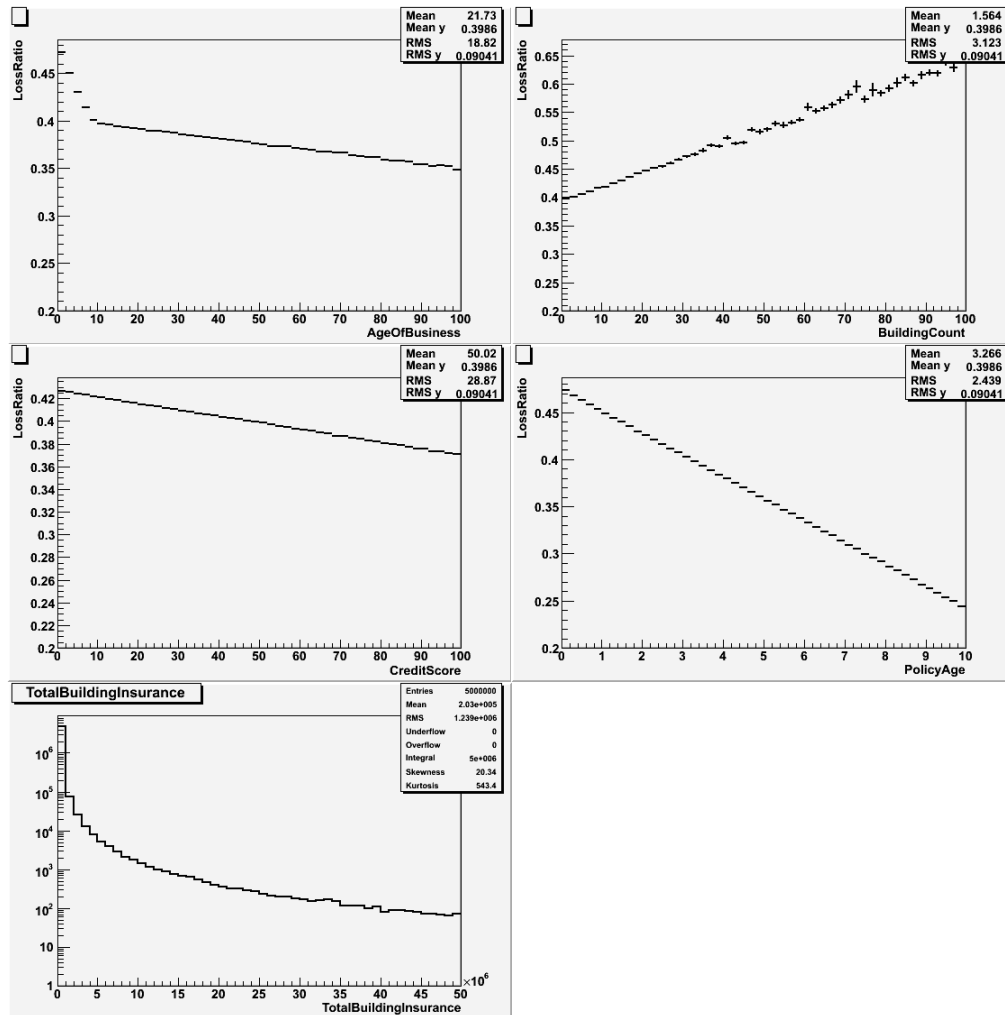


Figure 18: Profile histograms of LossRatio vs. various predictive variables. The Y-axis in each of these plots shows the average loss ratio for the corresponding value of the X-variable.

## SEARCHING FOR PREDICTIVE VARIABLES

The first step in building a predictive model is to identify which variables to use as predictors. The TMVA [18, 19] package provides tools to facilitate this process. TMVA is an acronym for “Toolkit for Multivariate Data Analysis.” It provides several classification algorithms, mainly designed to tackle the task of separating signal from background in complex high-energy physics data. However, it can be used in any environment where classification of data is needed. TMVA already comes bundled with ROOT.

Several of the algorithms in TMVA also provide the ranking of the predictive variables in terms of their discriminating power. These algorithms are: Likelihood, H-Matrix and

Fisher discriminants, Neural Network, Boosted Decision Trees, and RuleFit. The reader is referred to the TMVA user’s guide [18, 19] for more details of TMVA and the available classification algorithms.

For our simulated data, we will use Fisher Discriminants, as an example, to find out the ranking of the discriminating power of the various variables. First, we need to determine the classification we are interested in. A natural classification of interest for us is based on loss ratio—we would like to use as predictors the variables that can effectively separate policies with low loss ratio from those with high loss ratio. Once we have identified the variables that are most effective in providing this kind of discrimination, then we can use them, for example, in multiple regression, to come up with the regression equation that will predict loss ratio from these predictive variables.

For our analysis, we define (arbitrarily) policies with loss ratio less than 0.4 as “good policies” (signal), and those with loss ratio greater than 0.4 as “bad policies” (background). Our goal is to find the ranking of predictive variables, based on how well they can distinguish between these two kinds of policies. Borrowing from high-energy physics terminology, we will use the words signal (loss ratio  $< 0.4$ ) and background (loss ratio  $\geq 0.4$ ) from now on to represent these two classes of data.

We fed all the five predictive variables (shown in equation 1), and the three random variables (ran1, ran2, ran3) to the Fisher classifier. The random variables were included to test the effectiveness of the classifier—it should be able to identify these as the least predictive. The data sample used consisted of 100,000 observations of each type (signal and background, as defined above) for training and testing. So a total of 400,000 observations were used in this analysis.

Rank	Variable	Discriminating Power
1	PolicyAge	1.20E-01
2	CreditScore	1.51E-02
3	AgeOfBusiness	1.06E-02
4	TotalBuildingInsurance	4.56E-05
5	BuildingCount	2.24E-05
6	ran3	7.15E-06
7	ran2	8.95E-07
8	ran1	1.36E-07

Figure 19: The results of variable ranking using Fisher discriminant in TMVA.

Figure 19 shows the results of this analysis. We can see that the classifier has correctly identified ran1, ran2, and ran3 as least predictive. This is just a simple example of how

TMVA can be used to search for predictive variables. The meaning of rankings can be found in the TMVA user’s guide.

From now on, we will just keep the five predictive variables for fitting and discard the three random variables.

## FITTING

Once we have identified the predictive variables we want to use to make the prediction, we need to perform a fit to the data using these variables to come up with the optimal relationship between the target variable and the predictive variables. In our case, we are interested in obtaining the relationship between LossRatio and the five predictive variables shown in equation 1. Obviously, several techniques can be used to solve this problem, depending on the characteristics of the data at hand. ROOT provides several tools to achieve this, including least squares regression, method of maximum likelihood, neural networks, etc. In this section, we will show the results from using just a couple of these tools.

### 10.1 The TLinearFitter Class

We will use TLinearFitter [20] class in ROOT to fit LossRatio, with the five predictive variables mentioned above. We used the first 1 million observations from our simulated data set for the purpose of this fit. The total CPU time taken for this analysis was 11 seconds.

Figure 20 below shows the values of the parameters obtained by the fit, and also their true value. Since we simulated the relationship between the target variable, and the predictive variables, we know exactly what the correct parameter estimates should be. This allows us to do an *absolute end-to-end calibration/verification* of the analysis/regression chain.

Variable Name	Parameter Estimate from the Fit	True Value of the Parameter
<i>Intercept</i>	0.50	0.5
AgeOfBusiness	-5.31E-4	-5.3E-4
BuildingCount	2.47E-3	2.5E-3
CreditScore	-5.70E-4	-5.7E-4
PolicyAge	-2.27E2	-2.27E-2
TotalBuildingInsurance	4.37E-08	4.37E-08

Figure 20: Results of the fit to the simulated insurance data. Also shown are the true values of the parameters.

In Figure 20, if we compare the parameter estimates from the fit with their true values, we see that the two agree quite well. This comparison gives us confidence that our analysis chain used here, from data preparation to fitting, is correct.

In order to test the performance of TLinearFitter on a much bigger data set, we used it to perform a fit on a data set consisting of 49 variables and 9 Million (9E6) observations. This analysis took 16.3 CPU Minutes, and 20.5 Real Minutes.

## **10.2 Non-parametric Fitting: Neural Network**

In the previous section, we fit the simulated data using a linear relationship between the target variable (LossRatio) and the five predictive variables. This was appropriate, since the simulated data indeed follows such a linear relationship.

In several real-life situations, we don't know beforehand what the "true" functional form of the relationship between the target variable and the predictive variables is. In such situations, non-parametric fitting techniques might be more effective.

In this section, we will use ROOT's neural network package [21] to fit the data already simulated.

The network used in this analysis consisted of five input nodes (the five predictive variables used to simulate the data), two hidden layers with 8 and 5 nodes respectively, and a single output node (LossRatio). The network was trained on 2000 data points from the simulated data set, over 300 epochs. All the input and output nodes were normalized so that they take on values between 0 and 1.

Figure 21 shows the structure of the network, with the thickness of the connecting lines (synapses) being proportional to the corresponding weights.

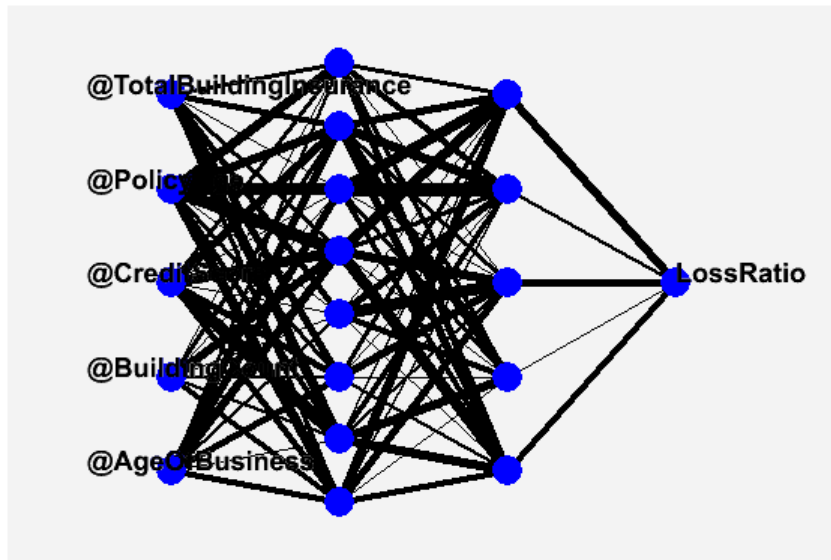


Figure 21: The structure of the neural network obtained by training it on the simulated data set. The thickness of lines is proportional to the weight for the corresponding connection.

Figure 22 shows a profile of the residuals for different values of the output (LossRatio) for the independent test data set.

The optimized neural network can be automatically exported to a C++ class, to be used elsewhere, e.g., for implementation.

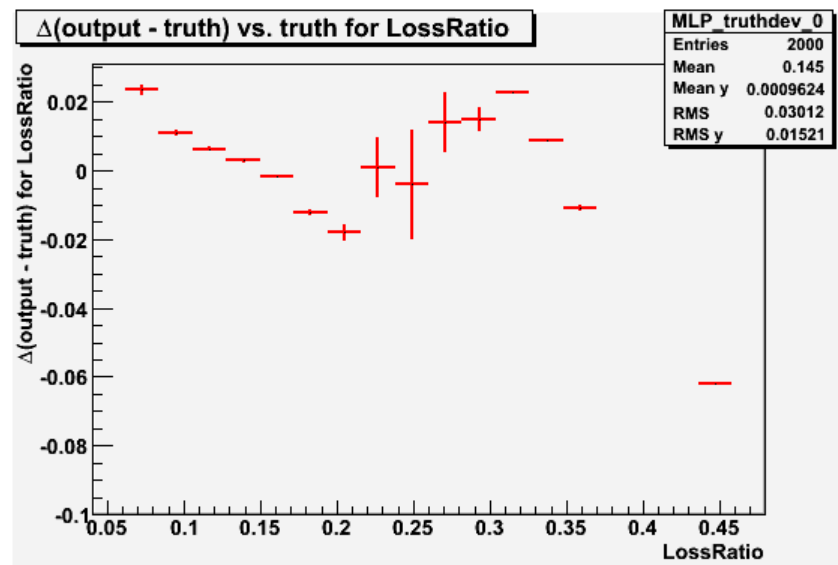


Figure 22: Residual profile for various values of the neural network (NN) output (LossRatio) on the test data set. The Y-axis is the mean value of the NN output minus the true LossRatio, for a given value of the true LossRatio. The true loss ratio has been normalized, so that it takes on values between 0 and 1.

### 10.3 Other statistical analysis packages in ROOT

The previous two examples use just two of the several statistical analysis classes available in ROOT. Here we list some other classes that might be of interest to anyone interested in statistical analysis of data in ROOT.

- The TMinuit [22] and TFumili [23] minimization packages can be used to do either least-squares or maximum likelihood fitting. Also see chapter 5 of the user's guide.
- The TPrincipal [24] class can be used for principal components analysis.
- The TMultiDimFit [25] is another non-parametric fitting package, which can be used to approximate the relationship between target and predictive variables in terms of simple basis functions.
- The TMath [26] class contains several useful mathematical functions. See chapter 13 of the user's guide.
- In addition to some of the algorithms mentioned earlier in this note, the TMVA package [18, 19] provides several other algorithms for data classification.
- The one-dimensional, two-dimensional, three-dimensional, and Profile histogram classes provide a convenient and interactive way to visualize, explore and fit data. See chapter 3 of the user's guide for more detail.
- The linear algebra package. See chapter 14 of the user's guide.

### 11. SOME OTHER USEFUL FEATURES OF ROOT

Previous sections demonstrated, using simple simulated data, some simple application of ROOT. In this section, we will list some more tools available in ROOT that might be relevant to us.

- Animation facilities, which can again be used to generate animations of some of the analysis chain, to be used in presentations.
- Efficient random number generators, with large periodicity =  $2^{19937}-1$ . Also, one can easily generate random numbers either following any analytical distribution, or following any empirical distribution provided by the data.
- A complete, self-contained GUI toolkit, including a GUI builder, which can be used to develop customized GUIs for specific tasks. See chapter 25 of the user's guide.
- ROOT comes with CINT, the C/C++-interpreter, which allows one to script an analysis in C++, and quickly execute it on the command line, without having to compile. This provides a quick way to prototype an analysis, which can later be compiled for better performance. See chapter 7 of the user's guide for more details.
- All the libraries available in standard C++, of course, can easily be used and integrated with a ROOT analysis.
- Interfaces to Ruby and Python scripting languages. See chapter 19 of the user's guide for details.
- Parallel processing. See chapter 24 of the user's guide for details.
- Networking. See chapter 22 of the user's guide.
- Automatic HTML documentation generation. See chapter 27 of the user's guide for details.

- Three-dimensional graphics package.

## 12. SUMMARY AND CONCLUSIONS

In this note, we have briefly introduced ROOT via some simple examples. Hopefully this will give the reader a feel of how to get started with using ROOT in insurance environment. We have also provided links below that a curious reader can follow to get a more detailed and advanced understanding of this tool. The user's guide, online tutorials, and the how-to's pages provide a wealth of information and several working examples that one can leverage to get started with any kind of analysis.

## 13. REFERENCES AND LINKS:

- [1] Rene Brun and Fons Rademakers, *ROOT – An Object Oriented Data Analysis Framework*, Proceedings of AIHENP'96 Workshop, Lausanne, Sep 1996, Nucl. Inst. & Meth. in Phys. Rev. A 389 (1997) 81-86.  
See also <http://root.cern.ch/>
- [2] <http://public.web.cern.ch/Public/Welcome.html>
- [3] <http://info.cern.ch/>
- [4] <http://public.web.cern.ch/Public/Content/Chapters/AboutCERN/Achievements/Achievements-en.html>
- [5] <http://atlasexperiment.org/>
- [6] [http://en.wikipedia.org/wiki/ATLAS\\_experiment](http://en.wikipedia.org/wiki/ATLAS_experiment)
- [7] <http://lhc.web.cern.ch/lhc/>
- [8] <http://root.cern.ch/root/ExApplications.html>
- [9] <http://root.cern.ch/root/License.html>
- [10] <http://root.cern.ch/twiki/bin/view/ROOT/Download>
- [11] <http://root.cern.ch/root/doc/RootDoc.html>
- [12] <http://root.cern.ch/root/Reference.html>
- [13] <http://root.cern.ch/root/Tutorials.html>
- [14] <http://root.cern.ch/root/HowTo.html>
- [15] <http://root.cern.ch/phpBB2/>
- [16] <http://root.cern.ch/root/html516/TTreeViewer.html>
- [17] <http://root.cern.ch/root/html516/TBrowser.html>
- [18] A. Hocker et al., *TMVA – Toolkit for Multivariate Data Analysis*, <http://arxiv.org/abs/physics/0703039>. Also available at <http://tmva.sourceforge.net/docu/TMVAUsersGuide.pdf>
- [19] <http://tmva.sourceforge.net/>
- [20] <http://root.cern.ch/root/html514/TLinearFitter.html>
- [21] <http://root.cern.ch/root/html514/TMLayerPerceptron.html>
- [22] <http://root.cern.ch/root/html516/TMinit.html>
- [23] <http://root.cern.ch/root/html516/TFumili.html>
- [24] <http://root.cern.ch/root/html516/TPrincipal.html>
- [25] <http://root.cern.ch/root/html514/TMultiDimFit.html>
- [26] <http://root.cern.ch/root/html516/TMath.html>

## Appendix A

### SAMPLE DATA LOAD PROGRAM

Below is a simple data load program, which reads a tab-delimited text file, and stores them in a ROOT tree. This tree is then written to an output ROOT file.

The contents of the file to be read are produced below. This is a simple, space-delimited file, which contains 10 records. Each record consists of a Policy Number, Policy Effective Year and Month, State, and the Total Premium collected. The purpose of this exercise is to give a simple example of how one can read data into ROOT. Note that any line starting with “#” is ignored by the ReadFile method, which we will use to read the sample data file below. Therefore, the first line in this sample data set is just for our convenience to tell us what the various fields are.

#PolNum	EffYear	EffMonth	State	TotalPremium
123456789	2006	10	CA	5000
123456790	2007	01	NY	6000
123456791	2005	12	KS	1500
123456792	2007	08	CA	3500
123456793	2007	05	AZ	2000
123456794	2006	11	CA	3500
123456795	2006	04	NY	5500
123456796	2007	02	AZ	1850
123456797	2006	12	CA	2560
123456798	2007	03	KS	1250

### The Program to Read the Data

This program reads in the above file, and stores its contents into a ROOT tree. This tree is then written to disk in a ROOT file. This ROOT file then can be use for data visualization, exploration, and analysis.

This program is a ROOT macro, which means it can be executed from the ROOT window. It will not run as a stand-alone program, since it does not contain the “main()” code block. In order to see how to generate stand-alone, compiled ROOT code, refer to the ROOT user’s guide.

Assuming this program is saved under the name “C:/Documents and Settings/temp/ReadSampleData.cpp” on disk, the user can run it by typing the following command from the ROOT window:

**.x C:/Documents and Settings/temp/ReadSampleData.cpp**

This command will execute the program, which will result in the input file being read, and saved as a ROOT tree on disk. The program is reproduced below.

```
// Reads a simple data file
#include <TFile.h>
#include <TTree.h>
#include <string>
#include <iostream>
using namespace std;
void ReadSampleData() {
    // Location of the directory for input and output files
    string dirName = "C:/Documents and Settings/temp/";
    // Name of the input and output files
    string inFileName = dirName + "SampleData.txt" ;
    string outFileName = dirName + "SampleData.root" ;
    // Create a tree to store the data
    TTree *tree = new TTree("dataTree","Sample Data");
    // Open the output file to write to
    TFile *fout = new TFile(outFileName.c_str(),"RECREATE");
    fout->cd();
    // Read in the data from the text file
    int nentries = tree->ReadFile(inFileName.c_str(),
"PolNum/I:EffYear/I:EffMonth/I:State/C:TotalPremium/D");
```

```
cout << "Read " << nentries << " entries from the input file " << inFileName << endl;

// Write the ROOT tree to output file

tree->Write();

fout->Close();

// Cleanup

delete tree;

delete fout;

}
```

## Appendix B

In this appendix, we reproduce the ROOT macro we used to generate the simulated data used in this paper. TNtuple class (<http://root.cern.ch/root/html516/TNtuple.html>) is used to store the simulated data. TNtuple class inherits from TTree class, and is useful when one is only storing numeric information. In order to run this macro, follow these steps:

1. Copy the following program (see below), and save it on your computer.
2. Change the value of variable Nloop (highlighted in bold face) to the number of observations you want to simulate.
3. Change the value of the variable DirName (highlighted in bold face) to the name of folder where you saved this program.
4. From ROOT console, type:  
**.L Program-File-Name** (where Program-File-Name is the name of the file, including complete folder path, in which you saved this program).
5. Next, still in the ROOT console, type:  
**SimulateData()**

This will generate the chosen number of simulated observations. The resulting ROOT file (SimulatedData.root) will be written in the folder specified by DirName.

Figure B1 below shows the screen shot of the ROOT console after executing the above steps. In this case, the name of the file containing the simulation code (reproduced below) was called SimulateData\_Simple.cpp; and it was located in the folder C:/Documents and Settings/temp.

```

ROOT 5.17.02
*****
*          W E L C O M E  t o  R O O T          *
*          *          *          *          *   *
*   Version   5.17/02   29 August 2007         *
*          *          *          *          *   *
*   You are welcome to visit our Web site     *
*          http://root.cern.ch                *
*          *          *          *          *   *
*****

Compiled on 30 August 2007 for win32 with thread support.

CINT/ROOT C/C++ Interpreter version 5.16.24, July 26, 2007
Type ? for help. Commands must be C++ statements.
Enclose multiple statements between { }.
root [0] .L C:/documents and settings/temp/SimulatedData_Simple.cxx
root [1] SimulatedData()
Generating Event 5000
Generating Event 10000
Total CPU Seconds Consumed = 22.5
Total Real Seconds Consumed = 29
root [2] _
    
```

Figure B1: Screen shot of the ROOT console after executing the macro to generate 10,000 simulated observations.

```

//////////////////// Beginning of the Data Simulation Code////////////////////
// This is a ROOT macro to generate simulated data for a simple model.
// This model relates the loss ratio to five independent variables.

#include "TRandom2.h"
#include "TStopwatch.h"
#include "TFile.h"
#include "TNTuple.h"

#include <string>

using namespace std;

// How many events do we want to generate
Int_t Nloop = 10 000 ;

// Directory where we will be writing to or reading from
string DirName = "C:/Documents and Settings/temp/";

// Generate a random number from a Landau distribution
Double_t GetRandomFromLandau(Double_t mean, Double_t sigma, Double_t
xmin, Double_t xmax) {
    Double_t x = -999.0;
    while (( x < xmin) || (x > xmax)) {
        x = gRandom->Landau(mean, sigma);
    }
    return x;
}
    
```

```

// Generate a random number from a Uniform distribution
Double_t GetRandomFromUniform(Double_t xmin, Double_t xmax) {
    Double_t x = -999.0;
    while ((x < xmin) || (x > xmax)) {
        x = xmin + (gRandom->Rndm() * (xmax - xmin));
    }
    return x;
}

// Generate a random number based on any formula
Double_t GetRandomFromFormula(const char* formula, Double_t xmin,
Double_t xmax) {
    TF1 *f1 = new TF1("f1",formula, xmin, xmax);
    Double_t xx = -999.0;
    while ((xx < xmin) || (xx > xmax)) {
        xx = f1->GetRandom();
    }

    delete f1;
    return xx;
}

void SimulateData() {

    // Instantiate and start a stop watch

    TStopwatch StopWatch;

    StopWatch.Reset();
    StopWatch.Start();

    // First, open a file to output the data set
    string OutFileName = DirName + "SimulatedData.root";
    TFile *fout = new TFile(OutFileName.c_str(),"RECREATE");

    if (!fout) {
        cout << " Error opening input file" << endl;
        return;
    }

    // Next, book an ntuple
    TNtuple *nt = new TNtuple("nt", "
", "AgeOfBusiness:BuildingCount:CreditScore:PolicyAge:TotalBuildingInsur
ance:PolicyNumber:LossRatio:Premium:Loss:EventNum:ran1:ran2:ran3");

    // Create an array to store the simulated numbers
    Float_t VarList[13];

    // Now start generating the variables. Loop desired number of times.
    int counter = 0;
    for (Int_t i = 0; i < Nloop; i++) {
        counter++;
    }
}

```

ROOT: A Data Analysis and Data Mining Tool from CERN

```
if (counter%5000 == 0) cout << "Generating Event " << counter << endl;

VarList[0] = (Float_t) GetRandomFromLandau(10.0, 5.0, 0, 100) ;
//AgeOfBusiness is based on Landau distribution
VarList[1] = (Float_t) GetRandomFromLandau(0.4, 0.1, 0, 100) ;
//BuildingCount is based on Landau distribution
VarList[2] = (Float_t) GetRandomFromUniform(0, 100);
//CreditScore is based on Uniform distribution
VarList[3] = (Float_t) GetRandomFromFormula("85.57-7.7948*x", 0, 10); //PolicyAge is based on a linear distribution

VarList[4] = (Float_t) GetRandomFromLandau(-3.3736e4, 5.4577e2 , 0, 50000000); //TotalBuildingInsurance is based on Landau distribution
VarList[5] = counter; //PolicyNumber
VarList[6] = (Float_t) 0.5 + VarList[0]*(-0.0053)*0.1 + VarList[1]*0.025*0.1 + VarList[2]*(-0.0057)*0.1 + VarList[3]*(-0.0227) + VarList[4]*(0.0437)*(1.0e-6) + gRandom->Gaus(0.0, 0.04); //LossRatio is a linear combination of other variables plus an error term
VarList[7] = (Float_t) gRandom->Landau(5000, 500); //Premium
VarList[8] = (Float_t) VarList[7] * VarList[6]; //Loss
VarList[9] = (Float_t) counter; //EventNum

VarList[10] = (Float_t) 10.0*(gRandom->Rndm()); //ran1 is just a random number
VarList[11] = (Float_t) 10.0*(gRandom->Rndm()); //ran2 is just another random number
VarList[12] = (Float_t) 10.0*(gRandom->Rndm()); //ran3 is just another random number

nt->Fill(VarList);
} // End of loop over Nloop

// Write the ntuple to output file
fout->cd();
nt->Write();
fout->Close();

// Stop the stop watch, and report the time taken to run this macro

StopWatch.Stop();

cout << " Total CPU Seconds Consumed = " << StopWatch.CpuTime() << endl;
cout << " Total Real Seconds Consumed = " << StopWatch.RealTime() << endl;

// Clean up
delete fout;
}
```

////////////////////////////////// END of the Data Simulation Code //////////////////////////////////////

# Staying Ahead of the Analytical Competitive Curve: Integrating the Broad Range Applications of Predictive Modeling in a Competitive Market Environment

Jun Yan, Ph.D., Mo Masud, and Cheng-sheng Peter Wu, FCAS, ASA, MAAA

---

## Abstract

In this paper, we describe a general process on how to integrate different types of predictive models within an organization to fully leverage the benefits of predictive modeling. The three major predictive modeling applications discussed in this paper are marketing, pricing, and underwriting models. These applications have been well applied and published over the past several years for the property and casualty (P&C) industry, but the literatures and discussions focused on their individual application. We believe that significant value can be realized if they are fully integrated, offering P&C companies the opportunity to take an enterprise wide view of managing their business through analytics. Therefore, the paper will discuss a general process on how they can be integrated and how the integrated result can assist insurance companies with managing the complex insurance business, such as minimizing the underwriting cycle and achieving profitable growth and reacting to external market forces faster than their competition.

---

## 1. INTRODUCTION

In recent years, predictive modeling has been widely used as a new strategic tool for P&C insurance companies to compete in the market place. Originally introduced in personal auto insurance to improve pricing precision [1], predictive modeling has been extended to homeowner's and small commercial lines as well [2]. Predictive modeling and the use of generalized linear models (GLM) have been individually applied widely in three key areas of insurance operations: underwriting, pricing, and marketing. In this paper, we will discuss the value in integrating results from three traditionally distinct predictive modeling applications and the additional strategic and tactical benefits companies can achieve by taking an enterprise-wide view of predictive analytics. Through the integration of predictive modeling results across multiple business operations, insurance companies can maximize their benefit and differentiate themselves in a competitive market environment where everyone seems to be using predictive modeling in some fashion. For instance, the integration of predictive modeling could enable existing underwriting and marketing predictive model results to drive enhancements to pricing models and to align pricing with the underwriting market cycle.

## 2. THREE TYPES OF P&C PREDICTIVE MODELING APPLICATIONS

In this section, we will discuss the similarities and differences as to how predictive models are built and applied to three different types of insurance business applications—underwriting, pricing, and marketing. We will also discuss the data and modeling issues associated with each application.

## **2.1 Pricing Models**

In predictive models for pricing, the main focus is on predicting loss cost, determining premium to charge, evaluating rating adequacy, or determining rating class plan factors. One typical result developed from a pricing model is a rating plan, which displays the rating variables, factors, and loss cost relativities across the rating variables.

In developing the rating plans, actuaries often use the standard GLM frequency and severity approach, where the Poisson distribution is used to fit frequency data and the Gamma distribution is used to fit severity data. Recently, it has become more popular to combine the frequency and severity models into a pure premium model, where the Tweedie distribution, a Poisson-Gamma compound distribution, is used to fit the pure premium data directly.

For pricing models, the source data files used to build the models need to be set up at a detailed exposure level. For example, for private personal auto (PPA), a pricing predictive model is generally set up at the vehicle and coverage level (i.e., lowest form modeling data level).

With regards to the rating variables, they are very different from one line of business to another, within the line of business, and can also differ from one coverage to another. Some complicated PPA rating plans may allow policy level variables across coverages and interaction between rating variables.

Perhaps, the most significant development for personal lines rating plans in recent years is the usage of personal financial credit score [3]. Some states allow the usage of credit scores in class plans or tiering, others allow credit scores for underwriting or target marketing activities only, while few states completely ban the use of credit scores. In addition to credit scores, other regulatory restrictions for pricing models include using not-at-fault accidents, capping the factors for youthful drivers or economic disadvantage territories, or enforcing forgiveness rules of prior years' loss and violation records, to name a few.

In the past several years, there has been a wealth of research, literature, seminars, and training classes in the Casualty Actuarial Society (CAS) community on using GLM to build pricing models [4, 5]. Therefore, we will not repeat these theoretical discussions for GLM pricing models. Instead, we would like to discuss, based on our past experience, several typical data and modeling issues that arise when building the pricing models.

- First, the commonly known data issues, such as missing data, miscoding information, information not captured in a insurance company data repositories, and unavailability of historical data due to purge, will hinder the development of predictive models.
- Compared to personal lines data, commercial lines data posts an even greater challenge during the development of pricing models:
  - Due to less regulation and scrutiny of commercial lines business operations, commercial lines data typically has much more commonly known data issues, as stated above, than personal lines data with regards to missing information, miscoding, and information availability.
  - For personal lines, the exposure is well-defined and fairly homogeneous: car-month for auto and home-year for homeowners. On the other hand, the exposure base for commercial lines is less defined and can even vary within the same line of business. For example, for general liability (GL), some classes use sales and revenue for exposure, while other classes use payroll for exposure. Given the complexity associated with exposure, applying the pure premium approach for pricing within commercial lines is fairly difficult.
  - For commercial lines, their data structure is heavily driven by rating bureau requirements. Therefore, the data is typically kept at the “industry class code” level, not at the exposure level. For example, for a commercial auto policy with multiple classes and multiple vehicles, the premium and loss information may be coded at the class level but not at the vehicle coverage level.
  - For commercial lines, more data credibility issues exist than they do with personal lines. Even for a mid-size regional personal carrier, it is fairly easy to collect millions of records for building up personal auto and homeowner’s models. However, for commercial lines, there poses significant challenges regarding the availability of unique data points and it is very common that the data size is at least 10 times less than what is available with personal lines.
- In general, some major pricing variables are excluded in a company’s analysis due to complex data structures, issues with data credibility, market competitiveness, or other business reasons. For example, “territory” and “vehicle symbol” are typically excluded from a modeling process of a PPA rating plan development. For these two variables, there exists many different values and therefore it is rare that a single company’s data can provide fully credible data to evaluate these two rating variables. Another example for commercial lines is that most of the business, such as commercial auto, GL, property, commercial multi-peril

(CMP), and workers compensation (WC), will follow the industry class loss cost by ISO or National Council on Compensation Insurance, Inc. (NCCI). There exist hundreds of industry classes for each line of business. One way to appropriately consider their impacts on the model results is to adjust the exposure or pure premium by their indicated relativity. Another way is to use the GLM offset options, and this approach is discussed in a separate paper [6].

- One data issue that needs to be considered for pricing model development is catastrophic (CAT) losses for property lines, such as fire or hurricane loss, and extreme large losses for liability lines. Therefore, it is prudent to exclude CAT losses or cap large losses and then build the long-term estimates for large loss loads or CAT loads back to the modeling data set.
- For property coverages, the losses are net of the deductible. For liability coverages, the losses are capped by the liability limit. Therefore, we do not have the “complete” loss information to establish the entire severity distribution curve. This is a challenge in building up the severity models.
- Another issue for building up the severity models is that for some of the segments in pricing, the severity data can be very thin and the modeling results can be extremely volatile with a great deal of “noise.” The issue is significant for low-frequency and high-severity coverages, such as BI for PPA, and GL. This is why the pure premium models based on a Tweedie distribution have attracted more and more interest in recent years.

## **2.2 Underwriting Models**

The major business objective of an underwriting model is to assess the risk quality for an insured on a prospective basis. One difference between underwriting models and pricing models is that pricing models focus on determining the final class rates, while underwriting models focus on evaluating risk quality beyond the class rating and the currently charged rate. The underwriting models can assist underwriters or product managers with their underwriting decision making, such as company placement, crediting or debiting, limitation of coverage, payment plan selection, new business acceptance or rejection, renewal business referral and cancellation, and customer service and marketing activities. Regarding the modeling design, one difference is that pricing models use the pure premium approach at the exposure and coverage level, while underwriting models use the loss ratio approach at a policy level.

Ideally, if a perfect rating plan exists, all risks are priced at their adequate rate level and there is no need for underwriting models or even underwriting because, generally speaking, underwriting

models sit on top of pricing models and are designed to address pricing inadequacy through improved underwriting precision. However, ideal rating plans do not exist due to various internal and external restrictions, including regulatory constraint, dynamic changes in the external economic environment, long delays for filing approvals, inability of using certain variables in rating plans, and limitation on rating structure (e.g non-linear pattern, interaction between rating variables, interaction between exposures at a policy level, etc.). Therefore, underwriting models are used to evaluate the risk quality by identifying potential deficiencies in the rating plan.

The information used by underwriters can vary widely and is sometimes highly subjective. Also, underwriting actions are not always truly risk-based, but instead are influenced by the market, subjective decision making and external competition. This issue of a “market-driven” price is a more prevailing concern for commercial lines than for personal lines. Therefore, predictive modeling can be used to build up objective underwriting models to assist underwriters with making consistent and fact-based underwriting actions each and every time and ensuring alignment with external market cycles.

Another advantage of underwriting models is that the models can help insurance companies improve their underwriting efficiency. This is because the models can segment “good risks” versus “poor risks,” and with such segmentation, underwriters can spend their major time and effort on poor risks, while good risks can flow through the process with minimum underwriting touch. In addition, underwriting models can be used to segment good and bad risks within classes of business, which is a significant improvement over traditional pricing and underwriting decisions that are made on a class basis.

In general, the target variable of an underwriting predictive model is the loss and allocated loss adjustment expense ratio. Since underwriting is mostly performed on a policy basis, the predictive variables and the data files used for developing an underwriting model are at the policy level. For predictive variables, there are many more candidate variables: rating versus non-rating, internal versus external, credit and territorial, among others. There is less restriction for underwriting models than pricing models. For example, there is a trend in the industry with using insured’s premium payment records from historical billing data, such as late payments and bad checks, as underwriting variables. The trend of using billing information makes logical sense, since an insured’s premium billing records are essentially a proxy for personal financial credit data and an insured’s ability to pay bills on time.

For underwriting models, the potential data and modeling issues are as follows:

- Several data issues stated before for pricing model development are equally applicable to underwriting model development, such as data quality and data availability and data completeness issues.
- Many candidate variables can be included in underwriting models that generally cannot be included in pricing models. Creating and selecting the candidate variables demands a look at the availability of the underlying information, internal or external, to insurance companies and the ease of implementing these variables and gaining underwriting acceptance on their use. Here are several examples:
  - While there is a trend with using billing information for underwriting models, some companies may purge their billing data on a frequent basis; therefore, such information is not available in the historical data. Over a long term, companies need to devise a master data quality initiative to maintain and update historical data in their corporate data repositories to support these underwriting models and devise mechanisms to ensure that these data elements are available to be extracted. The role of data quality and data governance as a key strategy to successfully maintaining and gaining value from predictive modeling applications is taking on even greater significance in the P&C industry as more companies seek new ways to differentiate themselves in today's market.
  - Another example is that some underwriting information is kept on paper instead of in electronic files or in back-end data repositories. For example, for new business underwriting, while many insurance companies ask for prior loss experience or other external data, such as motor vehicle records (MVR) for commercial auto, rarely do they store this information in their back-end data repositories. Therefore, it is difficult to use such information during the development of underwriting models, even though it is common for underwriters to use prior loss information in underwriting new business.
- When loss ratio is used as the target variable for modeling, we need to apply due actuarial consideration to adjust the data, such as rate on-leveling, loss development, and trending. By applying the appropriate actuarial adjustments, the underwriters can have a higher level of confidence so that when they use the underwriting model, the indicated results on the quality of the risk as derived from the model are based on up-to-date information with the appropriate longitudinal adjustments made.
- Since underwriting models are constructed at the policy level, whether the results can be carried, or how the results can be carried, to the underlying pricing, is a difficult question. For

example, driver age is commonly used as an underwriting factor even though it is used for pricing already. If an underwriting model indicates that youthful driver policies are worse than average, it may not suggest that the underlying youthful pricing factors are wrong, but rather it may indicate the inadequacy of the pricing structure, such as purely multiplicative structure, or potential interaction of youthful drivers with other variables, such as vehicle type. The answer can be difficult to find without in-depth research and analysis.

- Sometimes, underwriting is not only performed on a policy level, but also on an account level. For example, it is very common for personal line carriers to cross-sell auto and homeowner's policies, and for commercial line carriers to cross-sell all the major small commercial lines of business, including BOP, commercial package, auto, and WC. Therefore, the full value of underwriting models may not be realized until they are built for all lines of business, for account-driven companies and underwriting models take a holistic view of assessing the quality of a risk.

## **2.3 Marketing Models**

The earliest, classical business application for predictive modeling is for marketing and sales operations, such as mail solicitation and response models. In general, the purposes of marketing and sales predictive models include identifying prospective customers, increasing the hit rate for solicitation, and assisting with retaining existing customers [7]. This is not for the P&C industry alone but historically predictive modeling has been used for marketing and consumer business related applications across multiple industries.

In general, the main focus of these marketing models is on the "success or failure" of converting or retaining a risk, so the target variable is typically a binary one. Whether the risk is profitable or not is not a consideration for these models but rather the probability that the risk will be acquired as a new policy or retained as a renewal policy.

Depending on the final usage of the marketing and sales models for insurance, there is wide variation in the types of models with regards to the predictive variables and the design of the target variable. For insurance applications, the marketing and sales models can be grouped into four main categories: new business qualification and targeting, new business conversion, renewal business retention, and renewal business conversion models. The details for these four types of models are as follows:

- For new business qualification and target models, the purpose is to identify a list of potential prospects for targeting. This list can be used for phone or mail solicitation campaigns. The data and variables used for the models are fairly limited, and are mostly from data sources

external to insurance companies. There are numerous data vendors who sell consumer databases, and insurance companies can use the data for these models. Since there is a cost associated with the solicitation campaign, such as phone call cost or mailing postage fee, it is important to measure the cost versus return benefit, that is, the response rate, after the models are implemented.

- For new business conversion models, the key is to increase the new business hit rate when an insurance company has an opportunity to offer a quote to an insured. Insurance companies are very interested in knowing the overall hit (or conversion) rate, how the hit rate varies by different segments of the book (for new business), and how to increase the hit rate. Many insurance companies do capture certain information in their insurance quote files, such as name, address, number of quotes, quoted prices, etc. For the conversion models, we can expect that one critical, if not the most important, factor that will influence the hit rate is how competitive the company's quoted price is compared to its competitors. The relationship between the hit rate and the quote price can be expressed through the "elasticity curve" commonly used for classical economic supply-demand theory. Without such price elasticity information, the value of new business conversion models will be significantly limited.
- For renewal business retention models, the main purpose is to understand the probability of an existing insured to stay for the next renewal term [8]. The reason that an existing insured does not stay for the next renewal term may be due to the insured's action, such as mid-term cancellation, non-response to renewal request, or non-payment of premium, or the insurer's action, such as non-renewal. Therefore, the renewal retention models will focus on understanding how an insured's characteristics correlate with the retention rate.
- For renewal conversion models, the model will measure the probability of the policy to be converted to the next term at the point of renewal for the existing policy. Therefore, these models exclude the mid-term cancelled policies. Similar to the new business conversion models, the renewal price offered and how it compares to the competitors will play an important role on the outcome.

Obviously, for renewal models, much more information, especially information from the company's internal data sources, can be used. For new business models, the predictive variables are very limited, and sometimes the models may completely rely on external data sources. In the end, these marketing models may not be as accurate as underwriting and pricing models but they do offer an opportunity to improve resource allocation and efficiency in the sales process by allowing

insurance companies to focus their marketing and sales efforts on the risks that are most likely to be bound or retained.

In the remaining sections of the paper, we will focus on the new business conversion models because they are the most challenging ones to build, and they are very critical for insurance companies to sustain long-term profitable growth. For the new business conversion models, predictive modeling techniques can be employed to find certain segments with a higher likelihood for responding to the quote (i.e., the response rate) and purchasing after taking quotes (i.e., the hit or conversion rate), as well as segments with a higher or lower sensitivity with respect to the price. Similar to underwriting models, the marketing models are often created on a policy level and sometimes even on a household or account level.

As mentioned earlier, in order to analyze the response rate and hit rate, it is important to capture the price competitiveness for the quote, that is, the price differentiation between the company and its competitors. The competitors' pricing information can be obtained in published rating manuals, company's quote files, or industry competitive information vendors' database. If the competitors' prices are well captured in the quote files, the core information of the price elasticity curve can then be established for the models.

The typical data issues for building up marketing and sales models are:

- Since quote files are not required for financial reporting or bureau reporting, the quality of the files are much worse than other files and data sources. In addition, insurance companies often purge their quote files after one or two years, therefore little historical quote data is available for analysis. Once again this highlights the importance of corporate data quality and governance as a key strategy to maximize predictive modeling benefits
- Typically, there is very limited information captured in the quote files and often only includes the following:
  - Name and address of an insured
  - Basic and key rating information
  - Agent information
  - Competitiveness information including prior carrier's name and price

Insurance companies rarely capture information other than the above and therefore the number of variables that can be derived is very limited.

- For the renewal retention process, insurance companies rarely follow up their non-renewal risks and find out the reasons for their non-renewal decision, the new company they took

their business to, or the new price that they received from their new company. Without such competitive information, the value of the marketing and sales models will be significantly limited. It also minimizes a company's opportunity to gain market intelligence and assess its own competitive position because there is valuable business insight that can be gained from understanding why a company's customers are leaving.

### **3. INTEGRATING THE THREE MODELS IN A COMPETITIVE MARKET ENVIRONMENT**

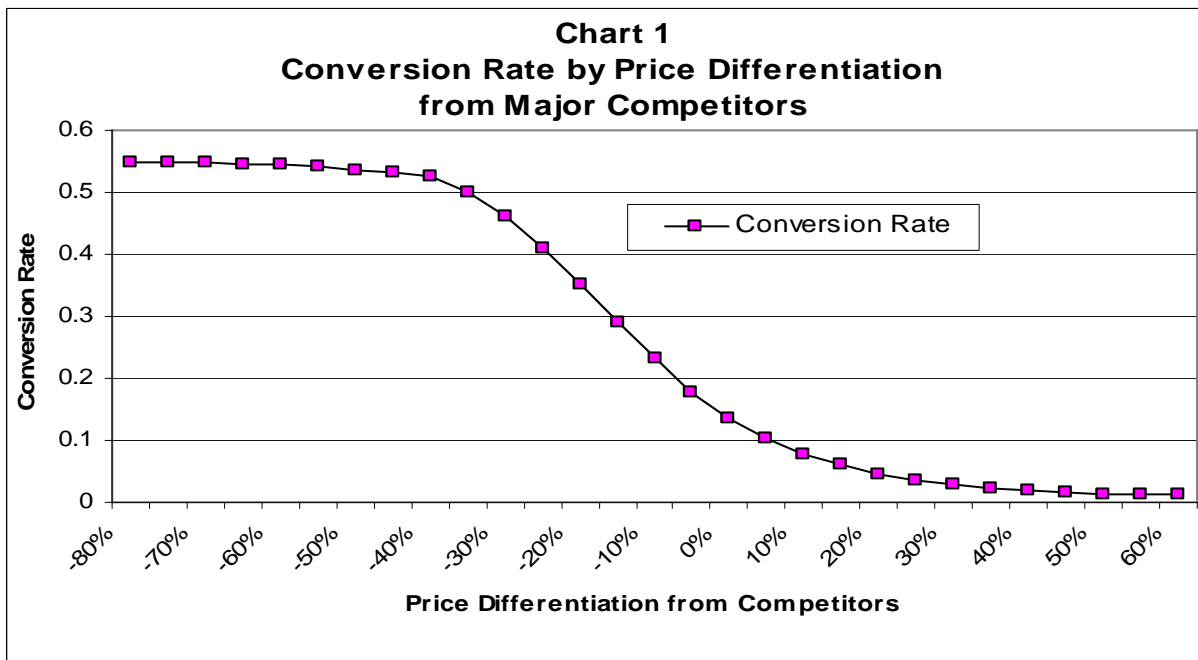
The U.S. insurance market is a highly regulated industry. There are more regulatory constraints for personal lines than for commercial lines. There is also a practical limitation for pricing due to long rate filings and the overall approval process. Therefore, underwriting and marketing models can be more flexible in assisting insurance companies in dealing with the dynamic external environment that they operate in. Another advantage, as mentioned earlier, is that with underwriting models in place, the subjective judgment by underwriters can be largely eliminated and the computer-generated model results can be consistently documented in the underwriting files for regulatory review.

The U.S. insurance market is very complex, dynamic, and competitive. One significant challenge for insurance companies is how to effectively manage their business through the ups and downs of an underwriting cycle. For example, one typical approach when the market is turning soft (i.e., increasing profit and declining price) is to reduce rates or increase the credits offered to insureds across the board in order to maintain market share. However, a blanket approach of reducing rates or increasing credits assumes that the market competitiveness, rate adequacy, and sensitivity of retention to price are the same across different segments of the market. In reality, we know that such assumptions mostly likely are not valid. Insurance companies should study and adjust the pricing as well as underwrite based on how price elasticity, pricing, and underwriting interact with each other [6]. Therefore, integrating the three predictive modeling solutions can assist insurance companies in dealing with the dynamic market conditions effectively.

The following is an approach to how the three models can be integrated:

- The first step of the integration process is to develop an "adequate" rating plan using the standard GLM approach. The GLM rating plan would assume that the rate is adequate with regards to the rating variables and the structure of the rating plan.
- The second step after the completion of the GLM rating plan is to develop a new business conversion model by studying the sensitivity of how insurance buyers react to price difference, such as the price elasticity curve in Chart 1. In Chart 1, the graph is based on the quote file described before, and is used to link the price level and the conversion rate. In the

chart, we can see that the overall conversion rate is between 0 and 50%. This means that no matter how low the company's price is, the maximum conversion rate is 50%, and if the company's price is too high, the chance to get new business is 0. Also, the conversion rate will change in the S-shape region when the company's price is between 30% below and 10% above the competitors' price. It is in this S-shape region that the conversion rate is most sensitive to the price change. The graph can be generated across the whole book, or it can be further broken down by different segments of the book, such as by age group, household profiles, territory, etc. With such elasticity information at hand, the company will know not only the trade-offs between the price change and conversion rate, but also where it will get the most benefit in new business growth from the price change.



- The third step, now that we have identified the key range for the rate adjustment-conversion rate relationship, is to use the results to adjust the GLM rating plan so that the parameters can be re-optimized with different adjustments. This step can be tedious and involves an iterative process but the benefits can be significant. At this step, the company's historical data is employed in the pricing model development. At the same time the marketing information is used along with the pricing information to improve the overall performance for the company's operation by striking a balance between profitability and growth.
- The last step is to build the underwriting models on top of the pricing and marketing models. There are several reasons that the underwriting model is important to use along with the pricing and marketing models.

First, the GLM pricing model may still be far from addressing the overall rate adequacy because many significant variables are not used in the pricing models. Such information may include agent's performance data, credit score, and demographic and territorial information on a more refined level. A great deal of non-rating information can be used to enhance the segmentation of an insured's profitability.

Our experience indicates that for commercial lines, such underwriting models are very important, since most of the commercial line carriers follow the bureau loss cost and rate structures for most of the major lines of business. They do not have their own GLM-based pricing models.

The second reason is that, as the result of adjustments for the rating plan are due to conversion consideration, it is likely that some segments can turn unprofitable because of the trade off for growth and retention. The decrease in profitability can be minimized with additional underwriting information. For example, if it is determined that youthful policyholder factors need to be tempered to increase the conversion rate, the potential profitability impact can be minimized through the application of the underwriting models by allowing profitable agents to write more youthful risks than unprofitable agents write (i.e., offsetting the risk of youthful risks by focusing on youthful risks with favorable credit scores).

Finally, it is very important to note that, when developing the underwriting models, the underlying premium should be based on the final pricing structures and rating factors. All historical premium data should be adjusted to the final selected pricing level.

When these three applications are integrated, modelers should be conscientious about the data and modeling issues and problems described in previous sections for each application. In addition, there exist unique, challenging data and modeling issues during the integration process.

- The first unique challenge is due to the fact that the data levels are different between the pricing model and the underwriting and marketing models. Therefore, how to “accurately” profile the policies identified by the conversion model and link the model results to the subsequent pricing model is a challenge. For example, a youthful policy may have all of or part of its drivers as youthful drivers. When the marketing model profiles youthful driver policies to be targeted or not targeted, it needs to be very specific in defining whether the profile is partial (if partial, the percentage of youthful drivers on the policy) or all youthful driver policies. In other words, how to “roll up” exposure-based pricing information from the pricing model to the policy level information for the underwriting and marketing models needs to be prudently considered.
- Another challenge for integration is that the marketing application is “forward-looking” while the pricing and underwriting applications are based on “historical” information. Due

to constant changes associated with the internal and external environments for insurance operation, the historical data distribution and composition may not serve well for the forward-looking integration application. For example, if a national insurance company would like to expand its business in certain a geographic region, such as in the northeast, it is possible that the northeast risks behave differently than the risks in other regions. Therefore, modelers need to make extra efforts as to how to prepare the data for the integration analysis, and, for this example, may want to use data in the northeast region only. Other considerations include the distribution change in industry class, affinity programs, or premium size.

- As discussed in the previous sections, different applications may have different data available. In general, data is more sparsely available for the marketing application than for the underwriting or pricing applications. For example, driver and vehicle details are fairly populated in the pricing and underwriting data sources, but not for the marketing data sources. When the details are available in the marketing data sources, it is possible that they are more available for certain regions, branch offices, agents, or programs than for others. The inconsistency in data availability may lead to “bias” in the analysis results.
- By combining a comprehensive underwriting model with a pricing model, a company can more accurately estimate loss cost and profitability than by using the pricing model alone. Previously, we illustrate how to use the pricing model and the marketing model together first, and then develop an underwriting model second. In theory, there is no limitation for the sequence of integration, and the underwriting model can be used alone with the pricing model to fine-tune the marketing model. Of course, the challenge for this approach is that the underwriting model is on the policy level, while the pricing model is on the exposure level.

#### **4. SUMMARY**

Several years ago, merely using predictive models in some fashion to support underwriting, pricing, and marketing gave insurance companies a competitive edge. However, in today’s competitive market, predictive modeling is not limited to just personal lines but is used widely in commercial lines as well. Therefore the first mover advantage no longer exists and insurance companies must find new ways to maximize the benefits of their predictive modeling investment and stay ahead of their competition.

Our paper illustrates the strategic and tactical approach of taking an enterprise-wide view of predictive modeling and integrating the results from pricing, underwriting, and marketing models to

support business decisions across multiple business operations. In today's market, companies that will succeed are the ones that incorporate analytics as a core business strategy and align multiple business operations with a single unified view of analytics.

From a tactical perspective, our approach to integrating pricing, underwriting, and marketing predictive models is a four step integration process as outlined below:

- Step 1: Develop the GLM-based rating plan and pricing model.
- Step 2: Develop retention or conversion models to study the price elasticity behavior of insurance buyers.
- Step 3: Adjust the rating plan and class plan factors based on the retention and conversion models to strike a balance between rate adequacy and conversion rate.
- Step 4: Build up a series of underwriting rules based on underwriting models in conjunction with the pricing and market models to maintain the overall competitiveness.

By integrating the three types of predictive models seamlessly, insurance companies can gain two major benefits. First, instead of adjusting their rates across the board for growth, insurance companies can "target" the segments to gain a high return on growth with minimum price changes. Second, the potential profitability issue associated with rate cutting for growth can be minimized with underwriting models. We believe that with such integration, the full value of predictive modeling can be realized. It can provide insurance companies with an effective way to deal with the key business challenges of achieving profitable growth and minimizing the impact of the underwriting cycle. History tells us that companies that are successful and regarded as market leaders are the ones that can process information and make sound business decisions faster than their competition can. The P&C insurance industry should be no different and an integrated approach to predictive modeling gives P&C companies an opportunity to realize the full value of their predictive modeling investment and stay a step ahead of the competition.

## 5. REFERENCES

- [1] 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, pp. 457-543, (1992).
- [2] “The Top 10 Casualty Actuarial Stories of 2006,” *Actuarial Review*, Vol. 34, No. 1, Casualty Actuarial Society (2007).
- [3] Wu, C. P., Guszczka, J., “Does Credit Score Really Explain Insurance Losses?—Multivariate Analysis from a Data Mining Point of View,” *2003 CAS Winter Forum*, Casualty Actuarial Society (2003).
- [4] Mildenhall, S. J., “A Systematic Relationship Between Minimum Bias and Generalized Linear Models,” *Proceedings of Casualty Actuarial Society*, Vol. LXXXVI, Casualty Actuarial Society (1999).
- [5] Feldblum, S. and Brosius, J. E., “The Minimum Bias Procedure—A Practitioner’s Guide,” *Proceedings of Casualty Actuarial Society*, Vol. XC, Casualty Actuarial Society (2003).
- [6] Jun, Y., Flynn, M., Wu, C. P., Guszczka, J., “Offset Techniques for Property and Casualty Insurance Predictive Modeling,” Submitted to 2009 CAS Ratemaking Seminar Call Paper Program.
- [7] Duncan, A., “Modeling Policyholder Retention,” Presented at the CAS Seminar on Predictive Modeling, Casualty Actuarial Society (2006).
- [8] Moore, B. D., “Direct Marketing of Insurance Integration of Marketing, Pricing and Underwriting,” *1998 CAS Discussion Paper Program*, Casualty Actuarial Society (1998).

# Actuarial I.Q. (Information Quality)

CAS Data Management Educational Materials Working Party

---

## Abstract:

**Motivation.** Provide an introduction to data quality and data management directed at actuaries.

**Method.** Expand on the concepts in Actuarial Standard of Practice No. 23 (Data Quality), then introduce practical methods that actuaries, actuarial analysts, and management can apply to improve their situation, with references for more information.

**Results.** Information quality is about more than coding data: processes affect quality. There are many principles and practices an actuarial department can employ immediately to improve the quality of the information it deals with. Actuaries have a unique role to play in the bigger arena of improving their organizations' information for decision making and it is in their interests to do so.

**Conclusions.** What every actuary should know about data quality and data management.

**Availability.** Code for creating Box Plots in Excel is a link with this paper at <http://www.casact.org/pubs/forum/08wforum/>.

**Keywords.** Actuarial Systems; Data Administration, Warehousing and Design; Data Quality; Data Visualization; Exploratory Data Analysis; Software Testing.

---

## 1. INTRODUCTION

Data quality is a significant concern for most actuaries. In Britain, a GIRO Data Quality working party survey [1] found that about 25% of actuaries' time is expended on data quality issues. The survey also found that about 30% of actuarial analyses are adversely affected by data quality problems. Poor data quality is sometimes viewed as an inescapable fact of life by actuaries and other insurance industry analysts. However, actuaries, as both key consumers and providers of information, are uniquely well-positioned to deal with the pervasiveness of poor data quality in insurance.

Some think data quality is merely the accuracy of data. This paper identifies and discusses other characteristics (such as completeness and timeliness) and then broadens the perspective to information quality, which considers the broader picture of how information is processed and communicated. This includes not only data accuracy but other pitfalls that can result in users

misunderstanding information. Strategically, data quality is more important today, given easy access to an unprecedented level of detail and the proliferation of new tools and analysis techniques. Consequently, actuaries can add value by broadening how they think about data:

1. Data is a corporate asset that needs to be managed and actuaries have a role to play.
2. Data needs to be appropriate for all of its intended uses, not just the analysis at hand.

This paper contains tools, concepts, and references to support and facilitate this expanded perspective in order to help actuaries transform data into more useful information to make better decisions.

## **1.1 Research Context**

The actuarial literature on data quality is relatively sparse. In North America, the Actuarial Standards Board (ASB) Actuarial Standard of Practice No. 23 on Data Quality (ASOP No. 23) [2] provides guidelines to actuaries when selecting data, relying on data supplied by others, reviewing and using data, and making disclosures about data quality.

The Casualty Actuarial Society (CAS) Committee on Management Data and Information and the Insurance Data Management Association (IDMA) produced a white paper on data quality [3]. This CAS committee also promotes periodic calls for papers on data management and data quality which are published in the *CAS Forum*. The CAS online database (DARE) taxonomy can help users narrow their searches to papers on specific topics such as actuarial systems, data organization, and exploratory data analysis.

In response to one such call for papers, Francis [4] provided guidance for specific techniques which can be used to screen data for quality errors. Francis pointed out that 80% or more of time spent on large modeling projects is spent on data issues. However, the focus of the paper was on detecting errors after the fact, and not on techniques for preventing them.

The subject of data quality is also of interest internationally. A working party of the UK General Insurance Research Organization (GIRO) developed recommendations for improving the quality of reserve estimates. The Reserving (GRIT) working party report [5] recommended more focus on data quality and suggested that UK professional guidance notes incorporate standards from ASOP No. 23. Furthermore, the GRIT survey found that many respondents expressed concern over data quality.

In researching this paper, the working party reviewed seven books recommended by the IDMA,

## *Actuarial IQ*

as well as two more recommended by a working party member. Many books talk about data management as a means to achieve data quality, and some deal specifically with data quality. However, these books tend to be written for information technology professionals to apply to any organization. Since our goal is an introduction for actuaries, these texts are only cited occasionally. The collection of reviews of these books was published in the Winter 2007 CAS *Forum* [6].

### **1.2 Objective**

ASOP No. 23 sets standards for data quality that address a number of key areas but there are times when an actuary might want to go further. For example, if a reasonableness check reveals some data shortcomings, ASOP No. 23 outlines the ramifications for the analysis at hand. However, the actuary may be in a position to prevent data quality issues in source databases from arising by advocating improvements in data management and data quality practices. This paper discusses some of the practices and options available.

Other papers published by the CAS tend to focus on particular data management subjects: there is no broad introduction to the subject. Conversely, it is difficult for actuaries to apply nonactuarial texts on data management and data quality since these texts often presume the reader has a working knowledge of related IT concepts and unrestricted access to an organization's data centers.

This paper is a data quality introduction and reference for actuaries and actuarial analysts. As such it attempts to bridge the gap between ASOP No. 23 and the literature available for people in the actuarial profession who want or need more information. It is also the authors' hope that actuaries and actuarial analysts will become advocates for information quality once they see the business value information quality provides in:

- More accurate analyses (and hence smaller margins of error),
- Ability to focus on higher value activities once significant data issues are resolved,
- Increased impact of their analyses by increasing transparency and legibility of results.

### **1.3 Disclaimer**

While this paper is the product of a CAS working party, its findings do not represent the official view of the Casualty Actuarial Society or the employers of the authors. Nor is anything in this paper intended as a standard of practice nor an interpretation or guidance of existing standards. Moreover, while we believe the approaches we describe provide sound guidance on how to address the issue of information quality, we do not claim they are the only acceptable ones. Similarly, we believe the

textbooks and papers cited here are good sources of educational material on data management and data quality issues, but we do not claim they are the only appropriate ones. Finally, we have illustrated various concepts and methods with examples. The particular software used to illustrate examples is not necessarily the only or the best software for the purpose.

## 1.4 Outline

Section 2 will discuss concepts, whereas section 3 will focus on techniques. In brief, section 2 discusses the motivation for data quality and describes characteristics of quality data. It then expands the scope to a discussion of metadata and a common example of metadata in property and casualty insurance: statistical plans. Section 3 begins with techniques for improving the quality of data (exploratory data analysis and data audits) then turns to information quality in processing (models and presentations). It concludes with a discussion of the organizational and management issues: data quality measurement (as a tool to track improvement), improvement strategies, and data management. Section 4 reiterates the main topics of the paper.

## 2. BACKGROUND AND THEORY

Quality issues have come to forefront recently due to several key developments:

- **(Unprecedented) level of detail.** Computerization and cheap data storage along with changes in regulatory requirements have led to extraordinary amounts of data being captured, stored, and provided to actuaries. Consequently, enormous amounts of data can amass enormous numbers of errors and inconsistencies.
- **Availability of new tools.** Recent years have seen the proliferation of powerful data analysis packages and technologies: from XML-enhanced data exchange to object-oriented databases to servers enabled with On Line Analytical Processing.
- **Competition.** Competition encourages pricing techniques to be more and more precise – witness the growth of predictive modeling. In this environment, requirements for quality of data used in pricing algorithms grow immeasurably.
- **The growing data management skill set of actuaries.** Modern actuaries are more technically prepared for the challenges of dealing with huge amounts of data using contemporary tools and techniques. They should be able to tackle data quality issues with aplomb.

## *Actuarial IQ*

In their work, actuaries rely on vast amounts of data: claims loss runs, premium bordereaux, interest rates, and industry statistics, just to name a few. All of these data originate outside of actuarial reach and their collection and accumulation generally occur without actuarial control. Before it reaches actuaries, every piece of data passes through several stages: data are collected by a TPA, MGA, or some other source; then they get transferred to the insurance company system; and, after that, they can be grouped, accumulated, and mapped to a suitable structure. At each of these stages, data are processed and modified by people of different professions and qualifications who inevitably introduce errors into the data. The longer the data pipeline, the more errors accumulate and can compound one another. Multiple data sources also tend to multiply data problems.

As data progresses from input to information to decisions, the actuary's role changes from consumer to provider. This position is almost unique in the insurance data life cycle: indeed, as information providers for decision makers, actuaries are held to the highest standards of work quality; but, as consumers, actuaries depend on someone else. Better than any other professionals in the insurance industry, actuaries can become data quality protectors: they have knowledge of the data content, expertise to develop sophisticated data testing tools, and high stakes in the quality of the data.

Whereas ASOP No. 23 focuses on data's suitability for a particular actuarial analysis, we will present a broader introduction to data and information quality. A schematic overview of the development and usage of insurance data with respect to actuarial work is provided by [Figure 2.0.1](#). The schematic outlines the data life cycle for insurance. The goal of each major step and the function within the organization most responsible are given. These are followed by some examples of the types of errors that can be introduced in each step. The "Topics" column identifies the sections in this paper most pertinent to each step. As such, the figure is a helpful roadmap identifying where to find more information in this paper and providing the general context. Note that metadata (section [2.3](#)), data quality measurement (section [3.5](#)), data quality improvement strategies (section [3.6](#)), and data management (section [3.7](#)) considerations permeate the entire process. The multiple references to actuaries illustrates actuaries' broader opportunity to improve information quality not just for the analysis at hand, but for better decision making in the organization as a whole.

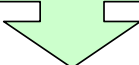
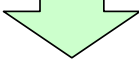



Steps	Purpose	Responsibility	Examples of Errors	Topics
<p>Step 0 <b>Data Requirements</b></p> 	Determine the intended use of data and required data elements	Data managers and actuaries at the source and the destination	Specification errors, granularity mismatches	Data quality (2.1) and its principles (2.2)
<p>Step 1 <b>Data Collection</b></p> 	Collect data and satisfy statistical reporting	Data managers at primary sources: TPAs, MGAs, insurers, statistical agents	Input errors	Statistical plans (2.4), data audits (3.2)
<p>Step 2 <b>Transformations Aggregations</b></p> 	Make data available to users in the necessary format and level of detail	Data managers at the source and the destination	Missing values, duplicate records, mapping errors	EDA (3.1), data audits (3.2)
<p>Step 3 <b>Analysis</b></p> 	Extract useful information from raw data	Actuaries (perhaps with the help of others)	Wrong model choice, censorship, over-fitting, calculation errors	IQ in Models (3.3)
<p>Step 4 <b>Presentation of Results</b></p> 	Help management make right decisions	Actuaries	Inconsistencies, mislabeling, inadequate labeling	Data Presentation (Reports) Quality (3.4)
<p>Final Step <b>Decisions</b></p>	Make profit, customer care, public welfare	Management	Interpretation errors, wrong conclusions	Not addressed in this paper

Fig. 2.0.1

## **2.1 What is Data Quality?**

Generally speaking, something is of high quality if it is particularly appropriate for its purpose. According to ASOP No. 23, “for purposes of data quality, data are appropriate if they are suitable for the intended purpose of an analysis and relevant to the system or process being analyzed” ([2], page 2). ASOP No. 23 advises the actuary to obtain a definition of data elements in the data, to identify questionable values and to compare data to the data used in a prior analysis. The actuary is also advised to judge whether the data is adequate for the analysis, requires enhancement or correction, requires subjective adjustment, or is so inadequate that the analysis cannot be performed. In making this judgment, ASOP No. 23 lists six things actuaries should consider when selecting data (discussed in section [2.2](#) below). ASOP No. 23 is often considered only with respect to the analysis at hand. However, if the analysis is repeated periodically or the same data is used for multiple purposes, it may be advantageous to address some of the recurring data quality issues.

A key component of this bigger picture is the concept of metadata. “Metadata” is simply information about data. As such, it helps determine if particular data are suitable for a particular purpose and insures that it is used appropriately. Metadata can help identify invalid entries, facilitates transferring data among systems, can improve the interpretation of analyses, and can prevent blunders due to misinterpretation of data. It is described more fully in section [2.3](#).

The key idea is that quality data is appropriate for its intended purpose. Note that this makes quality a relative, not absolute, concept: data may be of adequate quality for one analysis while being inappropriate for another purpose. For example, data that is appropriate for an annual overall rate adequacy study may not be appropriate for a relativity analysis or even for a midyear overall rate indication. This is particularly an issue in predictive modeling, where the analyst attempts to find better predictors (of losses, for example): promising variables may not have been coded or processed with the intent of using them for this purpose.

### **2.1.1 Data quality versus information quality**

Everyone has heard the well known IT adage “garbage in – garbage out”: it says that poor quality inputs will lead to poor quality outputs. Put another way, it says that processing or analysis cannot completely correct bad input. This consideration of processing distinguishes information quality from data quality. Dasu and Johnson [7] talk about “end-to-end-data-quality.” That is, there are many stages in the data assembly process where data quality needs to be monitored and improved, such as during data collection, transformation and aggregation, data storage, and data analysis. Their equation:

DATA + ANALYSIS = RESULTS

highlights that quality results depend not only on quality data, but also on quality analysis. The quality of the final product is not only affected by the quality of the data itself, but also by how the data is processed (e.g., how it is transformed, aggregated, analyzed and presented).

This consideration of processing leads to a larger concept of metadata: the initial definition of metadata could be restricted to a particular database, but it can also be expanded to integrate information across applications, as new data is created with each application. Metadata is discussed more fully in section [2.3](#).

Information quality does not have a commonly accepted definition. It is used in this paper to remind readers that data quality is about more than just correct coding: quality is affected by how data is stored, processed, and analyzed, and how results are presented.

From a data manager's perspective, it also includes what facts are captured as data and how they are captured.

## 2.2 Principles of Data Quality

When evaluating the quality of a dataset for a particular analysis, ASOP No. 23 advises actuaries to “select the data with due consideration of the following”:

- **Appropriateness** for the intended purpose of the analysis, including whether the data are sufficiently current;
- **Reasonableness** and **comprehensiveness** of the necessary data elements, with particular attention to internal and external consistency;
- Any known, material **limitations** of the data;
- The cost and feasibility of obtaining **alternative data**, including the availability to obtain the information in a reasonable time frame;
- The benefit to be gained from an **alternative data set** or data source as balanced against its availability and the time and cost to collect and compile it; and
- **Sampling methods**, if used to collect the data. ([2], page 3)

Similarly, the CAS Management Data and Information Committee “White Paper on Data Quality” [3] states that evaluating the quality of data consists of examining the data for:

### *Actuarial IQ*

- **Validity:** “the value of a given data element is one of all allowable ones” ([3], page 155)
- **Accuracy:** “each data transaction record or code is a true and accurate representation of what it’s intended to represent” ([3], page 156)
- **Reasonableness:** “is the data reasonable compared to our prior and current knowledge?” ([3], page 157), and
- **Completeness:** each record contains all the data necessary for business needs and every step in data collection and processing handles it correctly, without duplication.

The white paper goes on to note that there are three levels of accuracy for usable data:

- **Absolute:** data is 100% correct for every data element and every transaction,
- **Effective:** there are some errors but they should have no material impact on the results of the analysis,
- **Relative:** data is “inaccurate but consistent over time” ([3], page 158).

#### **2.2.1 Validity versus accuracy**

One misconception is that if data is valid, then it is accurate. To see why this is not true, consider, for example, the ZIP Code. The recorded ZIP Code may be one of the possible ZIP Codes in the state (valid) but it may not be the correct one associated with the particular risk’s address. Standalone edits in policy administration systems can check the validity of the data while more complex relationship edits and audits can be used to check for accuracy.

#### **2.2.2 Data quality through data management**

Now that data quality is defined, how is it achieved? Section 3 describes a number of options actuarial analysts can pursue to improve their information quality, but the most holistic way is by good data management. This is because good data management broadens the point of view from the data for the analysis at hand to the entire process that gave rise to the data as well as other potential applications and users of the data. There are some additional data quality principles from this broader perspective.

Various authors of data quality literature describe the dimensions of data quality. A comprehensive list is provided in *Data Management: Databases and Organization* [8], by Richard T. Watson. Watson defines eighteen dimensions of data quality. Some of these dimensions are the key principles described above. Others describe ways of storing data such as:

*Actuarial IQ*

Dimension	Conditions for high quality data
Representational consistency	Values for a particular data attribute have the same representation across all tables (e.g., dates)
Organizational consistency	There is one organization-wide table for each data element or entity and one organization-wide data domain for each data attribute
Record consistency	The values in a record are internally consistent (e.g., a home phone number's area code is consistent with a city's location)
Flexibility	The content and format of presentations can be readily altered to meet changing circumstances
Precision	Data values can be conveniently formatted to the required degree of accuracy (e.g., in cents or in thousands)
Granularity	Data are represented at the lowest level necessary to support all uses (e.g., hourly sales)

*Table 2.2.1*

Notice how these dimensions support the key principles of validity, accuracy, reasonableness, and completeness.

Watson's list goes beyond data characteristics to processing and management principles, such as:

Dimension	Conditions for high quality data
Stewardship	Responsibility has been assigned for managing data
Sharing	Data sharing is widespread across organizational units
Timeliness	A value's recentness matches the needs of the most time critical application requiring it. Values remain up to date.
Interpretation	Clients correctly interpret the meaning of data elements

*Table 2.2.2*

Other key concerns for data managers are the proprietary nature of data and the privacy issues. An insurer's data contains much information about its business: who it insures, the premium it charges, the claim it has paid. Many insurers consider this information to be a trade secret. As such, data managers and the users of the data (e.g., actuaries) must be careful to protect the data of their employer or client from being divulged to their competitors. Likewise, insurance data may contain

## *Actuarial IQ*

data elements about an individual person, such as their social security number, FICO scores, and health records that from an ethical and legal perspective should remain confidential.

Data management is discussed more fully in section [3.7](#) below. The next section, metadata, is the key to the interpretation dimension.

### **2.3 Metadata**

#### **2.3.1 What is metadata?**

Metadata is a term used by data management and data quality professionals to denote the data that describes the data, e.g., the documentation of the contents of a database. In addition to information about the data itself, metadata contains information about business rules and data processing. Examples of metadata in insurance are the ISO and NCCI statistical plans.

Good metadata serves as a roadmap to the business processes of the entire organization and as such needs to be shared with the entire organization. As a result, actuaries should take an active role in understanding and developing metadata. The actuary's role in metadata will be discussed in section [2.3.2](#) and the sharing of metadata across an organization will be discussed more in section [2.3.3](#).

At a minimum, metadata will include a listing of all data elements in a database, along with a description of what is contained in each data element. Each data element listed should be defined clearly, and the data that is in the data element described. For example, the data element "pol\_eff\_date" may be defined to contain the policy effective date and should contain only date values. Furthermore, the date format may be specified, such as mm/dd/yyyy. The permissible ranges of the values (e.g., 1/1/2000 to present) should be specified. Any default values (e.g., 1/1/2000) should be documented. Similarly, metadata should define the values in categorical data.

Metadata should also identify when and how a data element is processed. As an example, Table [2.3.1](#) shows seven values in the data for the marital status data element, including a value for the case when marital status is missing. If multiple sources of data are used to populate a database, then the source of the data should be listed. Any transformations done to the data need to be documented as well. The documentation should also describe how frequently the data is updated from the sources.

*Actuarial IQ*

<b>Marital Status Value</b>	<b>Description</b>
1	Married, data from source 1, straight move of field ms_code
2	Single, data from source 1, straight move of field ms_code
4	Divorced, data from source 1, straight move of field ms_code
D	Divorced, data from source 2, straight move of mstatus
M	Married, data from source 2, straight move of mstatus
S	Single, data from source 2, straight move of mstatus
Blank	Marital status is missing

*Table 2.3.1*

Metadata can also exist on the compilation or extraction processes. It should include information on such items as fiscal period definitions and how evaluation dates are determined.

Ideally, metadata should also include business rules, such as how reported claims are defined. It should also document interdependencies with other data elements. For example, the date of birth for a driver should be at least 15 years earlier than the date the driver received their license.

The inclusion of documentation on the quality of data can enhance the metadata. For example, to really understand the data, a general narrative on the quality checks and controls of the data is necessary. Other useful metadata include a data quality matrix for each data element. This would describe the quality checks done on the data element, how frequently the checks are done, and where in the process the check occurs.

Better process documentation can also enhance metadata. For example, a high-level data process flow diagram that shows each initial data feed (source) and any data stores (databases) associated with the data will give users and developers better insight into the processing. Another example of enhanced process documentation is a glossary of terms that provides definitions specific to the data and systems under consideration.

Finally, some sort of versioning is helpful to identify when changes take place. For example, when did the claims department change the average reserves? When did rating territory begin being derived from zip code instead of input? When did a new product or alternative distribution channel go live?

## *Actuarial IQ*

A complete description of the contents of a database is important for the appropriate use of the data. Good metadata will assist the analyst in avoiding misunderstandings that result in revisions of the analysis when the contents of a data element or variable are discovered to be other than what it was assumed to be. As a result, metadata is an important tool for actuaries to use when planning their analyses. Problems can arise for actuaries when metadata is either nonexistent or is inaccessible to actuaries. Metadata that is incomplete, inaccurate, or out of date can also lead to problems.

Creating quality metadata at an organizational level is a large undertaking and really requires commitment from all levels of the organization. The next section talks about the actuary's role in metadata and some suggestions that can be used in any organization to get started and perhaps build the necessary commitment.

### **2.3.2 The actuary's role in creating and sharing metadata**

Maintenance of adequate documentation describing data can help avoid problems associated with relying exclusively on people's memories of what is contained in the data. As actuaries, we can help persuade our business and data management partners that system documentation is vital to the actuarial work product.

At the same time, we can employ the same standards of metadata and documentation to the actuarial work product. After all, actuarial work is a source of data and information for others in the insurance industry, so it follows that the same principles of metadata should be applied. Metadata from actuarial projects can be shared with appropriate data management and system colleagues to ensure that the data is being properly used. Sharing of metadata within the user community (actuarial, data management, finance, etc.) is a vital activity for the organization. To quote the *Corporate Information Factory* ([9], page 170): "Metadata is the glue that holds the architecture together. Through metadata, one component of the architecture is able to interpret and make sense of what another component is trying to communicate."

Documenting anything from a basic actuarial project to a complex information system can be a daunting task. The following sets of considerations can be used to help test existing metadata or get started on putting together new metadata.

Minimum considerations:

- Are all the data elements listed?

### *Actuarial IQ*

- Has the source of each data element been provided?
- Is there a special value that is used to indicate missing data?
- Are any transformations being applied to data? (Note: data cleanup such as filling in missing values should be considered a data transformation.)

More advanced considerations:

- Have the contents and use of each data element been properly described?
- Have all the categorical values of each data element been properly described?
- In the case of numeric data, has the range of possible values for each data element been provided?
- Has the valuation date of all data been provided?
- Has a schedule of planned updates to the data been provided?
- Has the business process changed during the experience period?
- Have any of the data definitions changed during the experience period?

As was noted above, a good place to start is with our own actuarial work product. In many instances, we may produce or maintain databases that underlie our analyses. How well documented are these systems? How well understood are the sources that feed the actuarial systems? Once the actuarial systems are understood, one can start to drill back into the source systems. Along the way, missing metadata can be identified. The benefits and costs of producing the metadata can be weighed and ownership could be assigned.

As metadata is developed, it needs to be shared across the organization. That is the topic of the next section.

#### **2.3.3 Sharing metadata across an organization**

Actuaries can also face the problem of access to metadata (or at least to the most up-to-date metadata). Just like data, metadata can exist in multiple forms, such as word processing documents, printed documents, spreadsheets, and databases. It can also be stored in multiple locations, including file servers, paper files and within the documented system itself. Keeping track of and sharing all that metadata can be difficult.

Technology can provide answers to these types of collaboration issues. It is worthwhile for

## *Actuarial IQ*

actuaries to be plugged into the collaboration technologies that are available within their organizations. Examples include intranets, quick places, hyperlinks, comment boxes, and the emerging wiki technologies.

The *Corporate Information Factory* [9] addresses this issue by introducing the concept of autonomous versus shared metadata. The key issue is that “metadata has a need to be shared, and a propensity to be managed and used in an autonomous manner. Unfortunately, these propensities are in direct conflict with each other” ([9], page 170). Consequently, each component of a system, such as a table or database, should have its own metadata and metadata should be split into autonomous and shared groups. Autonomous metadata is only used (or applicable) within the component. “Sharable metadata must be able to be replicated from one architectural component to another” ([9], page 174). Splitting metadata into these groups need to be mutually exclusive and exhaustive. The rule of thumb is that “very commonly used metadata needs to be shared” ([9], page 175).

At the end of the day, access to metadata is as simple (and as difficult) as building and maintaining good relationships between the actuarial and data management communities.

### **2.4 Statistical Plans**

Some of the most widespread examples of metadata are the statistical plans used for the collection of property-casualty insurance statistical data. Regulators in the various jurisdictions are charged with ensuring that rates meet statutory standards – that rates are not inadequate, excessive, or unfairly discriminatory. One of the tools the regulators use to fulfill this function is the collection of data by line of insurance by statistical agents that aggregate the data and report it to regulators. A statistical agent is an organization that helps insurers satisfy legal requirements for reporting data to regulators. The statistical agent processes data submitted by insurers, performs data quality checks on the data, consolidates the data across insurers, and provides aggregate data compilations to state insurance departments on the behalf of the insurers. The well-known statistical agents in the United States are:

- The four that collect data for the major property/casualty lines of insurance, except workers compensation and health. These include the American Association of Insurance Services ("AAIS"), the ISO Data, Inc.<sup>TM</sup> (a wholly owned subsidiary of Insurance Services Office, Inc. or ISO), the Independent Statistical Services ("ISS"), and National Independent Statistical Services ("NISS").
- For workers compensation, the dominant statistical agent is the National Council on

### *Actuarial IQ*

Compensation Insurance (“NCCI”). In some US jurisdictions, workers compensation data is collected by an independent state bureau such as the New York Compensation Insurance Rating Board (“NYCIRB”).

Other statistical agents exist in the United States for more specialized lines of business such crop-hail (National Crop Insurance Services) and surety (Surety & Fidelity Association of America) insurance. In addition there are some state-specific/line-of-insurance-specific agencies that collect industry data. An example of this is the Texas Insurance Checking Office (“TICO”) which collects data for private passenger automobile, residential property, and farm and ranch insurance in Texas under Texas Department of Insurance (“TDI”) statistical plans.

Among these statistical agents, numerous statistical plans have been developed in each of the US jurisdictions. Statistical plans also exist outside of America. In general, the statistical plans are organized around one or more lines of insurance. For example, the ISO has three statistical plans<sup>1</sup> – the personal auto statistical plan (“PASP”), the personal lines statistical plan (other than auto) (“PLSP(OTA)”) and the commercial statistical plan (“CSP”). Each of these plans then has subparts or modules devoted to a particular line of insurance. For workers compensation, the underwriting experience (premiums and losses) is collected through the unit statistical plan (“USP”). Additional unique data collection requirements exist for workers compensation. For a more complete discussion of workers compensation see the study note “NCCI Data Collection Calls and Statistical Plans” by Richard Moncher [10].

In general, the statistical plans contain information or metadata – general reporting requirements and specific, detailed definitions for each data element – that describe the information to be collected. In the sections below, these items are explained further, followed by an example of these instructions and definitions excerpted from the homeowners module of the ISO personal lines statistical plan (other than auto).

#### **2.4.1 Reporting instructions**

Reporting instructions describe the overall scope of the plan such as:

- To which jurisdictions the plan applies,
- To which lines of business the plan applies,

---

<sup>1</sup> ISO also has separate plans for those companies with very limited market share in a line of insurance.

### *Actuarial IQ*

- Instructions on specific situations such as mid-term endorsements to policies and cancellations.

#### **1. GENERAL REPORTING REQUIREMENTS**

##### **A. Premiums**

Premiums must be reported separately for each policy and each unique set of codes in the Coding Section of this module.

When a policy insures more than one dwelling, each dwelling must be reported separately.

When Water Back-Up Damage coverage is attached to a policy, this coverage must be reported separately.

When a policy includes additional coverage which requires coding under a separate module of this Plan, the premium and amount of insurance reported under this module must **not** be increased.

*Table 2.4.1*

#### **2.4.2 Data element definitions**

Each element to be collected on the premium and loss records needs to be defined. In some cases the same data elements are collected on both the premium and loss records. These definitions cover multiple dimensions, including:

- A text description of the element to be collected,
- Field length or field position on the record,

##### **2. Transaction Type Code (Field: Position 5)**

Report the appropriate Transaction Type Code.

- Valid codes or attributes for the data element,

## Actuarial IO

### TRANSACTION TYPE CODES

DESCRIPTION	CODE
PREMIUM TRANSACTIONS	
Fully Coded (including "Exception" coded)	1
Limited Coded	2
LOSS TRANSACTIONS	
Paid Losses	6
Outstanding Losses	7
Paid Allocated Loss Adjustment Expense	8+
Outstanding Allocated Loss Adjustment Expense	9+
Salvage (Paid Losses)	4
Subrogation (Paid Losses)	5
+ Applicable to Section II (Liability) Losses only.	

- Record layouts that show the exact position and field length on the statistical plan record,
- Examples of coding and interpretations of the coding,
- Due dates for reporting to the statistical agent,
- Quality requirements.

Quality requirements for the submission would address the error tolerances that may be allowed.

For more information on statistical plans in the United States, the reader should refer to "Statistical Plans for Property/Casualty Insurers," by Virginia R. Prevosto [11], published in the 1997 Casualty Actuarial Society Discussion Paper Program and the study notes by Richard Moncher and Virginia R. Prevosto on the NCCI [10] and the ISO [12] statistical plans, respectively.

### 3. TECHNIQUES AND APPLICATIONS

Section 2 introduced key concepts of information quality. In this section, we present procedures and processes designed to improve information quality.

#### 3.1 Exploratory Data Analysis

A common approach to detecting data quality problems in a dataset is to perform a preliminary screening of the data elements. These data elements are treated as variables for the purpose of statistical analysis. Exploratory data analysis ("EDA") is a family of techniques that use graphs and descriptive statistics to explore the structure of a dataset and to identify outliers. (Data errors are often found by detecting outliers and then investigating the outliers for validity.) These techniques were pioneered and the practice given its name by John Tukey (see "exploratory data analysis" at [www.wikipedia.org](http://www.wikipedia.org)). These techniques are widely accepted in the statistical community as a key

activity within any statistical project, and they are widely implemented in statistical software.

Data quality problems can take several forms including:

- **Missing data and null values**, which impair the analyst's ability to use the affected variables and may render some variables useless for analyses,
- **Data errors** such as a paid amount of \$1,000 coded as \$1,000,000 or the state NY coded as NJ,
- **Default values** may be coded rather than actual values (e.g., for convenience), and
- **Duplicate transactions**: it is not uncommon for duplicates of the same claim, same transaction, etc. to be in a database.

Being mindful of the sources of data errors, one can detect, remediate, and most importantly, prevent them. Dasu and Johnson [7], whose book on data quality and data cleaning is considered a key reference by data mining professionals, detail many mishaps affecting data that create quality problems. Some of the sources of data quality problems are: unreported changes in layout, unreported changes in measurement, and temporary reversions to defaults, missing values, inappropriate default values, and gaps in time series.

The following subsections introduce several EDA techniques to deal with data quality issues in a given dataset. For more information, see Francis [4] or Dasu and Johnson [7].

### 3.1.1 Data cubes

A data cube is a one-way or multiway summarization of key statistics for the variable(s). Cross-tabulations and pivot tables are examples of data cubes. For instance cross-tabulations or two-way tabulations of the frequencies for two variables are widely used in statistics and most statistical software such as SAS, S-PLUS, SPSS, and Access have the capability of quickly producing cross-tabulations.

For example, one can tabulate the frequency of records in the data containing each value of a categorical variable. Table 3.1.1 displays the frequencies of injuries for each of the 6 injury codes in a Massachusetts Private Passenger Auto database.<sup>1</sup> The table was created using Microsoft Excel's pivot table capability. Note that there are two codes where only a small number of records contain

---

<sup>1</sup> The data was supplied by the Automobile Insurers Bureau of Massachusetts and is from a database used to do fraud research.

### Actuarial IQ

the code. The results from pivot table summaries need to be compared to a document defining which codes are valid values for the data element. These tabulations can be performed over multiple dimensions at once, although it is most common to perform one dimensional (variable by variable) frequency analysis.

Massachusetts Auto PIP	
Injury Type Code	Count of Injury Type Code
1	793
2	197
3	2
4	250
5	151
6	7
<b>Grand Total</b>	<b>1400</b>

Table 3.1.1

#### 3.1.2 Identifying missing data

As noted by Francis [4], missing data is the rule rather than the exception in large insurance databases. Missing data complicates an analysis by reducing the number of data records with completely valid information. At a minimum, the uncertainty about parameter estimates will be increased, even when measures can be taken to adjust the data elements containing missing values. It is not uncommon for the majority of data records to be missing data on variables that are presumably in the database and available to the analyst. If a sufficient percentage of records on a given variable are missing values, that variable may have to be discarded from the analysis. In some extreme circumstances, the missing data problem may be so severe that an analysis cannot be undertaken. Tabulations of missing values should be compiled for each variable in the database.

Analysts must also be alert to missing values they create by their data manipulations. For instance, division by zero will create a missing or not available value that can affect further analyses if not detected. Most statistical software produces a log which records the history of calculations completed and their results. Cody [13] recommends reviewing the logs of the statistical software for statements that missing values are being created as a result of transformations performed.

### Actuarial IQ

Data cubes can be used in the detection of missing values and in screening categorical variables for data glitches ([7] page 74). Table 3.1.2 presents an example of a report that can be produced within most statistical packages. The report displays number of valid, invalid, and missing records for each variable specified:

	<b>Age</b>	<b>Model Year</b>	<b>Incurred Losses</b>	<b>Gender</b>	<b>Marital Status</b>
<b>Valid</b>	41,000	35,000	50,000	45,000	46,000
<b>Invalid</b>	100	1,000	-	500	1,200
<b>Missing</b>	9,000	15,000	-	5,000	4,000

Table 3.1.2

Note that it is not uncommon for missing values to be recorded as blanks. This situation will not be detected by procedures summarizing missing values. However, procedures used to tabulate all the values of a variable (e.g., data cubes, Microsoft Excel's AutoFilter) can be used to summarize the number of blanks on these variables. This is shown in Table 3.1.3:

<b>Value</b>	<b>Gender</b>
M	25,000
F	20,000
	5,000
<b>Total</b>	<b>50,000</b>

Table 3.1.3

Descriptive statistics can also be used to identify the presence of null values in numeric data.

### 3.1.3 Descriptive statistics

Descriptive statistics include such statistics as the mean, median, minimum, maximum, and standard deviation. Table 3.1.4 displays descriptive statistics, produced with the Microsoft Excel Analysis ToolPak, for an illustrative sample of general liability claims. The descriptive statistics summarize key information about the paid allocated expenses in the data. Looking at the minimum and maximum values can quickly inform us as to whether any values appear to be outliers or to have

### Actuarial IQ

unusual values. In this example, the minimum paid expense is a negative value. The table also indicates that the second smallest value is also negative. Both of these numbers indicate data records that may need to be reviewed further before using in any analysis.

Allocated Loss Adjustment Expenses	
Mean	1,323
Standard Error	252
Median	611
Mode	0
Standard Deviation	8,217
Sample Variance	67,513,031
Kurtosis	207
Skewness	13
Range	170,668
Minimum	(19)
Maximum	170,649
Sum	1,411,246
Count	1,067
Largest(2)	99,206
Smallest(2)	(11)

Table 3.1.4

#### 3.1.4 Box and whisker plots

A box and whisker plot is a one dimensional visualization of the distribution of a variable. The box plot, a predecessor of the box and whisker plot, can be programmed into Microsoft Excel. It displays a 5-point summary of a variable's distribution. The 5 points are: minimum, 25th percentile, median, 75th percentile, and maximum. A box is placed around the edges encompassing the 25th through 75th percentiles and lines extend from the box to the minimum and maximum values. The box and whisker plot modifies the box plot by displaying lines from the box to a specified distance (e.g., two standard deviations from the mean) from the box and by individually displaying

## Actuarial IQ

observations outside these lines.

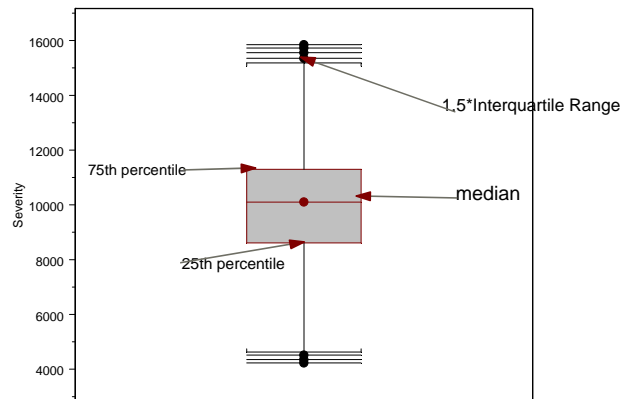


Fig. 3.1.1

Figure 3.1.1 displays a box and whisker plot. The top and bottom of the box are defined by the 75th and 25th percentiles of the distribution plotted. A line through the middle of the box denotes the 50th percentile (i.e., median) value. The width of the box carries no meaning. Lines extend from both the top and bottom of the box. These lines are referred to as the whiskers. For this graph, the lines denote the points 1.5 interquartile ranges<sup>1</sup> above and below the box edges. Points beyond this boundary are individually displayed (the circles with lines through them). These points may be considered outliers; they depict data records that the analyst might want to investigate.

Figure 3.1.2 displays the box and whisker plot for data containing an intentionally introduced error (the first number was replaced with ten times its value):

---

<sup>1</sup> The interquartile range is the difference between the 75th and 25th percentile

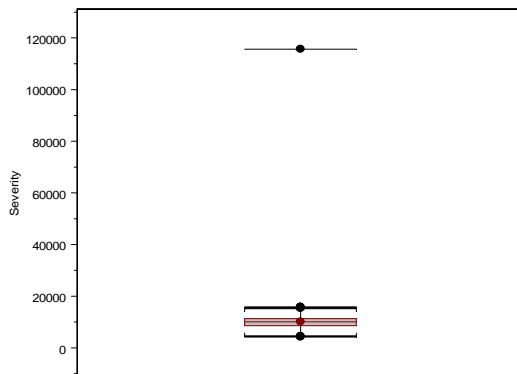


Fig. 3.1.2

In this paper we provide only a basic introduction to the methods of exploratory data analysis. A number of excellent references are available on this topic for those wanting a more thorough exposure to the topic. Hartwig and Dearing [14] provide an easy-to-understand introduction to the methods of exploratory data analysis, and Dasu and Johnson [7] introduce EDA within the context of its application to data cleaning.

### 3.2 Auditing Data

Whereas EDA cleans a dataset, auditing influences the process that generates the data. As such, auditing for data quality is a tool to help both assess and monitor data quality. While ASOP No. 23 does not require actuaries to audit data ([2], sections 1.2 and 3.6), knowing how audits are conducted can improve actuarial practice in at least two ways:

- First, it produces a more informed basis to assess what kind of reliance should be placed on audited versus unaudited data, and
- Second, the procedures and concepts used in auditing can be applied to resolve data issues without having to do a full-scale audit.

The main idea of data auditing is to compare the data intended for use to its original source(s), such as policy applications or notices of loss. This is done using both top-down and bottom-up approaches. The top-down approach is reconciliation: checking that totals from one source match

### *Actuarial IQ*

the totals from another (usually more reliable) source. These totals are usually dollars, but counts and records can also be reconciled. Auditors will often not only do their own reconciliations, but also review an organization's reconciliation procedures. Obviously, making sure totals match is one way to assess the reasonableness and comprehensiveness of a data set, so reconciliation can be useful to actuaries both on its own as well as when it is part of an audit.

The bottom-up approach takes a sample of input records and follows them through all the processing to the final report. Any good sampling textbook should provide the theoretic basis to address sampling issues. One such book is *Elementary Survey Sampling* by Scheaffer, Mendenhall and Ott [15]. Defining accuracy ratios can make results comparable from one audit to the next. An example of an accuracy ratio is the number of occurrences a given data element is correct divided by the number of occurrences reviewed. The number, type and rigor of these statistics are determined by the intended use of the data. Note that ratios of record counts can provide different information than dollar ratios, so sometimes it can be helpful to include both for phenomena of particular interest.

The following summary of major steps in a data quality audit is based on ISO's *Strength in Numbers* pamphlet [16]:

1. **Test the preparation of the data:** Measure how correctly and completely data is coded. Also measure how current it is.
2. **Test the data entry and data transfers:** How much of the data reaches its final destination intact? How much of this takes place in an acceptable period of time?
3. **Test the program controls:** Measure the extent that "only authorized data is entered for processing and that data is processed completely, accurately, and in a controlled environment" ([16], page 6). A controlled processing environment will have procedures and checks to ensure that computer jobs are run in the right order, computer jobs are not accidentally run twice, total outputs equal total inputs, users are aware when software programs end abnormally and so forth.
4. **Test the output controls:** Measure the accuracy, timeliness and correct distribution of reports.
5. **Test error procedures:** Measure the extent that the system detects and corrects errors in a timely manner.

"Performing periodic [data] audits will indicate:

## *Actuarial IQ*

- The accuracy and completeness of the picture... [which the] data gives of the insured risks,
- The timeliness of data processing,
- Any differences between statistical and other insurance data to be reconciled,
- Problems or potential problems related to collecting, coding, and reporting your data” ([16], page 3).

More information on data audits can be found in the Insurance Services Office’s (ISO) *Quality of Data Audit Guide* [17]. Accounting professional organizations may also publish information on auditing.

### **3.3 Information Quality in Models**

We now turn our attention from a strict focus on data to broader information quality issues. With the broader perspective of information quality, it becomes clear that actuaries are active participants in the data life cycle of an organization. They take data as an input, analyze it, and produce output that is used in decision making. The quality of this analytical step is thus a crucial contributor to the overall quality of information used in the company.

Analysis is about building models to explain or predict phenomena. As such, analysis behaves like software in some respects: it is a set of steps to manipulate data. Software quality is a function of design, implementation and testing. Good design decisions may improve not only the functionality and usefulness of the application, but also simplify quality assessment and ensure easy modifications and updates. Testing, especially if integrated with implementation, may improve the quality of the resulting software product. Any actuary involved in design or modification of spreadsheets and other analytical applications will clearly benefit from knowing the main principles of good software design.

#### **3.3.1 Quality design**

To use a manufacturing metaphor, the quality of actuarial work products depends on the choice and quality of the tools actuaries use to process incoming data. The tools should be good and suitable: the actuarial methods used should be appropriate for the data at hand. Quality of the method relies heavily on 1) model selection and validation, 2) model’s parameters estimation and 3) model’s verification (see [18], chapter 2.9 for detailed explanation). To understand the difference between validation and verification one should consider two questions: “did I use the right model?”

versus “did I use the model right?”

Some actuarial methods are designed only for data with particular properties, i.e., it is assumed that the data satisfy some preliminary conditions. Thus, before applying an actuarial method to a set of data, it would be prudent to test the method’s assumptions on that specific dataset.

Failed assumptions may either indicate inappropriateness of this particular method or uncover hidden data problems. In this sense, assumption testing may also serve as data quality tests.

An aspect of an analysis’s quality is **model performance**. Many actuarial methods for pricing and reserving predict some events that can be observed. Comparison of predicted and actual values may lead to method improvements, recalibration, or even rejection. Note that any of these outcomes leads to improvements in the model’s quality.

### **3.3.2 Implementation (software) quality**

In the actuarial toolbox, the spreadsheet occupies a special, quite dominant place. However, while tomes are written about C++ or VBA programming techniques and SQL optimizations, it is very hard to find practical advice on effective spreadsheet design.

The deceptive simplicity of a spreadsheet’s grid makes many users think of a spreadsheet as a single user’s ad hoc advanced calculator that can also chart and print. Users don’t even think there could be design recommendations and do not look for them. Indeed, a spreadsheet created by a user for a single use is quite disposable, but if there are multiple users or repetitive usage, spreadsheets become applications and should be treated as such.

An application is a part of the data flow of an organization and therefore subject to quality control. It has to be well-designed and documented to simplify 1) usage, 2) testing and 3) modification. What can be done on this front? Experience shows that one of the most effective techniques is separation of data and algorithms. Calculations (formulae and VBA code) should be stored in one file or spreadsheet tab (called the **template**), while data should be loaded from an external source (e.g., spreadsheet tab, file, or database). In practice, actuaries usually realize that *input* data like loss triangles, premiums and industry factors do not belong in a calculation template. What they rarely realize is that **output** data such as predicted ultimates or fitted distribution parameters do not belong in the template either: results have to be stored outside just like inputs.

## Actuarial IQ

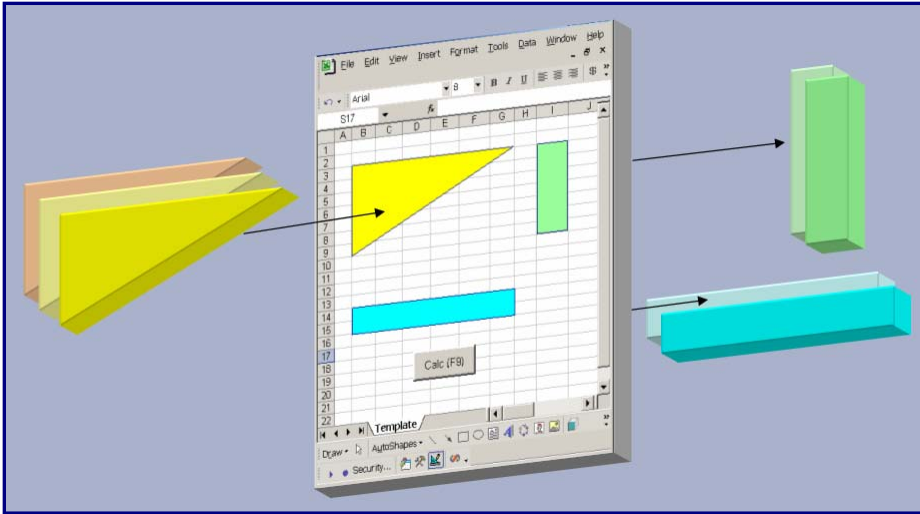


Fig. 3.3.1

Such a setup 1) brings consistency to calculations, 2) simplifies housekeeping, 3) allows versioning, and (combined with access rights) 4) improves control over modifications. An additional benefit for quality pursuers is the fact that separation of data and algorithms facilitates checking calculations with different data samples, thus enormously improving the quality of testing.

Another useful technique which extends the notion of separation is **layering**. Both users and designers may benefit when data (input placeholders), reconciliation, calculation, user interface (scenarios, selections, and assumptions), and presentation (results and charts) layers are located on separate spreadsheet tabs or worksheets. Such a layout not only simplifies navigation, it also shortens the learning curve for users, allows designers to better understand workflow and provides better documentation.

## Actuarial IQ

For example, below is a hypothetical layering scheme for a rate review application:

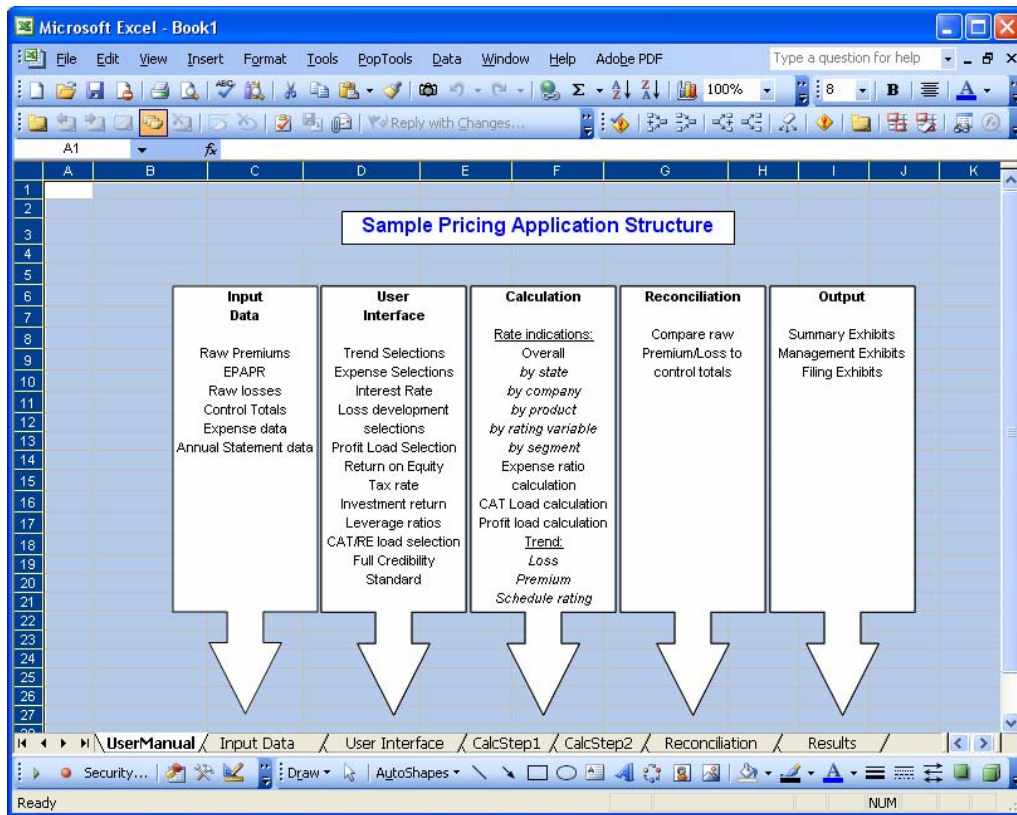


Fig. 3.3.2

Good documentation is a centerpiece of quality design. Every application should have a "header" identifying inputs, outputs, purpose, and contacts. Spreadsheets also generally provide adequate facilities for file versioning, VBA code commentaries and cell comments. As noted in [29], one can use built-in document properties or create custom ones and link them to cells inside spreadsheets. The trick is to remember to update documentation with **every** modification or improvement made to a template.

### 3.3.3 Testing

Testing is critical to the good design of successful applications. Indeed without testing, a spreadsheet (or query or notebook) may never become an application (i.e., a reusable tool) – there

### *Actuarial IQ*

would be no assurance that it could handle different situations correctly. Once an application becomes successful, thus widely used, testing becomes even more important.

The majority of books on testing deal with higher languages (C++ and the like): very few publications give practical advice on spreadsheet development. Some of the main thoughts from these books, however, will be of interest to actuaries.

Testing, according to Edward Kit's *Software Testing in the Real World* [19] should start with **specifications**, end with **final product evaluation**, and should be performed by an independent party. The main testing techniques are verification and validation, i.e., checking the code and examining final product outcomes. In the actuarial paradigm, examples of the final product could be Excel spreadsheets, Mathematica notebooks or Oracle stored procedures. Similarly, specifications could be a reserve test or pricing method, and the "code" could be formulae in cells, VBA subroutines or SQL statements.

Some of Kit's verification testing techniques can be applied to spreadsheets. For example, checking programming code against a list of common mistakes applies mostly to those who use Visual Basic. However, the recent addition of "Formula Evaluation" and "Watch Window" tools make Excel much friendlier for debugging. These tools allow users to validate formulae placed in cells by displaying results of all intermediate calculations and by monitoring values in "watched" cells. They bring debugging power previously available behind the scenes (to VBA coders) to the forefront (to cell formulae designers).

The most common testing technique is validation: checking that calculations produce expected results for different (and not necessarily correct) data. Validation treats an algorithm as a black box, feeding it with different inputs and observing results. On the one hand, validation checks algorithm limitations (e.g., whether it can work with negative amounts, strings, missing values). On the other hand, it also checks the accuracy of calculations on the datasets with known results. In either case, validation feeds the algorithm with different datasets so this process may benefit from separating data from calculations as described above. Indeed, using Excel's "scenarios" functionality, one can create a library of test datasets and recall them by selecting the corresponding scenario. Similarly, with assumption sets, if they are separated from the rest of the application, then it is easier to test algorithm results against various assumption sets.

Testing is very repetitive by nature, so it makes sense to accumulate testing tools for future reuse. It is very easy to build libraries of "bad" and "benchmark" datasets for testing actuarial methods. Testing routines and functions could also be accumulated into a library available to every tester or

designer.

Kit's suggestions that 1) testing should be an integral part of development, and 2) testing should be performed by people outside of the development team should definitely be implemented for any application that is part of a company's data flow.

In conclusion, the keys to quality models are:

- **Good design,**
- **Accurate implementation,** and
- **Thorough testing** of everything: from methods and assumptions to auditing spreadsheet formulae and query results.

### **3.4 Data Presentation (Reports) Quality**

If data reaching the presentation stage are accurate, reasonable, complete, and have been analyzed in a high quality model, what can go wrong with the presentation? Unfortunately, a lot:

- Data can be **mislabeled** or incompletely labeled. "Total Loss" may refer to "loss net of recoveries" or "loss and ALAE net of reinsurance" or "unlimited loss before deductibles."
- Data can be **incorrectly related** to other information, producing wrong calculations. Date mismatches in losses and premiums may produce erroneous loss ratios.
- Data may be arranged in such a way that the essential information it is supposed to convey may be **overlooked**. A good report should emphasize the message and guide the reader to the most important information.
- Data can be **misinterpreted** and the message they deliver may be misunderstood. It is not unusual to witness confusion and misuse of such notions as reserve range, expected shortfall, confidence interval, or risk transfer.

To avoid costly mistakes from wrong decisions based on poor data presentation, crucial reports should be prepared with the involvement of someone who understands the data (e.g., an actuary). Therefore actuaries should be familiar with some tools and techniques to improve the quality of reports.

There is an enormous variety of reporting tools of different capabilities and complexities, but the

most versatile, familiar, and readily available is a spreadsheet. Modern spreadsheets provide enough features for building quality reports.

[Appendix A](#) contains some practical solutions to address:

- Unambiguous labeling,
- Consistent calculations,
- Focusing attention, and
- Minimizing misinterpretation.

### 3.5 Measuring Data Quality

The first four subsections of section 3 addressed individual steps of the data life cycle ([Fig. 2.0.1](#)). The remaining three subsections of section 3 address general issues that apply to the entire life cycle.

Many data quality authors (e.g., Redman [20], Dasu and Johnson [7]) are strong proponents of measuring data quality. These authors believe that in order to motivate improvements in data quality, it is imperative that data quality be measured, even when the measures are somewhat subjective. The following is a brief introduction.

A key concept in measuring data quality is the data's "conformance to constraints." Dasu and Johnson describe both static and dynamic constraints ([7], page 131). Static constraints relate to properties of the data itself, such as its validity. For example, for the constraint "value should be present and be only from a fixed list of correct values," the corresponding measure would be "the number or percentage of missing or invalid values in a variable." Dynamic constraints relate to the processes used in the flow of data from its source to the different databases. Examples of dynamic constraints would be 1) "a reserve change is added to prior cumulative reserves (not to cumulative losses)" and 2) "incurred losses can never be less than the sum of the amount paid." Thus, dynamic constraints capture business rules.

Some of the key data quality measures recommended by Dasu and Johnson ([7], pages 131 - 134) are:

1. **Extent of automation:** sample some transactions, follow them through the database creation process, and tabulate the number of manual interventions
2. **Successful completion of end-to-end process:** the number of processes that have the outcome they are expected to have. For instance, a sample of claims can be followed through

### *Actuarial IQ*

closure and it can be determined how soon after the final payment is made that the claim is closed.

3. **Impact on analyses:** measure how many errors in analyses result from errors in the data. Using sampling, the number of analyses adversely affected by data quality problems can be tabulated. Both the frequency and severity of the problems should be measured.
4. **Accessibility:** how easily can the data be accessed? For example, the time between a request for data and access to the data can be measured.
5. **Interpretability:** how understandable is the data? The quality of the metadata determines how interpretable the data is to users. The interpretability of data should be based on 1) the availability of metadata and 2) the extent to which the data adheres to the definitions in the metadata.
6. **Conformance to business rules:** how well does the data adhere to insurance business rules? For instance, how often are negative paid losses recorded in lines where losses should always be positive (i.e., no salvage and subrogation)?
7. **Conformance to structure:** Select important constraints that the data must follow and measure how well the data conforms to those constraints.
8. **Accuracy:** what proportion of the data contains valid values? This can be expensive to measure, so measures based on samples or based on proxies such as complaints or surveys are recommended.
9. **Consistency:** how often do databases at different points in time or data in different databases and tables within the company agree with one another?
10. **Uniqueness:** certain data elements should only have one observation in the dataset. For instance, a claimant level database should have only one record for each claimant. Measuring this amounts to identifying duplicates, which is discussed in section [3.1](#).
11. **Timeliness:** how often is the data updated and what proportion of it is available on schedule? Dasu and Johnson also mention that data should have an accurate time stamp.
12. **Completeness:** to what extent does the data contain all the data elements relevant to the analyses and reports a company undertakes? Thus, a database that is accurate and timely may be of low quality because it contains only a few variables or only a few years of history.

The different metrics are weighted together into an overall data quality index using business

### Actuarial IQ

considerations and the analysts' goals to develop weights. For example, if improvement in the database itself is considered most important, the static measures (e.g., accuracy, completeness, timeliness) might be given greater weight than dynamic measures (e.g., successful completion of end-to-end processes).

Table 3.5.1 illustrates a simple data quality measurement for a company beginning a data quality initiative (i.e., these are simple not comprehensive measures). All audits, sample findings, and survey results have been converted into scores between one and ten, where one is low and ten is high. Weights have been assigned subjectively.

Measure	Score	Weight
Extent of Automation	4	0.1
Accuracy	3	0.2
Glitches in Analyses	3	0.2
Completeness	6	0.2
Interpretability	7	0.3
<b>Total</b>	<b>4.9</b>	

Table 3.5.1

The data quality variables can be measured periodically after a data quality initiative is undertaken. Over time, the score should improve. Dasu and Johnson note that when used as a tool for quality improvement, it is the **direction** of the data quality measure over time that is of interest ([7], page 134). A number of other authors (e.g., Loshin [21], Redman [20]) offer additional advice as well as some alternative measures of data quality.

### 3.6 Data Quality Improvement Strategies

Two strategies to improve data quality are data cleansing and re-engineering. The objective of the first strategy is to take defective data and correct, reformat, consolidate, and standardize it so that standards are met and maximum value can be achieved from the data. The objective of re-engineering is to proactively eliminate the causes of poor quality data by changing processes. Note that data cleansing is an ongoing cost-added process. The overall objective is to eliminate the need to perform error correction, but the most effective approach is to couple data cleansing with re-engineering. While the former attacks specific defects in the data, the latter focuses on the root

## Actuarial IQ

causes of the defects. Note, however, that the profiling process advocated by Olsen would often require process changes such as a full-time team dedicated to data quality, as well as recommended changes resulting from the team's investigations.

What follows are some alternative strategies based on *Data Quality, the Accuracy Dimension* [22] and *Improving Data Warehouse and Business Information Quality* [23]. Note that the costs of applying a particular data improvement technique need to be weighed against the benefits. For example, it may be too expensive to correct lost, missing, or incorrect data if the source data is not readily accessible.

### 3.6.1 Data cleansing

The objective of data cleansing is to improve the data quality in existing files to maximize its value and to minimize the cost due to poor quality information. This includes correcting wrong data, standardizing nonstandard data values, filling in for missing data, and consolidating duplicate occurrences.

In *Data Quality, the Accuracy Dimension*, Olsen introduces a proactive data quality assurance program for detecting and addressing data inaccuracies throughout the many databases used by an organization. The system has two basic approaches, denoted **inside out** (a ground up, detailed, data-dependent approach) and **outside-in** (essentially an outcomes-based, business-driven approach

The **inside-out** approach can be summarized as follows:

1. Build the organization's metadata to have a complete and correct set of rules that define data accuracy for a particular dataset,
2. Gather inaccurate data evidence, i.e., collect facts about data shortcomings,
3. Aggregate the inaccurate data evidence into issues,
4. Analyze the issues to determine the external impact,
5. Set the priority of each issue based on its external impact, and then
6. Rectify the issues.

The **inside-out** approach can detect many data inaccuracies that are routinely missed by users working with aggregated data.

In contrast, the **outside-in** method "identifies facts that suggest that data quality problems are having an impact on the business" ([22], page 73). Such facts might be reworks, returned merchandise, or customer complaints, for example. The facts are then evaluated to determine the

### *Actuarial IQ*

degree of culpability attributable to defects in the data. The advantage of the outside-in approach is that it automatically focuses on issues that have a noticeable external impact. One of its disadvantages is that it may miss issues with still larger, but unnoticed, impacts. That is, by using the outside-in approach alone, only those data quality problems that have already manifested as business issues will be detected. It is also less likely to discover the full scope of related issues that interact to produce the observed impact. This approach requires the participation of business analysts along with a dedicated data quality analyst. Olsen's recommendation is that both approaches need to be applied.

The **outside-in** approach can be summarized as follows:

1. Identify information indicating a data quality problem
  - a. investigate customer complaints
  - b. investigate business user complaints
  - c. interview users of data to assess their level of satisfaction
2. Determine the extent to which data accuracy issues contribute to the problem.

The following chart summarizes the two approaches to data quality improvement programs:

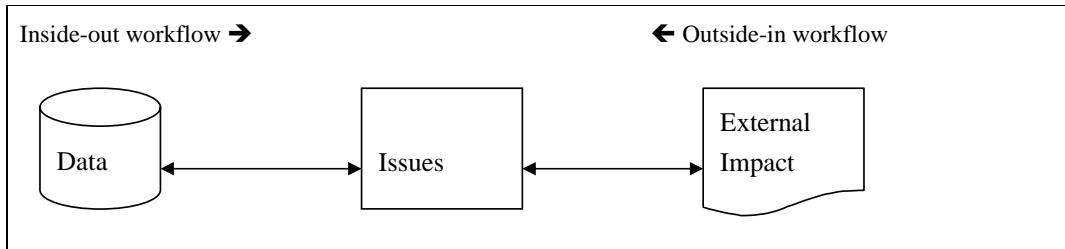


Fig. 3.6.1

For comparison, English's general steps for data cleansing are outlined in [Appendix B](#).

### 3.6.2 Re-engineering, a.k.a. process improvements

This strategy improves business processes by eliminating the causes of poor quality data.<sup>1</sup> It is a proactive method that analyzes the cause of problems and eliminates them. The rationale is that it is much less expensive in the longer term to prevent errors than to repeatedly screen for them and repair them in different databases every period. Therefore the long-term solution to data quality is not to fix the data but to fix the processes that produce the defective data. Data cleansing fixes the problems after they have occurred, whereas process improvements eliminate the causes. English ([23], pp. 285-310) also provides a data defect prevention approach, which is described in [Appendix C](#).

## 3.7 Actuarial Data Management

Actuaries are among the most prominent users of an organization's data. Thus, they have a natural vested interest in ensuring that their organization's data is of the highest quality. Over time, data management has evolved as a unique specialty within the actuarial community. In some insurers, though, the role of data manager is not held by an actuary, but by a person trained in this field that understands the data needs of the actuary and other users of data.

In performing any analysis, the actuary must consider many things, but the starting point for the analysis is the historical premium, exposure, loss and expense experience for the type of insurance under review. "This experience is relevant if it provides a basis for developing a reasonable indication of the future. Other relevant data may supplement historical experience. These other

---

<sup>1</sup> Note that the data cleansing process as describes by Olsen [22] and Redman [20] is intended to also affect the processes generating the errors once the errors are uncovered and thus may entail some re-engineering.

### *Actuarial IQ*

data may be external to the company or to the insurance industry and may indicate the general direction of trends in insurance claim costs, claim frequencies, expenses, and premiums” ([24], page 7).

The data management actuary provides a bridge between those who are responsible for the collection and repository of the organization's data and the pricing or reserving actuary who will use the data in analyses. Thus, two critical areas for the actuarial data manager are:

- The appropriateness of the collected data elements for the analysis to be done, and
- The quality of the collected statistical experience for the analysis to be done.

Some of the activities performed by the data management actuary include:

- Reviewing the various data compilations for reasonableness. This includes comparing the current data compilation against the previous data compilation to ensure that the change in the data for overlapping years is as expected. For example, the losses as of 24 months versus as of 12 months have grown as expected for the line of business under review.
- Reviewing the growth patterns by year within a compilation.
- Reviewing the distribution of data within a data element. For example, reviewing the written premium distribution by geographic location to make sure it accurately reflects the book of business in the compilation and that it has not been erroneously coded to one location.
- Ensuring that any definitional changes in the data elements are accounted for and notifying the actuary who will use the report of this situation.
- Reviewing the data compilation for completeness – that is, only the data that was supposed to be included is included and it is included only once.

As data management actuaries grow in responsibility, they should also take a more proactive role in understanding the data processing stream from source through transformations, data base repositories, and data extraction and compilation; ensuring that the organization is following data management best practices at every step in the process. Thus, they will have a complete understanding of the data that has been extracted and will ensure its proper use in analyses done by themselves or another actuary to whom they are supplying the data.

The insurance data management profession has established a set of guiding principles and best practices for data management [25]. Some of these key principles and practices are listed in bold

### *Actuarial IQ*

below, followed by further explanation and where appropriate, a cross-reference to ASOP No. 23 [2].

1. **Data must be fit for the intended business use.** This principle is in sync with ASOP No. 23 that states "For purposes of data quality, data are appropriate if they are suitable for the intended purpose of an analysis and relevant to the system or process being analyzed" ([2], page 2). Data should be collected in the level of detail (breadth and depth of the data elements) and at a level of quality that are sufficient for the intended applications or analyses to be performed.
2. **Data should be obtained from the authoritative and appropriate source.** Data should flow from the underlying business process, whether it is the underwriting and rating of the risk or other processes such as claim reserving, accounting of payments received or claim paid out, or litigation metrics. For example, insurance statistical data for a risk related to the premiums charged should be collected in a level of detail consistent with how the risk is underwritten and rated. That is, which data elements are collected and the depth of the detail (or attributes) within the data elements should be consistent with how the risk is underwritten or rated. The actuary using data received from others is required by ASOP No. 23 to "take into account the extent of any checking, verification, or auditing that has already been performed on the data, the purpose and nature of the assignment, and relevant constraints" ([2], page 4). It is also important that data be supplied by a source that understands the data. For example, detailed data regarding the nature of an injury should be supplied by the health care provider who understands the nature of the injury rather than a claims coder.
3. **Common data elements must have a single documented definition and be supported by documented business rules.** As ASOP No. 23 notes "The actuary should make a reasonable effort to determine the definition of each data element used in the analysis" ([2], page 4).
4. **Metadata must be readily available to all authorized users of the data.** The actuarial data manager should ensure that data, systems, and reporting mechanisms are designed and maintained in a manner that promotes good data management and data quality. This includes a robust, comprehensive business data dictionary that provides a clear, unambiguous definition of each data element that is consistent with the underlying business process.
5. **Data standards are key building blocks of data quality.** To promote consistency in the data collected, increase efficiency of the data collection process, and maximize utility of the data, organizations must foster the development and adoption of data standards and data quality standards. Industry standards must be consulted and reviewed before a new data element is

created.

6. **Data should have a steward** responsible for defining the data, identifying and enforcing the business rules, reconciling the data to the benchmark source, assuring completeness, and managing data quality.
7. **Data should be input only once and edited, validated, and corrected at the point of entry.** Data quality should be managed as close to the source as possible. This includes defining the data quality standards for the data to be collected. Processing steps between the data source and the data capture increase the likelihood there will be errors and often increase the cost of correcting those errors.
8. **Data should be captured and stored as informational values, not codes.** For example, if age of driver is a desired date element, the birth date of the driver should be captured and stored rather than the driver's age. By following this principle, misinterpretation of the data will be reduced, and serious errors in business decisions can be avoided. The data will also be more complete and more likely to be useful in answering unanticipated questions. Following this principle also facilitates reviews of the data for reasonableness and consistency.
9. **Data must be readily available to all appropriate users and protected against inappropriate access and use.** Insurance statistical data is the life blood of the property-casualty insurance industry and much of the data is considered a trade secret or is highly personal in nature (see 2.2.2). Data managers must balance access to data against inappropriate access or use. The actuarial data manager should ensure the actuaries' repository data base meets current and future business and analytical needs by partnering with the IT professionals in designing it.

For more information regarding data management best practices, see the Insurance Data Management Association website, <http://idma.org/productsDMBestPractices.htm>.

#### 4. CONCLUSIONS

Data quality is a core issue affecting the quality and usefulness of the actuarial work product. Data quality is often perceived as a mundane issue with less recognition and attention devoted to it than other issues, such as actuarial models and methodologies. However, data exists to fulfill a need: the need for optimal decisions. To the authors' knowledge, this is the first paper to provide a general introduction to data quality and data management directed specifically at actuaries since the CAS Committee on Data Management and Information White Paper of 1997.

#### **4.1 Pragmatically**

[Figure 2.0.1](#) outlines the steps in the insurance data life cycle, the kinds of errors that can occur in each, and references to relevant sections of this paper. As such, figure 2.0.1 forms a handy reference both to trace where a particular error may be occurring and which section of this paper may be most relevant.

Several tools to help actuaries improve their information quality are:

1. Exploratory data analysis to identify outliers and explore the structure of a dataset ([3.1](#)),
2. Improving the quality of actuarial models ([3.3](#)),
3. Improving actuarial presentations and reports ([3.4](#)),
4. Measuring data quality to track progress ([3.5](#)) and awareness of quality audits ([3.2](#))
5. Strategies to improve data quality ([3.6](#)), and
6. Guiding principles and best practices ([3.7](#)).

Each section has references to books and/or CAS papers for readers who need more information.

## 4.2 Conceptually

We began by drawing attention to the increased importance of data quality given easy access to an unprecedented level of detail and the proliferation of new tools and techniques to analyze such data. The actuarial frame of reference (2.1) was broadened beyond the scope of ASOP No. 23 in three ways:

1. **Data is a corporate asset that needs to be managed and actuaries can play a role.** Actuaries have the ability and motivation to influence the processes that give rise to the data they use (3.7)
2. **Data needs to be appropriate for all of its intended uses.** Actuaries have a unique role to play in achieving this goal here too: actuaries can expand their concerns for data beyond the analysis at hand. Finally
3. **Expansion of data quality principles (2.2 and 2.3)** to support these broader perspectives.
4. It should be noted that these expansions are those of the working party; not interpretations of the standard.

Data quality is not just about how data is coded: we have coined the phrase “information quality” to emphasize the impact of processes on the quality of the final product(s). Metadata (2.3), information about the data, is critical to actuaries correctly interpreting their data and the glue that holds an organization’s data structures together. Statistical plans (2.4) were introduced as a form of metadata. Data management best practices (3.7) embrace and support all of the above.

Ultimately, empowering actuaries to improve the quality of information in their organizations can increase the efficiency, effectiveness and impact of actuaries on their organizations by turning data into more useful information to make better decisions.

### Acknowledgment

The working party thanks our IDMA liaisons for their feedback and support throughout this project.

The working party also thanks the Insurance Services Office, Inc., for the use of excerpts from their homeowners module of the ISO personal lines statistical plan (other than auto).

### Supplementary Material

Code for creating Box Plots in Excel (described in section 3.1.4) can be found at [www.data-mines.com](http://www.data-mines.com).

Presentation template with live charts (described in Appendix A, section 4) can be downloaded from [www.casact.org/research/drmwp/DRM%20presentation.ppt](http://www.casact.org/research/drmwp/DRM%20presentation.ppt).

**Appendix A: Practical solutions for addressing some problems with presentation quality (expansion on section 3.4)**

**1. Practical solutions: unambiguous labeling**

Unambiguous labeling requires first that the label is consistent with the content and, second, that the label is descriptive enough to avoid ambiguity.

Consistency of labels and content can be achieved by examining every transformation<sup>1</sup>, transfer<sup>2</sup>, and calculation data goes through while keeping track of data sources and formulae applied to the data. Spreadsheets provide some assistance with this: a table created by importing external data keeps query information available for examining and editing. This SQL text helps to identify sources and clarify the nature of the extracted data. A spreadsheet’s ability to name ranges gives users an ability to create readable and, thus, traceable calculations. The next logical step in readability is to use labels within formulas. The “using labels in formulas” feature allows the user to create a quite traceable expression like “=Case Reserves + IBNR” using field names “Case Reserves” and “IBNR” in the formula for “Reserves.” Another useful facility in spreadsheets is “commenting:” a descriptive tag attached to an upper left corner of the triangle will “travel” with the data during copy and paste operations and will help the user to avoid obvious errors, such as making sure that paid losses

AY\Age		36	48	60
1994	\$ ...	107,847	\$ 115,288	\$ 124,592
1995	\$ Shape --> <b>Triangle</b>	110,271	\$ 112,562	
1996	\$ Amount--> <b>Losses</b>	104,029		
1997	\$ Cumulative- <b>True</b>			
1998	\$ 105,647			

wouldn’t end up in a calculation intended for claim counts.

The second type of labeling problem, which we will call “disambiguation of labels,” presents a different challenge. Readability and aesthetics considerations advocate short labels, while the need for quality and precision requires labels to be quite detailed and relatively long. The solution seems to be in hiding less necessary details until needed. The user would still need to be able to display the

<sup>1</sup> Data transformation step – edits, rearrangements and conversions from one format to another.

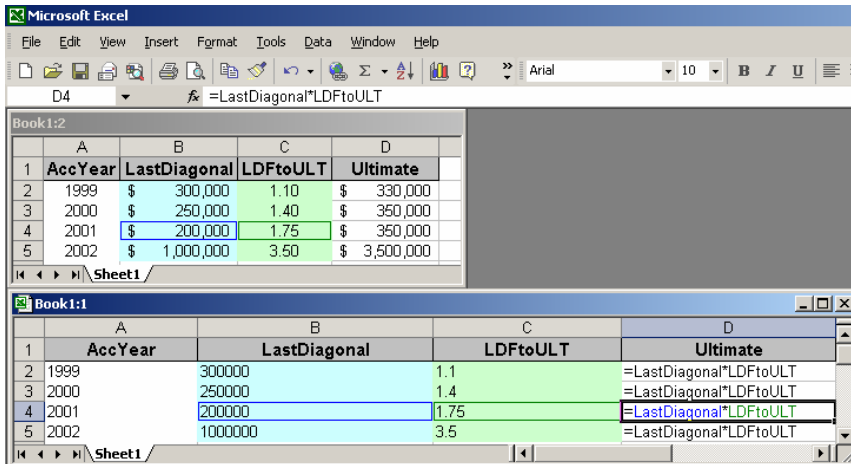
<sup>2</sup> Data transfer step – extraction from one system, transportation and upload to another system.

## Actuarial IQ

detailed labels on demand. Spreadsheet “comments” satisfy such design requirements: they are hidden until the computer’s mouse moves over a cell with a label in question. This technique, while convenient, is not very reliable because it doesn’t firmly relate short and long labels. More reliable, but much more involved, is SmartTag technology that allows a spreadsheet to recognize certain labels, lookup for their longer descriptions in the metadata table, and display long labels on demand. SmartTags may ensure enterprise-wide label consistency, but may not help in the creation of an ad-hoc report with the new labels. Another hide-display technique available to spreadsheet users is the use of an outline. Short labels can be placed on a higher outline level and additional (clarifying) meta-information can be placed on a lower outline level and collapsed. A collapsible outline view is a convenient arrangement for other meta-information regarding reports: for example, lists of formulae used in reports or lists of data sources and analysis methods.

### 2. Practical solutions: calculation consistency

It is very hard to prevent users from adding “doctors” to “hospital beds” to obtain “total exposure,” but some precautions could be made to prevent embarrassing mistakes. One can borrow



	A	B	C	D
1	<b>AccYear</b>	<b>LastDiagonal</b>	<b>LDFtoULT</b>	<b>Ultimate</b>
2	1999	\$ 300,000	1.10	\$ 330,000
3	2000	\$ 250,000	1.40	\$ 350,000
4	2001	\$ 200,000	1.75	\$ 350,000
5	2002	\$ 1,000,000	3.50	\$ 3,500,000

	A	B	C	D
1	<b>AccYear</b>	<b>LastDiagonal</b>	<b>LDFtoULT</b>	<b>Ultimate</b>
2	1999	300000	1.1	=LastDiagonal*LDFtoULT
3	2000	250000	1.4	=LastDiagonal*LDFtoULT
4	2001	200000	1.75	=LastDiagonal*LDFtoULT
5	2002	1000000	3.5	=LastDiagonal*LDFtoULT

an idea from programming languages that enforce so-called “strong typing”: every piece of data has a type associated with it and no operations between incongruent types are allowed. To mimic “strong typing” in a spreadsheet situation one has to keep “type” information associated with data elements, bring it (“type” information) to the report along with the data and use it for “type checking” in the formulae. For example, the formula for a loss ratio should first check that both the numerator and denominator belong to the same year *and the same kind of year* in order to avoid “underwriting year” vs. “accident year” mismatch. Spreadsheets don’t have built-in “typing”

### Actuarial IQ

enforcement tools; however, they provide mechanisms that may help avoid some simple errors. In a columnar report, one can write formulae with column labels rather than with nondescriptive cell references. Assuming that labels correspond to column content, this is a much more reliable way to refer to particular data. Additional information on this topic and implementation ideas can be found in [26].

More accurate solutions would involve storing results of the actuarial analysis in a well-designed relational database and creating reports from it. Assuming that database integrity is intact, the database engine would ensure proper relationships between data elements from different tables.

### 3. Practical solutions: focusing attention

There are many techniques for attracting a report reader’s attention to important information. We will mention just three of them: adaptive reporting, visualization, and alarm systems. All three are within a reach of any spreadsheet user and can be used to improve the informational value of reports (see [27]).

**Adaptive or data-driven** are reports whose size, shape, and format adapts to the data. Placing these reports in an interactive environment such as a spreadsheet allows the user to interact dynamically with the report (effectively creating a whole family of reports rather than a single one), shaping it to the level of detail that suits the user.

A partial list of data-driven implementations found in spreadsheets includes:

- **Filtering** (reduces amount of data displayed).
- **Outlining** (hierarchically organizes data with an ability to hide and display data on different levels of the hierarchy).

1	2	3	4	5	A	B	C	D	E
	1				<b>Reinsured</b>	<b>LOB</b>	<b>State</b>	<b>AccYear</b>	<b>UltNetLoss</b>
	2				ABC	WC	NY	1996	1,712,201
	3				ABC	WC	NY	1997	1,730,918
	4				<b>ABC WC NY Total</b>				3,443,119
	5				ABC	WC	CT	1996	1,944,502
	6				ABC	WC	CT	1997	1,975,489
	7				<b>ABC WC CT Total</b>				3,919,991
	8				ABC	WC	NJ	1996	2,172,041
	9				ABC	WC	NJ	1997	2,227,708
	10				<b>ABC WC NJ Total</b>				4,399,750
	11				<b>ABC WC Total</b>				11,762,860
	14				<b>ABC GL Total</b>				14,245,270
	17				<b>ABC AL Total</b>				7,249,632
	18				<b>ABC Total</b>				33,257,762
	48				<b>XYZ Total</b>				32,809,931
	49				<b>Grand Total</b>				<b>66,067,693</b>

### Actuarial IQ

- **Sorting** (does not reduce amount of data displayed, but brings the most important information to the top or bottom).
- **Conditional Formatting** (defines color, font, size and other formatting attributes of a cell as a function of the values in it or in other cells).

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		12-24	24-36	36-48	48-60	60-72	72-84	84-96	96-108				
2	1992	2.441	1.423	1.140	1.239	1.101	1.087	1.069	1.085				
3	1993	2.373	1.406	1.429	1.110	1.060	1.091	1.061					
4	1994	2.387	1.567	1.143	1.133	1.091	1.074						
5	1995	2.420	1.356	1.138	1.112	1.087							
6	1996	2.322	1.374	1.162	1.166								
7	1997	2.365	1.310	1.198									
8	1998	2.237	1.368										
9	1999	2.371											

- **OLAP-enabled tools** (provide an ability to display cross-sections or aggregations of multi-dimensional data in 2-D). OLAP-enabled tool (such as Excel's Pivot Table) is the ultimate adaptive reporting mechanism which supports filtering, sorting, outlining and conditional formatting and as such should become a preferred choice for any report designer.

Comment [d1]: Should be bold?

**Alarm system** is a technological solution whose purpose is to warn about undesired development. Alarm system usually triggers some action when the problem is found. The actions range from passive (paint some cells differently in the report) to interactive (display a warning dialog, send an e-mail requiring a response) to autonomous (launch a software program to fix the problem). At different stages of the data workflow, alarm messages can be aimed at different recipients: data integrity issues could be addressed to data managers, model's assumption test failures should be directed to actuaries, and sudden reserve increases should be presented to the management. Correspondingly, determination of which events under what conditions trigger an alarm is up to professionals responsible for the information quality on every given stage. In particular, actuaries should define what is acceptable and (on the other hand) what constitutes error or warning for data

## *Actuarial IQ*

suitability testing, actuarial analysis and presentation of actuarial results.

**Visualization** is the process of exploring, transforming, and viewing data as images to gain understanding and insight into the data. Images have unparalleled power to convey information and ideas. Consequently, visualization is a primary tool for communicating complex and/or voluminous information.

There exist a multitude of visualization approaches: mapping scalars to colors, contouring (iso-surfaces), glyphs (arrows of different color, length, direction), warping (display of different stages in the motion), displacement plots, time animations, streamlines (particle traces), and tensor algorithms. For the majority of actuaries, the most convenient and familiar visualization tool is a chart. From a presentation quality perspective, the report designer should be most concerned with the chart type, axis scaling, and the clarity and accuracy of the legend.

While there are numerous **chart types** available in spreadsheets, their add-ins, and other reporting packages, only a few are usually suitable for displaying each particular type of data. Percentages and shares are best presented by a pie chart, while XY-scatter is better suited for dependencies or comparisons (i.e., “Risk vs. Return”). Discrete values (i.e., “Total Premium per year”) are easy to present as a bar chart, while continuous variables (i.e., “Payment pattern”) are better displayed as lines. One shouldn’t use stacked bars for nonadditive values (i.e., “Incurred Loss” stacked on top of “Paid Loss”) or radar chart other than for comparison of several sets of data in multiple categories (criteria).

**Axis scaling** is very important for the readability of the chart, especially when displaying several data series. Sometimes in situations when one set of data (i.e., “Premium in dollars”) dwarfs another (i.e., “Exposure in number of cars”) it is necessary to create a second axis (with different scale) for the second set of data. Choosing an axis to be “time-scaled” automatically adds a capability to display monthly and annual aggregations of the data by selecting corresponding axis step (so called “axis base unit”). Occasionally, automatic scaling provided by a spreadsheet makes a wrong guess or is not as illustrational as desired. Sometimes data are better viewed in a logarithmic scale or in reverse order or with preset maximum and minimum. For example, displaying “Inception-to-date payments” on a chart with the maximum preset to “Aggregate Limit” could be more informative than just using automatically scaled axis<sup>1</sup>.

---

<sup>1</sup> To set up Maximum Value for the Axis click on the chart, right-click on the Axis and select “Format Axis...” menu option. In the “Format Axis” dialogue on the “Scale” tab uncheck “Maximum” checkbox and type desired value in the corresponding edit box.

## Actuarial IQ

The importance of **clarity and accuracy** of the chart title and legend cannot be overstated. Even the most primitive chart needs a precise description of the data displayed. Even more important, the user should provide accurate axes definitions and data series descriptions for the charts with multiple data series, dual axes, or of mixed types (i.e., bar and line on one chart). Without that, a chart may become a source of confusion instead of being a source of information.

### 4. Practical solutions: fighting misinterpretation

Actuaries deal with more and more sophisticated notions that are easy to misinterpret, misunderstand, and misuse. Three of the most difficult notions (as identified in [28]) for decision makers, regulators, accountants, and auditors are uncertainty, development, and multidimensional ranking. Attempts to explain and illustrate these concepts can result in confusion and wrong decisions. The problem is fundamental, given that accountants, performance measurers, and lawmakers operate with numbers rather than with distributions of random values. For example, an attempt to represent the distribution of possible aggregate losses with just one (“Reserve”) or two (“Reserve Range”) numbers inevitably leads to shortcuts in understanding and may create the impression that any value within a range is equally probable. The misinterpretation may be reinforced by a chart with reserve ranges shown as solid bars.

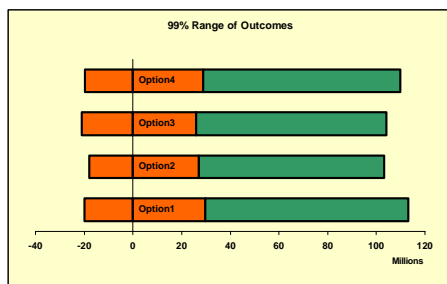


Fig. A.4.1

In reality, aggregate losses are not uniformly distributed and deserve more sophisticated graphical representation.

## Actuarial IQ

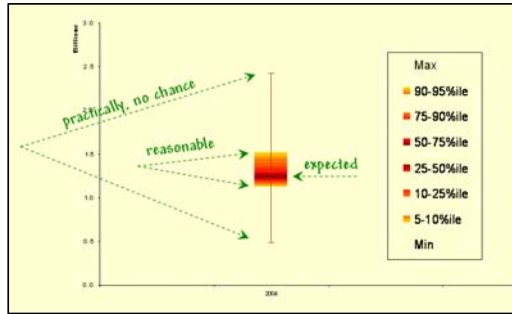


Fig. A.4.2

Some visual cues like gradients<sup>1</sup> or properly shaded areas should assist in visualizing uncertainty. Indeed, vanishing color is supposed to emphasize diminishing probabilities of extreme outcomes. Thus, Fig. A.4.3 may give better representation of the reserve ranges than Fig. A.4.1.

---

<sup>1</sup> To set up chart attributes such as gradient and borders, right-click on the chart element (i.e., bar), choose “Format %chart element%...” menu option and select tab “Pattern”. For gradient click on “Fill Effects...” button in the “Area” section, for borders make proper selections in the “Borders” section of the dialog.

## Actuarial IQ

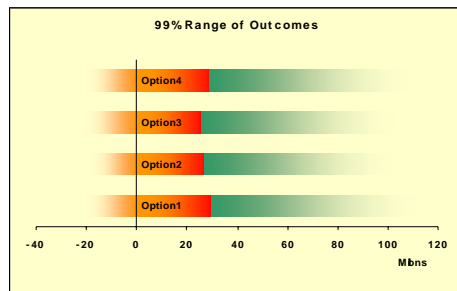


Fig. A.4.3

Another hurdle for the end users of reports to overcome is the concept of **development**. Combined with uncertainty (which itself changes over time), it creates a lot of opportunities for misunderstanding. Numbers in a spreadsheet or on a printed page do little to demystify trends, while standard chart options produce misleading results. With some effort, however, it is possible to illustrate development of random values in a spreadsheet chart (for example, charts on the Fig. A.4.4 below utilize vertical dimension (width of the curve in one case and height of the line in another) to illustrate the size of uncertainty which changes over time).

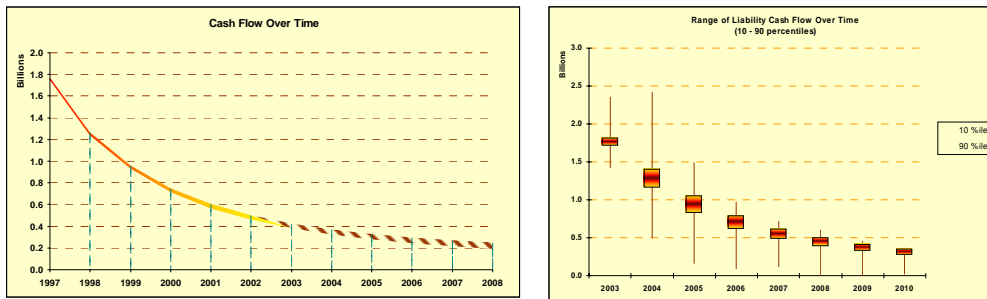


Fig. A.4.4

Decision makers rarely have their options conveniently ranked for them in one numeric dimension (i.e., "Net Profit"); they usually have to take into account multiple considerations (i.e., "Profit vs. Risk"), attempting to do **multidimensional ranking**. Geometrically speaking, their challenge is to say which one of the several points on the 2-D plane is "the best" one.

### Actuarial IQ

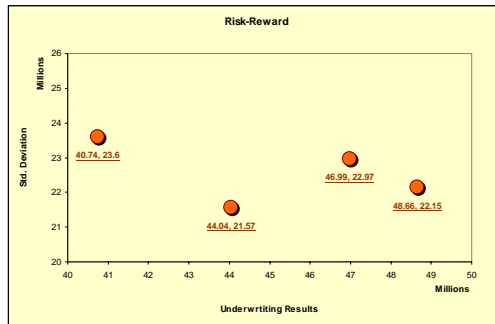


Fig. A.4.5

If the decision maker can formalize his preferences and express them as a so-called “goal function” (i.e., “the goal is to maximize risk to return ratio”), then display of the data can be optimized for that goal-seeking purpose. Taking a cue from a geographical map, the report designer may draw isolines (where goal function remains constant) and shade areas in between differently (for different values of these constants).

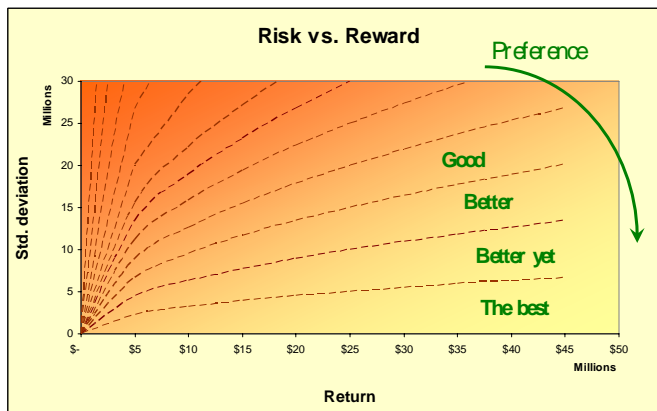


Fig. A.4.6

Placing 2-D points on such a map may significantly assist in selecting “the best” option.

## Actuarial IQ

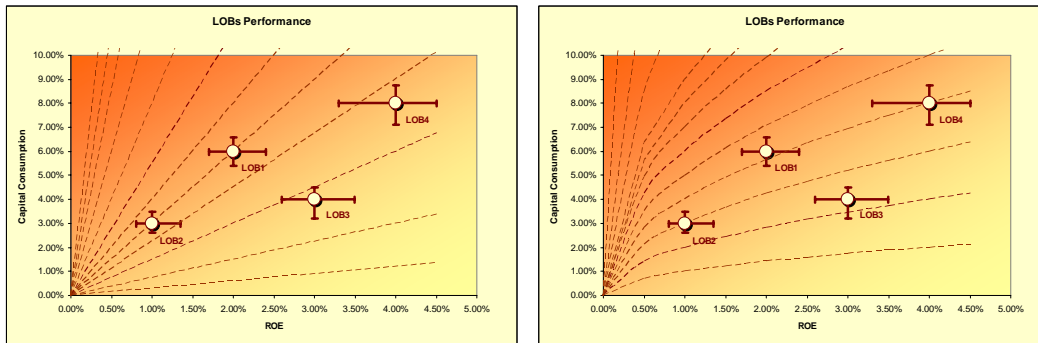


Fig. A.4.7

Fig. A.4.7 places points from the Fig A.4.5 on the grid of iso-lines from Fig. A.4.6. With a visual cue from the gradient (reddish areas are bad, yellowish areas are better) it is evident that lower right point labeled “LOB3” should be ranked #1 for a given choice of goal functions.

While use of visualization techniques requires some effort from the report designers and some training for the report readers, the payoff in interpretation quality (and, consequently, in decision accuracy) is considerable.

### Appendix B: Data cleansing steps (addendum to section 3.6.1)

The following is based on *Improving Data Warehouse and Business Information Quality* ([23], pp. 237-282).

1. **Identify data sources** and select the most authoritative source (i.e., company policy database, company claims database, company bill review database, etc.). Note that this may not be as easy as selecting a single file: the most reliable version of different data elements could come from different files. Similarly, different sources may be more authoritative for a single data element in different circumstances. Best data are coming from the sources and processes that have the largest stake in the correctness of data (e.g., the accounting department may treat payments info more accurately than their claims management colleagues and vice versa for the case reserves data). Frequency and timeliness of updates may also serve as indicator of reliability (more recently updated data probably has been looked at and corrected, so it could be more accurate). Metadata (2.3) can help to identify the most authoritative files.
2. Extract and **analyze** source data **for anomalies**.

### *Actuarial IQ*

- a. Analyze the meaning of the data with source data subject matter experts. For example, confirm that “AccDate” is indeed a date of the accident and find out what “GrossNetPrem” means.
  - b. Document the definitions, domain value sets, and business rules for each data element as used in its source file.
  - c. Extract a representative data sample and analyze it to confirm that the actual data is consistent with its definition and to discover any anomalies in how the data was used and what these incorrect entries mean. The objective is to discover undocumented values and their meanings.
3. **Standardize** the contents of data attributes: the definition and domain value sets for each standardized data attribute become the authoritative enterprise definition. Format nonstandardized data into standardized data elements with standardized domain value sets. For example, if certain files were using “2” for “married” but the enterprise definition is “M” for “married,” then replace the “2”s with “M”s.
4. **Correct and Complete Data.** Improve the quality of the existing data by correcting inaccurate or nonstandard data values and finding and capturing missing data values. The objective is to improve the quality of the data to the highest level.
- a. Identify missing data and obviously incorrect or suspect data (using, for example, EDA techniques described in [3.1](#))
  - b. Prioritize data to be cleansed based on value of correct data compared to correction costs.
  - c. Determine how to handle suspect data. The most efficient approach for simple but massive cleansing is to use automated transformation routines that can modify data according to business rules. However, where the suspect data is critical and the investigation is economically feasible, the best handling of suspect data is investigation and request for correct data from the source. Alternatives:
    - i. Reject the data.
    - ii. Accept the data without change.
    - iii. Accept the data without change but document that it is suspect.
    - iv. Accept the data but estimate the correct or approximate values based on other

### *Actuarial IQ*

related attributes. In this case, make sure that data are flagged as “estimated” and that impact on the intended use of data is tolerable.

- v. The best handling of suspect data is investigation and request for correct data from the source.
  - d. Implement the selected approach(s) for cleansing.
  - e. Document what was done and why.
5. **Eliminate duplicates.**
- a. Establish criteria to identify duplicate data records.
  - b. Determine impact of incorrectly consolidating multiple different records into one.
  - c. Determine matching techniques to use.
  - d. Look for intra-file duplicate records.
  - e. Look for inter-file duplicate records.
  - f. Investigate duplicates to make sure they are in fact duplicate records.
  - g. Document the matching and merge rules in the data map of source to target.
  - h. Establish a control mechanism to cross-reference duplicate occurrences in multiple files when primary key cannot be kept identical across files.
  - i. Examine and re-relate data related to the old records being consolidated to the new record.
  - j. Maintain an archive of the original source data for an appropriate length of time for error recovery purposes.
6. Analyze data for **patterns of errors**. The objective here is to leverage the knowledge of the data cleansing work to discover patterns of data errors and eliminate the most significant problems caused by data errors, as well as the most significant causes of errors. Analyze results to understand the kinds of errors, frequencies, and the cost impacts of the errors on the business.
- a. List and analyze examples of various kinds of data anomalies.
  - b. List two or three representative examples of each type of data defect.
  - c. Categorize the information quality problems and patterns.

### *Actuarial IQ*

- d. Estimate the frequency of each information quality problem.
  - e. Estimate the relative costs or impacts of each information quality problem, if possible.
  - f. Summarize the impact by data defect type.
7. **Map** the corrected data into its data file. Prepare the data for loading into the warehouse or target database, and include converting or formatting the cleansed consolidated data into the new data architecture. This step requires:
- a. Cleansed and standardized data.
  - b. Data from external information sources for integration with internal data.
  - c. Business rules governing the source data.
  - d. Business rules governing the target data warehouse data.
  - e. Transformation rules governing the transformation process.
8. **Optimize** data warehouse performance by determining and storing derived data (like triangles or other pre-aggregations) for the most frequently asked queries requiring complex calculations.
9. **Audit** and control data extraction, transformation, and loading. **Update** these procedures as necessary. Once the above steps are completed, this step is a matter of implementing procedures to assure the processes are performed as specified and kept up to date. See [3.2](#) for more information on data audits.

#### **Appendix C: Data defect prevention (addendum to section [3.6.2](#))**

The following is based on *Improving Data Warehouse and Business Information Quality* ([23], pp. 285-310). The systematic approach for preventing data defects from recurring contains 6 organizational steps:

0. Analysis and identification of all data processes and procedures in the company with particular focus on those processes associated with defective data.
1. Selecting a particular process for improvement.
2. Brainstorming and developing an improvement plan.

### *Actuarial IQ*

3. Implementing improvements in a controlled manner and confirming that improvements do solve a real problem.
4. Evaluation of impact against preset success criteria.
5. Rollout of improvement through the entire company along with training and documentation.

As an example, consider a process of building loss triangles for actuarial analysis of a company's reserves using accurate data.

Step 1 would first involve identifying and selecting those processes associated with the most significant payoff based on the impact of the data errors. Then identify the data sources (company's warehouse) and data owners (data management team) along with data consumers (reserving actuaries). This step ends with an assignment of improvement project sponsor and a team accountable for process changes.

Step 2 is the essence of the defect prevention activity. It requires identification of the root of the problem (for example, miscoding of "Line of Business" attribute), a feasible technical solution (integrity check, link to line-subline table, update of incorrect values), a plan to implement a solution (build line-subline table, write SQL queries, test on a sample of data, deploy), a measure of success (% of incorrect records) and costs associated with fixing the problem versus not fixing it (programming time versus errors in reserves and insolvency).

Step 3 includes not just implementation itself but also testing, documentation of the changes, and training of the personnel.

Step 4 consists of measurement of success defined in Step 2 or analysis of failure with possible repeat of Steps 2 and 3.

Step 5 involves generalization of the improvement with an attempt to apply it to all applicable areas in the company (fixing triangles used for pricing).

Making improvements in a systematic rather than haphazard manner will help to prevent more errors more effectively and more successfully with less effort and lower cost.

## 5. REFERENCES

- [1] Campbell, R.; Francis, L.; Prevosto, V.; Rothwell, M; Sheaf, S., “Report of the Data Quality Working Party” 2006, <http://www.actuaries.org.uk/files/pdf/proceedings/giro2006/Francis.pdf>
- [2] Actuarial Standards Board. *Actuarial Standard of Practice No. 23: Data Quality, revised edition*. Schaumburg, Illinois: American Academy of Actuaries, 2004.
- [3] CAS Committee on Management Data and Information. “White Paper on Data Quality.” *CAS Winter 1997 Forum*. 145-168.
- [4] Francis, Louise A. “Dancing with Dirty Data: Methods for Exploring and Cleaning Data.” *CAS Forum Winter 2005*: 198-254.
- [5] Copeman, P.; Gibson, L; Jones, T.; Line, N.; Lowe, J.; Martin, P.; Mathews, P.; Powell, D., “A Change Agenda for Reserving: A Report of the General Insurance Reserving Issues Task Force” 2006, [www.actuaries.org.uk](http://www.actuaries.org.uk)
- [6] CAS Data Management Educational Materials Working Party. “Survey of Data Management and Data Quality Texts.” *CAS Winter 2007 Forum*. 273-306.
- [7] Dasu, Tamraprni and Theodore Johnson. *Exploratory Data Mining and Data Cleaning*. Wiley, 2003.
- [8] Watson, Richard T. *Data Management: Databases and Organization, fifth edition*. New Jersey: Wiley, 2005.
- [9] Immon, William and Claudia Imhoff and Ryan Sousa. *Corporate Information Factory, second edition*. New Jersey: Wiley, 2000.
- [10] Moncher, Richard B. “Study Note: NCCI Data Collection Calls and Statistical Plans.” *CAS Exam Study Note Casualty Actuarial Society - Arlington, Virginia*, 1-17
- [11] Prevosto, Virginia R. “Statistical Plans for Property/Casualty Insurers.” *Casualty Actuarial Society Discussion Paper Program Casualty Actuarial Society - Arlington, Virginia*, May 1997, 201-216
- [12] Prevosto, Virginia R. “Study Note: ISO Statistical Plans.” *CAS Exam Study Note Casualty Actuarial Society - Arlington, Virginia*, 1997, 1-21
- [13] Cody, R. *Cody's Data Cleaning Techniques Using the SAS Software*, SAS Institute, 1999.
- [14] Hartwig, Frederick and Brian E. Dearing. *Exploratory Data Analysis*. Beverly Hills: Sage Publications, 1979.
- [15] Richard Scheaffer, William Mendenhall III and Lyman Ott. *Elementary Survey Sampling, Fifth Edition*. Wadsworth. 1996.
- [16] Insurance Services Office, Inc. *Strength in Numbers – A Total Data Quality Audit Program for Your Company*. Jersey City, New Jersey, 1989.
- [17] Insurance Services Office, Inc. *Quality of Data Audit Guide*. Jersey City, New Jersey, 1978.
- [18] Stuart A. Klugman, Harry H. Panjer, Gordon E. Willmot. *Loss Models: from data to decisions*. Wiley, 1998
- [19] Kit, Edward. *Software Testing in the Real World*. New York: Addison-Wesley, 1995.
- [20] Redman, Thomas C. *Data Quality, the Field Guide*. Boston: Digital Press, 2001.
- [21] Loshin, David. *Enterprise Knowledge Management*. Morgan Kaufman, 2001.
- [22] Olson, Jack E. *Data Quality: the Accuracy Dimension*. Morgan Kaufman, 2003.
- [23] English, Larry P. *Improving Data Warehouse and Business Information Quality*. New York: Wiley, 1999.
- [24] Casualty Actuarial Society. *Statement of Principles Regarding Property and Casualty Insurance Ratemaking*. 1988.
- [25] Insurance Services Office, Inc. *Data Management Best Practices*. Jersey City, New Jersey, 2003.
- [26] Popelyukhin, Aleksey S. “On Hierarchy of Actuarial Objects: Data Processing from the Actuarial Point of View”, *CAS Forum Winter 1999*: 219-237.
- [27] Popelyukhin, Aleksey S. “Let Me See: Visualizing Actuarial Information.” *CAS Forum Winter 2001*: 399-425.
- [28] Popelyukhin, Aleksey S. “Presenting DRM Results: Helping Executives Make Sense of DRM.” A report of CAS Working Party on Executive Level Decision Making Using DRM, 2005.
- [29] Popelyukhin, Aleksey S. “Rainy Day: Actuarial Software and Disaster Recovery.” *CAS Forum Winter 2003*: 55-73.

Abbreviations and Glossary

**ASB**, Actuarial Standard Board

**ASOP**, Actuarial Standard of Practice

**ASOP No. 23**, Actuarial Standard of Practice No. 23

**CAS**, Casualty Actuarial Society

**Categorical data**, (as opposed to numerical data) data whose values correspond to a specific category or label. Examples include alphanumeric data such as claimant state or NCCI injury code.

**Data attribute** is a characteristic of an object or an observation. Data attribute consists of a name and a value and is usually stored in a field in a data record. For example, attribute's name: "Date of Accident", value: "January 1, 2000".

**Data cube**: a multi-dimensional representation of the data. Dimensions are usually constructed from the categorical data, while cube content is usually some aggregate function (sum, count, max) of numerical data. For example, Excel's pivot table is a 2-dimensional projection of the data cube.

**Data domain**, the set of values valid for a given data element. For example, data domain for the "Gender" data element is a pair {"Male"; "Female"}.

**Data element** or **data entity**, the smallest unit of data record that has meaning to a knowledgeable worker. Data element is usually a value of a data attribute or a reference to another record in a (more detailed) table. For example, a loss record may contain the following data elements: values of the "Date of Accident" and "Line of Business" attributes and "Policy ID" reference to a record in "Policies" table.

**Data record** or **database record** is a (structured) row in a database table that represents a single object or observation as a collection of related data elements (stored as fields). For example, a record for insurance policy may consist of "Policy ID," "Inception Date," "Expiration Date," and "Premium" data elements.

**EDA**, Exploratory Data Analysis

**Field**, a column in a database table that stores a value of a single data attribute or a reference (key) to a record in another table.

**GIRO**, General Insurance Research Organization

**GRIT**, General insurance Reserving Issues Taskforce

**IDMA**, Insurance Data Management Association

**IT**, Information Technology

**ISO**, Insurance Services Office, Inc.

**MDDDB**, Multi-dimensional Database

**MGA**, Managing General Agency

**NAII**, National Association of Independent Insurers

**NCCI**, National Council on Compensation Insurance

**OLAP**, On-Line Analytical Processing, a mechanism for efficient analytical queries. OLAP heavily relies on data cubes as data structure and pre-aggregations as a way to speed up queries.

**Regulator**, Insurance is regulated by state insurance departments. Financial statements, rates, licenses to write business, etc. are monitored by regulators, including actuaries, who work for insurance departments.

**SQL**, Structured Query Language, a computer language to retrieve (place and modify) information from a (relational) database.

**Statistical Agent**, an organization that helps insurers satisfy legal requirements for reporting data to regulators. The statistical agent processes data submitted by insurers, performs data quality checks on the data, consolidates the data across insurers, and provides aggregate data compilations to state insurance departments on the behalf of the insurers.

**TPA**, Third Party Administrator, a company managing insurance claims, one of main sources of actuarial data.

**VBA**, Visual Basic for Applications, a programming language implemented in many applications, most notably in Microsoft Office.

**XML**, eXtensible Markup Language, a language that combines text with descriptive information about that text. For example, XML would store Excel's cell value along with the formula that generated that value.

## **Biographies of Working Party Contributors**

**Keith Allen** is the associate actuary for United Educators and is responsible for underwriting duties within the public school sector and general corporate actuarial issues. Allen has 13 years of experience in the insurance industry as an underwriter, claims adjuster, and actuary. Keith previously worked for Tillinghast-Towers Perrin as an actuarial specialist where he did reserving, pricing, and forecasting for various public and private entities. Prior to that, Allen worked as a claims adjuster and underwriter for State Farm Insurance where he helped develop the “Reinspection Program” used to assess coastal risks. Before joining the insurance industry, Allen was a teacher at Bellaire High School in Houston, TX. Allen holds a bachelor’s degree in mathematics from the University of Texas and is an Associate of the Casualty Actuarial Society.

**Robert Campbell** is Assistant Vice President, Actuarial Services at Lombard Canada in Toronto, Canada. He has a Bachelor of Mathematics in Business Administration from the University of Waterloo. He is a Fellow of the CAS and a Fellow of the Canadian Institute of Actuaries. He is chair of the Data Management Educational Materials working party, participates on the CAS Committee on Data Management and Information, and was a participant on the 2006 GIRO Data Quality working party.

**Louise Francis** is a Consulting Principal at Francis Analytics and Actuarial Data Mining, Inc. She is involved in data mining projects as well as conventional actuarial analyses. She has a BA degree from William Smith College and an MS in Health Sciences from SUNY at Stony Brook. She is a Fellow of the CAS and a Member of the American Academy of Actuaries. She serves on several CAS committees /working parties and is a frequent presenter at actuarial and industry symposia. She is a four-time winner of the Data Quality, Management and Technology call paper prize, including one for “Dancing with Dirty Data: Methods for Exploring and Cleaning Data (2005).”

**Dave Hudson** is an Actuary for The Travelers in Hartford, CT. He has a MS degree in Mathematics from Washington State University in Pullman, WA. He is a Fellow of the CAS and a Member of the American Academy of Actuaries. He is also a member of the CAS Committee on Data Management and Information.

**Gary W. Knoble** is a consultant for insurance data management and professional education. He serves as a senior advisor to the Insurance and Finance Professional Education Consulting (Beijing) Co. Ltd . (Bao Rong), and the US Asia Business and Financial Services group assisting them in providing educational services to the Chinese insurance industry. Prior to his retirement in January 2006 he was a Vice President of The Hartford Financial Services Group where he directed the Data Management Division of the Actuarial Department. He is a past president of the Insurance Accounting and Systems Association (IASA). He is a founding member and past president of the Insurance Data Management Association and currently serves on the Board as Vice President of Professionalism. He has served on the ACORD P&C Steering Committee, the International Association of Industrial Accident Boards and Commissions (IAIABC) Electronic Data Interchange Council and Associate Member Council, ISO’s Ad Hoc Data Reporting Group, and for many years chaired the Surety Association of America’s Statistical Committee. He serves as a Visiting Professor at Nanjing Audit University and as an advisor to the Actuarial Alumni Association of the University of Science and Technology of China in Hefei. Mr. Knoble is a recipient of several distinguished industry awards including the IASA President’s award in 2002, an outstanding achievement award from the IAIABC in 2004, and a special award for creativity, diplomacy, fidelity and vision from the IDMA in 2005. In 2005, IDMA announced the creation of the Gary Knoble Award that will be given periodically to an individual who has made an outstanding career contribution to the field of Data Management. A native of Salt Lake City, Utah, Mr. Knoble is a graduate of Yale University with a major in International Political and Economic Institutions.

**Rudy Palenik** is the Commercial Actuary at Westfield Insurance Group in Westfield Center, Ohio. He is responsible for the development of rates for all the commercial lines of business. He has a degree in Math from Marquette University in Milwaukee, Wisconsin and is a Fellow of the Casualty Actuarial Society and a member of the American Academy of Actuaries. Rudy participates on a number of CAS committees including: Data Management and Information, Actuarial Education and Research Foundation, Research Paper Classifier and University Liaison.

**Aleksey Popelyukhin** is a Vice-President of Information Systems with the 2 Wings Risk Services and a Head of

### *Actuarial IQ*

Quantitative Analytics Group with the Wall Street North Consulting in Stamford, Connecticut. He holds a Ph.D. in Mathematics and Mathematical Statistics from Moscow University (1989). Aleksey actively participates in CAS research and is frequent presenter on CAS conferences. CAS recognized Aleksey's contributions by awarding him the very first prize in "Data Management" papers competition and inviting him to the very first Working Party (on presentation of DFA/DRM results). In addition to numerous publications Aleksey helps to advance actuarial science by building convenient software tools for actuaries such as Triangle Maker®, Affinity and Actuarial Toolchest™. For those actuaries having troubles explaining statistics to the management, Aleksey built a DRM presentation template available from CAS website. And for those who have troubles fitting clean models to dirty data Aleksey developed advanced data quality service called Data Quality Shield<sup>SM</sup>. Aleksey is currently developing an integrated pricing/reserving/DRM computer system for reinsurance called "SimActuary" and also an action/adventure computer game tentatively called "Actuarial Judgement."

**Virginia R. Prevosto** is a Vice President at Insurance Services Office, Inc. Ms. Prevosto is a Phi Beta Kappa graduate of the State University at Albany with a Bachelor of Science degree in Mathematics, *summa cum laude*. She is a Fellow of the CAS and a Member of the American Academy of Actuaries. She serves as General Officer of the CAS Examination Committee and as liaison to various other CAS admission committees. She also serves on the CAS Committee on Management Data and Information. In the past, Ms. Prevosto also served on the Data Quality Task Force of the Specialty Committee of the Actuarial Standards Board that wrote the first data quality standard of practice. Virginia has been a speaker at the Casualty Loss Reserve Seminar on the data quality standard and to various insurance departments on data management and data quality issues. Ms. Prevosto authored the paper "Statistical Plans for Property/Casualty Insurer" and "Study Note: ISO Statistical Plans" and co-authored "For Want of a Nail the Kingdom was Lost – Mother Goose was right: Profit by Best (Data Quality) Practices" for the IAIDQ.

# Capital Allocation by Percentile Layer

Neil M. Bodoff, FCAS, MAAA

---

## Abstract

**Motivation.** Capital allocation can have substantial ramifications upon measuring risk adjusted profitability as well as setting risk loads for pricing. Current allocation methods that emphasize the tail allocate too much capital to extreme events; “capital consumption” methods, which incorporate relative likelihood, tend to allocate insufficient capital to highly unlikely yet extremely severe losses.

**Method.** In this paper I develop a new formulation of the meaning of holding capital equal to the Value at Risk. The new formulation views the total capital of the firm as the sum of many percentile layers of capital. Thus capital allocation varies continuously by layer and the capital allocated to any particular loss scenario is the sum of allocated capital across many percentile layers.

**Results.** Capital allocation by percentile layer produces capital allocations that differ significantly from other common methods such as VaR, TVaR, and coTVaR.

**Conclusions.** Capital allocation by percentile layer has important advantages over existing methods. It highlights a new formulation of Value at Risk and other capital standards, recognizes the capital usage of losses that do not extend into the tail, and captures the disproportionate capital usage of severe losses.

**Availability.** To discuss further, please contact the author at [neil.bodoff@willis.com](mailto:neil.bodoff@willis.com) or [neil\\_bodoff@yahoo.com](mailto:neil_bodoff@yahoo.com)

**Keywords.** Capital Allocation; Percentile Layer of Capital; Value at Risk; Enterprise Risk Management; Risk Load; Risk Adjusted Profitability

---

## 1. REQUIRED CAPITAL, REQUIRED RATE OF RETURN, AND CAPITAL ALLOCATION

How much capital should an insurance firm hold? And what rate of return must the firm achieve on this capital? While these questions are of critical importance to the firm, external forces in the operating environment often dictate the answers. For example, regulators and rating agencies greatly influence the amount of capital the firm must hold; in addition, investors influence both the amount of capital the firm holds and the required rate of return on this capital. Therefore, the issues of the amount of capital and the required rate of return on capital are often ultimately beyond the decision making power of the company; rather, they are demands that the operating environment imposes upon the firm.

Given that a firm must hold a certain amount of capital, the firm essentially incurs a firm-wide “overhead” cost related to the required rate of return on this capital. Management often desires to allocate this cost, like other overhead costs, to subsets of the firm such as subsidiaries, business units, and product lines. How should the firm allocate the cost of required return on capital? This is the question of “capital allocation”.

## **1.1 Why is Capital Allocation Important?**

How a firm allocates capital, similar to other cost allocation decisions, can significantly affect the measured profitability of a particular line of business. Moreover, allocating capital can affect target pricing margins and the volume of business the company writes in each line of business and product type. As a result, the topic is critically important and often the subject of contentious debate among the heads of the firm's various business units.

## **1.2 Defining the Scope of the Problem**

We will restrict our discussion to the situation of a publicly traded insurance company that writes property catastrophe business, both insurance and reinsurance, covering several perils around the world; we will exclude long tail casualty business in an attempt to simplify our discussion to a single year time horizon problem. We will assume that investors require that the firm holds capital based upon the Value at Risk (VaR) at the 99th percentile and that the required return can be expressed as an annual percentage rate of return on this amount of capital. The issue we grapple with here relates only to allocation.

## **1.3 Allocating Capital to Users of Capital**

Mango [4] has stressed that the entire capital of the firm is available to pay the claim of any single policy. Thus, the required rate of return on capital is a cost that accrues on the total firm level, and Kreps [1] has clarified that capital allocation is really the allocation of the required rate of return on capital. Mango [3] also has highlighted the connection between allocating capital and broader issues of cost allocation. Therefore, similar to other cost allocation situations, we want to connect the firm-wide cost of capital to those subsets of the firm which require the company to incur this cost: essentially, to match the expenditure to its source. Namely, we desire to allocate the cost of capital to those business units, products, perils, reinsurance contracts, and individual insurance policies that contribute to the loss scenarios that "use" capital.

## **1.4 So Who "Uses" Capital? Investigating Value at Risk (VaR) and Tail Value at Risk (TVaR)**

In our situation, the company must hold capital based upon Value at Risk (VaR) at the 99th percentile. The traditional view of this requirement is that the firm is holding capital in order to pay for a catastrophically bad scenario (the 99th percentile loss), but is not concerned with other loss scenarios that are either greater than or less than this VaR (99%) scenario. Thus Kreps [1] and



### *Capital Allocation by Percentile Layer*

“Wind + EQ” event; using Kreps’s “co-measures” approach, we can then further allocate the capital for the “Wind + EQ” event to its components: Wind [ $= 49.75\% = 99 / (100 + 99)$ ] and EQ [ $= 50.25\% = 100 / (100 + 99)$ ]. In total EQ would receive approximately 90% [ $= 80\% + 50.25\% * 20\%$ ] and Wind would receive roughly 10% [ $= 49.75\% * 20\%$ ]. But again, the substantial possibility of a standalone Wind event of 99M has no significance?

Method #3 (“coTVaR”): Another approach might be to use the TVaR measure for loss events  $\geq 100M$  to allocate. Then the EQ event receives allocation proportional to  $80\% * 100M$  and the “Wind + EQ” event receives allocation proportional to  $20\% * 199M$ . Using Kreps’s co-measures again, ultimately EQ receives 83.5% and Wind 16.5%; but again, we will allocate zero capital based upon the “Only Wind” event of 99M, which is much more likely to use capital and nearly as large of a loss as the “EQ only” event!

It seems intuitively clear that Wind is not receiving the appropriate capital allocation in this situation. More broadly, tail based methods in general have been criticized for ignoring loss scenarios below the tail threshold (e.g., Wang [7]).

## **2. REFORMULATING AND CLARIFYING VALUE AT RISK (VaR)**

It therefore seems appropriate to reformulate and clarify what it means for a firm to hold capital at the 99<sup>th</sup> percentile, or VaR (99%). While the prior formulation suggests that the firm holds sufficient capital “**for** the 99th percentile loss”, I believe that a better formulation of the meaning of the VaR capital requirement is that the firm holds sufficient capital “**even for** the 99th percentile loss”. Once we focus on VaR requiring sufficient capital “**even for** the 99th percentile loss”, we can see that this capital amount is intended to also cover losses at lower percentiles as well; thus, we must allocate capital and its cost even to loss events that fall below the VaR threshold.

We can use an analogous argument to reformulate TVaR as well. Specifically, using TVaR (99%) to set capital means we are holding capital “**even for** the average loss beyond the 99th percentile”, but not “only for” these events. Beyond VaR and TVaR, the same line of reasoning may be appropriate when interpreting other capital benchmarks as well.

### **2.1 Ramifications of New Formulation of VaR**

What are some of the ramifications of our formulation that holding capital equal to VaR (99%) means holding sufficient capital “even for a 99th percentile loss”?

### *Capital Allocation by Percentile Layer*

It would appear to follow that we need to think about capital allocation by percentile layer. In other words, why does the firm hold capital equal to the 99th percentile loss rather than the lower amount of the 98th percentile loss? The difference between the required capital amounts at these two percentile losses can be attributed solely to those loss events that outstrip the 98th percentile. Similarly, the difference between the amount of capital at the 98th percentile loss and the 97th percentile loss can be attributed solely to those losses that exceed the 97th percentile. And so on...

Therefore, allocation of capital to loss scenarios would appear to require calculations that vary by layer of capital.

### **3. DEFINING A “PERCENTILE LAYER OF CAPITAL”**

Thus, we can define a “Percentile Layer of Capital” as follows. Define percentile  $\alpha$ , increment  $j$ , and percentile  $\alpha + j$  on the interval  $[0, 1]$ . Then

$$\text{Percentile Layer of Capital } (\alpha, \alpha + j) = \text{Required Capital at percentile } (\alpha + j) - \text{Required Capital at percentile } (\alpha) \quad (3.0)$$

We can also define a “Layer of Capital” as follows. Define amounts  $a$  and  $b$ , then

$$\text{Layer of Capital } (a, a + b) = \text{Capital equal to amount } (a + b) - \text{Capital equal to amount } (a) \quad (3.1)$$

For example, assume we have simulated 100 discrete loss events and the 78th loss (ordered from smallest to largest) is 59M and the 77th loss is 47M, then the percentile layer of capital (77%, 78%) = 59M – 47M = 12M.

#### **3.1 Refining the Percentile Layer of Capital**

Note that we can set Capital ( $\alpha$ ) = any function of (VaR ( $\alpha$ )). For example, if we want a 99th percentile loss to consume no more than 50% of capital, then

$$\text{VaR (99\%)} = 50\% * \text{Capital (99\%)} \text{ and}$$

$$\text{Capital (99\%)} = 2 * \text{VaR (99\%)}$$

### *Capital Allocation by Percentile Layer*

For ease of use, we will assume that the capital required at a loss percentile will equal that loss amount:

$$\text{Capital } (\alpha) = \text{VaR } (\alpha) = \text{loss percentile } (\alpha)$$

Also, we will assume that  $j$ , which equals the “width” or “increment” of a layer’s percentiles between lower and upper bounds, equals  $1/n$ , where  $n$  = number of available discrete values. For example, if we have 100 simulation outputs, then the layer increment  $j = 1\%$ , and if we have 1000 simulated values, then  $j = 0.1\%$ .

### **3.2 Allocating a Percentile Layer of Capital to Loss Events**

We can see that each layer of capital is potentially used or depleted (or “consumed” in Mango’s [4] terminology) by loss events that exceed the lower bound of the layer, but not by loss scenarios that fall short of the lower bound of the layer (i.e., those losses that do not penetrate or “hit” the layer). Thus, it is desirable to allocate each layer of capital only to those events that penetrate the layer. Another critical consideration is that some of the losses that penetrate the layer are more likely to do so than others. Therefore, each event (i) that penetrates the layer of capital receives an allocation based upon its conditional exceedance probability.

Conditional Exceedance Probability for event (i) = Probability of event (i) that penetrates the layer of capital / Probability of all events that penetrate the layer of capital

Thus, for any layer of capital, we take the amount of capital (or the “width” of the layer), we allocate this amount of capital only to loss events that penetrate the layer, and we calculate the allocation percentages based upon each loss event’s conditional probability of penetrating the layer. The allocation percentages, by definition, sum to 100% on any layer.

After performing the allocation of each layer of capital (from zero up to the required VaR capital amount - but not beyond it), we will have allocated 100% of the capital to loss events.

Many loss scenarios will penetrate several different percentile layers of capital and therefore receive varying allocations of capital from many layers of capital. The total capital allocated to any particular loss event is simply the total, summed over all layers of capital that the loss event penetrates, of the capital allocated on each individual layer. As an example, take the 83<sup>rd</sup> percentile loss event. On each layer of capital (from zero up to the 83<sup>rd</sup> percentile layer of capital but not beyond) it receives varying amounts of allocated capital; sum across all of these layers to calculate total capital allocated to this event. Of course, each loss “event” or “scenario” may be an

### *Capital Allocation by Percentile Layer*

accumulation of losses from several business units, policies, and/or perils. But as Kreps [1] has shown, once we have the total allocated capital for a loss scenario, we can then allocate to the subcomponents based upon their contributions to the total.

#### **3.2.1 Applying Capital Allocation by Percentile Layer to Thought Experiment #1**

In this section we will apply the procedure of capital allocation by percentile layer to the simplified numbers of Thought Experiment #1.

In Thought Experiment #1, there are 4 potential scenarios:

- 1) 76% neither peril occurs, loss = 0
- 2) 19% only Wind occurs, loss of 99M
- 3) 4% only EQ occurs, loss of 100M
- 4) 1% both Wind and EQ occur, loss of 199M

We hold capital equal to VaR (99%) = 100M. The layer of capital of 1M x 99M can only be penetrated (or “depleted” or “consumed”) by event #3 or #4. Event #3, the “Only EQ” event, has a conditional exceedance probability of 80% [ $4\% / (4\%+1\%)$ ]. Event #4, the “Wind and EQ” event, has conditional exceedance probability of 20%. Therefore, we allocate the 1M in layer capital (100M – 99M) as follows:

- 80% for EQ event,
- 20% for Wind + EQ event
- 0% for Wind only event

The next layer of capital, 99M x 0, can be used by all 3 loss events.

- “Only Wind” event has conditional exceedance probability of 79% [ $19\% / (19\%+4\%+1\%)$ ]
- “Only EQ” event has conditional exceedance probability of 17% [ $4\% / (19\%+4\%+1\%)$ ]
- “Wind and EQ” event has conditional exceedance probability of 4% [ $1\% / (19\%+4\%+1\%)$ ]

Therefore, the allocation of 99M in capital (99M – 0) is

- 79% for Wind
- 17% for EQ

### *Capital Allocation by Percentile Layer*

- 4% for Wind + EQ

The total capital allocation to loss event across both layers (namely, 1M x 99M and 99M x 0) is then

- “Only Wind” = 79% x 99M = 78.4M
- “Only EQ” = 17% x 99M + 80% x 1M = 17.3M
- “Wind + EQ” event = 4% x 99M + 20% x 1M = 4.3M

The total allocated capital = 78.4 + 17.3 + 4.3 = 100 = VaR(99%)

The loss event of “Wind + EQ” can then be allocated further to the underlying perils that contribute to the loss event (per Krepes [1]) as follows. In a “Wind + EQ” event, which receives a 4.3M allocation, Wind contributes 99M and EQ contributes 100M. Therefore, Wind % = (99/199) = 49.75%, EQ = (100/199) = 50.25%. The total allocation to peril is therefore

- Wind = 78.4M + 49.75% x 4.3M = 80.5M
- EQ = 17.3M + 50.25% x 4.3M = 19.5M

Comparing results of different methods at the 99th percentile, we see that

- Capital allocation by percentile layer = Wind 80.5%, EQ 19.5%
- coTVaR for all events  $\geq 100M$  = Wind 16.5%, EQ 83.5%

Thus, capital allocation by percentile layer creates a completely different allocation than coTVaR.

### **3.2.2 Thought Experiment #2**

In Thought Experiment #1, capital allocation by percentile layer produced allocations that are essentially proportional to the perils’ average loss. So does this imply that the procedure will always result in such an allocation? After all, it would seem problematic to always allocate capital in proportion to the average loss; catastrophic perils with the capability to produce severe losses should receive a greater allocation of capital, regardless of the “average” outcome. Thought Experiment #2 shows that capital allocation by percentile layer will respond appropriately in such a situation.

Again assume we are dealing with two perils:

- 1) Wind 20% chance of 50M loss, else zero
- 2) Earthquake (EQ) 5% chance of 100M loss, else zero

Note that for Wind the average loss = 10M and for EQ the average loss = 5M.

### *Capital Allocation by Percentile Layer*

Assume the perils are independent. Thus, the possible scenarios for portfolio loss are:

- 1) 76% probability that neither peril occurs, loss = 0
- 2) 19% probability that only Wind occurs, loss of 50M
- 3) 4% probability that only EQ occurs, loss of 100M
- 4) 1% probability that both Wind and EQ occur, loss of 150M

Using VaR (99%) as our capital requirement, we hold 100M of capital to pay for 99% of the loss events; only the rare, 1% chance of a Wind event plus an EQ event will exceed the capital. Applying capital allocation by percentile layer to the 50M x 50M layer of capital as well as the 50M x 0 layer of capital, we obtain the following allocation:

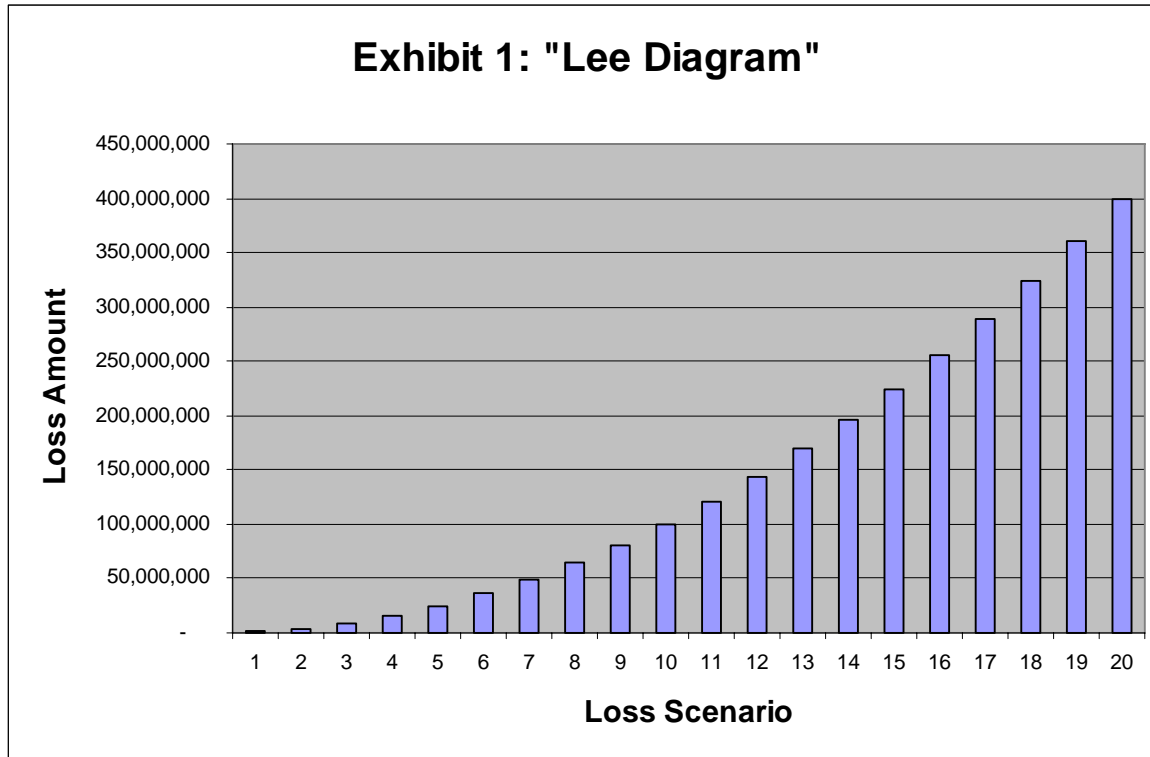
- Capital allocation by percentile layer = Wind 44%, EQ 56%
- Allocation in proportion to average loss = Wind 67%, EQ 33%

This example shows that capital allocation by percentile layer can produce unique allocations that are proportional neither to the average loss, nor to probability of occurrence, nor to standalone VaR.

## **4. GRAPHICAL DESCRIPTION OF CAPITAL ALLOCATION BY PERCENTILE LAYER - DISCRETE**

Let us view the “size of loss” distribution in graphical format to further clarify the approach; we will use sample numbers for simplicity. We will use “Lee Diagrams” (see Lee [2]), namely graphs where the loss scenario number (ordered in increasing size) is plotted on the X-axis and the loss amount is plotted on the Y-axis:

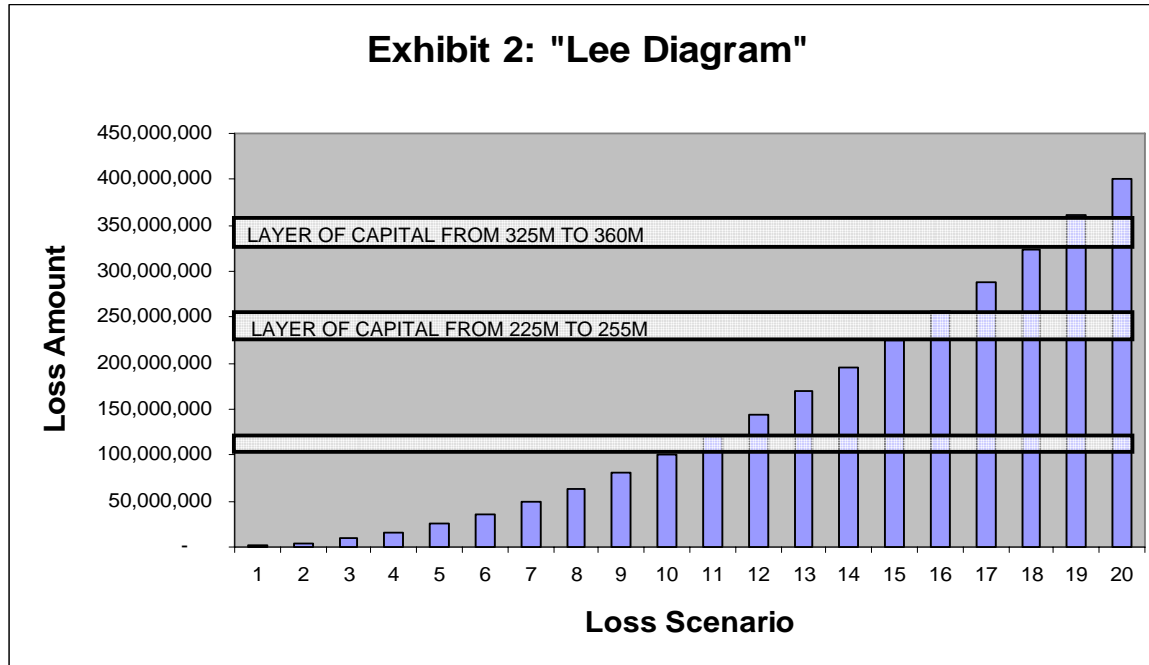
*Capital Allocation by Percentile Layer*



In this example (Exhibit 1) there are 20 loss scenarios; why is it that the firm needs to hold 360M of capital rather than just 100M of capital? It appears that loss scenarios 1 through 10, which are all less than or equal to 100M, do not require this "layer of capital". In contradistinction, loss scenarios 11 through 20, which exceed 100M, clearly do utilize this layer of capital in excess of 100M. Examining in further detail, we see that all of scenarios 11 through 20 utilize the 1M x 100M layer, but not all of them require the 1M x 200M layer, and even fewer require the 1M x 300M layer.

Thus, we must allocate each individual layer of capital to the loss events that penetrate the layer in proportion to the relative usage of the layer of capital; i.e., in proportion to the relative exceedance probability, as per Exhibit 2:

*Capital Allocation by Percentile Layer*



Numerical example:

- Loss scenario #19 is one of 2 events (scenarios 19 and 20) that require the 35M x 325M layer of capital.
  - Thus scenario #19 receives 1/2 allocation of this 35M of capital.
- Loss scenario #19 also is one of 5 events (scenarios 16 through 20) that require the firm to hold the 30M x 225M layer of capital.
  - Thus it receives 1/5 allocation of this 30M of capital.
- Apply the procedure to all layers; allocate to all loss events that exceed the lower bound of the layer via conditional exceedance probability.

Note that a loss event tends to receive a larger percentage allocation in the upper layers than in the lower layers for 2 reasons:

- 1) In the upper layers, we are allocating a full layer of capital to fewer loss events (i.e., the exceedance probability decreases as the loss amount increases); therefore, each event gets a larger share of the “overhead” of the total layer of capital.

### *Capital Allocation by Percentile Layer*

- 2) In the upper layers, we are allocating a wider layer of capital because the severity of each loss event tends to outstrip the prior loss event by a greater amount (i.e., the percentile layer of capital tends to widen as the loss amount increases). This behavior will depend, however, on the particular shape of the size of loss distribution.

## **5. GENERALIZATION OF CAPITAL ALLOCATION BY PERCENTILE LAYER TO DISCRETE LOSS EVENTS**

Let  $\text{VaR}(k) = \text{total required capital} = \sum [x(\alpha+j) - x(\alpha)]$

- $x(\alpha)$  is the loss amount at percentile  $\alpha$
- $j$  is selected percentile increment
- $\alpha$  sums from zero to  $(k - j)$

Allocation of capital **for each percentile layer of capital, across loss events**

- A Layer of Capital =  $[x(\alpha+j) - x(\alpha)]$
- Allocation of capital on layer  $[x(\alpha+j) - x(\alpha)]$  to loss event  $x(i) =$ 
  - $[x(\alpha+j) - x(\alpha)] * \text{Probability} (x = x(i)) / \text{Probability} (x > x(\alpha))$
- Sum across all loss events  $x(i)$  such that  $i > \alpha$

For an equivalent view, we can also look at the allocation of capital **for each loss event, across all percentile layers of capital =**

- A Layer of Capital =  $[x(\alpha+j) - x(\alpha)]$
- Allocation of capital on layer  $[x(\alpha+j) - x(\alpha)]$  to loss event  $x(i) =$ 
  - $[x(\alpha+j) - x(\alpha)] * \text{Probability} (x = x(i)) / \text{Probability} (x > x(\alpha))$
- Sum across all layers of capital such that  $\alpha \geq 0, (\alpha+j) \leq \min(i, k)$
- Note the  $\min(i, k)$  restriction. For any loss event, we sum across all layers of capital up to the amount of the given loss event, but not if the loss event exceeds the VaR threshold. In such a case, the loss beyond the VaR threshold does not generate additional allocated capital to the loss event.

## 6. GENERALIZATION OF CAPITAL ALLOCATION BY PERCENTILE LAYER TO CONTINUOUS LOSS FUNCTION

We can take the formulas for discrete loss events and generalize them into continuous versions.

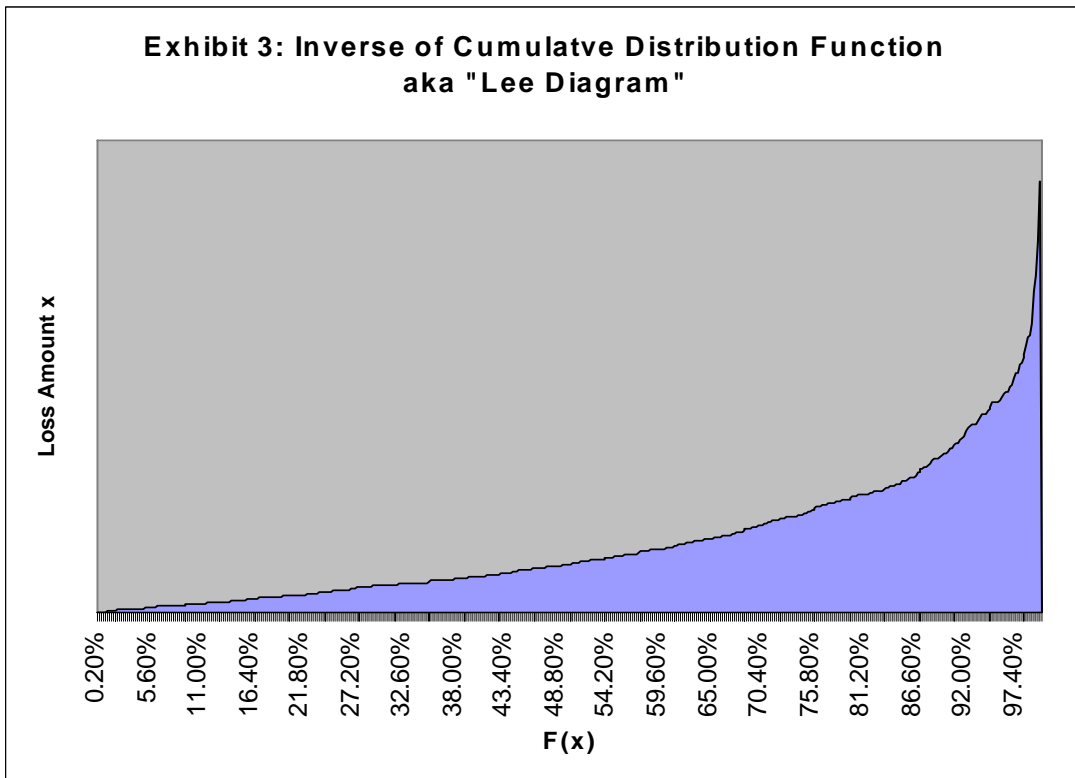
First, we will define the inverse function of  $F(x)$ , a function that accepts a percentile as input and returns the loss amount as output.

$$\text{Inverse function of } F(x) = F^{-1}(\alpha) = F^{-1}(F(x)) = x$$

$$\text{Derivative of } F^{-1}(F(x)) = dF^{-1}(F(x)) / dF(x) = dx / dF(x) = 1 / f(x)$$

$$\text{Incremental change in loss amount} = dx$$

$$\text{Incremental change in percentile} = dF(x)$$



In Exhibit 4, each horizontal bar is a layer of capital.

The length of the layer of capital, by definition, is 1.0.

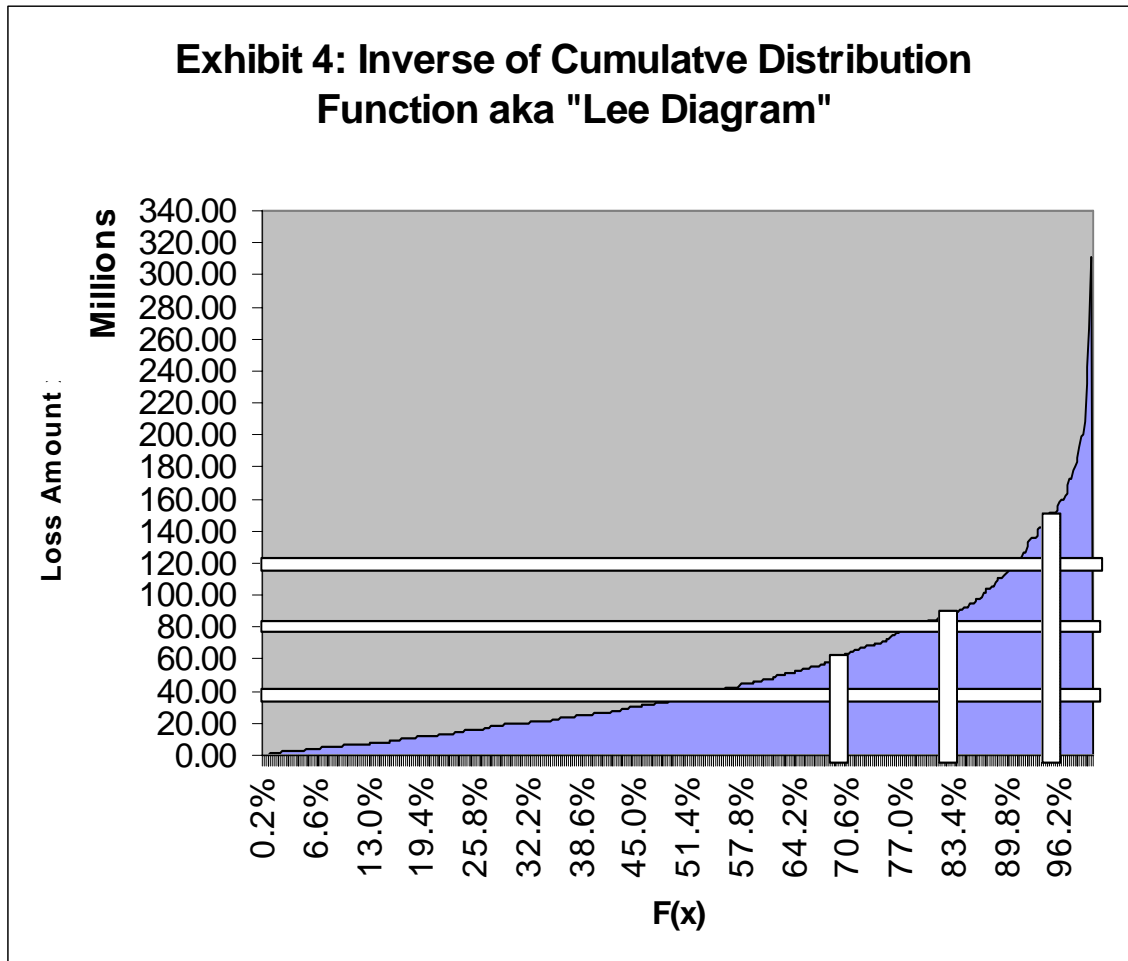
The infinitesimally small width of each layer of capital =  $dx$ .

### Capital Allocation by Percentile Layer

Each vertical bar represents a loss event.

The length = the loss amount =  $x$ .

The infinitesimally small width =  $dF(x) = f(x)dx$ .



#### 6.1 Two Alternative Views of Capital Allocation by Percentile Layer

We can view the capital allocation as a “horizontal procedure” which takes each layer of capital and allocates to all loss events which penetrate the layer.

We can also view the allocation as a “vertical procedure” which takes each loss event and allocates capital to it for all layers that it penetrates.

## 6.2 Approach #1: Horizontal then Vertical

Let  $x$  represent the loss amount and let  $y$  represent the capital.

First take an infinitesimally small layer of capital ( $y, y+dy$ ) and allocate it across loss events.

Integrate across all loss events  $x$  which penetrate the layer, from  $x = y$  to  $x = \infty$

$$\int_{x=y}^{x=\infty} f(x)/(1-F(y))dx \quad (6.0)$$

The allocation weights sum to 1 on each layer.

Then perform this procedure for all layers of capital:

$$\int_{y=0}^{y=VaR(99\%)} \int_{x=y}^{x=\infty} f(x)/(1-F(y))dxdy \quad (6.1)$$

Because capital is based upon the 99th percentile, there are no “layers of capital” above the 99th percentile to allocate, so we integrate  $y$  only up to  $VaR(99\%)$ .

The total allocated capital equals the total amount of capital, which is  $VaR(99\%)$ .

## 6.3 Approach #2: Vertical then Horizontal

Let  $x$  represent the loss amount and let  $y$  represent the capital.

Each loss event uses capital on many layers of capital ( $y, y+dy$ ).

Allocate to a loss event across each layer of capital:

$$\int_{y=0}^{y=x} f(x)/(1-F(y))dy \quad (6.2)$$

### Capital Allocation by Percentile Layer

Integrate  $y$  across all layers of capital less than or equal to the loss amount  $x$ .

If the loss amount  $x$  exceeds  $\text{VaR}(99\%)$ , we do not allocate additional layers of capital beyond  $\text{VaR}(99\%)$ ; in such a case when  $x > \text{VaR}(99\%)$ , we integrate as follows:

$$\int_{y=0}^{y=\text{VaR}(99\%)} f(x)/(1-F(y))dy \quad (6.3)$$

Then perform allocation across all loss events  $x$ :

$$\int_{x=x(0\%)}^{x=\infty} \int_{y=0}^{y=\min(x, \text{VaR}(99\%))} f(x)/(1-F(y))dydx \quad (6.4)$$

## 6.4 Formula for Allocating Capital to a Loss Event

The “vertical view” can provide some insight into the capital allocation to each loss event.

As we saw previously (equation (6.2)), for any loss event with amount  $x$  (assuming  $x$  is below the VaR threshold and therefore the allocated capital is not capped in any way), the Allocated Capital to loss event  $x = AC(x) =$

$$AC(x) = \int_{y=0}^{y=x} f(x)/(1-F(y))dy \quad (6.5)$$

Because we are integrating  $y$ , we can move  $f(x)$  outside the integral and rewrite the formula: Allocated Capital to loss event  $x = AC(x) =$

*Capital Allocation by Percentile Layer*

$$AC(x) = f(x) \int_{y=0}^{y=x} 1/(1-F(y))dy \quad (6.6)$$

For completeness, also recall that if the loss event is in the tail, namely  $x > \text{VaR}(99\%)$ , then

$$AC(x) = f(x) \int_{y=0}^{y=\text{VaR}(99\%)} 1/(1-F(y))dy \quad (6.7)$$

According to equation (6.6), the procedure of capital allocation by layer says that any loss event's allocated capital depends upon:

- 1) The probability of the event occurring (i.e.,  $f(x)$ ).
- 2) The severity of the loss event, or the extent to which the loss event penetrates layers of capital (i.e., the upper bound of integration is  $x$ , the loss amount).
- 3) The loss event's inability to share the burden of its required capital with other loss events (i.e.,  $\int 1 / (1-F(y)) dy$ ). We can think of this factor as the extent to which a loss event "sticks out" or is dissimilar in severity to other loss events.

#### **6.4.1 The Derivative of the Allocated Capital to Loss Event**

We can also use equation (6.6) to obtain the derivative of Allocated Capital to loss event with respect to the loss amount  $x$ :

$$d/dx\{AC(x)\} = d/dx\{f(x) \int_{y=0}^{y=x} 1/(1-F(y))dy\} \quad (6.8)$$

$$= f(x) * d/dx\{ \int_{y=0}^{y=x} 1/(1-F(y))dy\} + \int_{y=0}^{y=x} 1/(1-F(y))dy * d/dx\{f(x)\} \quad (6.9)$$

*Capital Allocation by Percentile Layer*

$$d / dx\{AC(x)\} = f(x)/(1 - F(x)) + f'(x) \int_{y=0}^{y=x} 1/(1 - F(y))dy \quad (6.10)$$

We can understand formula (6.10) as saying that as the loss amount  $x$  under consideration increases, 2 factors simultaneously affect the allocated capital:

- 1) The allocated capital increases to the extent that the loss amount receives allocation from an additional layer of capital based upon conditional probability [=  $f(x) / (1 - F(x))$ ].
- 2) The allocated capital changes (usually decreases) to the extent that the loss amount is less likely to occur and thus receives a lower allocation on the lower layers of capital [=  $d/dx (f(x)) * \int 1 / (1-F(y)) dy$ ].

Two observations about these 2 factors:

- 1) Usually, the derivative of  $f(x)$  is negative, so item #2 is usually negative, but can be positive when the derivative of  $f(x)$  is positive.
- 2) When dealing with simulation output of  $n$  discrete events, each discrete event has likelihood of  $1/n$  and thus is equally likely; therefore, the allocated capital to each larger event increases only with respect to factor #1, whereas factor #2 will equal zero.

### 6.4.2 Utility Function

Equation (6.6) also shows how we can use capital allocation by percentile layer to describe the disutility, or “pain”, given a particular loss event  $x$ .

Let  $r$  = required % rate of return on capital. Then the cost of capital associated with loss event  $x$   
=

$$r * f(x) \int_{y=0}^{y=x} 1/(1 - F(y))dy \quad (6.11)$$

The cost of capital of an event, given the loss event, is then

*Capital Allocation by Percentile Layer*

$$r \int_{y=0}^{y=x} 1/(1 - F(y)) dy \quad (6.12)$$

And the total cost, given the event, equals the loss amount  $x$  plus the cost of capital =

$$x + r \int_{y=0}^{y=x} 1/(1 - F(y)) dy \quad (6.13)$$

Equation (6.13) shows the disutility as an additive loading to the loss amount  $x$ . Rearranging terms, we can also show the disutility as a multiplicative factor as well:

$$x \left[ 1 + r \left( 1/x \right) \int_{y=0}^{y=x} 1/(1 - F(y)) dy \right] \quad (6.14)$$

## **7. INTERPRETATION, COMMENTS, AND EXTENSIONS**

The procedure for capital allocation by percentile layer outlined above generates allocations that are different than many other methods, with ramifications for measuring the relative risk and profitability of various lines of business. Some methods, such as coTVaR, tend to allocate the overwhelming amount of capital only to perils that contribute to the very worst scenarios; capital allocation by percentile layer, however, recognizes that when the firm holds capital even for an extremely catastrophic scenario, some of the capital also benefits other, more likely, more moderately severe downside events. On the other hand, some other methods (e.g., Mango's "capital consumption", XTVaR, etc.) allocate capital to a broader range of loss events that consume capital; the allocation varies proportionately based upon conditional probability. Because these methods fully account for relative probability, however, they may allocate insufficient capital to severe yet unlikely events. The potentially extreme loss of such events causes firms to hold an amount of

### *Capital Allocation by Percentile Layer*

capital that far outstrips the amount required by other loss events; although the actual occurrence of one of these events is very unlikely, the cost of holding precautionary capital is quite definite. Capital allocation by percentile layer appropriately allocates more capital cost to those unlikely, severe events that require the firm to hold additional capital.

Capital allocation by percentile layer as delineated above assumes that required capital is based upon VaR, but a similar model can also apply to TVaR. In other words, we can view TVaR as saying we want to hold enough capital “even for {the 99th percentile loss + the average amount by which losses above the 99th percentile tend to exceed the 99th percentile}”. In such a case, capital allocation by layer would be nearly the same, allocating capital up to the 99th percentile. The only additional step would then be to allocate one additional layer of capital (i.e.,  $TVaR - VaR$ ) to the losses that exceed the TVaR threshold. Consistent with TVaR’s meaning as well as the layer allocation approach, this additional layer of capital should be allocated to loss events in proportion to each event’s average amount of loss excess of the TVaR threshold.

## **7.1 Additional Areas of Application**

The application highlighted here focuses on property catastrophe risk, but the reformulation of the meaning of VaR should have similar ramifications to other sources of risk as well. Specifically, risk and capital for risky assets such as equities and fixed income securities have traditionally been defined based upon VaR metrics; as a result, methods that allocate capital among various asset classes and operating units may benefit from implementing capital allocation by percentile layer.

Capital allocation by percentile layer may also be germane when the firm’s total capital does not reside in one “indivisible bucket of equity capital” but rather is split into multiple tranches of capital. Because these tranches sustain capital depletion in a predetermined sequential order and, as a result, carry different cost of capital rates, it would seem appropriate to allocate capital with a procedure that explicitly accounts for the varying layers of capital and their costs. In addition, alternative forms of capital that apply on a “layered” basis (e.g., excess of loss reinsurance) and their costs (e.g., the amount of “risk load” or “margin” in the reinsurance price) would also appear to be candidates for capital allocation by percentile layer.

## **7.2 Implementation**

In many situations in which we want to implement capital allocation by percentile layer, we will be dealing with discrete output from a simulation model. By using the previously derived discrete

### *Capital Allocation by Percentile Layer*

formulas we can program a spreadsheet and achieve numerical results. Once capital amounts are allocated to each simulated loss event, we can then (per Mango, Kreps) further allocate the capital for the total loss to those individual components that contributed to the total.

#### **7.2.1 Contributions to Capital**

The main focus of the analysis until now has been on the allocation of capital with respect to loss without considering premium. When measuring the allocated cost of capital for a business unit or peril or individual contract, one must also recognize that the associated premium (net of expenses) is essentially a contribution to capital or “offset” to allocated capital. As a result, one should subtract collected premium net of expenses from the allocated capital before multiplying by the cost of capital rate.

## **8. IMPLICATIONS FOR RISK LOAD**

The discussion until now has related to a retrospective situation, when the price that the firm has charged for a certain transaction is a historical fact; the only question the firm asks is how to allocate capital costs in order to measure profitability. But what should the company do in a prospective situation? How does capital allocation affect what price the firm should charge? What does capital allocation by percentile layer imply about calculating risk load and determining the premium?

For the purposes of our discussion, we will ignore any provisions in the premium for expenses, parameter uncertainty, winner’s curse, or other loadings. Thus we will define

$$\text{Premium net of expenses} = \text{expected loss} + \text{cost of capital} \quad (8.0)$$

Let:

P = premium net of expenses

E[L] = expected loss

r = required % rate of return on capital

Then

*Capital Allocation by Percentile Layer*

$$P = E[L] + r * (\text{allocated capital} - \text{contributed capital}) \quad (8.1)$$

Let:

Contributed capital = premium net of expenses.

Then

$$P = E[L] + r * (\text{allocated capital} - P)$$

Rearranging terms, we derive:

$$P(1+r) = E[L] + r * (\text{allocated capital})$$

$$P = 1/(1+r) * E[L] + r/(1+r) * \text{allocated capital}$$

Let  $1/(1+r) = (1+r-r)/(1+r) = [(1+r)/(1+r) - (r/(1+r))] = [1-r/(1+r)]$ . Then

$$P = (1 - r/(1+r)) * E[L] + r/(1+r) * \text{allocated capital}.$$

Then

$$P = E[L] + r/(1+r) * (\text{allocated capital} - E[L]) \quad (8.2)$$

For any given loss event  $x$  (given it is below the VaR threshold), allocated capital is given by Equation (6.6) and  $E[L] = x * f(x)$ .

Then the Premium for any loss event  $x =$

$$P(x) = xf(x) + r/(1+r) \left[ f(x) \int_{y=0}^{y=x} 1/(1-F(y)) dy - xf(x) \right] \quad (8.3)$$

Rearranging terms, we derive

*Capital Allocation by Percentile Layer*

$$P(x) = f(x) \left\{ x + r/(1+r) \left[ \int_{y=0}^{y=x} 1/(1-F(y)) dy - x \right] \right\} \quad (8.4)$$

Equation (8.4) shows that the disutility function given loss event  $x$ , after taking into account its premium's contribution to capital, equals

$$x + r/(1+r) \left[ \int_{y=0}^{y=x} 1/(1-F(y)) dy - x \right] \quad (8.5)$$

We can also rearrange equation (8.3) to produce a multiplicative factor,

$$P(x) = xf(x) \left\{ 1 + r/(1+r) \left[ (1/x) \int_{y=0}^{y=x} 1/(1-F(y)) dy - 1 \right] \right\} \quad (8.6)$$

Equation (8.6) highlights that the required premium associated with loss event  $x$  is the expected value  $x \cdot f(x)$  multiplied by an adjustment factor. We can view the adjustment factor as either

- 1) an adjustment to the loss amount  $x$
- 2) an adjustment to the probability  $f(x)$

### **8.1 Properties of the Risk Load**

Equation (8.5) shows that given a loss event, the additive risk load amount =

$$r/(1+r) \left[ \int_{y=0}^{y=x} 1/(1-F(y)) dy - x \right] \quad (8.7)$$

### *Capital Allocation by Percentile Layer*

Equation (8.7) and its derivatives show that the risk load increases with respect to the loss amount  $x$  at an increasing rate. It also shows that even for very small values of the loss event  $x$  the risk load is strictly positive. This result suggests that capital allocation by percentile layer as applied above, in contradistinction to many common methods, requires that even small loss events that are less than the portfolio's mean receive an allocation of capital and a positive risk load.

Why should a loss event that is *less* than the average loss require an allocation of capital? In order to clarify this issue, we turn to thought experiment #3.

#### **8.1.1 Thought Experiment #3**

Again assume we are dealing with two perils:

- 1) Wind                                      20% chance of 5M loss, else zero
- 2) Earthquake (EQ)                      5% chance of 100M loss, else zero

Assume the perils are independent. Thus, the possible scenarios for portfolio loss are:

- 1) 76% probability that neither peril occurs, loss = 0
- 2) 19% probability that only Wind occurs, loss of 5M
- 3) 4% probability that only EQ occurs, loss of 100M
- 4) 1% probability that both Wind and EQ occur, loss of 105M

Note that the average loss for Wind =  $E[\text{Wind}] = 1\text{M}$  and  $E[\text{EQ}] = 5\text{M}$ . The two perils are independent so the portfolio expected loss = 6M. For simplicity assume that the premium for each peril equals the mean.

Now what happens when there is a “Wind only” loss of 5M? The Wind loss of 5M exceeds its 1M of premium, so it clearly needs capital. Yet overall, the portfolio has 6M of premium available and so the firm can use this money to pay the “Wind only” loss of 5M. Where, however, does this 6M of premium come from? While 1M comes from Wind, the majority, 5M, comes from the premium inflow from EQ. Thus it is clear that when a “Wind only” event occurs, the Wind subline “uses” or “consumes” capital, and the EQ subline “provides” capital by contributing its premium.

Therefore, this numerical example shows that even a loss event (e.g., Wind loss of 5M) that is *less* than the portfolio's mean loss (e.g. 6M) can consume capital and deserves allocation of capital. As a result, many common methods, which only allocate capital to loss events that exceed the mean,

may generate skewed allocations.

## 9. FINAL NUMERICAL EXAMPLE

Take the following situation involving 3 independent lines of business (LOB), corresponding to 3 perils

- LOB A: (e.g., Fire)
  - 25% chance of a loss;
  - If there is a loss, the amount is exponentially distributed
    - Exponential Mean = 4M
- LOB B: (e.g., Wind)
  - 5% chance of loss;
  - If there is a loss, the amount is exponentially distributed
    - Exponential Mean = 20M
- LOB C: (e.g., EQ)
  - 1% chance of loss;
  - If there is a loss, the amount is exponentially distributed
    - Exponential Mean = 100M

Each line of business has an annual average loss amount of 1M, but some lines have losses that are more infrequent and extreme than others.

We will run 10,000 simulations, set required capital equal to  $\text{VaR}(99\%)$ , and use capital allocation by percentile layer in order to calculate the allocated capital for each simulated loss event. Then we will take the amount of capital assigned to each loss event and allocate to the contributing perils; each peril will receive an allocation based upon the contribution of its loss to the total event loss. Finally, we will take allocated capital and subtract the amount of the mean loss (as a proxy for the contribution to capital from premium) from the allocated capital.

## 9.1 Final Numerical Example – Allocation Results

Method	Line of Business		
	A	B	C
Standalone TVaR @99th percentile	10%	30%	60%
coTVaR allocation @99th percentile	0%	24%	76%
coTVaR allocation @95th percentile	10%	42%	48%
coTVaR allocation @90th percentile	21%	39%	40%
coTVaR allocation @breakeven percentile	29%	35%	36%
Capital Allocation by Percentile Layer, VaR@99%	17%	53%	30%

Note that all of the tail-based methods such as VaR, TVaR, coTVaR, etc. allocate the greatest amount of capital to the severe yet extremely unlikely EQ event. Only capital allocation by percentile layer assigns the most capital to the more likely Wind event.

## 10. CONCLUSIONS

Capital allocation by percentile layer has several advantages, both conceptual and functional, over existing methods for allocating capital. It emerges organically from a new formulation of the meaning of holding Value at Risk capital; allocates capital to the entire range of loss events, not only the most extreme events in the tail of the distribution; tends to allocate more capital, all else equal, to those events that are more likely; tends to allocate disproportionately more capital to those loss events that are more severe; renders moot the question of which arbitrary percentile threshold to select for allocation purposes by using all relevant percentile thresholds; produces allocation weights that always add up to 100%; explicitly allocates the entire amount of the firm’s capital, in contrast to other methods that allocate based upon the last dollar of “marginal” capital; and provides a framework for allocating capital by layer and by tranche.

Capital allocation by percentile layer has the potential to generate significantly different allocations than existing methods, with ramifications for calculating risk load and for measuring risk adjusted profitability.

### Acknowledgment

The author would like to acknowledge Dave Clark, Vladimir Kremerman, Bill Panning, Ira Robbin, Marc Shamula, Walt Stewart, and Kyle Vrieze for commenting on earlier drafts of this paper.

### Appendix A: Calculating Results for an Exponential Distribution

If the loss distribution follows an exponential distribution,  $F(x) = 1 - \exp(-x / \theta)$ , we can solve formula (6.6) to derive a formula for allocated capital for loss event  $x$  (assuming  $x < \text{VaR}$ )

$$AC(x) = (1/\theta) \exp(-x/\theta) \int_{y=0}^{y=x} \exp(y/\theta) dy \quad (\text{A.1})$$

$$AC(x) = 1 - \exp(-x/\theta) \quad (\text{A.2})$$

We can also use formula (6.10) to calculate the derivative of allocated capital ( $x$ ) for an exponential distribution =

$$d/dx \{ AC(x) \} = (1/\theta) \exp(-x/\theta) \quad (\text{A.3})$$

= a positive number, confirming that allocated capital increases as the loss amount  $x$  gets larger. However, the second derivative is negative, so the rate of increase is decreasing.

We can also solve formula (6.13) to calculate the total cost (the loss amount plus the cost of allocated capital) given loss amount  $x$  =

$$x + r\theta(\exp(-x/\theta) - 1) \quad (\text{A.4})$$

We can also solve formula (6.14) to express the total cost given loss amount  $x$  as the product of the loss amount  $x$  and a multiplicative loading factor =

$$x[1 + r\theta(1/x)(\exp(-x/\theta) - 1)] \quad (\text{A.5})$$

## 11. REFERENCES

- [1] Kreps, Rodney, "Riskiness Leverage Models", Proceedings of the Casualty Actuarial Society (PCAS) XCII, 2005, 31-60
- [2] Lee, Yoong-Sin., "The Mathematics of Excess of Loss Coverages and Retrospective Rating-A Graphical Approach," PCAS LXXV, 1988, 49-77
- [3] Mango, Donald F., "An Application of Game Theory: Property Catastrophe Risk Load ", PCAS LXXXV, 1998, 157-186
- [4] Mango, Donald F., "Capital Consumption: An Alternative Methodology for Pricing Reinsurance", CAS 2003 Winter Forum, 351-378
- [5] Venter, Gary G., "Adapting Banking Models to Insurer ERM", ERM Symposium 2006
- [6] Venter, Gary G., "Discussion of 'Distribution-Based Pricing Formulas are Not Arbitrage Free' by David Ruhm", PCAS XCI, 2004, 25-33
- [7] Wang, Shaun, "A Set of New Methods and Tools for Enterprise Risk Capital Management and Portfolio Optimization", CAS 2002 Summer Forum, 43-78

### Biography of the Author

**Neil Bodoff** is Vice President and Actuary at Willis Re in New York. He currently focuses on property and casualty reinsurance analysis, stochastic modeling, and capital allocation. Neil is a Fellow of the Casualty Actuarial Society and a Member of the American Academy of Actuaries.

Please feel free to contact the author at [neil.bodoff@willis.com](mailto:neil.bodoff@willis.com) or [neil\\_bodoff@yahoo.com](mailto:neil_bodoff@yahoo.com)

# Uncertainty-Based Credibility and its Application to Excess-of-Loss Reinsurance

Pietro Parodi, PhD

Stephane Bonche, IA

---

## Abstract

This paper proposes a methodology to calculate the credibility risk premium based on the uncertainty on the risk premium, as estimated by the standard deviation of the risk premium estimator. An optimal estimator based on the uncertainties involved in the pricing process is constructed.

The methodology is then applied to pricing layers of excess-of-loss reinsurance, and the behaviour of the credibility factor as a function of layer excess is analysed. Results are obtained for both the general case and the significant special case where the severity distribution is the same for all clients, for which it is proved that credibility is broadly constant across the reinsurance layers. A real-world application to pricing motor reinsurance is also discussed.

Although the methodology is especially useful when applied to reinsurance, the underlying ideas are completely general and can be applied to all contexts where the uncertainties in the pricing process can be calculated.

**Keywords.** uncertainty-based credibility, pricing horizon, excess-of-loss reinsurance pricing, market heterogeneity, error propagation analysis

---

## 1. INTRODUCTION

The experience-based calculation of the risk premium for an insurance or reinsurance account is affected by several sources of uncertainty, the most obvious – and perhaps the best understood – of which is the limited size of the historical database of losses of the client.

To make up for such uncertainty the analyst may use average, or typical, information from the market (the market risk premium) to replace or complement the client risk premium. The problem with this is that the market experience is not fully relevant to a particular client. This is usually captured by the spread, or heterogeneity, of the client risk premiums around the standard market rate. As an added complication, although the market rate is typically computed from a larger data set than that of a client, it, too, is based on a loss database of limited size and is therefore affected by the same type of uncertainty.

The considerations above apply both to direct insurance and to reinsurance, but the problems are felt more acutely with excess-of-loss (XL) reinsurance, as the data on large losses are scarcer. The existence of a layer structure adds one obvious difficulty to the pricing process: the accuracy with which we price each layer will typically decrease rapidly as a function of the layer excess. This is a

consequence of relying on data from bottom layers to build a model that will be used to price the higher layers, beyond the limit for which our experience is relevant.

Given the scant supply of data that is typical of reinsurance, resorting to the market for an indication of rates is even more important. However, even at the market level the experience on large losses is limited and insufficient to price the higher layers of reinsurance accurately. Above a given layer excess the effect of uncertainty on the market reference rate may be comparable to the effects due to market heterogeneity.

The standard way to combine client and market information is credibility. The credibility risk premium is the convex combination of the client risk premium and the market risk premium:

$$\text{Credibility risk premium} = Z \times \text{Client risk premium} + (1-Z) \times \text{Market risk premium}$$

where  $Z$  is a real number between 0 and 1, reflecting the relative weight that we give to the client's experience.

The idea of this paper is to use the standard deviation of the client risk premium estimator ( $\sigma_c$ ) as a measure of (lack of) credibility, weighting this against the market heterogeneity ( $\sigma_h$ ) and the standard deviation of the market risk premium estimator ( $\sigma_m$ ). Furthermore, since the risk premium of the market is calculated based on data from the whole market, including in general the client itself, the two estimators for the market and the client are correlated ( $\rho_{m,c}$ ). The resulting formula for the credibility factor (Proposition 1):

$$Z = \frac{\sigma_h^2 + \sigma_m^2 - \rho_{m,c} \sigma_m \sigma_c}{\sigma_h^2 + \sigma_m^2 + \sigma_c^2 - 2\rho_{m,c} \sigma_m \sigma_c} \quad (1.1)$$

can be easily generalised to be used for XL reinsurance pricing, by considering the value of the parameters for a specific layer (formula (4.1)). As a consequence, the credibility factor will depend on the layer. However, in the important special case where the severity distributions of the different clients can be assumed to be the same, the credibility factor defined as above is broadly constant across the layers (Propositions 2 and 3).

This methodology was applied to pricing UK motor XL reinsurance, which can be performed by modelling the frequency of large losses as Poisson and the loss amounts above a certain threshold as a Generalised Pareto distribution (GPD). For this application, a hybrid approach to credibility was found to be adequate, using the general uncertainty-based credibility for the lower layers and a single-severity distribution model for the higher layers.

## **1.1 Research Context and Objective**

This paper presents a credibility methodology that we think is particularly appropriate for excess-of-loss reinsurance pricing, as it takes into account the uncertainty of the client and the market for different layers of reinsurance.

The modern approach to credibility – which stems from the works of Bühlmann and Straub (see Bühlmann [4]; Bühlmann & Straub [6] and the comprehensive book by Bühlmann & Gisler [5]) does not explicitly take the uncertainty on the market price into account in the formula for the credibility factor (see, e.g., theorem 3.7 in Bühlmann & Gisler [5], which gives results for both inhomogeneous and homogeneous credibility).

On the other hand Boor [3], who uses (as we do) uncertainty as a base for credibility, displays a credibility factor that contains an extra term for market uncertainty. This paper, however, focuses on a two-samples model (client v rest of the market) and attempts no analysis of the overall market heterogeneity/spread.

Credibility for excess-of-loss reinsurance was first examined by Straub [17]. This was extended by Patrik & Mashitz [13]. An implementation of this approach has been carried out by one of us for the UK motor reinsurance market [2].

All these works restrict their attention to the credibility of claim counts rather than considering aggregate losses, which are the real item of interest when pricing a reinsurance excess cover. Furthermore, these efforts have focused on the Poisson/Gamma credibility model applied to claim frequency.

An attempt to extend the ideas in [17] and [13] to provide a credibility formula for claim aggregate loss rather than claim frequency was made by Cockroft [7]. Cockroft provides a complex analytical solution involving infinite summations for the special case where the number of claims is Poisson with a Gamma prior distribution for the Poisson rate and the claim amounts are distributed according to a Pareto with a Gamma prior distribution for the power-law exponent.

Thus far, a simple general solution for calculating credibility for excess-of-loss reinsurance has not been provided in the literature. This paper argues that by using uncertainty as the main driver for credibility one is able to produce an intuitive and general method to calculate the credibility premium, which can be used both in insurance and in reinsurance.

## 1.2 Outline

Section 2 introduces a measure of uncertainty and outlines the various ways in which it can be calculated. Section 3 illustrates the methodology of uncertainty-based credibility in a general context, proving the basic result (Proposition 1) that gives the optimal value for the credibility factor. Section 4 illustrates the application of this result to reinsurance. Section 5 describes a real-world application of uncertainty-based credibility to pricing motor reinsurance. A detailed comparison with other methods is presented in Section 6. The limitations of the methodology are given in Section 7. Section 8 draws the conclusions.

## 2. THE RISK PREMIUM AND ITS UNCERTAINTY

### 2.1 Risk premium – definition and calculation

The risk premium  $\varphi$  is given by  $\varphi = \frac{E(S)}{w}$  where  $E(S)$  is the expected aggregate loss in a given period and  $w$  is the expected exposure in that same period.

Using the collective risk model assumption, the losses to an insurer in a given period can be modelled as a stochastic process  $S = \sum_{i=1}^N X_i$  where  $N$  represents the number of losses in the period and  $X_1, \dots, X_N$  represent their amounts. Both the number of losses and their amounts are random variables. The claims amounts  $X_1, \dots, X_N$  are i.i.d. and independent of  $N$ .

Using the collective risk model,  $E(S)$  can be written as  $E(S) = E(N)E(X)$  where  $E(N)$  is the expected number of claims and  $E(X)$  is the expected claim amount. To derive  $E(N)$  and  $E(X)$ , we need to know the underlying frequency and severity distributions with their exact parameter values (e.g.,  $N \sim Poi(\lambda w)$ ,  $X \sim Exp(\mu) \rightarrow E(S) = \lambda\mu w$ ,  $\varphi = \lambda\mu$ ).

However, the model is usually not so straightforward, since it is not always possible to express  $E(S)$  in a simple analytical form. This may be due to policy modifications (excesses, limits, reinstatements...) and to the effect of settlement delay and discounting. Therefore,  $E(S)$  will usually be appraised by a stochastic simulation or by an approximate formula.

### 2.2 Risk premium – sources and measures of uncertainty

In practice, we will only have an estimate of  $E(S)$  and therefore of the risk premium. This estimate will be affected by several sources of uncertainty: the models for frequency and severity will not replicate reality perfectly (model uncertainty); the values of the model parameters will only be

known approximately (parameter uncertainty); the data themselves are often reserve estimates rather than known quantities (data uncertainty).

Parameter uncertainty is the most important contribution to uncertainty and the one we will focus on in this paper. It depends on the fact that we only have a limited sample from which to estimate the parameters of the model. Data uncertainty has the effect of increasing parameter uncertainty; its effects, which can be studied by inspecting the IBNER distribution, will be analysed elsewhere [12]. Model uncertainty is difficult to quantify and will be usually dealt with in a low-profile fashion, by making sure that our models pass appropriate goodness-of-fit tests.

We will use the standard deviation as a measure of the uncertainty of an estimator. Although the standard deviation of an estimator is commonly denoted as “standard error”, we will stick to the expression “standard deviation of the estimator” to avoid the ambiguity surrounding the term “standard error” in the literature<sup>1</sup>.

We will refer to the standard deviation of the risk premium as shorthand for “the standard deviation of the estimator for the risk premium”. In general, the standard deviation of the risk premium will therefore depend on the process by which the risk premium is estimated. Notice that the standard deviation of the risk premium estimator should not be confused with the standard deviation of  $S/w$ , the aggregate loss per unit of exposure!

Section 3.4.2 will give examples of how the standard deviation of the risk premium estimator can be calculated in practice.

### **3. UNCERTAINTY-BASED CREDIBILITY**

Let  $\varphi_c$  be the “true” risk premium of the client. This is simply given by  $\varphi_c = \frac{E(S_c)}{w_c}$  where  $E(S_c)$  is the expected aggregate loss in a year and  $w_c$  is the exposure in the same year. According to the collective model,  $E(S_c)$  can be written as  $E(S_c) = E(N_c)E(X_c)$  where  $E(N_c)$  is the expected number of claims and  $E(X_c)$  is the expected claim amount. However, we will only have

---

<sup>1</sup> As an example, “standard error” is used as “standard deviation of the estimator” or “estimated standard deviation of the estimator” depending on the author.

an estimate of  $E(S_c)$ . The goodness of this estimate will be affected by data uncertainty, parameter uncertainty and model uncertainty.

Let  $\hat{\varphi}_c$  be the estimated risk premium of the client. This will typically be obtained by estimating the parameters of the frequency and severity distribution and by calculating the average frequency and severity based on those estimates. E.g., if frequency is a Poisson distribution:  $N \sim Poi(\lambda \cdot w_c)$  and severity is an exponential distribution with (true) mean  $\mu$ :  $X \sim Exp(\mu)$ , then the risk premium is given by  $\hat{\varphi}_c = \hat{\lambda} \cdot \hat{\mu}$ , where  $\hat{\lambda}$  is the estimated rate per unit of exposure and  $\hat{\mu}$  is the estimated mean of the exponential distribution.

We can also define  $\varphi_m$  (true risk premium) and  $\hat{\varphi}_m$  (estimated risk premium) for the market. The estimated risk premium  $\hat{\varphi}_m$  will be obtained in a similar fashion to  $\hat{\varphi}_c$  but it will use data from all participating clients, including the data used to calculate  $\hat{\varphi}_c$ .

Credibility is a standard technique by which the estimated risk premium of the client,  $\hat{\varphi}_c$ , and the estimated risk premium for the market,  $\hat{\varphi}_m$ , are combined to provide another estimate  $\hat{\varphi}$ , called the credibility estimate, of the client's risk premium  $\varphi_c$ , via a convex combination:

$$\hat{\varphi} = Z \cdot \hat{\varphi}_c + (1 - Z) \cdot \hat{\varphi}_m \quad (3.1)$$

where  $Z \in [0,1]$  is called the credibility factor.

In this section, we provide a means to calculate the credibility factor  $Z$  based on the uncertainty of the estimates  $\hat{\varphi}_c$ ,  $\hat{\varphi}_m$  and on the heterogeneity of the market. To do this we need an uncertainty model, i.e. a set of assumptions on how uncertainty affects the estimates.

### 3.1 The uncertainty model – Assumptions

1. The estimated risk premium of the market is described by a random variable  $\hat{\varphi}_m$  with expected value  $\varphi_m$  (the true risk premium for the overall market) and variance  $\sigma_m^2$ . For readability, we write this as

$$\hat{\varphi}_m = \varphi_m + \sigma_m \varepsilon_m \quad (3.2)$$

where  $\varepsilon_m$  is a random variable with zero mean and unit variance:  $E(\varepsilon_m) = 0$ ,  $E(\varepsilon_m^2) = 1$ . Notice that  $\varphi_m$  is *not* viewed as a random variable here. Despite the terminology above, which

resembles that used for Gaussian random noise, no other assumption is needed on the shape of the distribution of  $\varepsilon_m$ .

2. The true risk premium  $\varphi_c$  of the client is described by a random variable with mean  $E(\varphi_c) = \varphi_m$  (the true market risk premium) and variance  $Var(\varphi_c) = \sigma_h^2$ . In other terms,

$$\varphi_c = \varphi_m + \sigma_h \varepsilon_h \tag{3.3}$$

where  $\sigma_h$  measures the spread (or heterogeneity) of the different clients around the mean market value, and  $E(\varepsilon_h) = 0$ ,  $E(\varepsilon_h^2) = 1$ .

3. The estimated risk premium of the client,  $\hat{\varphi}_c$ , given the true risk premium,  $\varphi_c$ , is described by a random variable with mean  $E(\hat{\varphi}_c | \varphi_c) = \varphi_c$ ,  $Var(\hat{\varphi}_c | \varphi_c) = \sigma_c^2$ . In other words,

$$\hat{\varphi}_c | \varphi_c = \varphi_c + \sigma_c \varepsilon_c \quad (\hat{\varphi}_c = \varphi_m + \sigma_h \varepsilon_h + \sigma_c \varepsilon_c) \tag{3.4}$$

where  $\varepsilon_c$  is another random variable with zero mean and unit variance:  $E(\varepsilon_c) = 0$ ,  $E(\varepsilon_c^2) = 1$ . Again, no other assumption is made on the distribution of  $\varepsilon_c$ . Notice that in this case both  $\hat{\varphi}_c$  and  $\varphi_c$  are random variables.

4. Assume that  $\varepsilon_h$  is uncorrelated to both  $\varepsilon_m$  and  $\varepsilon_c$ :  $E(\varepsilon_m \varepsilon_h) = 0$ ,  $E(\varepsilon_c \varepsilon_h) = 0$ .

We are now in a position to prove the following result.

**Proposition 1.** *Given assumptions 1-4 above, the value of  $Z$  that minimises the mean squared error  $E_{m,c,h}((\hat{\varphi} - \varphi_c)^2) = E_{m,c,h}((Z \cdot \hat{\varphi}_c + (1-Z) \cdot \hat{\varphi}_m - \varphi_c)^2)$ , where the expected value is taken on the joint distribution of  $\varepsilon_m, \varepsilon_c, \varepsilon_h$ , is given by*

$$Z = \frac{\sigma_h^2 + \sigma_m^2 - \rho_{m,c} \sigma_m \sigma_c}{\sigma_h^2 + \sigma_m^2 + \sigma_c^2 - 2\rho_{m,c} \sigma_m \sigma_c} \tag{3.5}$$

where  $\rho_{m,c}$  is the correlation between  $\varepsilon_m$  and  $\varepsilon_c$ .

**Proof.** The result is straightforward once we express  $\hat{\varphi} - \varphi_c$  in terms of  $\varepsilon_m, \varepsilon_c, \varepsilon_h$  only. The mean squared error is given by

$$E_{m,c,h} \left( (Z \cdot \hat{\varphi}_c + (1-Z) \cdot \hat{\varphi}_m - \varphi_c)^2 \right) = E_{m,c,h} \left( ((Z-1)(\sigma_h \varepsilon_h - \sigma_m \varepsilon_m) + Z \cdot \sigma_c \varepsilon_c)^2 \right) \\ = (Z-1)^2 (\sigma_h^2 + \sigma_m^2) + Z^2 \sigma_c^2 - 2Z(Z-1) \rho_{m,c} \sigma_m \sigma_c$$

where  $\rho_{m,c} = E(\varepsilon_m \varepsilon_c)$ . By minimising with respect to  $Z$  one obtains equation (3.5).  $\square$

The following sections will go into more detail as to the meaning of the assumptions and of this result.

### 3.1.1 Explaining the assumptions

Assumption 2 tries to capture market heterogeneity: different clients will have different risk premiums, reflecting the different riskiness of the accounts. This is similar to the risk factor in Bayesian and Buhlmann's approach to credibility. We do not need to know what the prior distribution of the risk premiums is, as long as we know its variance. In practice, this will be determined empirically.

Assumptions 1 and 3 try to capture the uncertainty inherent in the process of estimating the risk premium. The quantities  $\sigma_m$  and  $\sigma_c$  should not be confused with the standard deviation of the underlying aggregate loss distribution for the market and the client.

The random variable  $\varepsilon_h$  gives the prior distribution of the client price around a market value, whereas  $\varepsilon_m, \varepsilon_c$  are parameter uncertainties on the market and the client. Therefore, assumption 4 ( $E(\varepsilon_m \varepsilon_h) = 0$ ,  $E(\varepsilon_c \varepsilon_h) = 0$ ) is quite sound. The correlation between  $\varepsilon_m$  and  $\varepsilon_c$ , however, cannot be ignored. The reason for this is that the estimated risk premium of the market is based on data collected from different clients, including client  $c$ .

### 3.2 Is $\hat{\varphi}$ an unbiased estimator for $\varphi_c$ ?

It is important to notice that the expected value  $E_{m,c,h} \left( (\hat{\varphi} - \varphi_c)^2 \right)$  is also taken over the distribution of  $\varepsilon_h$ . As a consequence, the mean squared error is not necessarily minimised for each individual client, but only over all possible clients.

For a given client  $c$ ,  $\hat{\varphi}$  is in general a biased estimator for  $\varphi_c$ . The bias is given by  $bias(\hat{\varphi} | \varphi_c) = E_{m,c}(\hat{\varphi} | \varphi_c) - \varphi_c = (1-Z)(\varphi_m - \varphi_c) = -(1-Z)\sigma_h \varepsilon_h$ . The expected value is in this case taken over the joint distribution of  $\varepsilon_m$  and  $\varepsilon_c$ . Averaging over  $\varepsilon_h$ , the bias disappears:  $E_h(bias(\hat{\varphi} | \varphi_c)) = 0$ .

Notice how the quest for an estimate  $\hat{\varphi}$  of  $\varphi_c$  that is *collectively* unbiased is a common feature of credibility theory (see for example Bühlmann's approach as described in the book by Klugman et al. [10]).

The meaning of the formula for the bias,  $bias(\hat{\varphi} | \varphi_c) = -(1-Z)\sigma_h\varepsilon_h$ , is that when credibility is close to 1, the credibility estimate for the risk premium will be close to the client estimated price,  $\hat{\varphi}_c$ , and the bias will be close to zero. On the other hand, if the credibility is close to 0, the credibility estimate of the risk premium will be close to  $\hat{\varphi}_m$ , and the bias will be about  $\sigma_h\varepsilon_h$  – i.e., the credibility estimate will be distributed randomly around the market risk premium with a standard deviation equal to  $\sigma_h$  – which is exactly what we expect to happen.

### 3.3 The effect of correlation

- In case the data for client  $c$  are not included in the market data set the correlation between  $\varepsilon_m$  and  $\varepsilon_c$  can be assumed to be zero. In this case the credibility factor simplifies to

$$Z = \frac{\sigma_h^2 + \sigma_m^2}{\sigma_h^2 + \sigma_m^2 + \sigma_c^2} \quad (3.6)$$

which is more intuitive than (3.5). This also suggests an alternative way to carry out the credibility calculations: for each client, first remove the client's data from the market database and then calculate  $\hat{\varphi}$  as  $\hat{\varphi} = Z \cdot \hat{\varphi}_c + (1-Z) \cdot \hat{\varphi}_{m-c}$ , where  $Z = \frac{\sigma_h^2 + \sigma_{m-c}^2}{\sigma_h^2 + \sigma_{m-c}^2 + \sigma_c^2}$  (notice how the market heterogeneity itself,  $\sigma_h$ , has to be recalculated). However, this methodology is more lengthy and awkward than that implied by (3.5), as the rest-of-the-market parameters need to be recalculated for each client.

The effect of a positive correlation between  $\varepsilon_m$  and  $\varepsilon_c$  is to increase the credibility factor. This makes sense intuitively, as a larger correlation indicates a larger participation of the client in the market loss database. As a consequence, the market data will provide less useful information to that client.

- Note that the condition  $Z \leq 1$  can be translated into  $\rho_{m,c} \frac{\sigma_m}{\sigma_c} \leq 1$ . As  $\rho_{m,c} \leq 1$ , this is automatically satisfied if  $\frac{\sigma_m}{\sigma_c} \leq 1$ , which will hold under non-pathological circumstances as it is normally the case that  $\sigma_m \ll \sigma_c$ , the market estimate being based on a far larger sample.

- Note also that if  $\rho_{m,c} < \min\left\{\frac{\sigma_m^2 + \sigma_h^2}{\sigma_m \sigma_c}, \frac{\sigma_m^2 + \sigma_h^2 + \sigma_c^2}{2\sigma_m \sigma_c}\right\}$  the credibility factor is guaranteed to be positive. Under non-pathological circumstances,  $\sigma_m < \sigma_c$  (the market has a larger sample than the client) and  $\sigma_c < \sigma_h$  (the uncertainty on the risk premium is smaller than the spread of prices across the market, otherwise it would make no sense to use the client risk premium at all). Therefore, both ratios inside the bracket are larger than 1 and the inequality above is automatically satisfied.

### 3.4 Practical considerations

In practice, the standard deviations  $\sigma_h$ ,  $\sigma_m$ ,  $\sigma_c$  and  $\rho_{m,c}$  are not known and they must be estimated from the data. Therefore the credibility factor can also be written as:

$$Z \approx \frac{s_h^2 + s_m^2 - r_{m,c}s_ms_c}{s_h^2 + s_m^2 + s_c^2 - 2r_{m,c}s_ms_c} \quad (3.7)$$

where  $s_h$  is the *estimated* market heterogeneity,  $r_{m,c}$  is the *estimated* correlation between the market and the client,  $s_m$  and  $s_c$  are the *estimated* standard deviations of the estimators for the market and client risk premiums.

#### 3.4.1 Estimating market heterogeneity

Market heterogeneity can be estimated as the empirical variance of the risk premium for all available clients. This may be done in a weighted or in a non-weighted fashion. If the market premium is calculated by collecting all data from all clients, larger clients will inevitably weigh more, and the weighted version of the variance will have to be used for consistency:

$$s_h^2 = \frac{\sum_c W_c}{\left(\sum_c W_c\right)^2 - \sum_c W_c^2} \sum_c W_c (\hat{\phi}_c - \hat{\phi}_m)^2 \quad (3.8)$$

where  $W_c = \sum_j w_c^j$  is the cumulative exposure of client  $c$  over all years  $j$  considered in the analysis.

#### 3.4.2 Estimating the standard deviation of the risk premium estimator

As mentioned in Section 2.2, the standard deviation on the risk premium depends on the process by which the risk premium is calculated. This is best explained with the following simple example.

Suppose the frequency distribution is modelled as a Poisson process whose estimated rate is  $\hat{\lambda}w_c$  and the severity distribution is modelled as an exponential distribution whose estimated mean is  $\hat{\mu}$ . The estimated risk premium will then be  $\hat{\phi} = \hat{\lambda}\hat{\mu}$ . The standard error on  $\hat{\phi}$  will depend on the standard deviation of the estimators  $\hat{\lambda}$  and  $\hat{\mu}$ : in this case we have the exact result  $\frac{Var(\hat{\phi})}{E(\hat{\phi})^2} = \frac{Var(\hat{\lambda})}{E(\hat{\lambda})^2} + \frac{Var(\hat{\mu})}{E(\hat{\mu})^2} + \frac{Var(\hat{\lambda})Var(\hat{\mu})}{E(\hat{\lambda})^2 E(\hat{\mu})^2}$ . The values of  $Var(\hat{\lambda})$  and  $Var(\hat{\mu})$  depend in turn

on how the distribution parameters are calculated, and the expected values in the denominators will usually be approximated by their estimated value:  $E(\hat{\phi})^2 \approx \hat{\phi}^2$ , etc. E.g., if the mean of the exponential distribution is calculated by MLE based on the data sample  $\{X_1, \dots, X_n\}$ , then

$$\hat{\mu} = \frac{\sum_{i=1}^n X_i}{n} \quad \text{and} \quad Var^{est}(\hat{\mu}) = \frac{\hat{\mu}^2}{n}$$

(there are two approximations here: one is the replacement of  $\mu$

with  $\hat{\mu}$  in the formula and the fact that this formula is only true asymptotically).

Usually, we cannot find an exact formula for  $Var(\hat{\phi})$ . This may happen for two reasons.

- Except for very simple cases such as that illustrated above, the formula linking  $\hat{\phi}$  to the severity and frequency parameters will be too complex to propagate the uncertainties on the parameters exactly. In this case the standard deviation of  $\hat{\phi}$  can be estimated by drawing at random from the distribution of the parameters, which in the case of MLE is asymptotically known to be normal (or rather, multivariate normal). The correlations between the parameters must be taken into account. See Section 5.2 for a detailed example of how this is achieved.
- There might not even exist an analytical formula for  $\hat{\phi}$ . This will often be the case when there are payment and settlement delays, complicated structures (excesses, limits, premium adjustments, premium reinstatements after a claim (reinsurance), etc). In this case  $\hat{\phi}$  may have to be estimated by a stochastic simulation. The stochastic simulation will then have to be run for different values of the parameters, according to the parameter distribution. As a consequence, the estimation of the standard deviation of  $\hat{\phi}$  will have a far larger computational complexity.

### 3.4.3 Estimating the correlation

How the correlation between the uncertainty on the client and on the market is calculated depends on the exact process to calculate the risk premium. The following is a simple example

assuming that the risk premium is calculated in a burning cost fashion. This works by dividing the estimated total losses over a base period (typically, at least 10 years of data for long-tail business such as liability) by the total exposure over that period. The estimated total losses over the base period are corrected for inflation, IBNR, IBNER, etc.

Notice that we are interested in calculating the correlation between the *uncertainty* on  $\varphi_m$  and  $\varphi_c$ , therefore we can assume that  $\varphi_m$  and  $\varphi_c$  are fixed. Let  $\hat{T}_m$  be the total losses for the market, and  $\hat{T}_c$  the total losses for the client. Since the client is part of the market, we shall have  $\hat{T}_m = \hat{T}_c + \hat{T}_{m-c}$  where  $\hat{T}_{m-c}$  represents the losses of the rest of the market. The risk

premium for the market and the client are defined respectively as  $\hat{\varphi}_m = \frac{\hat{T}_m}{\sum_j w_m^j}$  and  $\hat{\varphi}_c = \frac{\hat{T}_c}{\sum_j w_c^j}$ ,

where  $w_c^j$  and  $w_m^j$  are respectively the client and market exposures in year  $j$ , the year index ranging

over the base period. If we assume that<sup>2</sup>  $Cov(\hat{T}_c, \hat{T}_{m-c}) = 0$ , then  $Cov(\hat{\varphi}_c, \hat{\varphi}_m) = \frac{Var(\hat{T}_c)}{\sum_j w_m^j \sum_j w_c^j}$  and

$r_{m,c} = \sqrt{\frac{Var(\hat{T}_c)}{Var(\hat{T}_m)}}$ . In the case of a compound Poisson distribution, this translates into:

$$r_{m,c} = \sqrt{\frac{\hat{\lambda}_c \sum_j w_c^j \langle X_c^2 \rangle}{\hat{\lambda}_m \sum_j w_m^j \langle X_m^2 \rangle}} \quad (3.9)$$

Even when the risk premium is not obtained by the burning cost approach, this formula is still a good guidance as to the degree of correlation one may expect.

### 3.4.4 Updating the market statistics

Generally speaking, the client risk premium will have to be calculated at different times for different clients. Furthermore, once all contributors to the market have been priced there will still be a time lag between when the data for all clients are available and when the market-related statistics

---

<sup>2</sup> Notice that we are assuming  $\varphi_m$  and  $\varphi_c$  to be given (see Assumptions 1 and 3 of the credibility model in Section 3.1), therefore we can ignore the correlation of the aggregate losses of the client v the rest of the market, which of course exists (and motivates the credibility approach). We are focusing here on the correlation between the residual variations that exist because of parameter uncertainty.

(heterogeneity, uncertainty) are calculated. Typically, we will be comparing last year's market statistics with this year's clients. Therefore, the comparison will not be exactly like-for-like. At the very least, we will have to correct the market results for residual inflation.

#### 4. APPLICATION TO REINSURANCE

In the case of excess-of-loss reinsurance, the quantity to be estimated will be  $\varphi_c^{D,L}$ , the "true" risk premium for layer  $(D, D+L)$ . Equation (3.1) can be rewritten as:  $\hat{\varphi}^{D,L} = Z^{D,L} \cdot \hat{\varphi}_c^{D,L} + (1 - Z^{D,L}) \cdot \hat{\varphi}_m^{D,L}$ , where:

- $\hat{\varphi}_c^{D,L}$ ,  $\hat{\varphi}_m^{D,L}$  are the expected losses per unit of exposure for layer  $(D, D+L)$  for the client and the market respectively;
- $Z^{D,L} \in [0,1]$  is the credibility factor for the client for layer  $(D, D+L)$ .

The problem here is to determine the value of  $Z^{D,L}$  that minimises  $E\left(\left(\hat{\varphi}^{D,L} - \varphi_c^{D,L}\right)^2\right) = E\left(\left(Z^{D,L} \cdot \hat{\varphi}_c^{D,L} + (1 - Z^{D,L}) \cdot \hat{\varphi}_m^{D,L} - \varphi_c^{D,L}\right)^2\right)$ . By the same expansion of the mean squared error shown in the proof of Proposition 1, it is straightforward to show that the credibility factor for the layer  $(D, D+L)$  is

$$Z^{D,L} = \frac{(\sigma_h^{D,L})^2 + (\sigma_m^{D,L})^2 - \rho_{m,c}^{D,L} \sigma_m^{D,L} \sigma_c^{D,L}}{(\sigma_h^{D,L})^2 + (\sigma_m^{D,L})^2 + (\sigma_c^{D,L})^2 - 2\rho_{m,c}^{D,L} \sigma_m^{D,L} \sigma_c^{D,L}} \quad (4.1)$$

where  $Z^{D,L} \in [0,1]$ .

The crucial question about credibility applied to reinsurance is the behaviour of the credibility risk premium – and therefore of  $Z^{D,L}$  – as a function of the layer's characteristics. Since the most important dependency is that on the layer excess,  $D$ , the dependency on the layer limit can be removed by considering either infinitesimal layers ( $Z^{D,0} = \lim_{L \rightarrow 0^+} Z^{D,L}$ ) or infinite layers ( $Z^{D,\infty} = \lim_{L \rightarrow \infty} Z^{D,L}$ ).

The behaviour of  $Z^{D,L}$  as a function of  $D$  and  $L$  depends on a combination of factors including:

- a) The relative size of uncertainty for market and client;
- b) The asymptotic behaviour of market severity v client severity;
- c) How the market heterogeneity  $\sigma_h^{D,L}$  depends on  $D, L$ .

The effect of this combination is quite complex and must in general be investigated empirically, as a general analytical expression will not always be available. Furthermore, estimating market heterogeneity for the highest layers is difficult, because market heterogeneity for a given layer is calculated from the expected aggregate losses of each client to that layer (as in (3.8)), and these are themselves affected by a large error. There is, however, a special case that is significant for the practitioner for which this behaviour simplifies. This is illustrated in the next section.

### 4.1 The “single severity” model

An important special case is obtained under the hypothesis that – although the frequency of large losses depends on the risk profile of the insurance company (e.g. age, sex), the severity distribution is unaffected by it, and the market severity curve can be used instead. As a consequence, market heterogeneity will be mostly due to heterogeneity in frequency. Empirical evidence supports this hypothesis for some kinds of portfolio, e.g. for the motor reinsurance portfolio, at least above a certain threshold (see Section 5). This also reflects general reinsurance practice where a market reference curve is used for most lines of business to price the higher layers.

The basic result (Proposition 2) is for the case where the market severity curve is known with infinite accuracy. Proposition 3 will then consider the amendments in the case where the market severity curve is known with limited accuracy.

**Proposition 2 – Basic single severity model.** . Let  $\varphi_c^{D,L}$ ,  $\hat{\varphi}_c^{D,L}$ ,  $\hat{\varphi}_m^{D,L}$  and  $\hat{\varphi}^{D,L}$  be as above.

*Assume the validity of the collective risk model, and that:*

- i. The severity curve is the same for all clients (i.e., the market severity curve) above a threshold  $\mu$ ;*
- ii. the severity distribution of the market is known with infinite accuracy*

*Then the credibility factor  $Z^{D,L}$  is independent of the layer  $(D, D+L)$  and is equal to:*

$$Z = \frac{(\sigma_h^\lambda)^2 + (\sigma_m^\lambda)^2 - \rho_{m,c}^\lambda \sigma_c^\lambda \sigma_m^\lambda}{(\sigma_h^\lambda)^2 + (\sigma_m^\lambda)^2 + (\sigma_c^\lambda)^2 - 2\rho_{m,c}^\lambda \sigma_c^\lambda \sigma_m^\lambda} \quad (4.2)$$

*where  $\sigma_h^\lambda$  measures the heterogeneity of clients’ frequencies;  $\sigma_m^\lambda$  and  $\sigma_c^\lambda$  are the standard deviations of the estimators of the market and the client frequency respectively;  $\rho_{m,c}^\lambda$  is the correlation between the estimator of  $\lambda_m$  and that of  $\lambda_c$ .*

**Proof.** We need to calculate  $\sigma_m^{D,L}$ ,  $\sigma_c^{D,L}$ ,  $\sigma_h^{D,L}$  and  $\rho_{m,c}^{D,L}$  in general formula (4.1). Under the collective risk model applied to the losses for layer  $(D, D+L)$ , the mean aggregate loss is

$E(S^{D,L}) = E(N^{D,L})E(X^{D,L})$  where  $E(N^{D,L})$  is the expected number of losses to the layer  $(D, D+L)$  and  $E(X^{D,L})$  is the expected loss amount to the layer  $(D, D+L)$  given that a loss to that layer has occurred. As it is well known (see, e.g., [10]),  $E(N^{D,L}) = E(N) \cdot \text{Prob}(D \leq X < D+L)$  and  $E(X^{D,L}) = \frac{E(X \wedge (D+L)) - E(X \wedge D)}{\text{Prob}(D \leq X < D+L)}$ , where  $X \wedge a := \min(X, a)$ ;  $E(N)$  is the expected number of losses above  $\mu$ ;  $E(X)$  is the expected amount of those losses above  $\mu$ . One can then write  $E(S^{D,L}) = E(N) \cdot (E(X \wedge (D+L)) - E(X \wedge D))$ . The risk premium for layer  $(D, D+L)$  can then be written as  $\varphi^{D,L} = \lambda \cdot U^{D,L}$ , where  $\lambda = E(N)/w$  is the expected frequency per unit of exposure above  $\mu \leq D$ , and

$$U^{D,L} = E(X \wedge D+L) - E(X \wedge D) = \int_D^{D+L} S(x) dx \quad (4.3)$$

Since the severity curve above  $\mu \leq D$  is the same for the client and the market, the risk premium of the client and the market can be expressed respectively as  $\varphi_c^{D,L} = \lambda_c \cdot U^{D,L}$  and  $\varphi_m^{D,L} = \lambda_m \cdot U^{D,L}$ . The estimated risk premium for the client (the market) can be expressed as  $\hat{\varphi}_c^{D,L} = \hat{\lambda}_c \cdot U^{D,L}$  ( $\hat{\varphi}_m^{D,L} = \hat{\lambda}_m \cdot U^{D,L}$ ) respectively, where  $\hat{\lambda}_c$  ( $\hat{\lambda}_m$ ) is the estimated client (market) frequency.

The severity distribution is known with infinite accuracy. Therefore,

$$\begin{aligned} (\sigma_m^{D,L})^2 &= \text{Var}(\hat{\varphi}_m^{D,L}) = \text{Var}(U^{D,L} \cdot \hat{\lambda}_m) = (U^{D,L})^2 \cdot (\sigma_m^\lambda)^2 \\ (\sigma_c^{D,L})^2 &= \text{Var}(\hat{\varphi}_c^{D,L}) = \text{Var}(U^{D,L} \cdot \hat{\lambda}_c) = (U^{D,L})^2 \cdot (\sigma_c^\lambda)^2 \\ (\sigma_h^{D,L})^2 &= \text{Var}(\varphi_c^{D,L}) = \text{Var}(U^{D,L} \cdot \lambda_c) = (U^{D,L})^2 \cdot (\sigma_h^\lambda)^2 \\ \rho_{m,c}^{D,L} &= \text{Cov}(\hat{\varphi}_m^{D,L}, \hat{\varphi}_c^{D,L}) / \sqrt{\text{Var}(U^{D,L} \cdot \hat{\lambda}_m) \cdot \text{Var}(U^{D,L} \cdot \hat{\lambda}_c)} = \rho_{m,c}^\lambda \end{aligned} \quad (4.4)$$

and  $(U^{D,L})^2$  can be removed from both the numerator and the denominator of (4.1), yielding (4.2).

□

**Discussion of the assumptions.** The collective risk model is a standard assumption. The assumption (i) of a single severity curve for all clients above a certain threshold underlies common reinsurance practice for the pricing of high layers. This does not mean that it is always reasonable, and should be tested against available data when possible. In Section 5 the validity of this assumption will be illustrated in the case of UK motor reinsurance.

Assumption (ii) is not realistic as the severity curve of the market is always estimated based on a

finite set of data and therefore it is affected by model, data and parameter uncertainty. However, Assumption (ii) is often a useful approximation when the uncertainty for the market severity is small – i.e., for all but the top layers. Even if this assumption only holds approximately, it shows that when all clients follow the same single severity curve, the credibility factor is broadly independent of the specific layer being priced.

The following proposition explores what happens when Assumption (ii) is dropped and the inaccuracy of the market severity curve is taken into account.

**Proposition 3 – Single severity model with “inaccurate” severity distribution.** Let  $\varphi_c^{D,L}$ ,  $\hat{\varphi}_c^{D,L}$ ,  $\varphi_m^{D,L}$ ,  $\hat{\varphi}_m^{D,L}$ ,  $\hat{\varphi}^{D,L}$ ,  $\hat{\lambda}_c$ ,  $\hat{\lambda}_m$  and  $U^{D,L}$  be as in Proposition 2 and in its proof. Also, let  $\hat{U}^{D,L}$  be the estimate of  $U^{D,L}$ . Assume the validity of the collective risk model, and assume that the severity curve is the same for all clients (i.e., the market severity curve) above a threshold  $\mu \leq D$ .

Then the credibility factor  $Z^{D,L}$  is equal to:

$$Z = \frac{(\sigma_h^\lambda)^2 + (\sigma_m^\lambda)^2 - \rho_{m,c}^\lambda \sigma_c^\lambda \sigma_m^\lambda + \frac{\text{Var}(\hat{U}^{D,L})}{(U^{D,L})^2} (\lambda_m^2 - \lambda_m \lambda_c + (\sigma_m^\lambda)^2 - \rho_{m,c}^\lambda \sigma_h^\lambda \sigma_m^\lambda)}{(\sigma_h^\lambda)^2 + (\sigma_m^\lambda)^2 + (\sigma_c^\lambda)^2 - 2\rho_{m,c}^\lambda \sigma_c^\lambda \sigma_m^\lambda + \frac{\text{Var}(\hat{U}^{D,L})}{(U^{D,L})^2} (\lambda_m^2 + \lambda_c^2 - \lambda_m \lambda_c + (\sigma_m^\lambda)^2 + (\sigma_c^\lambda)^2 - \rho_{m,c}^\lambda \sigma_h^\lambda \sigma_m^\lambda)} \quad (4.5)$$

where  $\text{Var}(\hat{U}^{D,L})$  is the variance of the estimator  $\hat{U}^{D,L}$  for  $U^{D,L}$ , and  $\sigma_h^\lambda$ ,  $\sigma_m^\lambda$ ,  $\sigma_c^\lambda$ ,  $\rho_{m,c}^\lambda$  are as in Proposition 2.

**Proof (outline).** The proof goes as for Proposition 2, but remembering the relationship  $1 + CV^2(XY) = (1 + CV^2(X))(1 + CV^2(Y))$  ( $CV(X)$  being the coefficient of variation of  $X$ ) when expanding  $(\sigma_m^{D,L})^2$ ,  $(\sigma_c^{D,L})^2$ ,  $(\sigma_h^{D,L})^2$ : e.g.,  $\frac{(\sigma_m^{D,L})^2}{(\varphi_m^{D,L})^2} = \frac{\text{Var}(\hat{U}^{D,L})}{(U^{D,L})^2} + \frac{(\sigma_m^\lambda)^2}{\lambda_m^2} + \frac{\text{Var}(\hat{U}^{D,L})}{(U^{D,L})^2} \frac{(\sigma_m^\lambda)^2}{\lambda_m^2}$ .

Furthermore, it should be noticed that

$$\text{Cov}(\hat{\varphi}_m^{D,L}, \hat{\varphi}_c^{D,L}) = \text{Var}(\hat{U}^{D,L}) (\text{Cov}(\hat{\lambda}_m, \hat{\lambda}_c) + \lambda_m \lambda_c) + (U^{D,L})^2 \cdot \text{Cov}(\hat{\lambda}_m, \hat{\lambda}_c),$$

which is different from zero even when  $\rho_{m,c}^\lambda = 0$ . □

### Comments on Proposition 3.

- When  $\frac{\text{Var}(\hat{U}^{D,L})}{(U^{D,L})^2} \ll 1$  (bottom layers) the credibility factor will be roughly as predicted by Proposition 2.

- However, when the standard error on the estimator of the risk premium for a layer is comparable with the risk premium itself for that layer, the credibility factor is distorted. In the limit for which  $\frac{\text{Var}(\hat{U}^{D,L})}{(U^{D,L})^2} \rightarrow \infty$ , the credibility factor will tend to a limit independent of

frequency heterogeneity:  $Z^{D,L} \rightarrow \frac{\lambda_m^2 - \lambda_m \lambda_c + (\sigma_m^\lambda)^2 - \rho_{m,c}^\lambda \sigma_h^\lambda \sigma_m^\lambda}{\lambda_m^2 + \lambda_c^2 - \lambda_m \lambda_c + (\sigma_m^\lambda)^2 + (\sigma_c^\lambda)^2 - \rho_{m,c}^\lambda \sigma_h^\lambda \sigma_m^\lambda}$ . The exact value

of this limit is of little significance. What is important about this is the practical message that beyond a certain value  $D^*$  of the excess, which might arbitrarily be set to that for which  $\text{Var}(\hat{U}^{D^*,0}) = (U^{D^*,0})^2$  (we call this the *pricing horizon*), the uncertainty of both the client and the market estimates becomes overwhelming and the credibility estimate is of little relevance.

- As for Proposition 1, the credibility factors of Proposition 2 and 3 are theoretical credibility factors, and in practice the values of  $\sigma_h^\lambda, \sigma_m^\lambda, \sigma_c^\lambda, \rho_{m,c}^\lambda, \lambda_m, \lambda_c, U^{D,L}$  will have to be replaced by their estimated counterparts:  $s_h^\lambda, s_m^\lambda, s_c^\lambda, r_{m,c}^\lambda, \hat{\lambda}_m, \hat{\lambda}_c, \hat{U}^{D,L}$ . The estimation of  $s_h^\lambda, s_m^\lambda, s_c^\lambda, r_{m,c}^\lambda$  proceeds as explained in Section 3.4.1 to 3.4.3. Specifically, the correlation can be

written as  $r_{m,c}^\lambda = \sqrt{\frac{\hat{\lambda}_c \sum_j w_c^j}{\hat{\lambda}_m \sum_j w_m^j}}$ : notice how the term related to the average severity has

disappeared.

## 4.2 Hybrid models

In many cases, a hybrid model will be needed, which uses a full uncertainty model (as per Proposition 1) for the bottom layers and a single-severity model (Proposition 2 and 3) for the higher layers. There is no conceptual difficulty in doing this, but it is crucial to deal adequately with how the transition from one method to the other affects the uncertainties.

Specifically, assume the severity distribution of a client is given by

$$F(x) = \begin{cases} F_c(x) & \text{for } \mu \leq x \leq \mu' \\ F_c(\mu') + (1 - F_c(\mu')) \cdot F_m(x) & \text{for } x > \mu' \end{cases} \quad (4.6)$$

where  $F_c(x)$  is the cumulative distribution which is specific to the client and  $F_m(x)$  is the market severity curve, defined above  $\mu'$  and such that  $F_m(\mu') = 0$ ,  $F_m(\infty) = 1$ .

When this is the case, the risk premium for the layer  $(D, D + L)$  with  $D \geq \mu'$  is given by

$$\begin{aligned}\varphi_c^{D,L} &= \lambda_{\geq\mu} \cdot \int_D^{D+L} (1 - F_c(x)) dx = \\ &= \lambda_{\geq\mu} \cdot (1 - F_c(\mu')) \cdot \int_D^{D+L} (1 - F_m(x)) dx = \\ &= \lambda_{\geq\mu'} \cdot \int_D^{D+L} (1 - F_m(x)) dx\end{aligned}\tag{4.7}$$

where  $\lambda_{\geq\mu}$  is the frequency above  $\mu$  and  $\lambda_{\geq\mu'}$  is the frequency above  $\mu'$  (both per unit of exposure). As a consequence, the uncertainty on the risk premium  $\varphi_c^{D,L}$  depends on the uncertainty on  $\lambda_{\geq\mu}$  and on the uncertainty on the parameters of both  $F_c(x)$  and  $F_m(x)$ . The compound effect of these uncertainties is best estimated by stochastic simulation, except in the most trivial cases.

## 5. A REAL-WORLD APPLICATION: PRICING MOTOR REINSURANCE IN THE UK

We have applied the uncertainty-based credibility methodology to pricing motor reinsurance in the UK, based on a sample of 25 clients (about 70% of the UK market share in terms of premium).

### 5.1 The pricing process

Losses were first revalued according to an appropriate claim inflation rate (see, e.g., the study by Swiss Re [18]). Pricing was then carried out by considering a collective risk model where the frequency is a Poisson process and the severity is a Generalised Pareto distribution (GPD), with a cumulative distribution function equal to  $F(x) = 1 - (1 + \xi(x - \mu)/\sigma)^{-1/\xi}$ . The choice of the GPD as the distribution for modelling severity is justified by the Pickands-Balkema-de Haan theorem ([14], [1]), according to which under broad conditions the losses above a certain threshold converge in the distribution sense to a GPD.

Using this model, the risk premium for layer  $(D, D + L)$  is

$$\varphi_c^{D,L} = \frac{\lambda\sigma}{1-\xi} \left( \left( 1 + \frac{\xi}{\sigma} (D - \mu) \right)^{-1/\xi} - \left( 1 + \frac{\xi}{\sigma} (D + L - \mu) \right)^{-1/\xi} \right)\tag{5.1}$$

where  $\lambda = E(N)/w$  is the expected frequency per unit of exposure above  $\mu \leq D$ . This can be easily proven by writing  $\phi_c^{D,L} = \lambda \cdot \int_D^{D+L} (1 - F(x)) dx$  as in the proof of Proposition 3 and setting  $1 - F(x) = (1 + \xi(x - \mu)/\sigma)^{-1/\xi}$ .

## 5.2 Calculating the uncertainties

Determining an estimate  $\hat{\lambda}$  of the Poisson rate,  $\lambda$ , is quite complex as motor insurance has a long-tail component (bodily injury claims) and the number of claims above a certain threshold for a given year is known accurately only after all claims for that year have been settled. As a consequence, claim count projection techniques such as chain ladder or Bornhuetter-Ferguson must be used. The uncertainty on  $\lambda$ ,  $\sigma_\lambda$ , depends on the errors of the chain ladder estimates (Mack [11]; Renshaw & Verrall [15]) for each individual year and on the errors involved in the regression analysis to fit the results for the different years. The distribution of  $\hat{\lambda}$  can be roughly considered normal ( $\hat{\lambda} \sim N(\lambda, \sigma_\lambda)$ ) although positive-definite distributions such as Gamma may be more appropriate.

The values of  $\xi$ ,  $\sigma$  for the GPD can be estimated using maximum-likelihood based on the revalued losses database. The uncertainties are (asymptotically) normally distributed:  $(\hat{\xi}, \hat{\sigma}) \sim N((\xi, \sigma), \Sigma)$ , where the covariance matrix can be estimated as

$$\Sigma = \frac{(1 + \xi)}{n} \begin{pmatrix} (1 + \xi) & -\sigma \\ -\sigma & 2\sigma^2 \end{pmatrix} \approx \frac{(1 + \hat{\xi})}{n} \begin{pmatrix} (1 + \hat{\xi}) & -\hat{\sigma} \\ -\hat{\sigma} & 2\hat{\sigma}^2 \end{pmatrix} \quad (5.2)$$

$n$  being the size of the loss database (Smith [16]; Embrechts et al. [8]). Notice that the uncertainties on  $\xi$  and  $\sigma$  are (negatively) correlated.

By drawing random instances of  $\hat{\lambda}, \hat{\xi}, \hat{\sigma}$  from the distributions above, we obtain the (indirect) sampling distribution for  $\hat{\phi}_c^{D,L}$  and we can estimate the standard deviation  $\sigma_c^{D,L}$  of the risk premium estimator. (There are other uncertainties, such as that on claim inflation, that are not client-specific and are best addressed by sensitivity analysis.)

In practice, one finds that both the frequency estimation and the severity estimation are subject to very large parameter uncertainty and contribute significantly to the overall value of  $\sigma_c^{D,L}$ . One also finds that the distribution of  $\hat{\phi}_c^{D,L}$  is approximately normal for the bottom layers and significantly skewed for the higher layers.

## 5.3 Credibility pricing

A hybrid model for credibility pricing was adopted, which:

*Uncertainty-Based Credibility and its Application to Excess-of-Loss Reinsurance*

- uses the client severity distribution up to £2m, modelled as a GPD (using the GPD model is not critical in that region, and simpler models such as the single-parameter Pareto distribution can be also used);
- uses the market severity distribution above £2m, again modelled as a (different) GPD.

The rationale behind this model is described below. Notice that these results reflect the situation in 2007, with loss data from the latest 10 years.

The hypothesis that there is a single severity curve (the market severity curve) that fits the empirical data of all clients was tested for data above different thresholds: £1m, £2m, £3m. Goodness of fit was tested using the two-sample Kolmogorov-Smirnov test for each client. This test calculates the K-S distance between the empirical distribution of a client and the empirical distribution of the whole market after removing the client's data, and compares this distance with the critical value for a chosen confidence value (see, e.g., Gibbons & Chakraborti [9]). The results of this test are summarised in Table 1.

Analysis Level (£M)	No. of datasets failing test	No. of datasets in test
1	5	18
2	2	21
3	2	23

Table 1 – The number of data sets failing the two-sample KS test as compared to the total number of samples in the set. Notice that the total number of samples varies with the analysis level as for some of the clients the reporting level is too high for an analysis level of, say, £1m to be possible.

The results indicate that while the severity curve of different clients differ significantly above £1m, *the single severity curve hypothesis is valid for a threshold of £2m or above.*

This hybrid approach recognises that in the UK motor reinsurance market there are, broadly speaking, three regions of behaviour:

- I. A “bottom” region (from the lowest excess up to £2m) where clients are quite different as to frequency and severity, and credibility generally decreases with the layer excess (assuming infinitesimal layers).
- II. A “middle” region (from £2m to the market pricing horizon) where clients are assumed to have different frequencies but the same severity distribution. In this region, which extends up to the market pricing horizon (~£20-30m), credibility is broadly independent of excess.

- III. A “top” region that lies beyond the market pricing horizon. In this region, little can be said about the client price, except perhaps providing a broad upper bound to it, and credibility is of little help because both the client price and the market price are far too inaccurate to gain much accuracy by their combination.

## 5.4 Practical issues

In practice, since motor liability is a long-tail business for which bodily injury claims are reported and settled with considerable delay, the risk premium will usually be amended to take into account the time value of money. Specifically,

- losses are usually discounted to take into account the return on investment on the technical reserves between the accident date and the payment/settlement of claims;
- layers’ excesses and limits are usually indexed by earnings inflation. This mechanism is commonly used by reinsurers to avoid excessive gearing effects due to claims inflation.

The effect of these modifications will usually have to be assessed by running a stochastic model, as an exact formula such as (5.1) will not be available. However, the modified risk premium  $\tilde{\varphi}_c^{D,L}$  can be well approximated (errors of 2-5% using standard values of earnings inflation and investment discount rates for a typical reinsurance structure) by the following analytical formula:

$$\tilde{\varphi}_c^{D,L} \approx \frac{\lambda \sigma}{1 - \xi} \frac{\left( \left( 1 + \frac{\xi}{\sigma} (IC(D) - \mu) \right)^{1 - 1/\xi} - \left( 1 + \frac{\xi}{\sigma} (IC(D + L) - \mu) \right)^{1 - 1/\xi} \right)}{(1 + r_{INV})^{\tau - \tau_0}} \quad (5.3)$$

where:  $\lambda$  is the Poisson rate,  $\xi$ ,  $\sigma$ ,  $\mu$  are the GPD parameters,  $IC(X) = (1 + i_{FI})^{\tau - \tau_1} X$  is the layer level after full indexation with future inflation,  $i_{FI}$  is the expected future (earnings) inflation,  $r_{INV}$  is the investment discount rate,  $\tau$  is the mean settlement time, whereas  $\tau_0$  and  $\tau_1$  are offset values that depend on specific assumptions of the algorithm. This formula is an approximation in the sense that it assumes that all claims that happen at time  $t$  will be settled with a single payment at time  $t + \tau$ . This approximation is useful because it allows calculating the standard deviation of the estimator for the risk premium without running a stochastic simulation for every draw of the parameters, thereby reducing the computational complexity of determining credibility. A similar approximation can be used for the market risk premium,  $\tilde{\varphi}_m^{D,L}$ . Notice that the mean settlement time of the market and of the client may differ, which has some (minor) effect on the behaviour of the credibility factor as a function of  $D$ ,  $L$ .

Except for using  $\tilde{\varphi}_c^{D,L}$  and  $\tilde{\varphi}_m^{D,L}$  instead of  $\varphi_c^{D,L}$  and  $\varphi_m^{D,L}$ , the calculation of credibility for the UK motor reinsurance market is a standard application of the methods described in Section 4 – specifically, it is a hybrid model which uses the market severity curve above £2m. Other adjustments (aggregate deductible/limit, reinstatements) are not usually implemented in the UK motor reinsurance market and were therefore ignored in our study.

## 6. RELATIONSHIP WITH PREVIOUS WORK

We are now in a position to discuss at more length the considerations already touched upon in Section 1.1 on the relationship with other research.

The method is formally similar to other methods, in particular to the classical Bühlmann and Bühlmann-Straub methods [4][5][6]. By rearranging the formula for the Bühlmann credibility factor,  $Z = \frac{n}{n + v/a}$  ( $a = \text{Var}(\mu(\theta))$  is the variance of the means of the different clients;  $v = E(v(\theta))$  is the mean of the variances for each client;  $n$  is the number of years of experience), one obtains  $Z = \frac{a}{a + v/n}$ , which has the same form as formula (3.5), by interpreting  $v/n$  as a measure of the standard deviation  $\sigma_c$  of the estimator of the risk premium, and by assuming that the corresponding quantity for the market,  $\sigma_m$ , is zero.

Analogous considerations apply to the Bühlmann-Straub methodology [5][6]. The key difference between the Bühlmann method and the Bühlmann-Straub method is that the latter takes exposure into account – it gives more weight to years with greater exposure. In our case, this is taken into account implicitly, as the standard deviation of the estimator depends crucially on the overall exposure over all years of past experience.

Another similarity to the methods above is the use of a collectively unbiased estimator for the credibility premium (see discussion in Section 3.2).

A work that is closer in spirit to ours is that by Boor [3], where the two estimators  $X_1, X_2$  of the same random variable  $Y$  representing losses are credibility-weighted according to their accuracy and to the difference between them ( $X_1 - X_2$ ), to produce the credibility estimate  $Z \cdot X_1 + (1 - Z) \cdot X_2$ . The general formula for the credibility factor is then  $Z = (E((X_2 - Y)^2) - E((X_1 - Y)^2) + E((X_2 - X_1)^2))/2E((X_2 - X_1)^2)$ . When applied to the case of producing a rate for a subgroup  $\alpha$  ( $n$  elements) of a large group  $\Gamma = \alpha \cup \beta$  ( $n+m$  elements –

ideally, the whole market), this produces the following credibility estimate:  $\varphi = Z \cdot \mu_\alpha + (1 - Z) \cdot \mu_\beta$ , where

$$Z = \frac{\sigma_m^2 / m + (\mu_\alpha - \mu_\beta)^2}{\sigma_n^2 / n + \sigma_m^2 / m + (\mu_\alpha - \mu_\beta)^2} \quad (5.4)$$

and  $\mu_\alpha, \mu_\beta$  are the estimated means for  $\alpha$  and  $\beta$ . This has the same structure as formula (3.6), which holds when the client is compared to the *rest* of the market:

- $\sigma_m^2 / m$  and  $\sigma_n^2 / n$  in (5.4) are central-limit theorem approximations for the quantities  $\sigma_m^2$  and  $\sigma_c^2$  in (3.6);
- $(\mu_\alpha - \mu_\beta)^2$  is used rather than the spread of the market,  $\sigma_h^2$ . Obviously,  $\sigma_h^2$  can be seen as the average value of  $(\mu_\alpha - \mu_\beta)^2$  over all clients except  $\alpha$

Apart from this, the two models are different: [3] uses a two-samples model, whereas we use a collective model where a single measure of the market spread is used for all clients and the correlation between each client and the market is explicitly used.

– o –

Credibility for excess-of-loss reinsurance was first examined by Straub [17]. This was extended by Patrik & Mashitz [13]. An implementation of this approach has been carried out by Bonche [2] for the UK motor reinsurance market. All these efforts have focused on the Poisson/Gamma credibility model applied to claim frequency. The key idea in [17] is that the credibility factor for the Poisson/Gamma model, which is  $Z = \frac{k}{k + b}$  in the Bühlmann case (no excess of loss), becomes

$Z(D) = \frac{k}{k + b/P(X > D)}$  when applied to excess layers, with  $P(X > D)$  being the exceedance

probability, which depends on the severity distribution. The best linear estimate of  $\lambda$  (the Poisson

rate) in this context is therefore  $\lambda_{CREL}^{(1)}(> D) = Z(D) \frac{\sum_{j=1}^k n_j(D)}{k} + (1 - Z(D)) \frac{a}{b} P(X > D)$  where  $n_j(D)$  is the number of claims above  $D$  in year  $j$  for the client. Notice that  $Z(D)$  decreases as  $D$  increases – a property which conforms to the intuition that the client’s experience can be trusted to a lower degree for the higher layers.

Patrik & Mashitz [13] extended this work to the case where  $P(X > D)$  is not assumed to be known with certainty, thus recognising the need to take account of severity uncertainty (see Section 2.3.2 in their paper). This brings to a modified estimate of the credibility frequency, with the credibility factor becoming

$$Z(D) = \frac{k}{k + \frac{b}{E(P(X > D))(1 + (a+1)(CV(P(X > D)))^2)}}.$$

In this formula,  $E(P(X > D))$  and  $CV(P(X > D))$  are the expected value and the coefficient of variation of  $P(X > D)$ . In [13],  $CV(P(X > D))$  is selected so as to incorporate both parameter uncertainty and the subjective beliefs in the *a priori* estimates of the parameters of the severity distribution. Interestingly, in this case  $Z(D)$  is not guaranteed to decrease in  $D$ . Whether or not this is the case depends on the degree to which the increase in the term containing the coefficient of variation compensates the decrease in the expected value of the survival probability.

The main difference between our work and that by Straub [17] and Patrik & Mashitz [13] is that these authors have restricted their attention to claim counts rather than considering the uncertainties on aggregate loss, which is (to borrow an expression from Patrik and Mashitz) the real item of interest when pricing a reinsurance excess cover.

The other obvious difference emerges in the special case where we assume that all clients have the same severity distribution, that of the market. In our single-severity model the credibility factor is broadly constant across the layers, whereas the credibility factor decreases as a function of layer excess in the work of Straub, Patrik and Mashitz (ignoring for the moment the problem of the inaccuracy of the severity distribution). The underlying reason for this difference is that in the Straub-Patrik-Mashitz approach the client frequency above threshold  $D$  is taken as the empirical

mean above that threshold:  $\frac{\sum_{j=1}^k n_j(D)}{k}$ , and as such it is less credible if  $D$  increases; in our approach

the credibility factor is constant, but the client rate above  $D$  is based not on the empirical frequency measured separately for each excess, but on the empirical frequency  $\lambda_c$  at the lowest excess level ( $\mu$ ) projected according to the severity distribution:  $\lambda_c \cdot P(X > D)$ .

This explains the difference in the behaviour of  $Z$ . Our preference is for an approach that gives an approximately constant credibility factor because, if we really believe that the severity distribution is known with certainty, it is more accurate to use as an initial estimate of the number of losses

above  $D > \mu$  the quantity  $\lambda_c(> D) = \lambda_c \cdot P(X > D)$  rather than  $\frac{\sum_{j=1}^k n_j(D)}{k}$ , as the latter approach deliberately disregards the information below  $D$ .

The comparison becomes of course more complicated when the picture is completed considering errors in the severity curve. It is interesting to notice that both our work and [13], despite using different models, reach the conclusion that the uncertainty on the severity distribution ultimately corrupts the behaviour of the credibility estimate and does not guarantee *a priori* that the client will have decreasing credibility.

Cockroft [7] has extended the ideas in [17] and [13] to provide a credibility formula for claim aggregate loss rather than claim frequency. The formula for the credibility factor is still in the form  $Z = \frac{k}{k+b}$ , with  $b$  calculated analytically in terms of infinite series summations. Overall, Cockroft's solution is at this stage quite complex and relies on the assumption that the number of claims is  $\text{Poi}(\lambda)$  with a Gamma prior distribution for  $\lambda$ , and that the claim amounts are distributed according to a Pareto ( $F(x) = 1 - \left(\frac{\theta}{x+\theta}\right)^\alpha$ ) with a Gamma prior distribution for  $\alpha$ .

## 7. LIMITATIONS AND FUTURE RESEARCH

We now look into the limitations of this work and areas for improvement.

- The credibility estimate relies on second-order statistics only. This may not always be appropriate when errors on the parameters are large and the standard deviation may not in itself characterise the distortions on the risk premium in a sufficiently accurate way. More general estimates can be obtained by replacing the mean-squared error minimisation criterion used in Proposition 1 with more sophisticated criteria, perhaps based on the quantiles or the higher moments of the aggregate loss distribution. Further research is needed to explore these different criteria.
- In order to get sound results for the credibility factor a good knowledge of the pricing process and its uncertainties is required. Consider, however, that it is part of the actuary's job to acquire a sufficiently thorough knowledge of the uncertainties of the pricing process anyway. If this knowledge is available, the credibility estimate is simply a byproduct.

- For the method to work it is critical that the process by which the uncertainties are computed be fully automated and that its computational complexity be kept at bay, identifying the variables that have real financial significance. This is especially important if an analytical formula for the price is not available.
- Specifically for reinsurance:
  - o the credibility premium is calculated for each different layer *in isolation*, as if the reinsurance of each layer were bought/sold separately for each layer (this may or may not be the case). As a consequence, the credibility premium – the sum of the credibility premiums of the different layers – is in general not additive, in the sense that  $\varphi^{D, L_1+L_2} \neq \varphi^{D, L_1} + \varphi^{D+L_1, D+L_1+L_2}$ . The overall credibility premium paid for a reinsurance programme may in general depend on the details of the proposed layer structure. Further research is needed to understand what happens when additivity or other regularity conditions are imposed on the credibility premium. Notice that this problem only arises under the general uncertainty-based model, whereas additivity is automatically satisfied for the exact single-severity model (Proposition 2) and approximately satisfied for the single-severity model with inaccurate severity curve (Proposition 3);
  - o the credibility estimate does not give sensible results beyond the *pricing horizon* of the market. This, however, is not strictly a limitation of the method – it is rather the natural consequence of the intrinsic lack of adequate market experience about very high layers;
  - o the empirical calculation of market heterogeneity for the higher layers is quite difficult, due to the large errors involved (see introductory part in Section 4). This reduces the reliability of the credibility premium for those layers. One solution is to produce a realistic model for the behaviour of market heterogeneity as a function of layer excess, rather than relying on the empirical estimate only, much in the same way as we replace the empirical severity distribution with a continuous parametric distribution. We have carried out some preliminary work on this, which has shown that in the case where market heterogeneity becomes negligible in the limit  $D \rightarrow \infty$ , then – under quite general conditions – the credibility factor goes to zero. However, further evidence and research is needed to verify whether this “vanishing heterogeneity” hypothesis is realistic and supported by empirical evidence for some insurance classes. Incidentally, this hypothesis is at odds with the single severity hypothesis, which is strongly supported by empirical evidence in the case of motor XL reinsurance and is quite promising for other lines of

business, too.

## 8. CONCLUSIONS

This paper has presented a novel approach to calculating the credibility premium, called uncertainty-based credibility because it uses the standard deviation of the estimator of the risk premium (for both the client and the market) as the key to calculating the credibility factors.

This approach is especially useful for pricing XL reinsurance, where the balance of client uncertainty, market uncertainty and market heterogeneity is different for each layer of reinsurance. It has been used for pricing motor reinsurance in the UK market.

The methodology is in itself quite general and can be applied to many different problems, essentially to all situations where it is possible to compute the uncertainties of the pricing process and the heterogeneity of the market. Other examples include experience rating in direct insurance (possibly with different excesses) and combining exposure rating (as calculated by using exposure curves) and experience rating in property and liability reinsurance.

## Acknowledgments

This work has been done as part of the research and development activities of Aon Re Services, which is part of Aon Re Ltd.

We are grateful to Dr. Mary Lunn of St. Hugh's College, University of Oxford, for very helpful discussion on the proof of Proposition 1. Jane C. Weiss, FIA, has proposed and subsequently supervised the project. Jun Lin has helped us with many useful suggestions during the real-world implementation of the methodology. Dr. Stephen Mildenhall, FCAS, has reviewed the paper and given us crucial advice on how to restructure it. Warren Dresner, Tomasz Dudek, Matthew Eagle, Liza Gonzalez, David Maneval, Sophia Mealy, Mélodie Pollet-Villard, Jonathan Richardson, Jim Riley have given us helpful suggestions during the project, tested the software implementation of the methodology, and reviewed the paper. We would also like to thank Paul Weaver for support during the implementation of the project and for providing precious commercial feedback.

## 9. REFERENCES

- [1] Balkema, A., and L. de Haan (1974) Residual life time at great age, *Annals of Probability*, 2, 792-804.
- [2] Bonche, S. (2006) Credibility and Reinsurance, *Mémoire présenté devant l'Institut de Science Financière et d'Assurances pour l'obtention du diplôme d'Actuaire de l'Université de Lyon le 7 Décembre 21005* (in English).
- [3] Boor, J. (1992) Credibility based on accuracy, *CAS Proceedings*, LXXIX, Part 2, No. 151.
- [4] Bühlmann, H. (1967) Experience rating and credibility, *ASTIN Bulletin*, 4, 199-207.
- [5] Bühlmann, H., Gisler, A. (2005) *A Course in Credibility Theory and its Applications*, Springer, Berlin
- [6] Bühlmann, H., Straub, E. (1970) Glaubwürdigkeit für Schadensätze (credibility for loss ratios), *Mitteilungen der Vereinigung Schweizerischer Versicherungs-Mathematiker*, 70, 111-133.
- [7] Cockroft, M. (2004) Bayesian credibility for excess of loss reinsurance rating, *GIRO Conference*.

## *Uncertainty-Based Credibility and its Application to Excess-of-Loss Reinsurance*

- [8] Embrechts, P., Kluppelberg, C., Mikosch, T. (1997) *Modelling Extremal Events for Insurance and Finance*, Springer-Verlag, Berlin.
- [9] Gibbons, J.D. and Chakraborti, S. (2003), *Non-parametric statistical inference*, McGraw-Hill, New York
- [10] Klugmann, S.A., Panjer, H.H., and Willmot, G.E. (2004) *Loss Models: From Data to Decisions*, Second edition, John Wiley and Sons, Inc.
- [11] Mack, T. (1993) Distribution-Free Calculation of the Standard Error of Chain Ladder Reserve Estimates, *ASTIN Bulletin International Actuarial Association - Brussels, Belgium* 23:2, 213-225 (<http://www.casact.org/library/astin/vol23no2/213.pdf>).
- [12] Parodi, P. (2007) The role of data uncertainty in pricing reinsurance. (In preparation.)
- [13] Patrik, G. and Mashitz, I. (1990) Credibility for Treaty Reinsurance Excess Pricing, *CAS Discussion Papers*.
- [14] Pickands, J. (1975) Statistical inference using extreme order statistics, *Annals of Statistics*, 3, 119-131.
- [15] Renshaw, A.E. and Verrall, P. (1994) A stochastic model underlying the chain ladder technique, *Proceedings of the XXV ASTIN Colloquium, Cannes*.
- [16] Smith, R. (1987) Estimating Tails of Probability Distributions, *The Annals of Statistics*, 15, No. 3, pp. 1174-1207.
- [17] Straub, E. (1971) Estimation of the number of excess claims by means of the credibility theory, *ASTIN Bulletin*, 5, No. 3.
- [18] Swiss Re (2005), *European Motor Markets* ([http://www.swissre.com/INTERNET/pwsfilpr.nsf/Download?ReadForm&Redirect=../pwsfilpr.nsf/vwFilebyIDKEYLu/MBUI-699DWD/\\$FILE/Publ05\\_European\\_Motor\\_en.pdf](http://www.swissre.com/INTERNET/pwsfilpr.nsf/Download?ReadForm&Redirect=../pwsfilpr.nsf/vwFilebyIDKEYLu/MBUI-699DWD/$FILE/Publ05_European_Motor_en.pdf))

### **Abbreviations and notations**

GPD, Generalised Pareto distribution

XL, excess of loss

### **Biographies of the Authors**

**Pietro Parodi** is an actuarial consultant for Aon Re Ltd in London, UK, where he is involved in pricing and R&D. He has a PhD in Physics from the University of Genoa, Italy, and a Diploma of Actuarial Techniques from the UK Institute of Actuaries. He has held research positions at General Electric R&D in Schenectady, NY, at the Department of Computer Science of the University of Toronto and at the International School for Advanced Studies in Trieste, Italy. He has then moved to the insurance industry working as an IT project manager and a risk consultant for Willis Italy. His past research interests include artificial intelligence, neuroscience and computational complexity theory, on which he has published a number of papers.

Contact information: Aon Re Ltd, 8 Devonshire Square, London EC2M 4PL. E-mail: [pietro.parodi@aon.co.uk](mailto:pietro.parodi@aon.co.uk).

**Stephane Bonche** is an actuary for New Re in Geneva, where he is involved in pricing and complex cover solutions. He graduated at the Institut des Science Financière et d'Assurances in Lyon, with a dissertation on credibility. He has previously worked at Aon Re Ltd in London.

# Consideration of Bias in Chain Ladder Estimates

Rajesh Sahasrabuddhe, FCAS, MAAA

---

## Abstract

The chain ladder method may be the most commonly used and well-known approach for estimating ultimate claims. As it is most often employed, the same development pattern is used to project each accident year and its results are generally considered by practitioners to be valid for each accident year. It is the author's contention that, under this application, the chain ladder method will produce biased projections of the ultimate claims for a single accident year. This paper identifies the sources of the bias and provides the actuary with a tool to understand and compensate for a portion of the bias.

---

## Part 1: Notation, Properties and Relationships

This paper utilizes the following notation:

### Claims

$Y_{i,j}$  The random variable representing the incremental claims for accident period  $i$  and development interval  $j$ .

For example, a triangle of incremental claims may be represented by the following:

	Development Interval					
Accident Period	1	2	3	4	5	6
1	$Y_{1,1}$	$Y_{1,2}$	$Y_{1,3}$	$Y_{1,4}$	$Y_{1,5}$	$Y_{1,6}$
2	$Y_{2,1}$	$Y_{2,2}$	$Y_{2,3}$	$Y_{2,4}$	$Y_{2,5}$	
3	$Y_{3,1}$	$Y_{3,2}$	$Y_{3,3}$	$Y_{3,4}$		
4	$Y_{4,1}$	$Y_{4,2}$	$Y_{4,3}$			
5	$Y_{5,1}$	$Y_{5,2}$				
6	$Y_{6,1}$					

$\sum_j Y_{i,j}$  Cumulative claims for accident period  $i$  as at the end of development interval  $j$ .

$\mu(y)_{i,j}$  The mean of the distribution  $Y_{i,j}$ .

$\boldsymbol{\varepsilon}(\mathbf{y})_{i,j}$  The random error term for observed claims for accident period  $i$  and development interval  $j$ .

### Incremental Claims Development

$F_{i,j}$  The random variable, typically referred to as the “observed incremental development factor,” representing the percentage increase in cumulative claims during interval  $j$  for accident period  $i$ .

$f_{i,j}$  The quantity that actuaries will typically refer to as the “selected incremental claims development factor.” We include the subscript for accident period  $i$ ; however we recognize that, in practice, the selected development factor rarely differs by accident period. We also assume that this factor is determined based on an examination of  $F_{i,j}$  and various arithmetic averages of those observations.

$\boldsymbol{\mu}(\mathbf{f})_{i,j}$  The mean of the distribution  $f_{i,j}$ .

$\boldsymbol{\varepsilon}(\mathbf{f})_{i,j}$  The random error term for the development factor for accident period  $i$  and development interval  $j$ .

### Cumulative Claims Development

$C_{i,j}$  The quantity that actuaries will typically refer to as the “cumulative development factor” evaluated at the end of interval  $j$ . We include the subscript for accident period  $i$ ; however we recognize that, in practice, the selected claims development factor rarely differs by accident period.

### Projections of Ultimate Claims

$U_i$  The random variable representing the ultimate claims for accident period  $i$ .

$D_{i,j}$  The development method projection of ultimate claims for accident period  $i$  as of the end of interval  $j$ .

As a result, we have the following properties and relationships:

#### Claims

$$(1.1) \quad E[Y_{i,j}] = \boldsymbol{\mu}(\mathbf{y})_{i,j}$$

$$(2.1) \quad Y_{i,j} = \boldsymbol{\mu}(\mathbf{y})_{i,j} + \boldsymbol{\varepsilon}(\mathbf{y})_{i,j}$$

$$(3.1) \quad \sum_j Y_{i,j} = \sum_j \boldsymbol{\mu}_{i,j} + \sum_j \boldsymbol{\varepsilon}(\mathbf{y})_{i,j}$$

#### Claims Development

$$(4.1) \quad f_{i,j} \text{ is an estimator of } \boldsymbol{\mu}(\mathbf{f})_{i,j}$$

$$(5.1) \quad F_{i,j} = \boldsymbol{\mu}(\mathbf{f})_{i,j} + \boldsymbol{\varepsilon}(\mathbf{f})_{i,j}$$

$$(6.1) \quad C_{i,j} = \prod_{k=1}^j f_{i,k}$$

**Estimated Ultimate Claims**

(7.1)  $D_{i,j}$  is an estimator of  $U_i$

$$(8.1) \quad D_{i,j} = C_{i,j} \times \sum_j Y_{i,j}$$

$$(8.2) \quad D_{i,j} = C_{i,j} \times \left( \sum_j \mu_{i,j} + \sum_j \varepsilon(y)_{i,j} \right)$$

$$(8.3) \quad D_{i,j} = C_{i,j} \times \sum_j \mu_{i,j} + C_{i,j} \times \sum_j \varepsilon(y)_{i,j}$$

$$(8.4) \quad D_{i,j} = \prod_{j+1}^{\infty} f_{i,j} \times \sum_j \mu_{i,j} + \prod_{j+1}^{\infty} f_{i,j} \times \sum_j \varepsilon(y)_{i,j}$$

**Part 2: Bias in the Chain Ladder Method**

We should now recognize the following properties of the chain ladder method:

- » From 3.1, we recognize that cumulative claims are a function of the expectation of incremental claims for prior periods and the cumulative observed random errors in those prior periods. That is, cumulative claims are a function of all prior observations of incremental claims. From experience, we should recognize that the incremental error terms tend to be correlated. That is, years in which claims are developing adversely or favorably tend to continue to develop in the same manner.

More specifically, through summation of the correlated incremental error terms, there is correlation between the successive observations of cumulative claims. Therefore, we should now recognize that the development factors,  $F_{i,j}$ , within an accident year, are correlated. As a result, they are highly unlikely to be unbiased with respect to  $\mu(f)_{i,j}$  as that would require the sum of the error terms to have an expectation of 0. Although this may be true across multiple years, our experience shows that this is unlikely for a single accident period. This is demonstrated in Part 3 of this paper where we present an example that illustrates what most practitioners observe regularly: that certain accident years have “longer than average” development while others have “shorter than average” development. This occurs because of the correlation of the error terms produces actual development,  $F_{i,j}$ , that are consistently greater or less than the expectation of development,  $\mu(f)_{i,j}$ . Finally, since  $F_{i,j}$  is typically used to estimate  $f_{i,j}$ , it is unlikely that  $f_{i,j}$  is an unbiased estimate of  $\mu(f)_{i,j}$ .

- » Equation 8.4, provides the mathematical representation of  $D_{i,j}$ . In order for  $D_{i,j}$  to be unbiased, the underlying estimators in 8.4 must also be unbiased. The discussion above provides the rationale for  $f_{i,j}$  being considered biased.

Moreover, leaving aside the issue of bias in the development factors, for the chain ladder method to be unbiased, it would require the latest diagonal of observed losses to be “all signal, no noise.” This has the following important implications:

- > The expectation of the sum of  $\epsilon(\mathbf{y})_{ij}$  for accident year  $i$  would have to equal 0. Even if we relax this requirement and allow the sum of  $\epsilon(\mathbf{y})_{ij}$  to be “small,” we should know from experience and the discussion above that this is often not true.

What we should now recognize is that implementation of the chain ladder method ignores a fundamental truth of the claims emergence process, specifically:

1. the existence of correlations within an accident year, and
2. that the chain ladder method is almost certainly biased.

However, there is a method for consideration (though not elimination) of bias resulting from (1) the presence of error terms and (2) the correlation of error terms within an accident year. This method is the subject of the third part of this paper.

In the discussion above, readers should recognize that we have not yet even explored the impact of environmental factors on both  $\mu(\mathbf{f})_{ij}$  and  $\mu(\mathbf{y})_{ij}$ . These factors would include unexpected inflation, changes in limits, changes in case reserving, changes in payment practices and numerous other influences. It is hoped that readers recognize that real-world influences result in the virtual impossibility that development method estimators are unbiased<sup>1</sup>.

It is therefore incumbent on practitioners to evaluate whether its convenience is a sufficiently significant benefit to overcome its shortcomings. While this is true of other reserving methods as well, the goal of this paper is to raise the awareness of one particular shortcoming of the chain ladder method.

### **Part 3: Partial Correction for Bias**

Correction for bias in the development factors is beyond the scope of this paper. However, we do have a mechanism for (partially) addressing the bias created by both the presence and the correlation of (cumulative) error terms (the rightmost term of Equation 8.4). These conditions have the result that individual years will experience longer (more) or shorter (less) development than that implied by the selected development pattern. Additionally, also as demonstrated in Equation 8.4, the chain ladder method indiscriminately develops both the signal and noise component of the observed claims value. To address these issues we need to (1) use a tool that separates the “signal” from the “noise” and (2) employ a methodology that tracks the impact of the correlation.

---

<sup>1</sup> To correct for the bias resulting from changes in environmental factors, we would need to incorporate adjustment factors that would offset these biases. The author recognizes that it is likely not possible to calculate adjustment factors for all such changes regardless of the actuarial method selected. However, a frequency-severity method probably best allows for such adjustments as the parameters of that method (i.e. no of claims and value of claims) are specified at the same level of detail that the underlying changes would be expected to influence.

Regression is the typical tool used to isolate the signal from a series of observations of a random variable. We now turn to the question of how to apply principles of regression within the chain ladder method so as to also assess the bias created by the correlation of error terms. To do this, we should recognize that we need not apply the cumulative development factor solely to the last diagonal of the triangle. We can also apply development factors to all prior diagonals as well. We refer to this series of projected ultimate claims as the “retrospective estimates of ultimate claims.”

### **Benefits of Regression**

Use of regression in this context has multiple benefits:

- (1) Fitting a regression model to the series of projected ultimate claims will (partially) differentiate between the predictive component of  $D_{i,j}$  (the first term on the right side of Equation 8.4) and the “noise” (the second term on the right side of Equation 8.4). This will, in effect, reduce the impact of the error terms and therefore partially correct for the bias in the  $D_{i,j}$  that results from the noise / error terms.
- (2) Testing of the significance of the regression parameters will provide additional insight on the development applicable for any particular year. That is, the regression coefficient will be greater than 0 for years where the ultimate claims estimate is increasing; the regression coefficient will be less than 0 for years where the ultimate claims estimate is decreasing. More specifically, coefficients that are significant and greater than 0 would indicate that the development for a particular year was “longer” than average. Stated differently it would indicate the error terms,  $\epsilon(\mathbf{y})_{i,j}$ , were positive. Conversely, coefficients that are significant and less than 0 would indicate that the error terms,  $\epsilon(\mathbf{y})_{i,j}$  were negative. The value of the coefficient would also be an indicator of the strength of the correlation of incremental errors.
- (3) Finally, we could also create a statistic used to measure “net bias” for the development pattern. For example, regression coefficients significant and greater than 0 would contribute +1 to this statistic and coefficients significant and less than 0 would contribute -1 to the statistic. This would allow us to measure whether our development pattern was too long or too short with respect to the claims portfolio under review.

### **Description of Exhibits**

These calculations are demonstrated on the attached exhibits.

**Exhibit 1** - The data used in this example is based on the General Liability excluding Mass Torts (combined treaty and facultative) claims data as compiled by the Reinsurance Association of America. This data is presented on Exhibit 1. We also show the selected incremental claims development factors on this exhibit. For simplicity, this

presentation assumes that (1) the selected incremental factors are based on volume weighted averages and (2) there is no need for a tail factor.

**Exhibit 2** – Exhibit 2 presents the triangle of retrospective estimate of ultimate claims. Each of the entries in the triangle is calculated as the product of the observed claims and appropriate claims development factor.

**Exhibit 3** – Exhibit 3 presents the results of a regression model applied to the last five observations of the retrospective ultimate claims triangle. For simplicity we have used a linear regression model in order to conceptually demonstrate the approach. However, reader should recognize that alternative regressions (such as exponential or logarithmic) could also be used as the shape of the curve warranted.

**Exhibit 4** – Exhibit 4 presents estimates of ultimate loss as fitted by the regression model.

**Exhibit 5** – Exhibit 5 presents a graphical presentation of this model for accident year 1995.

#### **Part 4: Conclusions**

Readers should now realize that the chain ladder method is not simply an application of an algorithm to yield a deterministic result. Rather, it is a method that has implicit statistical underpinnings. With this knowledge, we can now turn to an evaluation of the methods from a statistical basis. With this analysis, it becomes apparent that the chain ladder as it is currently applied in practice is not unbiased. Unbiasedness is one of the qualities that we typically desire in statistical estimators – yet practitioners have (implicitly) chosen to ignore this property of the chain ladder method.

The paper then identifies the two primary sources of bias that result from the correlation of error terms in the cumulative observations of claims: (1) bias in the development factor estimators and (2) the bias created by the error terms themselves. The first is beyond the scope of the paper. For the latter, the paper provides a discussion of the use of retrospective estimates of ultimate claims and regression techniques that may be used to address the bias. However, even with these tools, we are not able to completely eliminate its impact.

#### **Part 5: Acknowledgements**

The author wishes to thank Bernard Chan, FCAS, MAAA, Katy Siu and the CAS Forum Committee for their reviews of this paper. Any errors that remain herein are the responsibility of the author. As with many research topics, the concepts presented herein are a “work-in-progress.” The author would welcome your comments. Please consult the CAS member directory for contact information.

1-December-2007