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 slyness… 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,"

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

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

quality. Both the IDMA and Olson cite an estimate that data quality problems cost companies 15% - 20% of operating profits.[1] The IDMA value proposition[2] also cites an estimate that poor data costs the U.S. economy $600 billion a year[3]. 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

---

[1] Olson, p9.
[2] This citation is apparently from a study done by The Data Warehouse Institute
[3] Based on information at econostats.com the 2006 gross domestic product of the U.S. was about $13,000 billion.

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

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 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[4] 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

[4] An association of state insurance regulators in the United States

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."[5]

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

[5] This finding is stated in the Executive Summary of the report.

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[6] 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

[6] In some cases, other quantitative analysts and systems people who work with and support actuaries were included in the survey.

- A sample of attendees at a WRG[7] 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.

[7] World Research Group, March 2007

Question 1: Percentage of Time Spent on Data Quality Issues

Employer	Number of Responses	Mean	Median	Minimum	Maximum
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

Employer	Number of Responses	Mean	Median	Minimum	Maximum
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.

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

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[8]. 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.

---

[8] 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[9]. 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.

[9] 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.

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

markers show the results based on the three datasets.

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

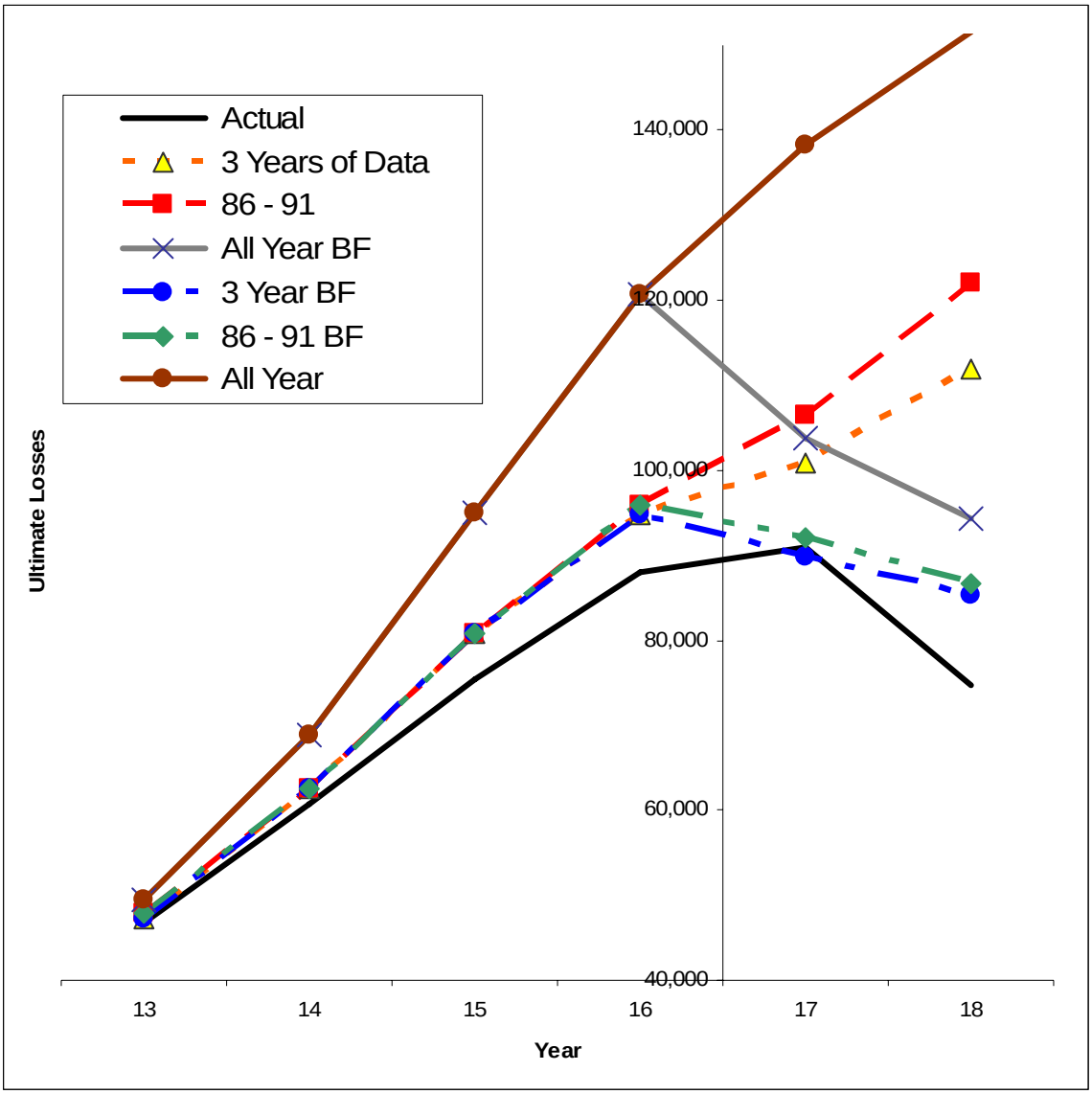
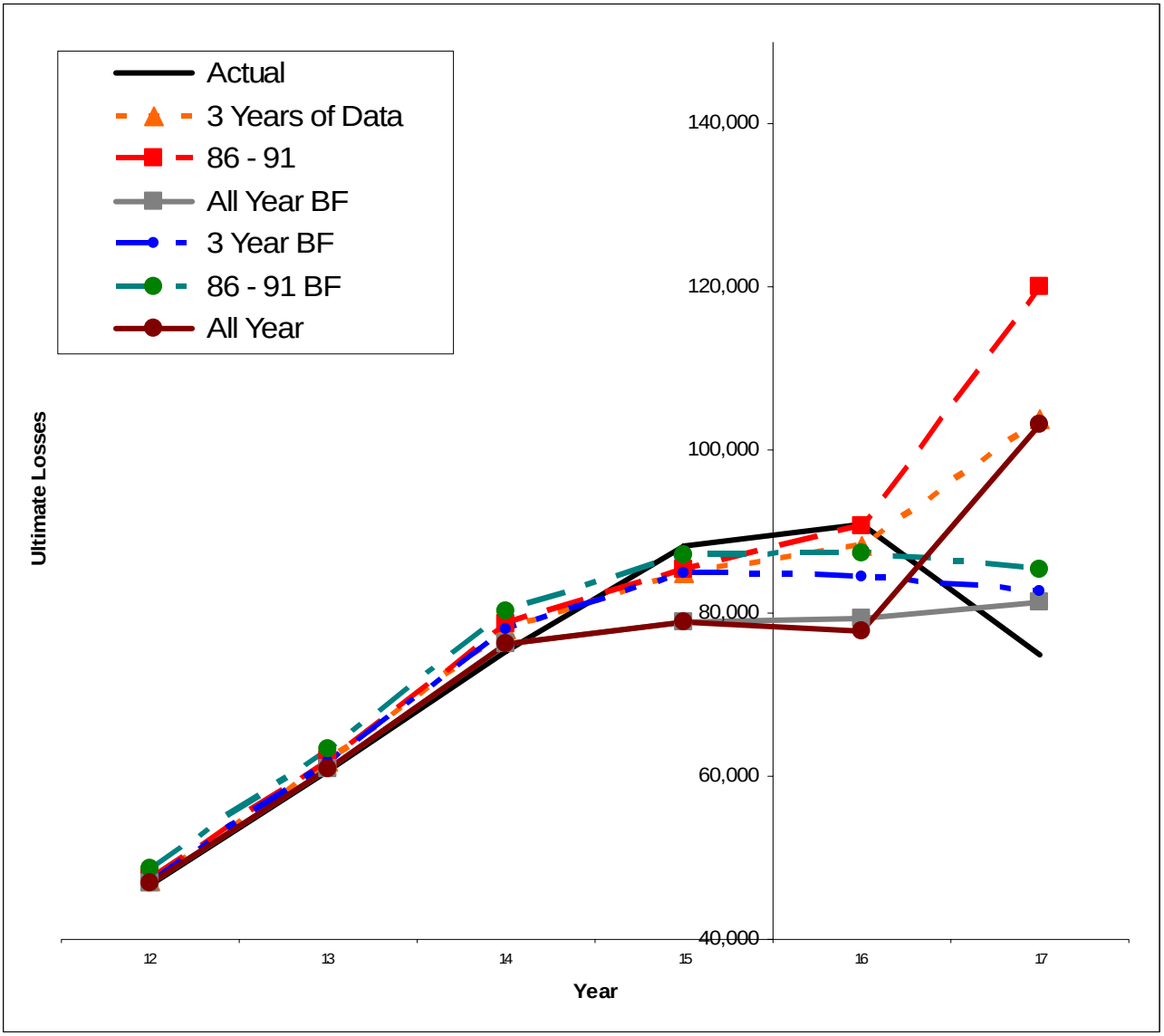


Figure 4.2: Estimated Ultimate Losses by Year Based on Adjusted Paid Data



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

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.[10] 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.

[10] 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

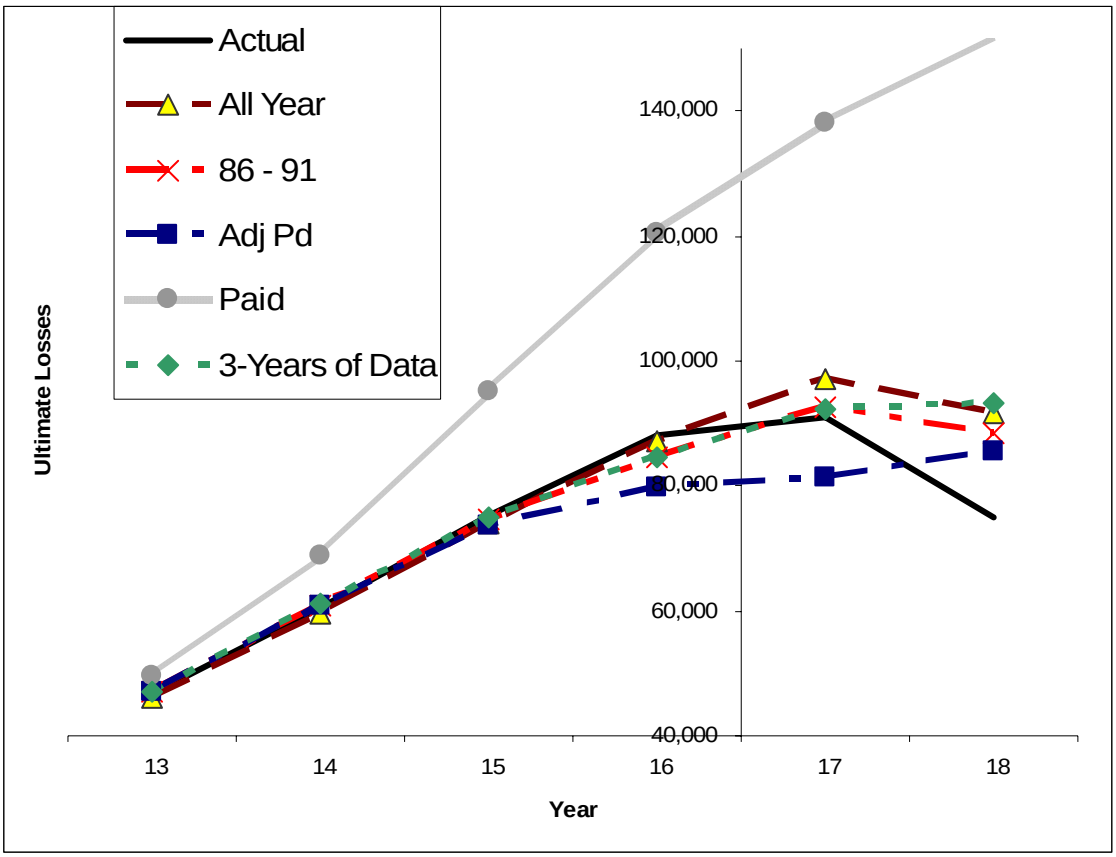


Figure 4.3 presents the results for estimated losses based on incurred loss data. For comparison, the graph also displays the ultimate losses from the all years paid and adjusted paid techniques. It is clear from this graph that the estimated ultimate losses based on incurred loss data are considerably 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)
$$se = \sqrt{\frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}{N-1}},\ Y_i \text{ is actual ultimate, } \hat{Y}_i \text{ is estimate}$$

Table 4.1: Bias of Estimation Methods and Datasets

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

Table 4.2: Standard Error of Estimation Methods and Datasets

	All Years	3-Years	86 - 91	All Year BF	3-Year BF	86 - 91 BF
Paid	5,460	2,197	3,137	2,525	765	1,098
Adjusted Paid	1,806	1,633	2,679	896	618	705
Incurred	1,003	1,048	566			
Adjusted Incurred	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.

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

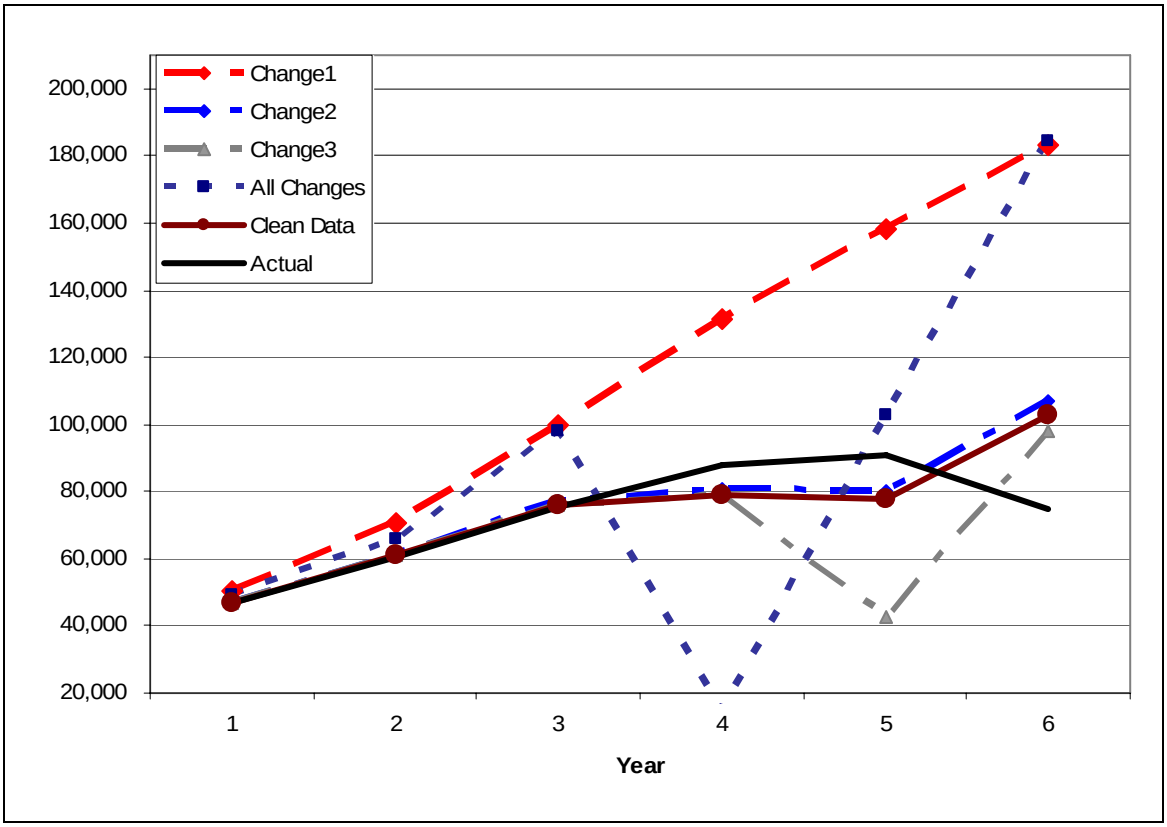
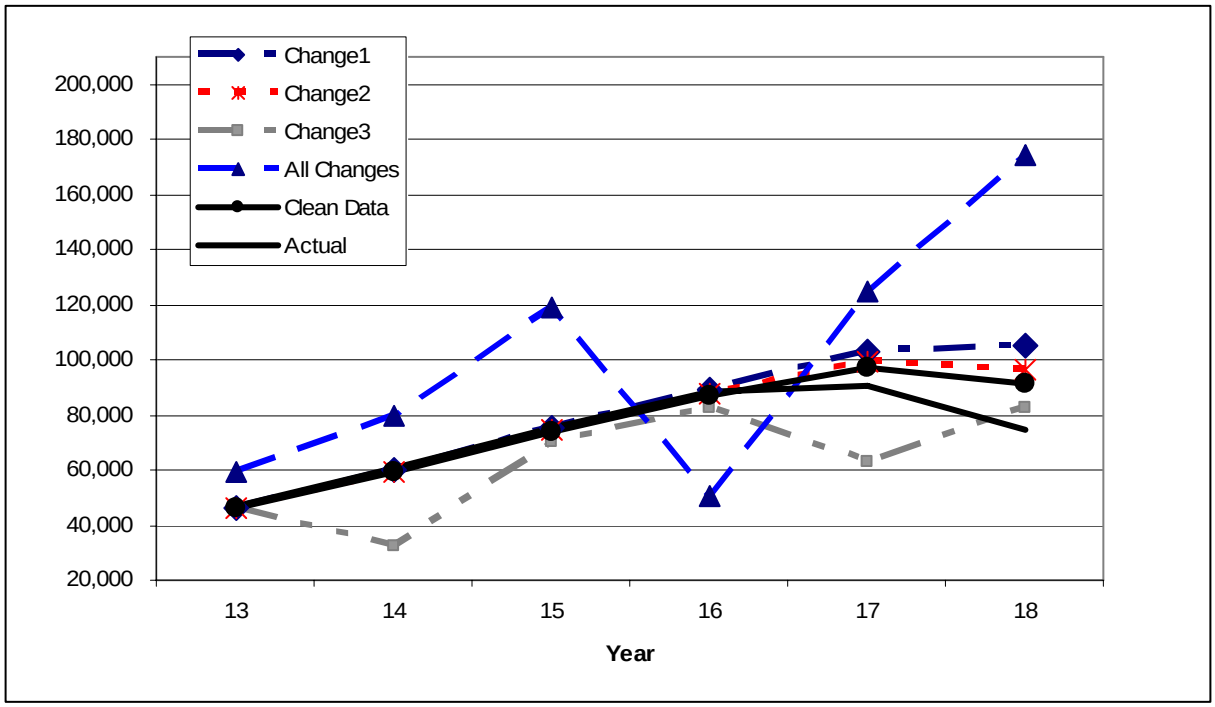


Figure 4.5: Comparison of Actual and Estimated Ultimate Losses Based on Error-Modified Incurred Loss Data



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).

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

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

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

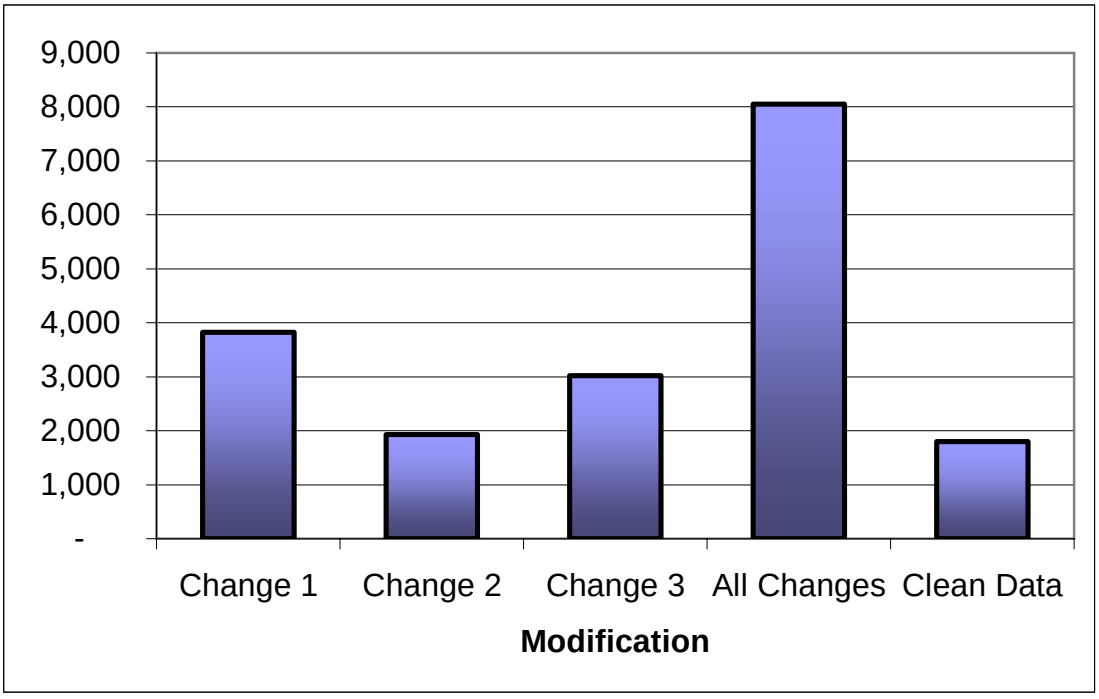
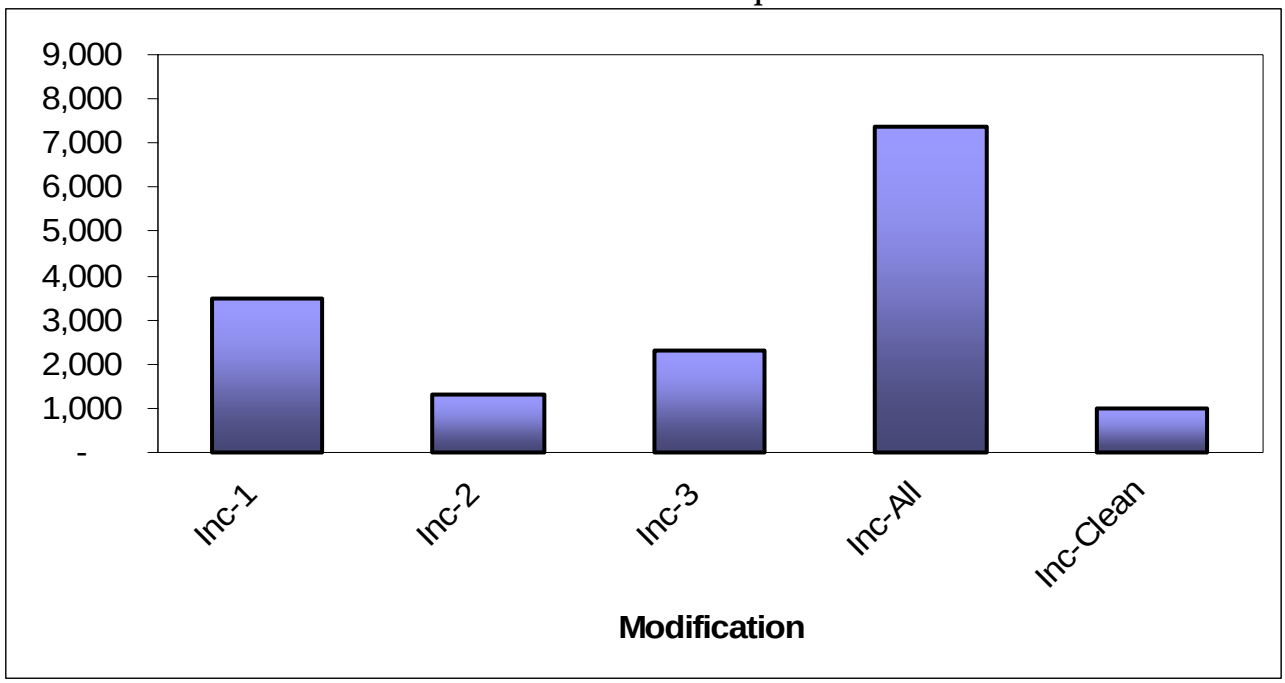


Figure 4.7: Standard Errors for Incurred Loss Data Modified to Incorporate Errors



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

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[11]. 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[12]:

(4.2) $$r = \frac{x - \mu}{\sqrt{Var(\mu)}},$$ x=actual value, μ 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

[11] 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.

[12] Following the procedures described by England and Verrall (England and Verrall, 1999), the variance is assumed to be proportional to the expected value.

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.

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).

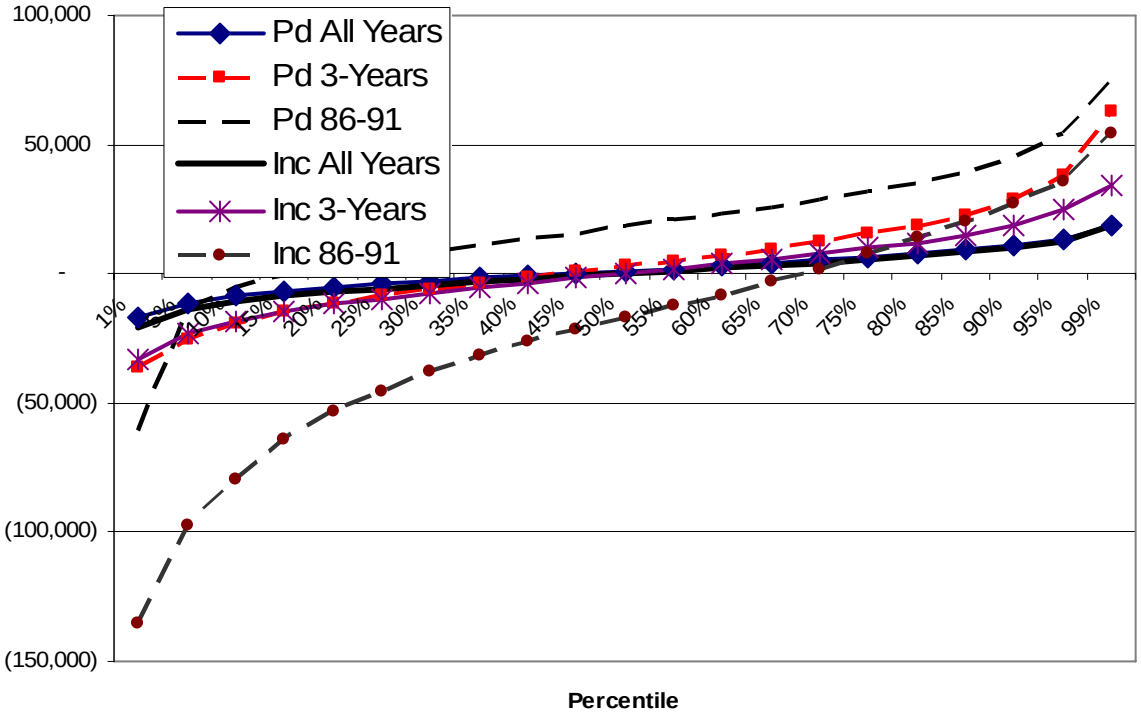
Table 4.4: Bootstrap Results Based on Incurred Chain Ladder Method

	Incurred Actual Reserve	Incurred Clean Data	Incurred 1986-1991 Data	Incurred Modified Data
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[13] was tabulated. Figure 4.8 displays the cumulative distribution of errors from the bootstrap experiment for the complete and incomplete data sets.

[13] 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.

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

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

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

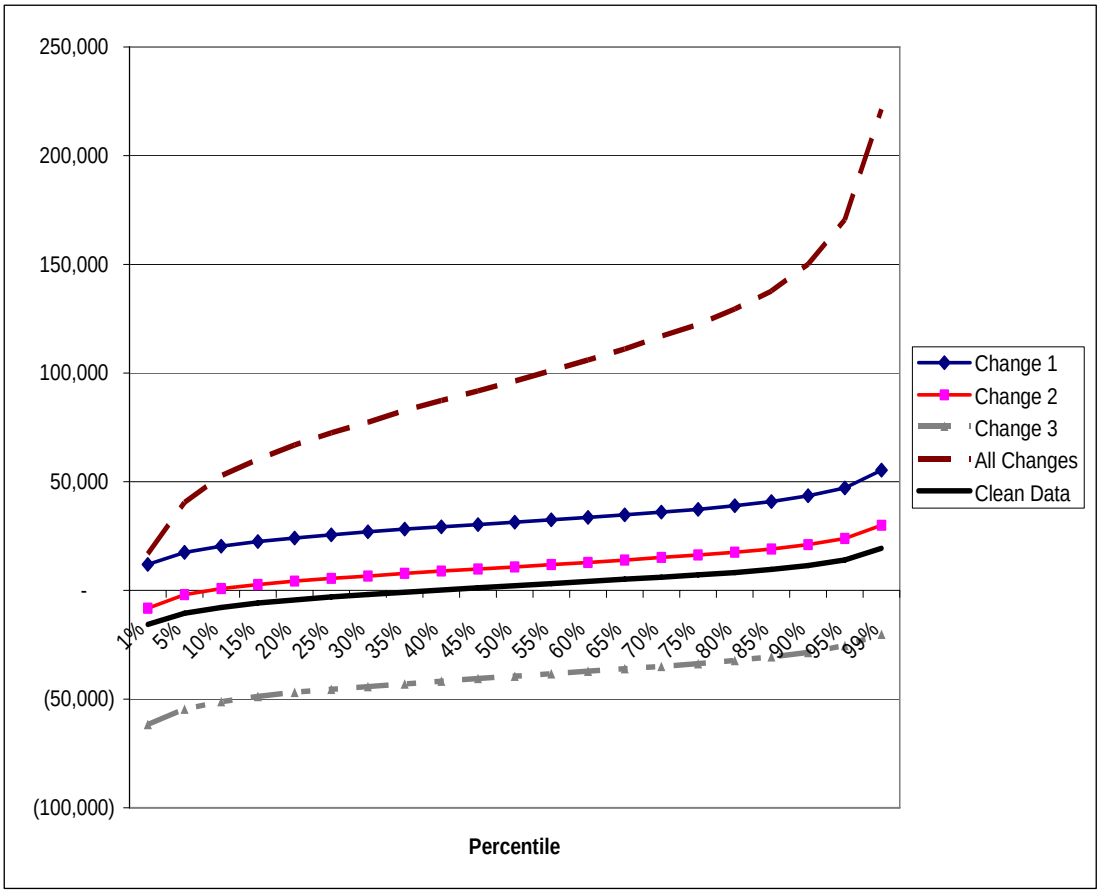
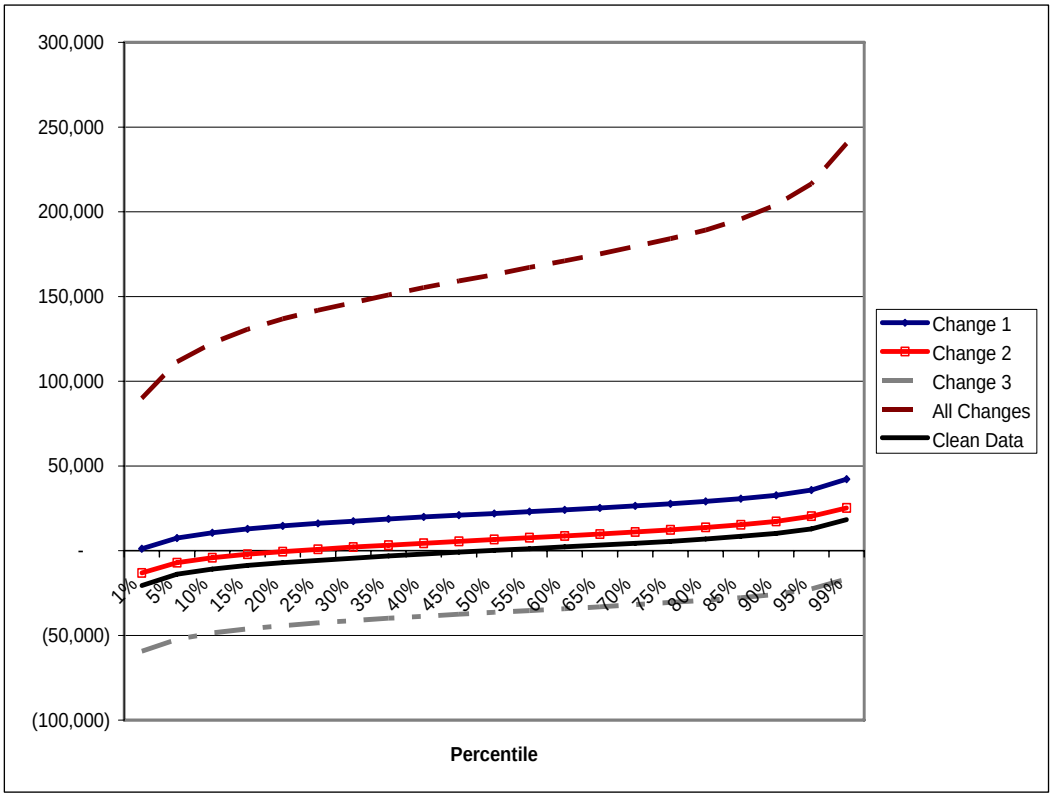


Figure 4.10: Distribution of Reserve Estimation Errors for Incurred Ultimates Based on Modified Data



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

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[14] 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.

[14] Redman, p. 107

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[15]:

- 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

[15] Redman, p108

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[16]

- 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."[17] On the contrary, Redman suggests that organizations do not provide adequate communication and feedback to their data suppliers. Thus Redman suggests[18]

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

[16] Redman, p.162
[17] Redman, p. 154
[18] Redman, p.155

- 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.

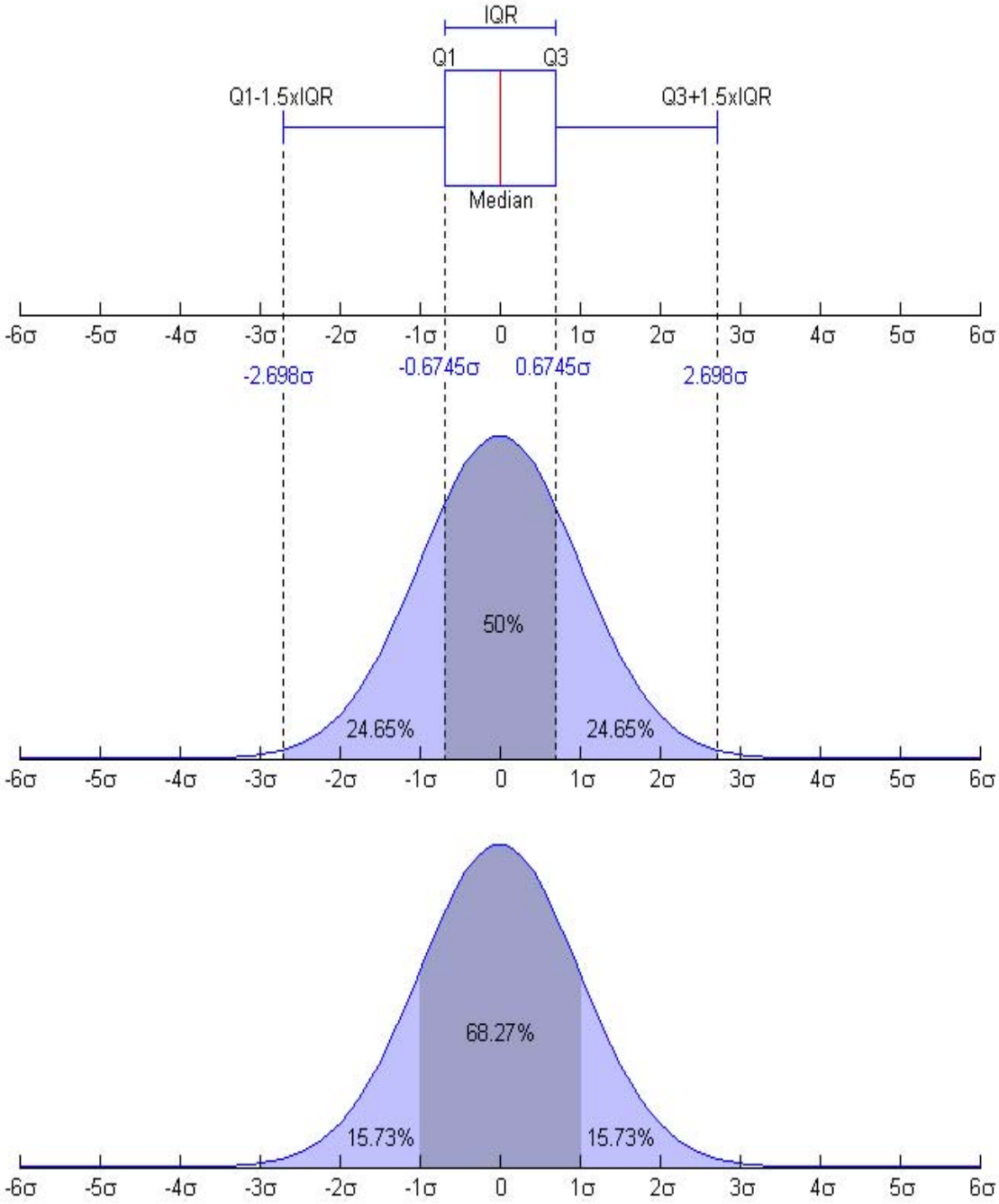
5.2.1 Primer on Box-and-whisker Plots

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[19] is presented below:

[19] 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

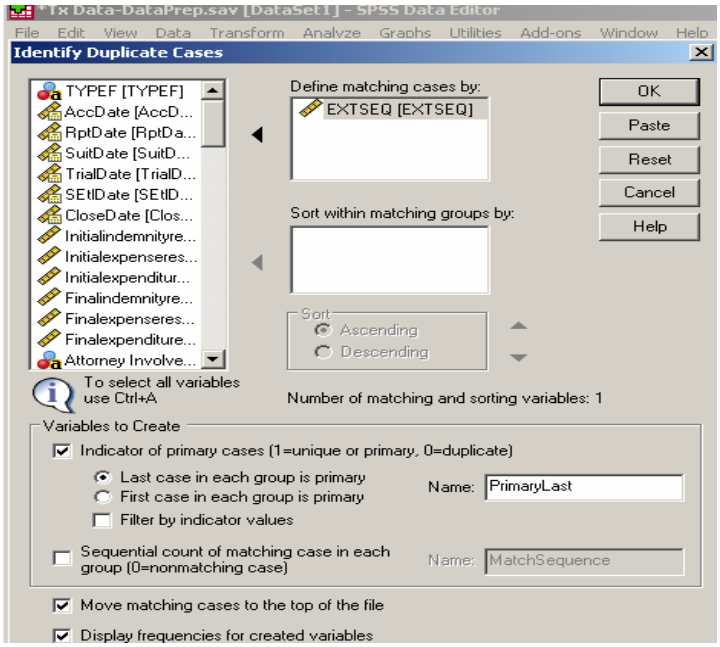


Table 5.1: Indicator of Duplicate Case

		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:

- 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.[20]
- 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.

[20] 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

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

- 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

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

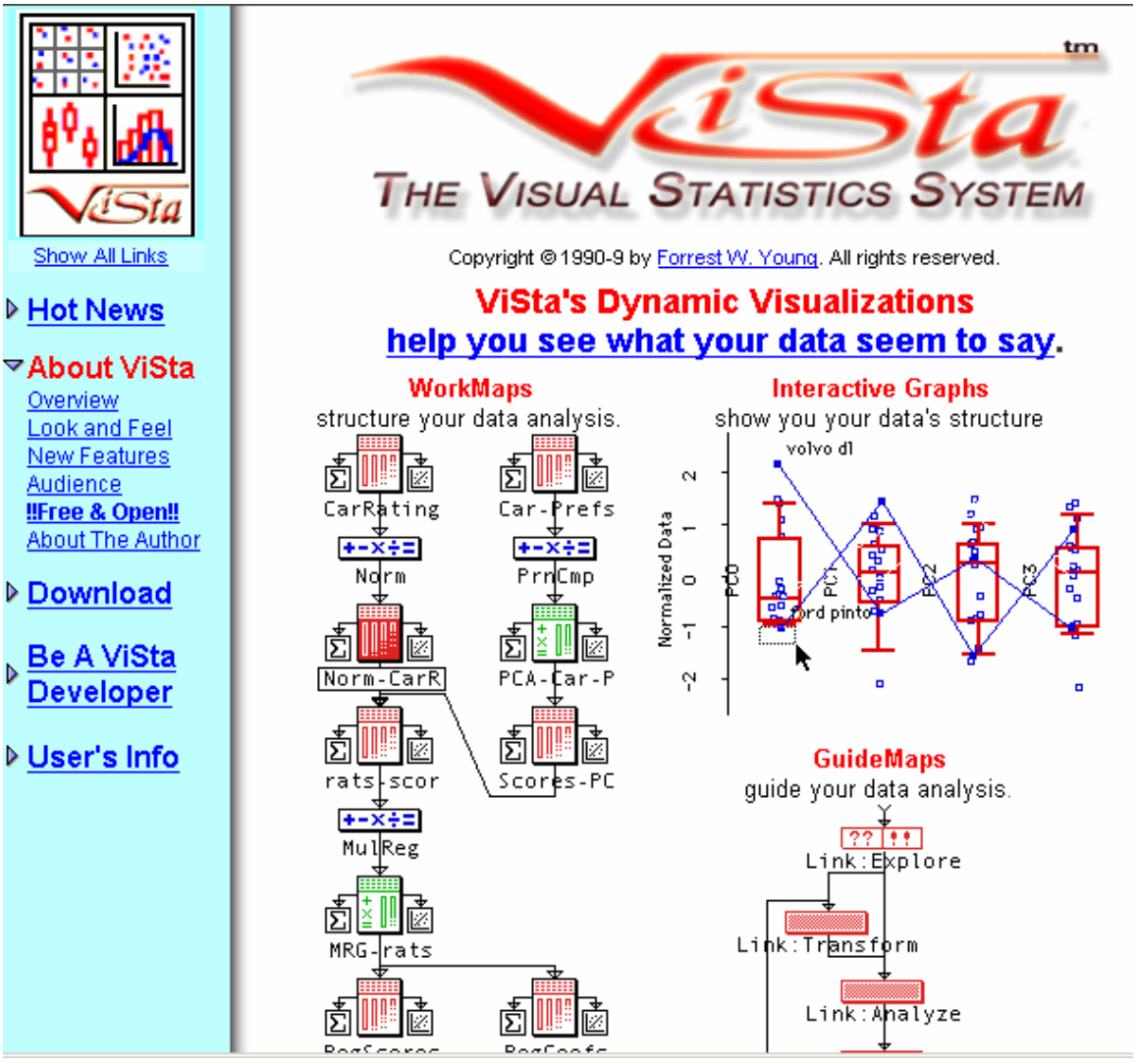
Appendix C: Experiment Projections based on Unmodified Data

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:

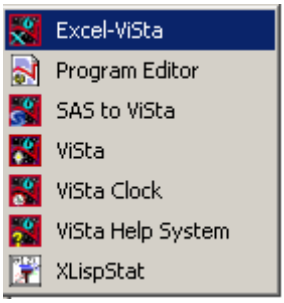


Going down the left side of the screen you will see several options offered including

"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.

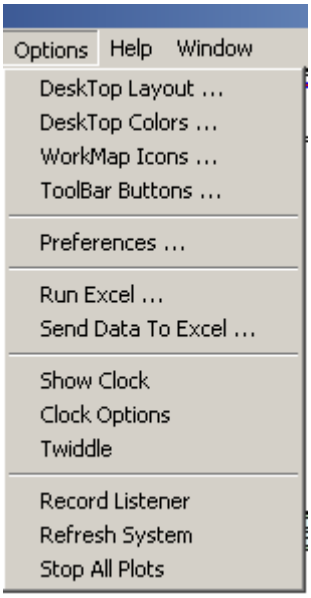
The first challenge to overcome is bringing data into ViSta. Because ViSta is programmed in the XLisp language, it reads Lisp files. However, it also has the capability of reading Excel files, text files and SAS files. Since a lot of actuarial analyses are done in Excel, it is relatively easy to read 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.



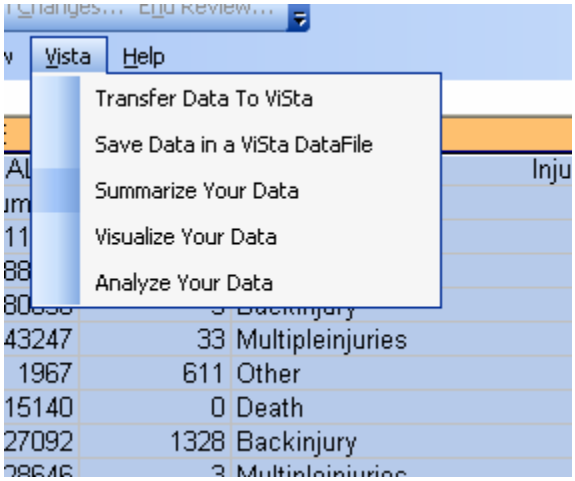
The first time the Excel-ViSta option is chosen, the user is asked to supply the location of the "Excel.exe" file. This typically resides in the Program/Office directory, but its location should be identified by using the search option of Windows explorer before attempting to use Excel and ViSta together.

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

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

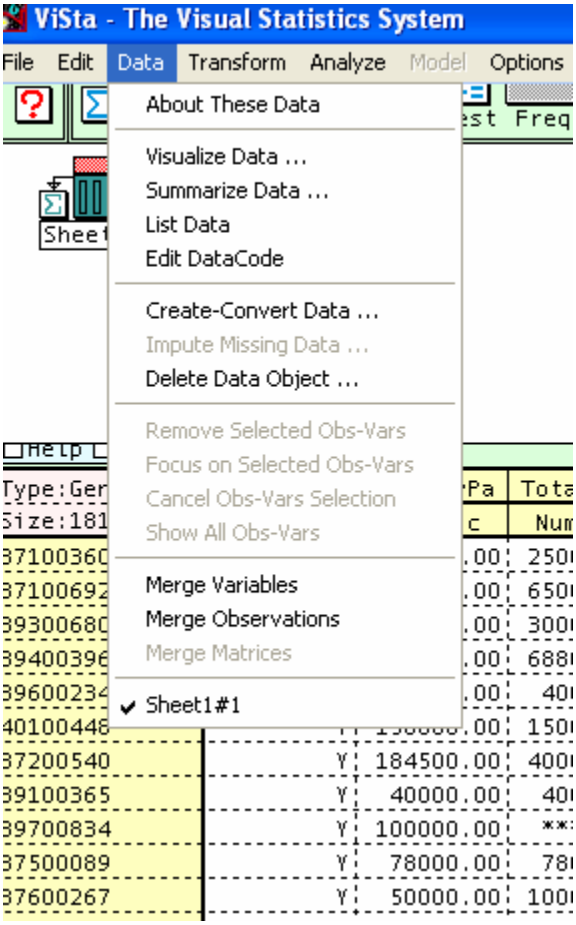


- When Excel is launched, make sure to enable macros.
- From within Excel open the database you want to analyze.
- Highlight all the data you want read into ViSta, while in Excel.
- In Excel, click on the ViSta tab on the tool bars at the top of the worksheet, then click "Transfer Data to ViSta."



- 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.

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[21]
  - Bar Plots (Figure5.6)

[21] 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

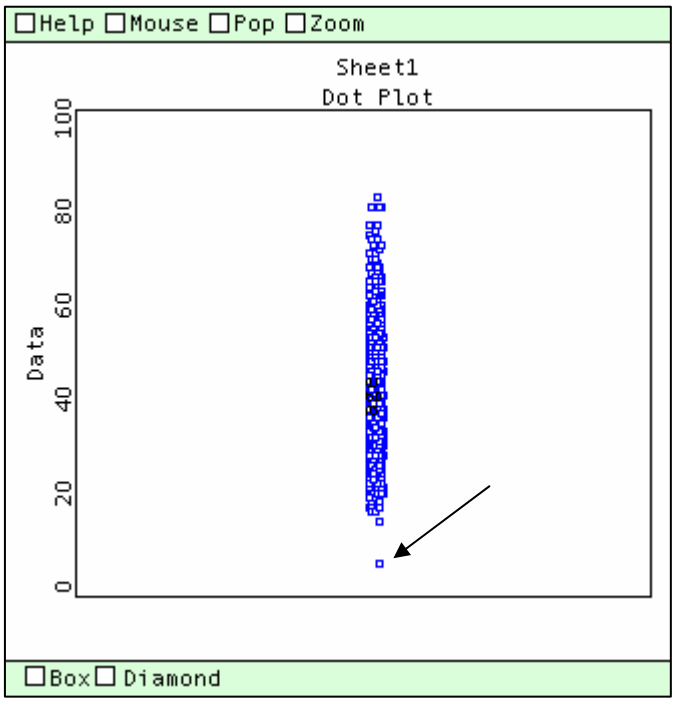


Figure A.2: Box-and-whisker Plot for Claimant Age with Error

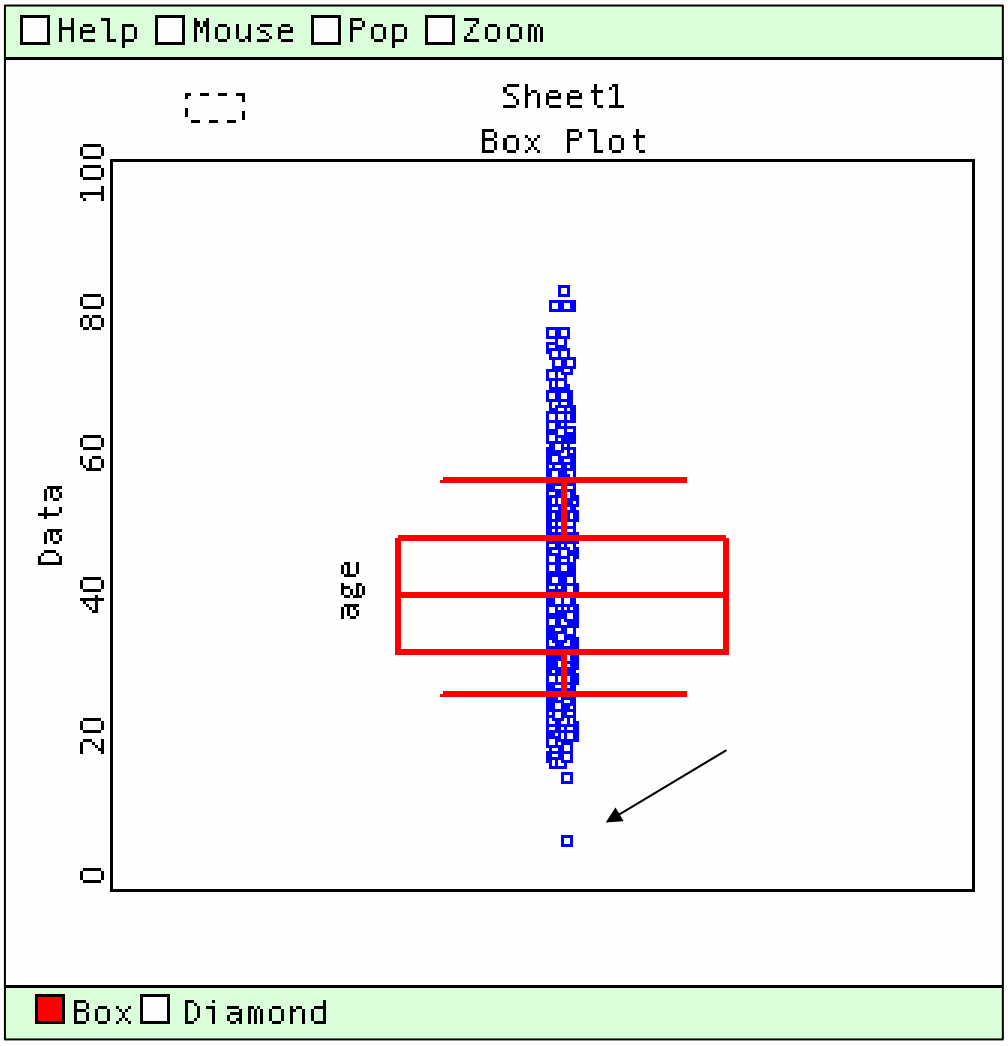


Figure A.3: Frequency Polygon of Log of Primary Paid Losses – No Errors in Data

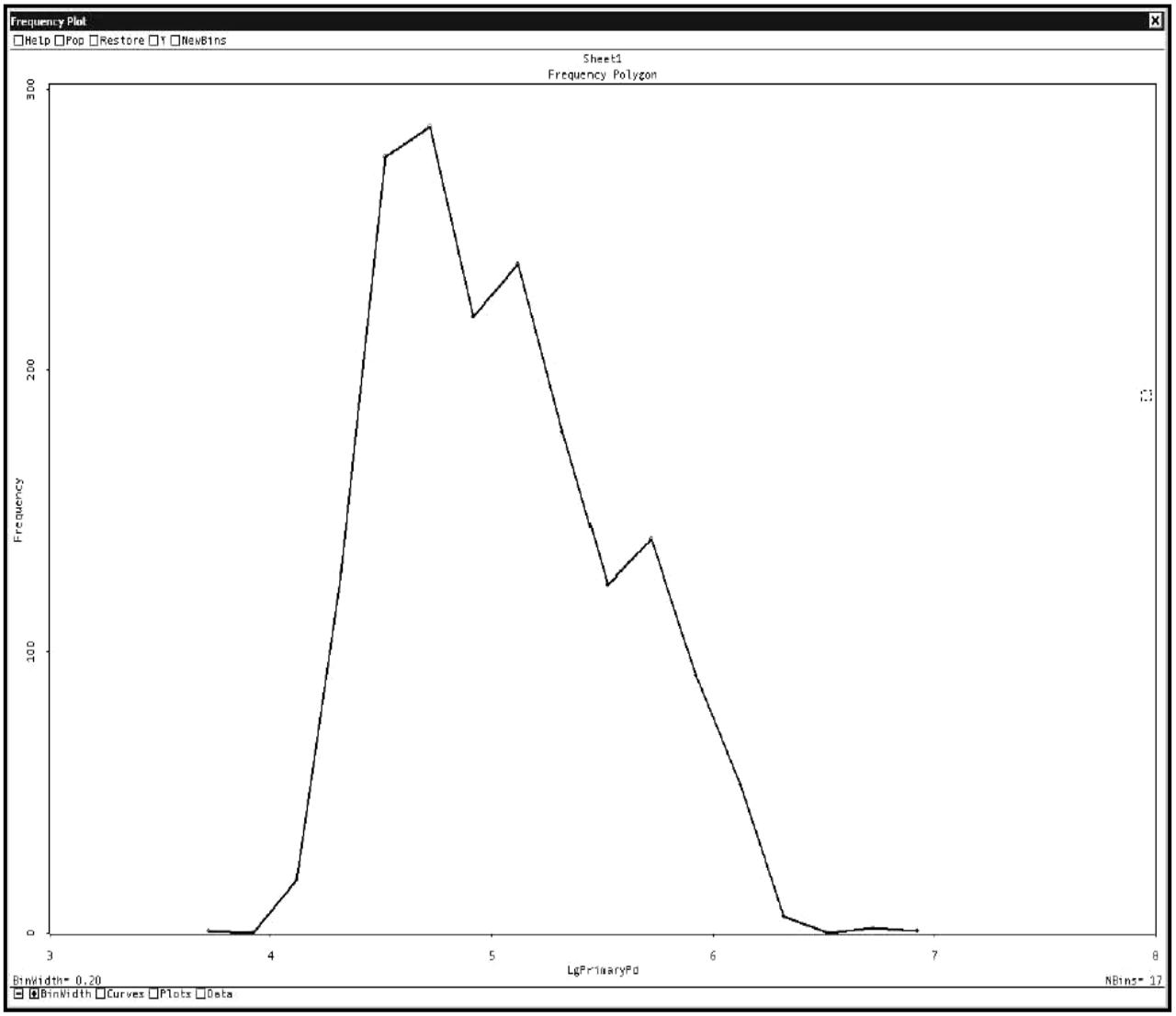


Figure A.4: Histogram of Log of Primary Paid Losses – Errors in Data

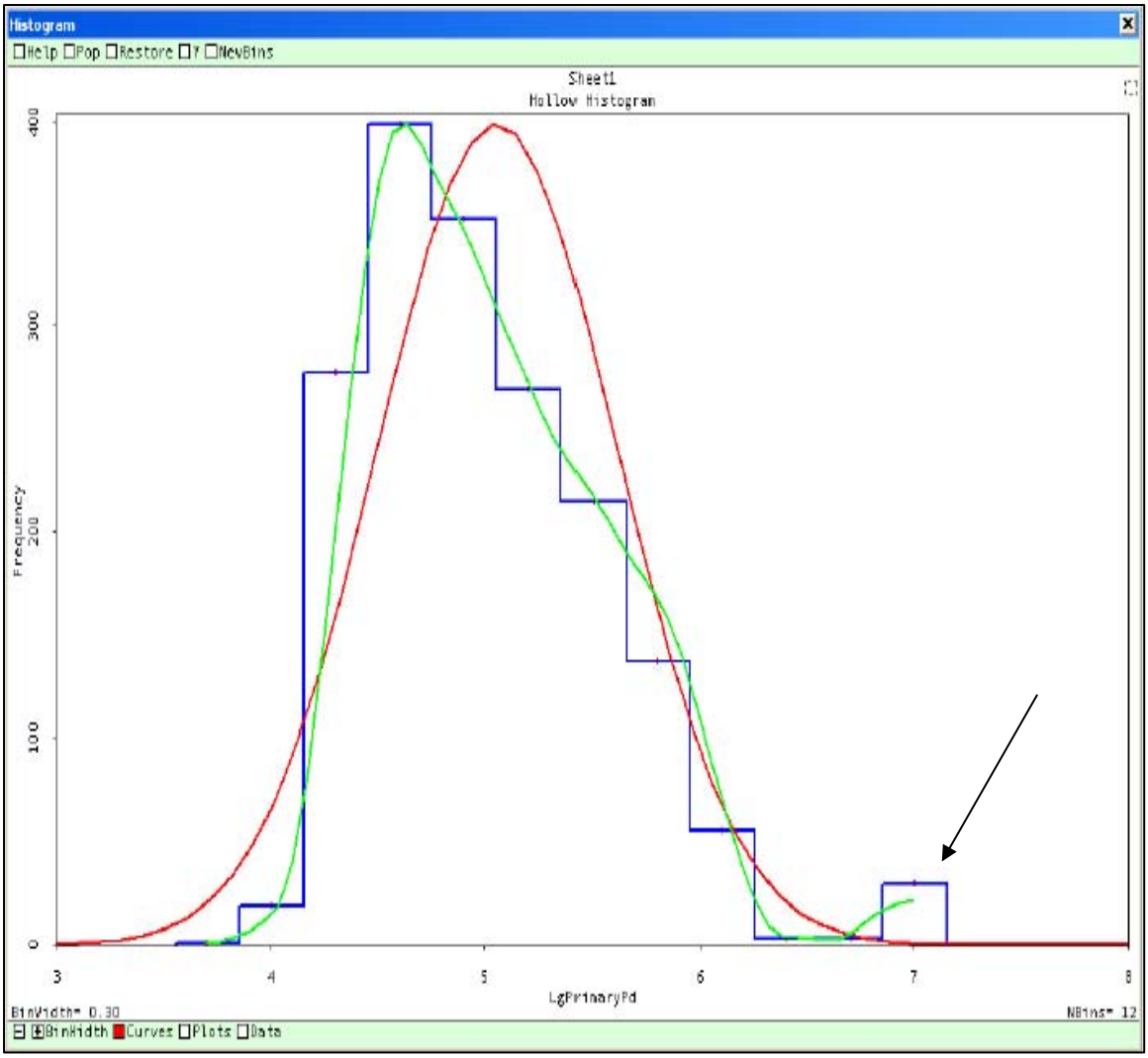
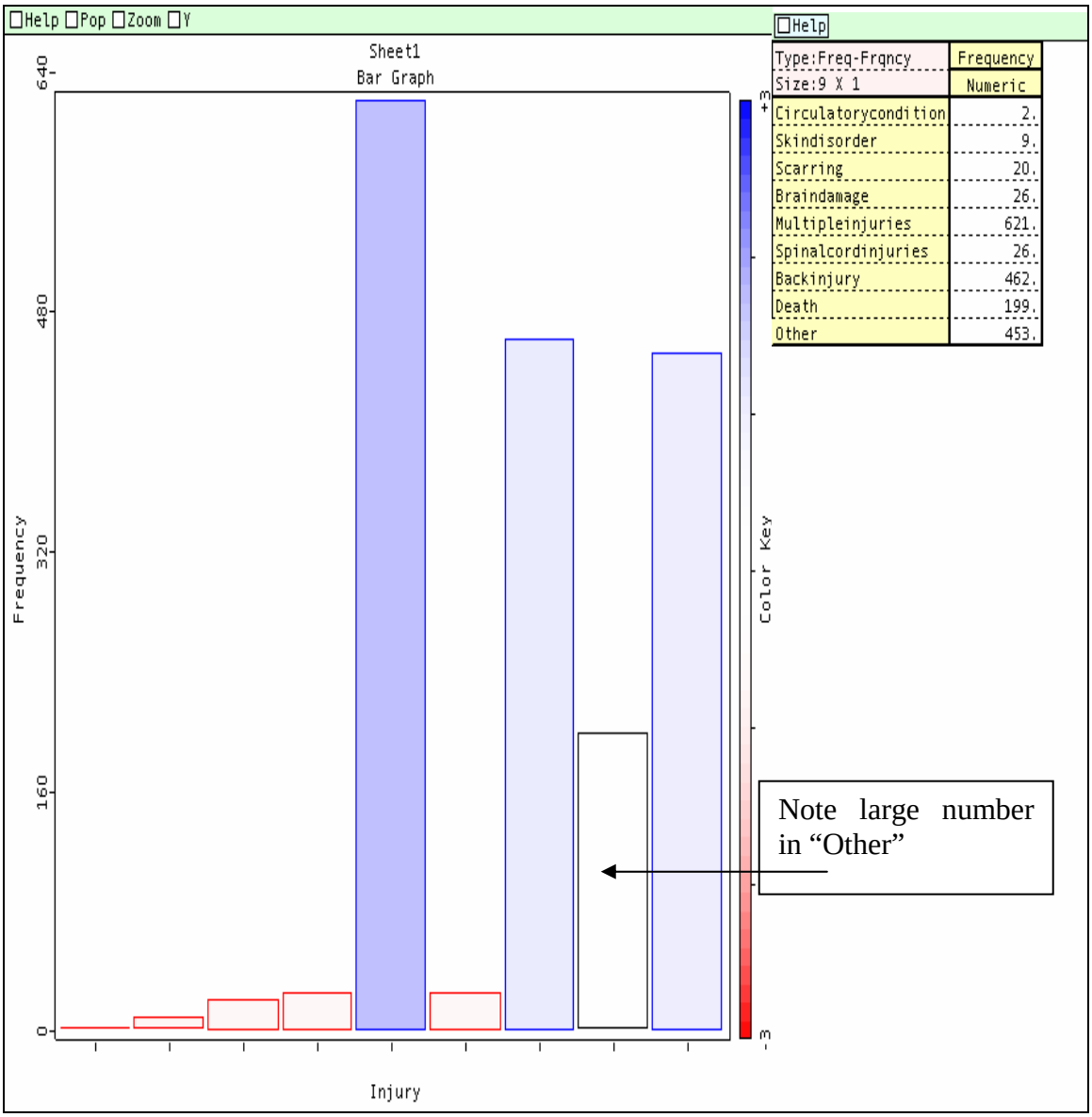


Figure A.5. Bar Plot for Categorical Data



7. REFERENCES

[1] Actuarial Standards Board, Actuarial Standard of Practice 23, Data Quality, Revised Edition, December 2004, www.actuaries.org.

[2] Anderson, Bolton, Callen, Cross, Howard, Mitchell, Murphy, Rakow, Sterling, Welsch, "General Insurance Premium Rating-The Way Foreward," 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.

[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, 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.

[20] Data Quality Solutions, www.dataqulaity solutions.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 , August 16, 2006.

[28] IQ Trainwrecks, www.iqtrianwrecks.com.

[29] Insurance Data Management Association, "Value Proposition," www.idma.org.

[30] Moore, L., "Data Business—No data quality control? Expect to count the cost," 2006, *Business Media Europe*.

[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, "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
DFA, dynamic financial analysis
General Insurance, non-life insurance
GIRO: General Insurance Research Organization
GLM, generalized linear models
GUI, Graphical user interface
GLM, generalized linear models
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

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.

Appendix B

Cumulative Paid Losses

Accident	Months of Development																	
Year	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	Months of Development																	
Year	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	2,902	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																	

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	6,348	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																	

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																	

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	Months of Development																	
Year	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																	

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 Incured 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

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 Incured 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