Back-Testing the ODP Bootstrap & Mack Bootstrap Models

Mark R. Shapland, FCAS, FSA, FIAI, MAAA

Abstract

Motivation. Distributions of unpaid claims are gaining importance within the actuarial community as management, regulators, and others look to the actuarial profession for a quantitative approach to evaluating risk. Actuaries have historically applied their judgment to determine if a best estimate is reasonable, but how do we know if the models used to produce distributions are reasonable? Determining if a distribution is reasonable is a much more complex task than for a point estimate. Is the model producing a reasonable estimate at the 95th percentile? Is it producing reasonable distribution shapes? In effect, actuarial judgment shifts focus from a single point estimate to the entire distribution and we must rely, at least in part, on the proposition that "if the theory is acceptable then the distribution is acceptable." Therefore, the purpose of this paper is to determine if the theory really holds up in practice.

There are five objectives of this research. First, by greatly expanding the database used to back-test models the testing can provide more evidence to validate (or not) prior research and address any weaknesses in the prior research. Second, all of the prior research focused only on the estimate of a single outcome (i.e., the ultimate for the current accident year), so this research expands the testing for every possible estimate, e.g., each accident period, each calendar period, each incremental cell, etc. Third, more models were tested and some of the model assumptions were tested in order to expand our understanding of the predictive value of different models. Fourth, recent proposals to address model weaknesses were examined to assess their viability. Fifth, a new proposal for using this research to benchmark unpaid claim estimates will be put forth.

Method. The estimated distribution of possible outcomes for various models based on the ODP Bootstrap model and the Mack Bootstrap model are saved and compared to the actual outcome up to 9 years later – i.e., a single back-test. While the result from a single data set is not indicative of the quality of the original estimate, comparing results for a large number of data sets does provide an indication of the quality of the model.

Results. Based on the back-testing, all tested models appear to underestimate the width of the "true" distribution but some of the models tested appeared to get closer to the "true" distribution than others and the tested adjustments to the model assumptions seem to improve the results, which is a desirable quality. Another key result is to show how the insurance underwriting cycle also impacts the results of the back testing.

Conclusions. The major results from prior similar research is confirmed, but the volume of this research has led to a new approach to benchmarking both deterministic and stochastic unpaid claim estimates in practice.

Keywords. Back-test, benchmark, bootstrap, chain ladder, Mack model, over-dispersed Poisson, reserve variability, systemic risk, underwriting cycle.

1. INTRODUCTION

Enterprise Risk Management has been at the leading edge of effectively managing insurance and other risk bearing operations for many years. Its use is expected to grow, and perhaps accelerate, into the foreseeable future as regulators and rating agencies focus on risk based approaches. One of the key metrics in any risk model used for ERM is the variability of unpaid claims as these are normally the largest liability in the balance sheet. While many

stochastic models provide the diagnostic tools for calibrating the assumptions of the model, there are no tools for gauging the quality of unpaid claim variability estimates. After vetting the theory underlying a model, the only way to gain valuable insights into the quality of the model is to back-test the results to see how well the models predicted the actual outcomes.

To calculate a distribution of possible outcomes and select unpaid claim estimates at a confidence level of say 75%, or to demonstrate that reserves are at such a level, is not a straight-forward process. Indeed, it is not a process that can be performed exactly or by using purely statistical approaches. Reasons for this include, but are not limited to, the following:

- Uncertainty in reserving can be attributable to three types of risk: process risk, parameter risk and model risk. Of these, process risk, and to a lesser extent parameter risk, can be assessed statistically and then only to the extent permitted by the volume and quality of available data. Model risk does not, in general, follow clear statistical patterns.
- Where process or parameter risk can be assessed statistically, the available historical data will not show the full breadth of the possible outcomes (i.e., variability). The resulting uncertainty in any outcomes will increase the further one moves away from the mean.
- New lines of business will have little or no data on which to assess variability due to process or parameter risk. The statistical credibility of the data for small volumes of business will also be limited.
- The assessment of process or parameter risk can be distorted by historic data including the effect of systemic risks, e.g., changes in case law that affect claim settlement amounts.

Therefore, any assessment of reserves at a particular confidence level will require the reserving actuary to exercise judgment to a significant degree. This is similar to how actuaries currently assess deterministic unpaid claim estimates, where actuaries use tools (such as the Chain Ladder ("CL") and the Bornhuetter-Ferguson ("BF") methods) to calculate a central estimate of the claims liabilities. Based in part on their knowledge of the strengths and weaknesses of the methods, they exercise considerable judgment in selecting factors and parameters, in adjusting for trends and for known or expected distortions, and in selecting the amounts to be booked.

For stochastic models, with sufficient data the process and parameter risk would usually be evaluated using stochastic tools applied to the historic data. Different data (e.g., paid data

and incurred data) and different models would generate different results and different Coefficients of Variation ("CoVs"). Judgment is needed in deciding which CoVs would be appropriate to address model risk in addition to process and parameter risk.

Given that prior research has shown that the ODP Bootstrap and other models tend to underestimate the "true" variability, the actuary will need support for helping to inform their judgments about estimates of possible outcomes. Similar to benchmarks for deterministic assumptions, benchmark CoVs would be a very useful addition to the actuary's toolkit as a means of sense checking the estimated distributions. Thus, a primary use of this research is to provide benchmarks for distributions of possible outcomes for insurance data.

Even with benchmarks of CoVs by line of business, the actuary would need to combine these across all business lines. By definition, process risk should be independent of other risk factors (and across lines of business) but there may well be some degree of contagion (i.e., large losses that affect multiple lines of business) and/or correlation between the other factors. In order to combine the CoVs, correlation matrices will be required. Again, judgment is required, but another key benchmark from this research is estimated correlations based on industry data.

1.1 Research Context

Because it is such a critical part of effective actuarial practice, it seems likely that understanding the effectiveness of a method has been part of the research from the early days of actuarial science. For deterministic reserving methods, one of the earlier papers on the effectiveness of methods is Skurnick [18] and more recent examples include Forray [6] and Jing, Lebens, and Lowe [9]. For deterministic methods it is often enough to focus on the theory to understand the strengths and weaknesses of a method. For example, all actuaries learn early in their career that the chain ladder method will tend to underestimate the current period when the initial development period outcome is lower than average, and tend to overestimate the current period when the initial development period outcome is higher than average.

If we consider a triangle of data as illustrated in Graph 1.1, the goal of estimating unpaid claims is to estimate the unpaid amounts, $u(w,d)$, by projecting the cumulative amounts, $c(w,d)$.[1] The total reserve for an accident period, $R(w)$, can be estimated directly or indirectly

[1] For ease of exposition, the notation $c(w,d)$ and $u(w,d)$ does not specify cumulative or incremental values. The reader can infer cumulative or incremental values depending on their use.

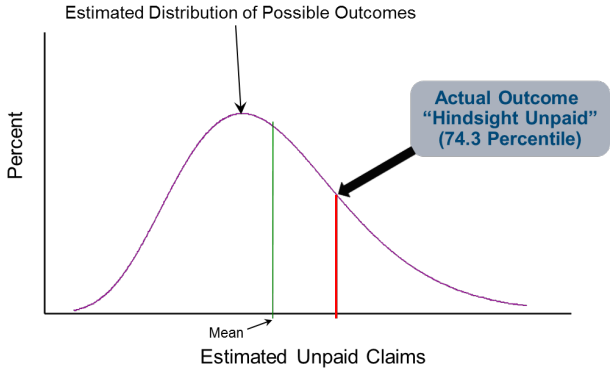
as a sum of the incremental unpaid amounts. For the chain ladder method, the estimation of *R(10)* is done using a factor times c(10,1), so it is easy to visualize how dependent this calculation is to the relative size of c(10,1).

Graph 1.1. Triangle of Data with Estimated Unpaid

	1	2	3	4	5	6	7	8	9	10	Total
1	c(1,1)	c(1,2)	c(1,3)	c(1,4)	c(1,5)	c(1,6)	c(1,7)	c(1,8)	c(1,9)	c(1,10)	
2	c(2,1)	c(2,2)	c(2,3)	c(2,4)	c(2,5)	c(2,6)	c(2,7)	c(2,8)	c(2,9)	*u(2,10)*	*R(2)*
3	c(3,1)	c(3,2)	c(3,3)	c(3,4)	c(3,5)	c(3,6)	c(3,7)	c(3,8)	*u(3,9)*	*u(3,10)*	*R(3)*
4	c(4,1)	c(4,2)	c(4,3)	c(4,4)	c(4,5)	c(4,6)	c(4,7)	*u(4,8)*	*u(4,9)*	*u(4,10)*	*R(4)*
5	c(5,1)	c(5,2)	c(5,3)	c(5,4)	c(5,5)	c(5,6)	*u(5,7)*	*u(5,8)*	*u(5,9)*	*u(5,10)*	*R(5)*
6	c(6,1)	c(6,2)	c(6,3)	c(6,4)	c(6,5)	*u(6,6)*	*u(6,7)*	*u(6,8)*	*u(6,9)*	*u(6,10)*	*R(6)*
7	c(7,1)	c(7,2)	c(7,3)	c(7,4)	*u(7,5)*	*u(7,6)*	*u(7,7)*	*u(7,8)*	*u(7,9)*	*u(7,10)*	*R(7)*
8	c(8,1)	c(8,2)	c(8,3)	*u(8,4)*	*u(8,5)*	*u(8,6)*	*u(8,7)*	*u(8,8)*	*u(8,9)*	*u(8,10)*	*R(8)*
9	c(9,1)	c(9,2)	*u(9,3)*	*u(9,4)*	*u(9,5)*	*u(9,6)*	*u(9,7)*	*u(9,8)*	*u(9,9)*	*u(9,10)*	*R(9)*
10	c(10,1)	*u(10,2)*	*u(10,3)*	*u(10,4)*	*u(10,5)*	*u(10,6)*	*u(10,7)*	*u(10,8)*	*u(10,9)*	*u(10,10)*	*R(10)*
Total											*R(T)*

This understanding of deterministic methods is largely possible because the focus of the method is a central estimate. For stochastic models, whose focus is the entire distribution, the same principles for the central estimate still apply, but understanding the entire distribution is impossible with a single observation. For example, it is common for a stochastic model to be used to simulate 10,000 possible outcomes for *R(10)*, but what if we later determine that the actual outcome was at the 74.3 percentile, as illustrated in Graph 1.2.

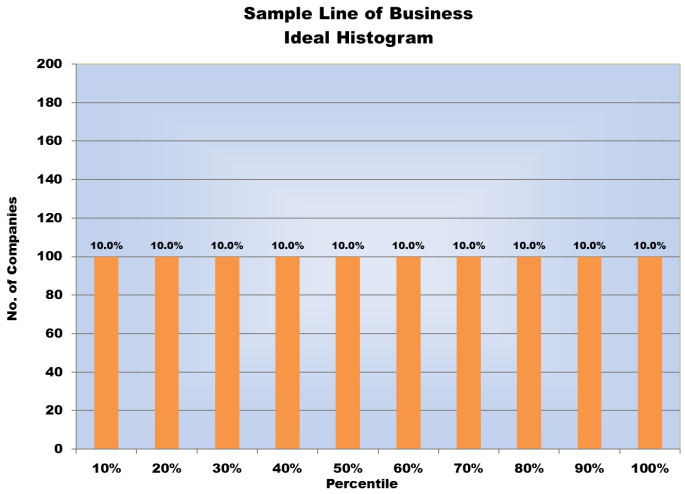
Graph 1.2. Back-Test of Estimated Distribution of Possible Outcomes



What does that tell us about the model? Did we get the mean wrong? What about the width of the distribution? We have no way to know with only one observation compared to our estimated distribution. Therefore, back-testing a large number of observations is essential to see if all parts of the distribution are represented in the outcomes. Another point to keep in mind is that when back-testing a model the mean of the estimated distribution is assumed to be the booked reserve even if the available data contains the actual booked reserve in order to test the efficacy of the model and not the judgment of the actuary selecting the reserve.

Testing all parts of a distribution can be illustrated graphically. For example, with 1,000 data sets, if the "true" distribution of the possible outcomes is fairly represented by the model then each decile group of the actual outcomes in a histogram should ideally contain 100 observations, as illustrated in Graph 1.3. For example, if the outcome from Graph 1.2 is included as one of the 1,000 datasets, then it would be one of the 100 in the bar labeled 80% (representing all outcomes greater than 70% and less than or equal to 80%) in Graph 1.3.

Graph 1.3. Ideal Histogram



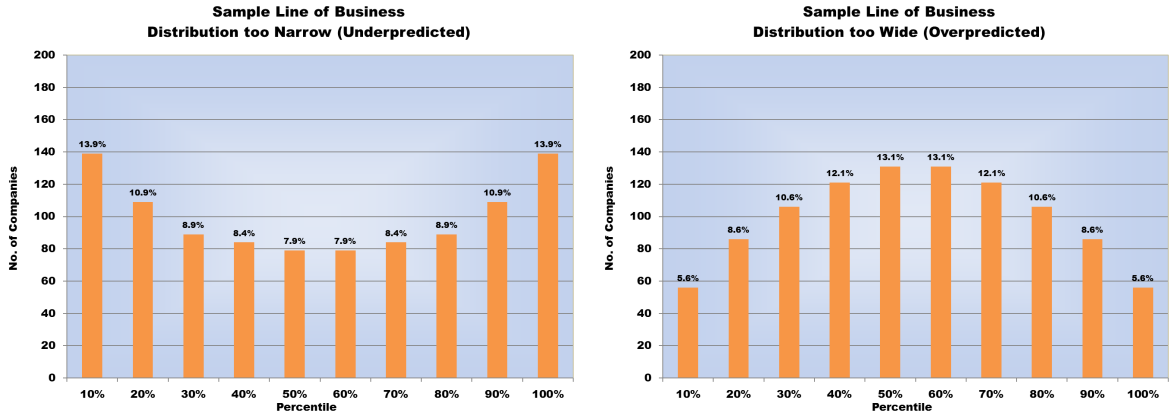
In Graph 1.3, like most of the similar graphs in the remainder of the paper, the percentiles along the X-axis are the decile groups for the percentile of the actual outcome compared to the estimated distribution. The Y-axis shows the number of companies or datasets in the bars, with the percent of the total number of companies or datasets as the bar labels.

If the model being back-tested is under predicting the "true" distribution then the histogram would show a higher than average number of observations at the extremes, say below the 20th percentile and above the 80th percentile, and it would show a lower than average number of observations in the middle percentiles. If the model being back-tested is over predicting the "true" distribution the histogram would show a lower than average number of observations at the extremes and higher than average observations in the middle percentiles. These two types of results are illustrated in Graph 1.4.

Of course, when back-testing real (or simulated) data the actual histograms will include random noise which could mask or partially mask the results, but typically the shape of the histogram will be indicative of the result even if random noise makes conclusions about a specific percentile problematic. The impact of random noise on the histogram can at least be

partially minimized by increasing the sample size to take advantage of the law of large numbers. This approach to understanding the effectiveness of stochastic models has been used by a number of researchers, but only a few key papers will be highlighted in Section 2.

Graph 1.4. Under & Over Prediction Histograms



1.2 Objectives

There are five objectives of this research. First, by expanding the database used to back-test models the testing can provide more evidence to validate (or not) prior research and address any weaknesses in the prior research. Second, all of the prior research focused only on the estimate of a single outcome, specifically the estimate of *R(10)* from Graph 1.1. For this research the outcomes for all possible estimates from Graph 1.1., e.g., each accident period, each calendar period, each incremental cell, etc., were included in the testing to see if any other insights can be gained by expanding the testing. Third, more models were tested and some of the model assumptions were tested in order to expand our understanding the predictive value of different models. Fourth, recent proposals to address model weaknesses were examined to assess their viability. Fifth, a new proposal for using this research to benchmark unpaid claim estimates will be put forth.

1.3 Outline

The remainder of the paper proceeds as follows. Section 2 will provide an overview of the prior research and proposed solutions. In Section 3, the data and the process used to validate it for the back-testing are described. Next, Section 4 will focus on the testing process. Then, in Section 5 the results of the back-testing are summarized, with additional details included as Appendix A. Finally, in Section 6 a process for using this research to benchmark unpaid claim estimates will be described.

2. OVERVIEW OF PRIOR TESTING

Other researchers have used back-testing to evaluate the quality of stochastic models, but providing an in depth review of prior work is beyond the scope of this paper. Since one of the objectives of this paper is to validate (or not) the prior research, some of the prior research is included in the References section for the interested reader and some highlights are included here. Note however, that the highlights discussed here are not intended to give a complete overview of these papers and other valuable insights could be gained by reading the original research papers.

Two early examples of back-testing stochastic models are the product of GIRO Working Parties [15, 16] in the U.K. in 2007 and 2008. The 2007 Working Party reviewed a number of models with a few real datasets, but also created simulated data (designed to meet all of the conditions/assumptions of the respective model) to more thoroughly test the ODP Bootstrap and Mack models. The 2008 Working Party expanded the simulation testing of the 2007 Working Party by creating a wider variety of simulated datasets (e.g., different triangle sizes). The back-testing was based on 10,000 samples of each simulated dataset for the ODP Bootstrap (paid chain ladder only) and closed form Mack models.

In theory at least, this testing was designed to see how well the model predicted outcomes for "perfect" data. The Working Parties also noted that simulated data was a good first step as it allows for controlled testing, but they also recognized that real data can include shocks and other anomalies which is likely to cause predicted results to be more inaccurate than simulated data. Interestingly, even with the "perfect" datasets the Working Parties concluded that:

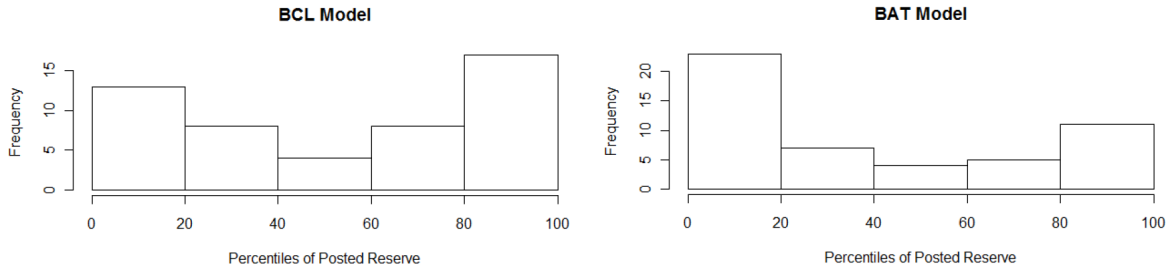
- The results for the Mack model exceeded the predicted 99th percentile 8.4% of the time for a 10 x 10 triangle, indicating the Mack model significantly under predicted the extreme outcomes. As the triangle size was increased to 100 x 100, the under prediction of the extreme outcomes reduced to 2.1% for the Mack model.
- The results for the ODP Bootstrap model exceeded the predicted 99th percentile 2.6% of the time, which also indicated an under prediction. As the triangle size increased for the ODP Bootstrap model the error rate stayed consistent.

In Meyers & Shi [12], the authors based their back-testing of the ODP Bootstrap model[2] on a database of 1997 Schedule P paid data from 50 companies. While the size of the

[2] The authors also proposed and tested a Bayesian Autoregressive Tweedie (BAT) model.

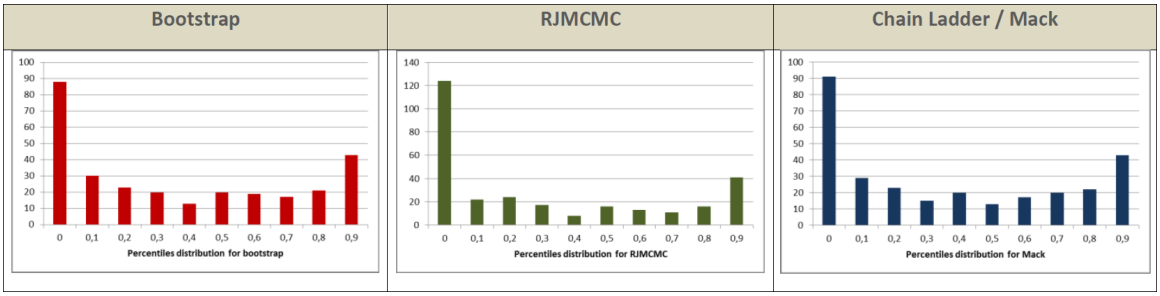
database was not sufficient to arrive at definitive conclusions, the authors recommended further testing and noted that their study "suggests that there might be environmental changes that no single model can identify" and "the actuarial profession cannot rely solely on stochastic loss reserve models to manage its reserve risk." To summarize the back-testing results, the authors included Graph 2.1, which show results for their tests of the Bootstrap Chain Ladder (BCL) model and the Bayesian Autoregressive Tweedie (BAT) model. Similar to the description above for Graph 1.3, the "frequency" label for the Y-axis represents the number of companies in each 20% group bar of the histograms. For the data as of 31 December 1997, only the current accident year was tested.

Graph 2.1. Percentile Results for Meyers & Shi



In Gremillet, Miehe & Zanón [7], the authors based their back-testing of the ODP Bootstrap[3] model on 296 triangles from four lines of business in the database created for the CAS by Meyers & Shi using 1997 Schedule P paid data. The authors concluded "it is core to have adjustments by actuaries prior to running the stochastic methods 'automatically'" and that "it seems that the 'actuary in the box' dream for stochastic reserves valuation is not yet happening…" To summarize the back-testing results, the authors included Graph 2.2, which show the results for the three models they tested. Similar to Meyers & Shi, only the current accident year was tested for the 1997 data.

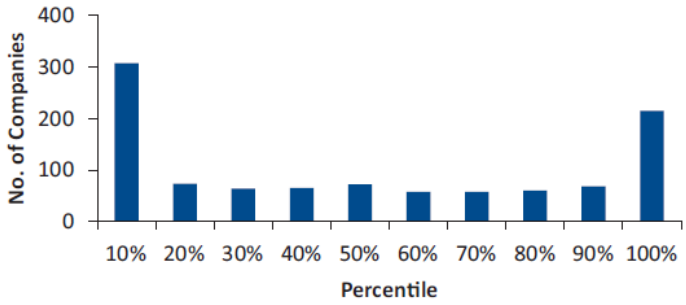
Graph 2.2. Percentile Results for Gremillet, Miehe & Zanón



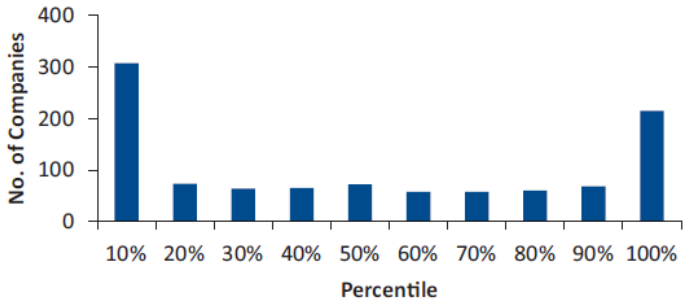
[3] The authors also tested the Reversible Jump Markov Chain Monte Carlo model and the Mack model.

In Leong, Wang & Chen [11], the authors based their back-testing of the ODP Bootstrap model on seven lines of business[4] for approximately 4,850 triangle datasets[5] from a database of Schedule P data from 1989 to 2002. The authors concluded "the popular ODP Bootstrap of the paid chain-ladder method is underestimating reserve risk" and that "it is because the bootstrap model does not consider systemic risk, or, to put it another way, the risk that future trends in the claims environment—such as inflation, trends in tort reform, legislative changes, etc.—may deviate from what we saw in the past". To summarize the back-testing results, the authors included Graph 2.3 showing the results for Homeowners and similar graphs for the other lines of business.

Graph 2.3. Results for Leong, Wang & Chen (HO – Paid CL – All Years)



Graph 2.4. Results for Leong, Wang & Chen (WC – Incurred CL – All Years)



In Leong, Wang & Chen [11], the authors then expanded their back-testing of the ODP Bootstrap to see if using the model for incurred data improved the models predictive power. The authors concluded "it appears that the incurred bootstrap model is also underestimating the risk of falling in these extreme percentiles" as illustrated in Graph 2.4 for Workers'

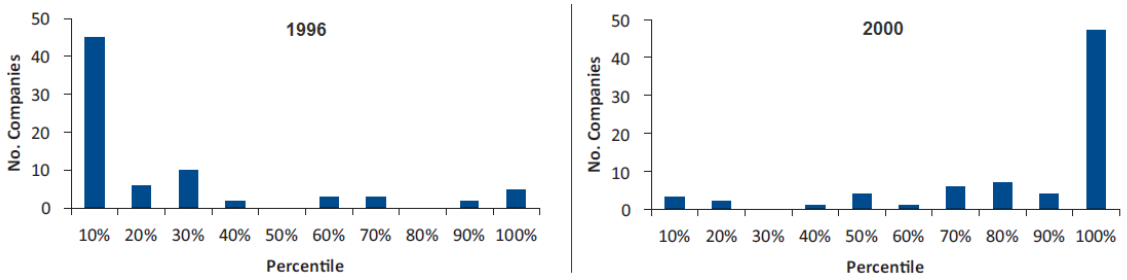
[4] For some years, data for Medical Professional Liability and Other Liability is split between Claims Made and Occurrence policies. In order to use this data consistently over all years, the parts are combined into one line of business. Thus, technically nine lines of business were included, but for four of the lines the splits were grouped to include both Claims Made and Occurrence.

[5] The authors do not state the actual number of datasets, but they do note that the line of business with the most data came from 78 companies and the line of business with the least data came from 21 companies. To estimate the total number of datasets, if we assume an average of 49.5 companies per line of business, 7 lines of business and 14 years, then 49.5 x 7 x 14 = 4,851.

Compensation.

An additional insight from the Leong, Wang & Chen [11] research was possible due to their use of multiple years to show that the reserving cycle has an impact on the results. The results in Graphs 2.3 and 2.4 are for all years combined, but results by year were also included by the authors such as Homeowners for 1996 and 2000 in Graph 2.5.

Graph 2.5. Results for Leong, Wang & Chen (HO – Paid CL – 1996 & 2000)



The authors also illustrated the reserving cycle for the industry in Graph 2.6 which shows that in 1996 the overall reserve level for the industry was too high and in 2000 it was too low. The left side histogram in Graph 2.5 corresponds to when the industry was over reserved and the back-testing resulted in a disproportionate number of outcomes less than 10%, which makes sense.[6] The right side histogram in Graph 2.5 also makes sense as the industry was under reserved in 2000, which leads to a disproportionate number of outcomes above 90%.[7]

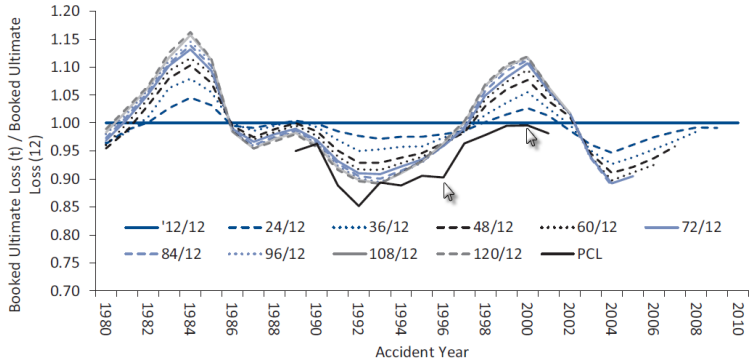
This insight led the authors to conclude that the ODP Bootstrap model only measures independent risk (arising from the randomness inherent in the insurance process) and not systemic risk (arising from the whole system). While there is a certain appeal to this conclusion, it seems the definition of systemic risk could be split into "internal systemic" risk (arising from within the modeling framework) and "external systemic" risk (arising from the outside the modeling framework). By using a broad definition of systemic risk the authors ignored weaknesses of the ODP Bootstrap model that contribute to this result. Their focus

[6] In Graph 2.6, the initial reserves at 12 months are the solid line for 1.00. For 1996, as the accident year matures the ultimate value gets lower and lower and at 120 months is a little over 90% of the ultimate at 12 months, i.e., the initial reserves were too high. In this case, if the initial mean of the simulated distributions was too high, say at 100, then when the final outcome is known if the mean should have been lower, say 90, then the odds that the actual random outcome is below 10% is increased, all else being equal.

[7] For 2000, as the accident year matures the ultimate value gets higher and higher and at 120 months is about 112% of the ultimate at 12 months, i.e., the initial reserves were too low. In this case, if the initial mean of the simulated distributions was too low, say at 100, then when the final outcome is known if the mean should have been higher, say 112, then the odds that the actual random outcome is above 90% is increased, all else being equal.

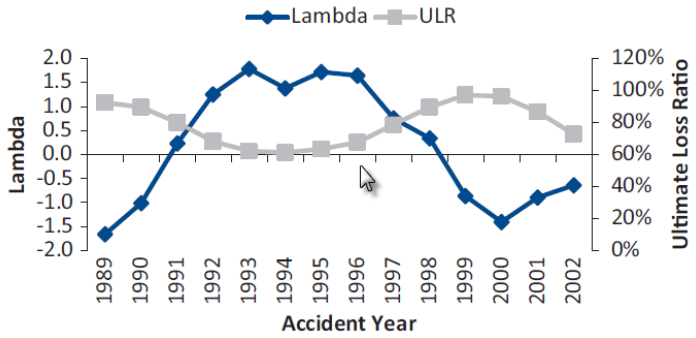
on systemic risk resulted in two methods to adjust for systemic risk in the ODP Bootstrap model, the systemic risk distribution method[8] and the Wang transform adjustment,[9] which allows the authors to show how the combination of both of these methods "fixed" the back-testing results.

Graph 2.6. Reserving Cycle in Leong, Wang & Chen



Digging deeper into the methods proposed by Leong, Wang & Chen [11] it seems that while their results do "correct" the back-testing results, the methods ignore a weakness of the ODP Bootstrap model, are backward looking only, and therefore should be used cautiously as a tool to adjust current ODP Bootstrap results. Starting with the variance adjustments, Graph 2.7 illustrates how when the ultimate loss ratio is less than the initial loss ratio (as in 1996) the variance is increased by Lambda, but this is not logical.

Graph 2.7. Comparison of Lambda and ULR in Leong, Wang & Chen



If the initial ultimate estimate is too high, a typical chain ladder method is likely to be

[8] The authors conclude that the ODP Bootstrap model only measures independent risk and not systemic risk. For the systemic risk distribution method a benchmark systemic risk distribution is estimated and combined with the independent risk distribution from the ODP Bootstrap model to obtain the total risk distribution.

[9] For the Wang transform adjustment, the authors note that the ODP Bootstrap estimate is biased (in their words "it is not assumed to be unbiased") so the adjustment tries to estimate the systemic bias over the course of the reserving cycle.

overestimating the central estimate due to a larger than average initial cumulative value in the current accident year. Extending this to the ODP Bootstrap, this overestimation from the chain ladder elements of the model also causes the variability of the future incremental values to be overestimated.[10] Logically, when the initial ultimate is overestimated the Lambda value should decrease the variance and vice versa. Thus, the Lambda proposed by the authors is above one when it should be below one and vice versa.[11]

Moving to the mean adjustments, the authors note that the Wang transform shifts the distribution to account for the estimation error. This implies that when the initial ultimate losses are overestimated (as in 1996) the shifting will reduce the mean of the distribution. Looking back at Graph 2.5 for 1996, the initial overestimation of the mean is a major contributor to why so many of the outcomes ended up in the lowest decile. Based on the combination of these two adjustments it makes sense that they "corrected" the historical biases in the model results. However, in order for this to have practical value looking forward the actuary would be required to guess at which part of the reserving cycle they are currently in and then select a Lambda which is opposite of what is indicated by the proposed formulas.

As noted above, this review of the Leong, Wang & Chen [11] paper indicates that their formulas should be used with caution when adjusting an estimated distribution from the ODP Bootstrap model, but this realization led to an alternative approach.[12] In summary, Leong, Wang & Chen [11] use a formula based approach for a single model to adjust an estimated distribution based solely on the data used for the estimated distribution. Alternatively, by using a very large database of outcomes from multiple models it becomes possible to create customizable benchmarks of unpaid claim distributions which can be used as a guide regardless of the model(s) being used by the actuary. Because of the cyclical bias in the mean noted above, another advantage of using benchmarks is that this approach assumes the actuary will address the bias in the mean and the benchmark can adjust for the remaining biases.

---

[10] To add process variance to the simulated outcomes, each future incremental value is assumed to be the mean and the variance is the mean times the Scale Parameter. Thus, if the future incremental values tend to be "too large" due to the chain ladder extrapolation of the first cell, then the variance of the sampled values will also tend to be "too large."

[11] The authors comment on the significant negative correlation between Lambda and the ultimate loss ratio at the end of Section 7.2.2.

[12] Of course the authors may refute this conclusion or perhaps use this review to revise their formulas to better address the issues driven by the reserve cycle.

3. DATA USED IN TESTING

The data used in this research includes net loss and ALAE data from nearly 31,000 real data sets (i.e., paid claim triangles, incurred claim triangles, earned premiums, etc.) for all 16 Schedule P lines of business, spanning 9 years from 1996 to 2004.[13] For each of these data sets, the actual results over the next nine years was also captured in the database used to back-test the efficacy of each model.

More specifically, data from 4,798 companies was downloaded from SNL for years spanning 1996 to 2013, but not all companies have data for all years as companies come and go over time. The data for all these companies was converted into 59,890 individual Company Files (i.e., CSV files by company by year), with each file containing Schedule P data triangles for all LOBs.[14] Processing all of this raw data to arrive at the data used for back-testing included several steps.

1. **Data Quality Tests** – In this step, each Company File was checked to determine which LOBs have complete data triangles for years spanning 1996 to 2004. For all key triangles, data quality tests include, but is not limited to, making sure there is non-zero data for each year and minimum data requirements of all models being tested are satisfied. Of the original 4,798 companies, only 2,716 had at least one LOB that passed this test in at least one of the years. For these 2,716 companies there were 79,573 "Data Quality" triangle sets, with the totals by LOB shown in Table 3.1.

2. **Data Validation Tests** – For each of the Data Quality triangle sets, additional tests were conducted to check the next 9 years to make sure none of the data in the original triangles changed over the next 9 years (i.e., to make sure pooling arrangements or other issues don't exist which would cause data to be invalid for testing purposes). The validation process reduced the total company count to 1,679 and for these remaining companies there were 30,707 "Valid Data" triangle sets, with the totals by LOB shown in Table 3.1.

3. **Create Complete Data** – For each of the Valid Data triangle sets, the data for the next 9 years was added to a new data file to speed up testing. Of course during simulation testing only the original triangles were used to parameterize the models, but having the

[13] The U.S. Annual Statement includes 22 lines of business in Schedule P, but there are only 16 lines of business containing 10 accident years of data. The remaining short tail lines are excluded from the research.

[14] If all 4,798 companies had data in all years there would be 86,364 (= 4,798 x 18) files, so there was no data at all about 30% of the time.

actual outcome speeds up the testing process.

4. **Save Diagnostics** – For each of the Valid Data triangle sets, the "optimal" hetero groups were found and diagnostics for all models were calculated and saved. These diagnostic tests were saved so that back-testing can include tests to determine the effectiveness of different diagnostics on assessing model parameters.[15]

Table 3.1. Summary of Datasets by LOB

Schedule P Line of Business	Quality	Valid	Ratio
Commercial Auto Liability	9,555	3,821	40.0%
Commercial Multi-Peril	9,955	4,130	41.5%
Homeowners & Farmowners	10,880	4,724	43.4%
International	317	123	38.8%
Medical Professional Liability - Claims Made	1,878	563	30.0%
Medical Professional Liability - Occurrence	1,465	481	32.8%
Other Liability - Claims Made	4,091	1,482	36.2%
Other Liability - Occurrence	10,923	4,160	38.1%
Products Liability - Claims Made	761	199	26.1%
Products Liability - Occurrence	3,996	1,220	30.5%
Private Passenger Auto Liability	10,075	3,962	39.3%
Reinsurance - Non-Proportional Assumed Financial	397	163	41.1%
Reinsurance - Non-Proportional Assumed Liability	1,758	611	34.8%
Reinsurance - Non-Proportional Assumed Property	2,123	989	46.6%
Special Lines	3,871	1,349	34.8%
Workers' Compensation	7,528	2,730	36.3%
Total All Lines	**79,573**	**30,707**	**38.6%**

For the 1,679 companies with at least one Valid Data triangle set, 1,182 of these companies had at least 2 LOBs with Valid Data for at least one year. For each company (and year) with 2 or more LOBs, the correlation between the residuals was also calculated and saved, both before and after the hetero group factor adjustments, for both paid and incurred data. This resulted in 195,228 pairs of LOBs with correlation values that were captured along with the P-Values and the Degrees of Freedom for all pairs for each company and year set of LOBs. A high level comparison of the data used in this research compared to prior research is shown in Table 3.2.

Table 3.2. Summary of Data by Author

Item	Meyers & Shi	Gremillet & Miehe	Leong, Wang & Chen	Shapland
Evaluation Periods	1	5	11	9
Models Tested	2	3	2	8
Lines of Business	1	4	9	16
Triangle Sets	50	296	~4,850	30,707

[15] Only limited back-testing related to the diagnostics has been completed to date. Future research will provide for more insights on the value of different diagnostic tests.

4. TESTING METHODOLOGY

Using each of the Valid Data triangle sets, the back-testing process starts by calculating the parameters for the six different ODP Bootstrap models[16] described in the Shapland [17] monograph and the Mack Bootstrap model as described in England & Verrall [5]. For all models, the residuals are based on the all year volume weighted average loss development factors, no tail factors were included, and no adjustments to the standard models were included. Because of the sheer volume of the test data, other than assumptions based on diagnostic tests it is nearly impossible to create assumptions tailored to the data in each data set. However, it is possible to use broad sets of assumptions that should be representative of what an analyst might select in practice in order to test how different broad sets of assumptions affects the results.[17]

For the Bornhuetter-Ferguson ODP Bootstrap models, the a priori loss ratios were based on the most recent ultimate loss ratios by year from Schedule P. While this does allow these models to benefit a bit from hindsight, one of the goals for these models was to remove as much of the cyclical bias as possible to see if this improved the accuracy of the models. As a counter to the foresight in the a priori loss ratios, the standard deviations were all set to zero for the preliminary tests.

For the Cape Cod ODP Bootstrap models, it is not possible to include rate level adjustment factors and trend factors based on the data are problematic without the ability to judgmentally review each factor or to set narrow ranges for the trend factors. Thus, all rate level factors were set to 1.0 and all trend factors were set to 2.5% per year.[18] For all tests a decay ratio of 90% was used and each accident year is given 100% weight so nothing is excluded. These assumptions for the Cape Cod models are not intended to be ideal in practice, but rather a reasonable baseline for which other broad sets of assumptions can be compared in future testing.

For the ODP Bootstrap family of models weighted results were also tested. For the weights by accident year, for the 7 oldest accident years the paid and incurred chain ladder

[16] As a technical note, the ODP Bootstrap modeling framework tested during all of the research described in Section 2 is from the original England & Verrall [3] paper that does not include various model enhancements introduced in subsequent papers. In addition, the incurred ODP Bootstrap tested in Leong, Wang & Chen [11] is essentially the paid ODP Bootstrap from England & Verrall [3] using incurred data and does not include the incurred to total unpaid steps described in Section 3.3.1 of Shapland [17].

[17] Only limited back-testing related to the broad sets of assumptions has been completed to date. Future research will provide for more insights on the value of different broad sets of assumptions.

[18] In other words, these assumptions assume there were no rate changes over the 10 years of history and all loss cost inflation is constant at 2.5%.

models were given equal weight. For the 3rd prior year, the paid and incurred chain ladder and Bornhuetter-Ferguson models were given equal weight. For the most recent 2 years, the paid and incurred Bornhuetter-Ferguson and Cape Cod models were given equal weight. While different weighting schemes by LOB would typically be used in practice, this weighting scheme was selected as being representative of a typical weighting scheme.

As a side note, it is also possible to test Aggregate results for each company with at least 2 LOBs of Valid Data, but the results from many different combinations of LOBs would not provide meaningful results without also segregating into groups with all the same LOBs. Instead of just testing the most recent accident year, i.e., only *R(10)* from Graph 1.1, the simulation output of these model tests was captured in great detail, i.e., by accident year, calendar year, calendar year runoff, loss ratios, and each incremental cell in Graph 1.1. Using all of the 10,000 iterations of simulated data, the final step is to compare the actual outcomes to the complete simulated distribution of possible outcomes to determine the percentile of actual outcome for each cell and combination of cells in Graph 1.1.

The companion files for the Shapland [17] monograph could be used to run all of the simulation tests, but those files are designed for educational purposes and not speed.[19] By way of comparison, the Excel model for just one ODP Bootstrap model takes about 15 minutes to run 10,000 iterations so even after completely automating the process it would take one computer over 7 years of continuous processing to finish all of the testing for all 8 models – i.e., the 6 ODP Bootstrap models, the weighted ODP Bootstrap and the Mack Bootstrap with paid data only.

In order to speed up this process commercial software was used, which reduced the total time for one computer from over 7 years to less than 43 days, much faster but still a long process. To reduce the elapsed time even further, the simulation tests were spread over 16 computers, which allowed the overall process to be effectively managed and cut the elapsed time to less than a week.[20]

The simulation back-tests with all of the standard assumptions noted above were considered the "Baseline" tests. Reviewing the baseline tests we found a significant number of simulations with extremely wide distributions. These extreme distributions are a

[19] The companion Excel files can be used to run each of the 6 ODP Bootstrap models and the weighted results but they do not include the Mack Bootstrap model. However, a similar Excel file could be created for the Mack Bootstrap model.

[20] In theory the total elapsed time is less than a week, but in actuality stopping each computer periodically to save results in case of a crash, freeze or other operating system issue and retesting after a data quality review of the output extended the total time to about 2-3 weeks.

somewhat common occurrence in practice and typically result from "small" sample values in the first column that lead to extreme 12-24 month ATA factors (both positive and negative), which in turn lead to some extreme iterations (i.e., a more extreme version of the chain ladder weakness noted above). To address these extreme distributions, a second round of testing included adding constraints to limit the sample outcomes to zero (i.e., to remove negative incremental values) for selected triangle sets (referred to as the "Baseline with Limits" tests).[21] The process used to select triangle sets for adding this limit constraint were based on whether the width of the distributions exceeded a threshold to approximate when an actuary might use these constraints in practice, rather than simply adding this constraint to all triangle sets.

A third round of back-testing was done using all of the "Baseline with Limits" assumptions plus for all of the ODP Bootstrap models the optimal hetero group factors were applied to the modeling framework to test the impact of this common modeling option. This third set of tests are referred to as "Baseline Limits & Hetero".

5. TESTING RESULTS

Starting with the "Baseline" tests, the results for the ODP Bootstrap paid chain ladder for the current accident year (i.e., *R(10)* in Graph 1.1), for all lines of business, and all evaluation periods combined[22] are illustrated in Graph 5.1.

From Graph 5.1 it is clear that the results using significantly more data are still consistent with prior research. Two additional elements of this, and later, graphs are the red "bars" in the lowest and highest decile groups and the average percentile. The red "bars" represent the portion of their respective groups that exceeded the smallest or largest simulated possible outcome, respectively. For example, for the 10% bar the red portion represents the number of tests where the percentile for the actual outcome was less than 0% (i.e., less than the smallest simulated possible outcome). The average percentile is the average over all samples[23] and helps give a sense of how close the simulated means were to the "true" mean on average.

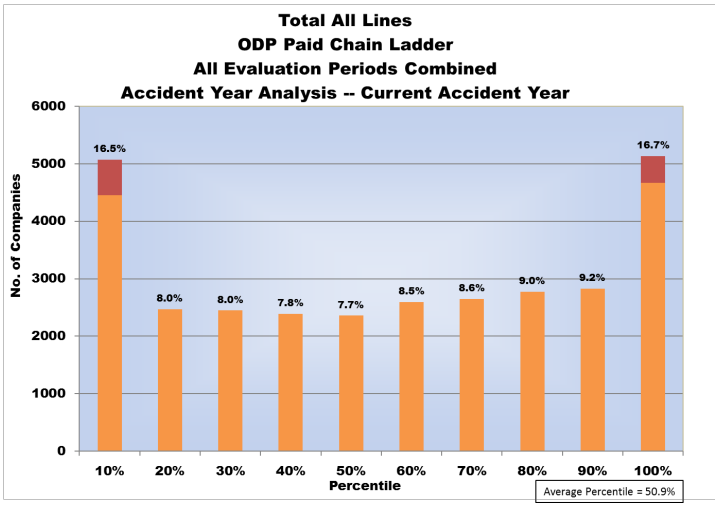
---

[21] This constraint on the simulation process is the third option described in section 4.1.1 of Shapland [17].

[22] For each evaluation period (e.g., 1996) the current accident year is always as of 12 months of development. Thus, while there are multiple evaluation dates the results for the current accident year for each evaluation date can be combined.

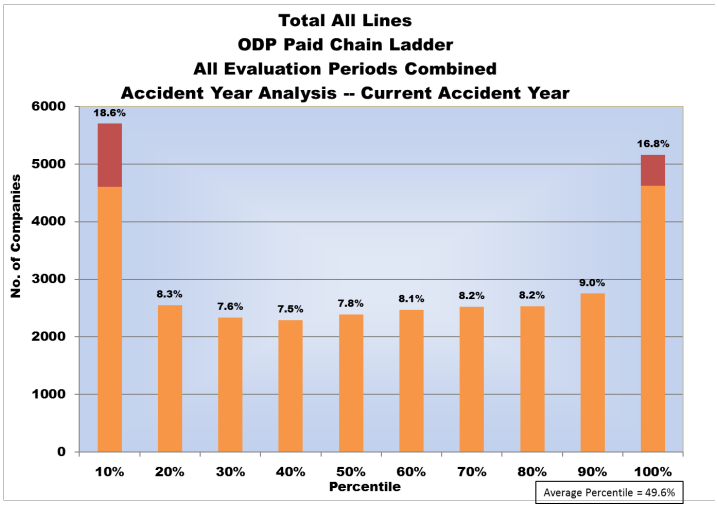
[23] For the samples below the minimum or above the maximum (i.e., represented by the red bars) the value used in the overall average percentile is 0% or 100%, respectively.

Graph 5.1. ODP Bootstrap Paid Chain Ladder – "Baseline"

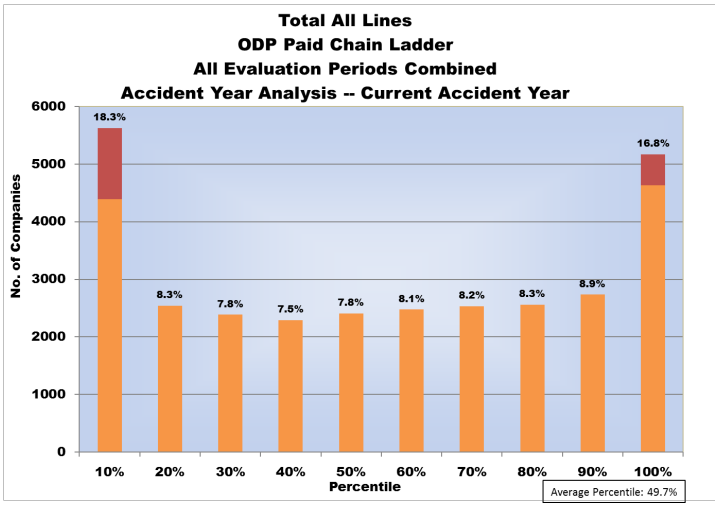


Moving to the "Baseline with Limits" tests the results for the ODP Bootstrap paid chain ladder for the current accident year, for all lines of business, and all evaluation periods combined are illustrated in Graph 5.2. Comparing Graph 5.2 with Graph 5.1 it makes sense that the "goal posts" at the extremes got higher, meaning the models further underestimated the "true" distributions, since the widest of the distributions in the "Baseline" tests were "narrowed" in the "Baseline with Limits" testing.

Graph 5.2. ODP Bootstrap Paid Chain Ladder – "Baseline with Limits"

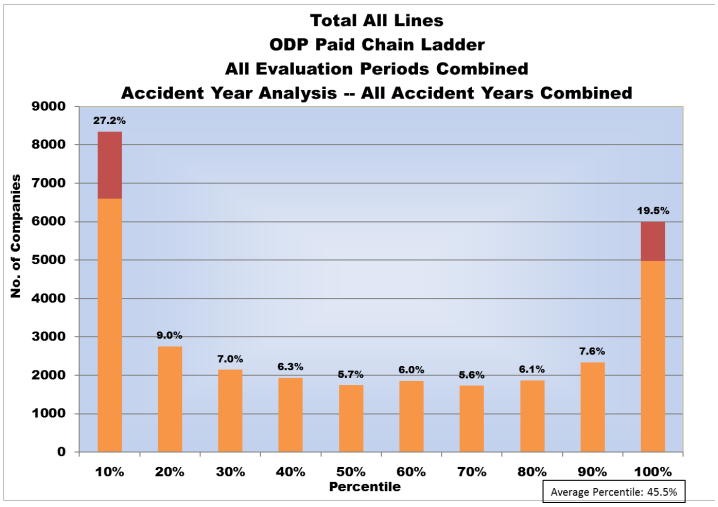


Graph 5.3. ODP Bootstrap Paid Chain Ladder – "Baseline Limits & Hetero"



At a high level the "Baseline Limits & Hetero" results for the ODP Bootstrap paid chain ladder for the current accident year, for all lines of business, and all evaluation periods combined are illustrated in Graph 5.3. The differences between Graph 5.3 and 5.2 are more subtle but a close inspection shows a slight improvement, which supports the use of heteroscedasticity adjustment factors in the ODP Bootstrap models. Admittedly, this support for using hetero factors is not strong but it is an improvement and rules out a negative conclusion (i.e., that hetero factors don't help). All of the results in the remainder of this paper are for the "Baseline Limits & Hetero" testing, but for simplicity this label is not included in any more graphs.

Graph 5.4. ODP Bootstrap Paid Chain Ladder – All Years Combined

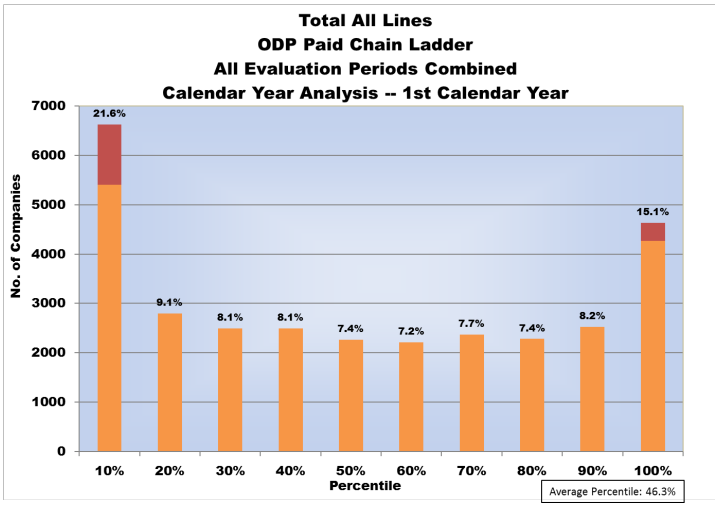


As we dig deeper into the back-testing results, a logical first dive would be to review results for prior accident years (i.e., *R(9)* to *R(2)* in Graph 1.1) to see if the estimation

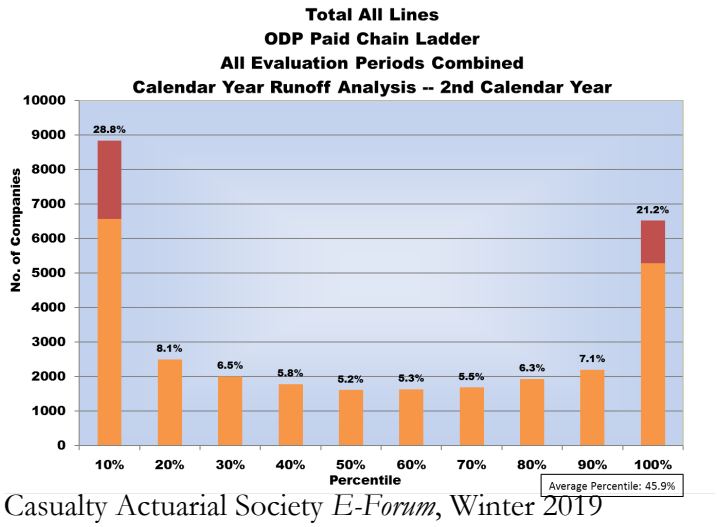
improves as the relative maturity of the accident year increases. The results by accident year are shown in Appendix A, but interestingly there is no improvement as the models predict fewer future periods. Similarly, combining all accident years (i.e., *R(T)* in Graph 1.1), as shown in Graph 5.4, does not improve the model predictions.

One of the insights from the Leong, Wang & Chen [11] paper was how the results were impacted by the reserving cycle. This impact was confirmed using this expanded database with the results by evaluation year shown in Appendix B. Consistent with the Leong, Wang & Chen results, the results by year show that the size of the "goal post" is predominantly in the lowest decile when the mean is being underestimated (e.g., in 1996) and shifts to being predominantly in the highest decile as the mean is overestimated. In addition, the average percentile shifts over the reserving cycle, which indicates how the estimates of the "true" mean change during the cycle.

Graph 5.5. ODP Bootstrap Paid Chain Ladder – First Calendar Year



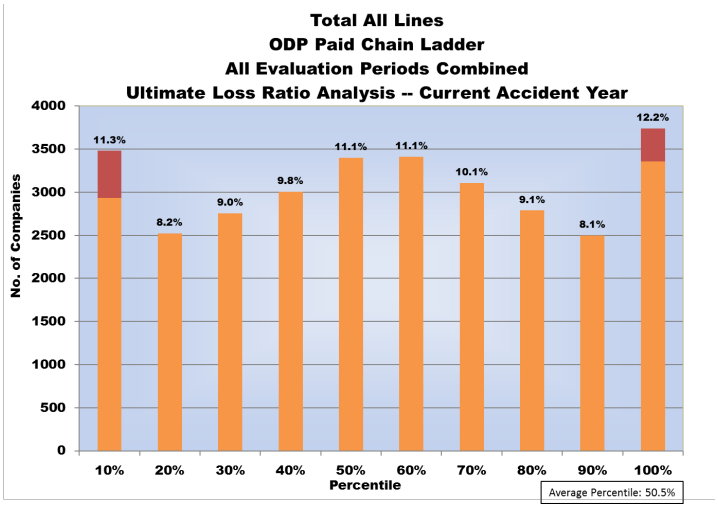
Graph 5.6. ODP Bootstrap Paid Chain Ladder – Calendar Year Runoff After 1 Year



In addition to looking at the predictions for the accident years, it was also possible to look at the calendar years (i.e., the sum of the diagonals in Graph 1.1), the calendar year runoff (i.e., the sum of all remaining diagonals as each diagonal is removed in Graph 1.1), and the time zero to ultimate loss ratios (i.e., the sum of an entire row in Graph 1.1). As might be expected after reviewing the accident year results, the calendar year results in Graph 5.5 and calendar year runoff results in Graph 5.6 are quite similar to the accident year results.[24]

For the time zero to ultimate loss ratio estimates by accident year shown in Graph 5.7 the predictions are much closer to the ideal histogram in Graph 1.3. This is an interesting result in the sense that the ODP Bootstrap predictions of the time zero to ultimate loss ratios appear to be more accurate than the predictions of the unpaid claims. To understand this we need to dig deeper into the results by incremental cell, which are shown in Appendix C. Interestingly, the results for the incremental cells reveals that the sampling of the incremental cells to create sample triangles for each iteration seems to produce more variability than observed in the data. On the other hand, since the model parameters are fit to the actual outcomes in the triangle perhaps seeing considerably more results in the middle decile groups is the expected result.

Graph 5.7. ODP Bootstrap Paid Chain Ladder – Ultimate Loss Ratio Current Year

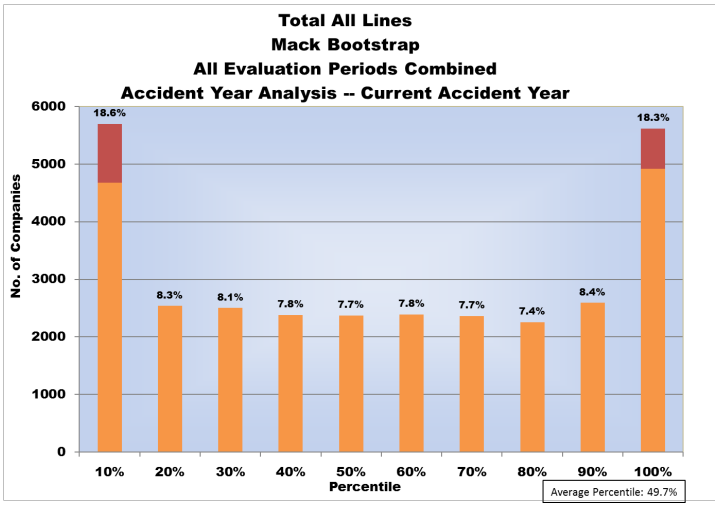


Now that we have dissected the ODP Bootstrap paid chain ladder model, we can compare this to the other models in the back-testing research. First, the results for the Mack Bootstrap paid chain ladder model are shown in Graph 5.8. Comparing Graph 5.8 with

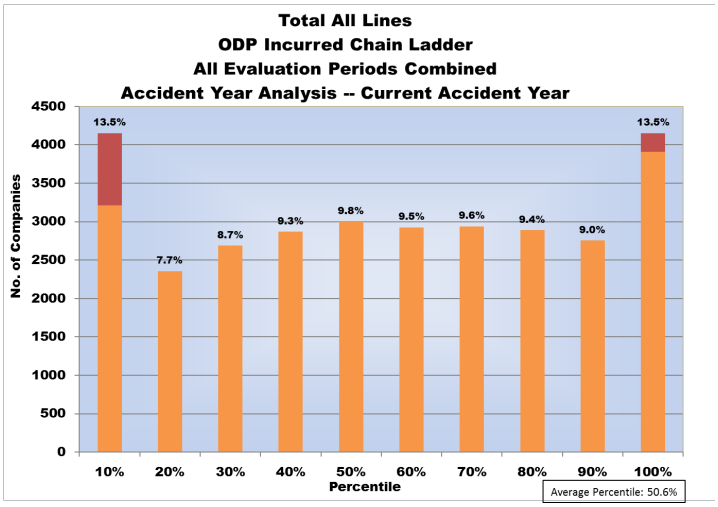
[24] Since Graph 5.5 is the first diagonal and Graph 5.6 is the sum of the remaining diagonals, the combination of these two graphs is the same as Graph 5.4.

Graph 5.3 shows that the Mack Bootstrap was a worse than the ODP Bootstrap, which is consistent with the findings of the GIRO Working Parties [15, 16].

Graph 5.8. Mack Bootstrap Paid Chain Ladder – Current Accident Year



Graph 5.9. ODP Bootstrap Incurred Chain Ladder – Current Accident Year



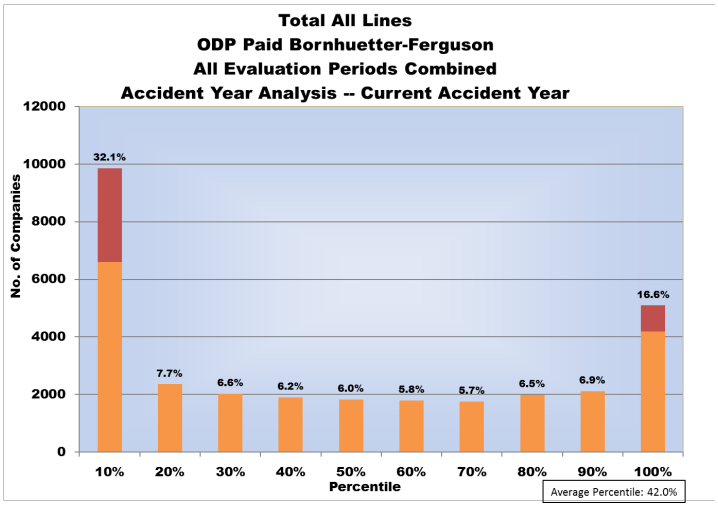
Next, the results for the ODP Bootstrap Incurred Chain Ladder model are shown in Graph 5.9. Comparing Graph 5.9 to Graph 5.3 there is a clear improvement in the predictive power of the incurred versus paid chain ladder versions of the ODP Bootstrap model.[25] Thinking about the mechanics of the ODP Bootstrap Incurred Chain Ladder model in Shapland [17] it seems fair to conclude that combining the variability of the paid and incurred data increases the relative variance of the unpaid estimates to come much closer to the ideal histogram in Graph 1.3. It is quite possible that the remaining "goal post" effect is

[25] This is inconsistent with the findings in Leong, Wang & Chen [11]. However, as mentioned in footnote 11, the algorithm being tested in this research is different.

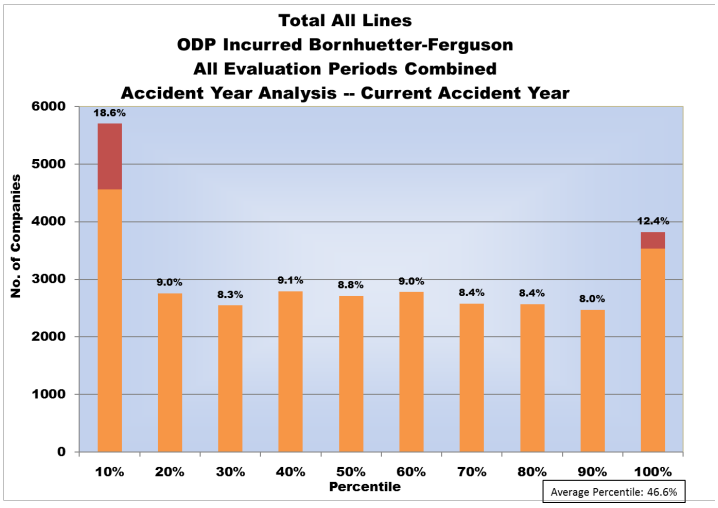
largely due to the mis-estimation of the mean during the reserving cycle.

Next, the results for the ODP Bootstrap Paid and Incurred Bornhuetter-Ferguson models are shown in Graph 5.10 and Graph 5.11, respectively. Comparing Graph 5.10 with Graph 5.3 and Graph 5.11 with Graph 5.9, respectively, it appears as though the Bornhuetter-Ferguson models are less predictive than their chain ladder counterparts are. This is inconclusive, however, since the variance assumption was set to zero during the current back-testing and it is easy to show that using a zero variance assumption will reduce the variability of the estimated unpaid claim distribution. Thus, conclusions about the predictive power of the ODP Bootstrap Bornhuetter-Ferguson models will need to wait until more testing can be completed.

Graph 5.10. ODP Bootstrap Paid Bornhuetter-Ferguson – Current Accident Year



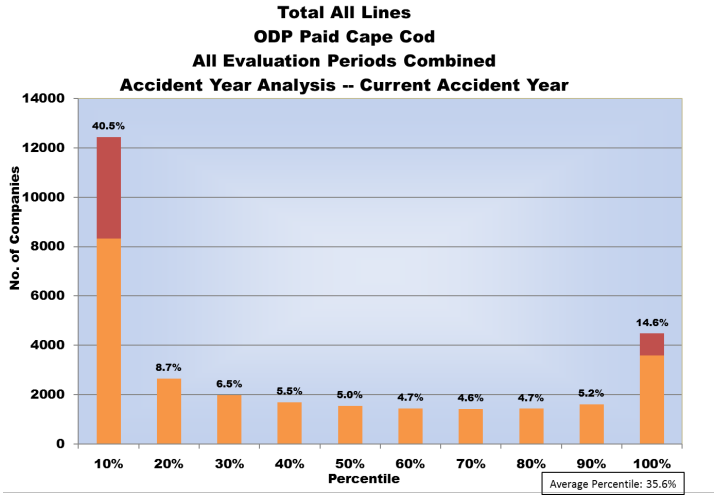
Graph 5.11. ODP Bootstrap Incurred Bornhuetter-Ferguson – Current Accident Year



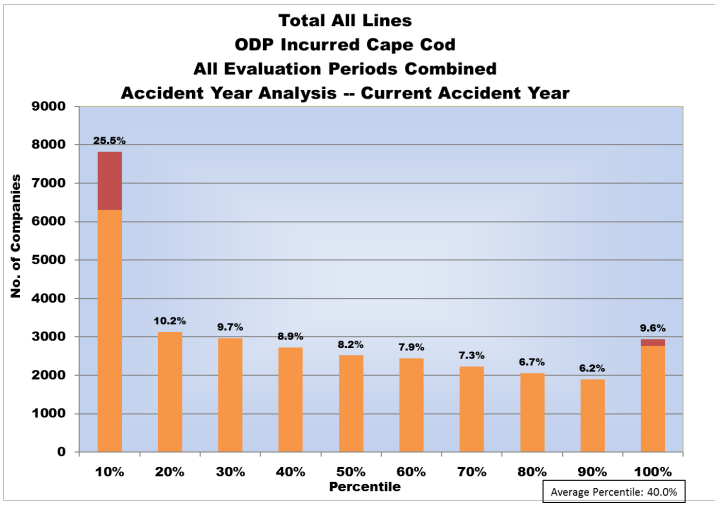
Next, the results for the ODP Bootstrap Paid and Incurred Cape Cod models are shown

in Graph 5.12 and Graph 5.13, respectively. Comparing Graph 5.12 with Graph 5.3 and Graph 5.13 with Graph 5.9, respectively, it appears as though the Cape Cod models are less predictive than their chain ladder counterparts are. This may also be inconclusive, however, since only one set of parameters has been tested so far. Thus, conclusions about the predictive power of the ODP Bootstrap Cape Cod models will need to wait until more testing can be completed.

Graph 5.12. ODP Bootstrap Paid Cape Cod – Current Accident Year



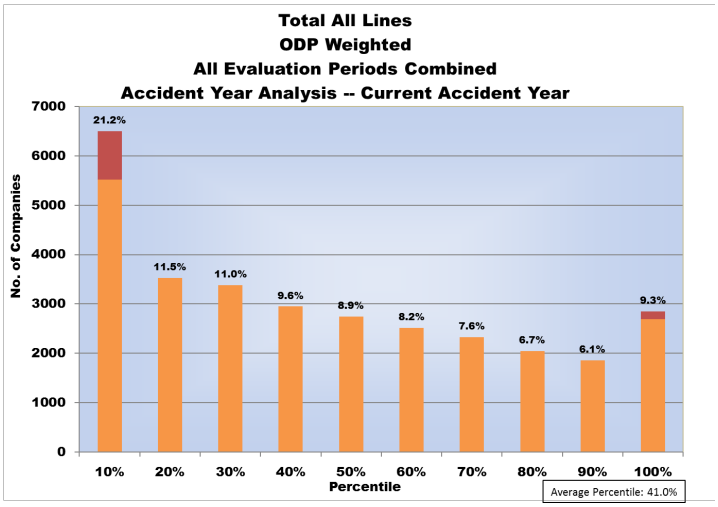
Graph 5.13. ODP Bootstrap Incurred Cape Cod – Current Accident Year



Finally, the results for the weighted combination of all six ODP Bootstrap models are shown in Graph 5.14. Comparing Graph 5.14 with Graphs 5.3, 5.9, 5.10, 5.11, 5.12, and 5.13, you can visualize how Graph 5.14 results from a combination of the other models. This seems promising as even with the deficiencies noted for each model individually the weighted results look like they are better that the sum of the parts. More importantly, this

demonstrates how weighting multiple models, to at least partially address model risk, can improve the results compared to a single model. As other assumptions for the Bornhuetter-Ferguson and Cape Cod models are tested, another avenue for future research will be considerations on how to apply Bayesian analysis to selecting the model weights.

Graph 5.14. ODP Bootstrap Weighted Models – Current Accident Year



All of the results presented in this Section and Appendices A, B, and C are for all lines of business combined. To show that the results are similar by line of business, the results by line of business for the ODP Bootstrap Paid Chain Ladder and Incurred Chain Ladder models are in Appendix D. It is possible to show many more details and combinations for all of these results, but this massive increase will be accompanied by an increase in random noise and will likely add little value beyond what we can already see at the higher level.

6. BENCHMARKS BASED ON TEST RESULTS

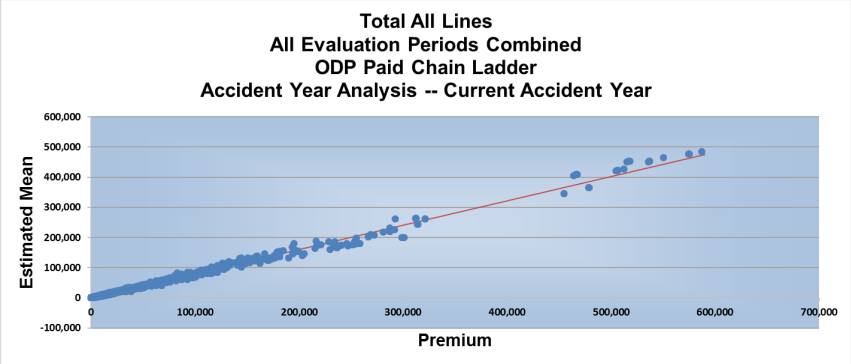
Even with the expansion of the research database, this research has confirmed the findings of prior authors. Thinking about the impact of the reserving cycle, it appears as though the results are strongly influenced by the internal systemic risks of the ODP Bootstrap modeling framework which, like the deterministic chain ladder, leads to the cycle of under and over estimation of the mean and in synch with this a lower and higher estimation of the variance. Even after potential corrections for the internal systemic risks, the ODP Bootstrap model is generally not accounting for the external systemic risks. On the other hand, it appears that some of the variations on the ODP Bootstrap framework may be significantly better at addressing the internal systemic risks.

In order to use this information in practice, one approach might be to consider how the formulas proposed by Leong, Wang & Chen [11] could be improved to separately address internal and external systemic risks. However, even with formula improvements, on a forward-looking basis the actuary is still faced with trying to understand which part of the reserving cycle they are currently in. Of course, knowing were one is in the reserving cycle is an issue no matter what the approach, but with a significantly larger database another way forward is possible.

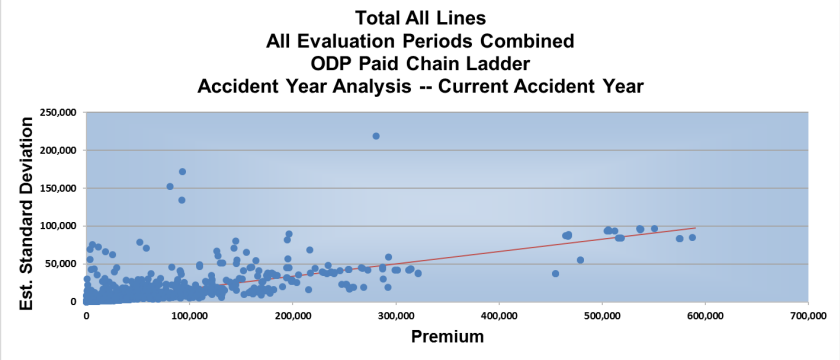
6.1. Unpaid Claim Benchmarks

Rather than try to create a precise formula for giving the "correct" distribution, we can take a page out of the deterministic reserving playbook and create benchmarks to help guide the judgment of the opining actuary. For example, consider Graphs 6.1 and 6.2, which illustrate the range of mean and standard deviation estimates from the ODP Bootstrap paid chain ladder model over the entire database for the most recent accident year. For Graph 6.1, it is not surprising that the mean unpaid is closely in line with the premium, with the deviations along the slope of the trend line representing differences in loss ratio by company.

Graph 6.1. ODP Bootstrap Means – Current Accident Year



Graph 6.2. ODP Bootstrap Standard Deviations – Current Accident Year

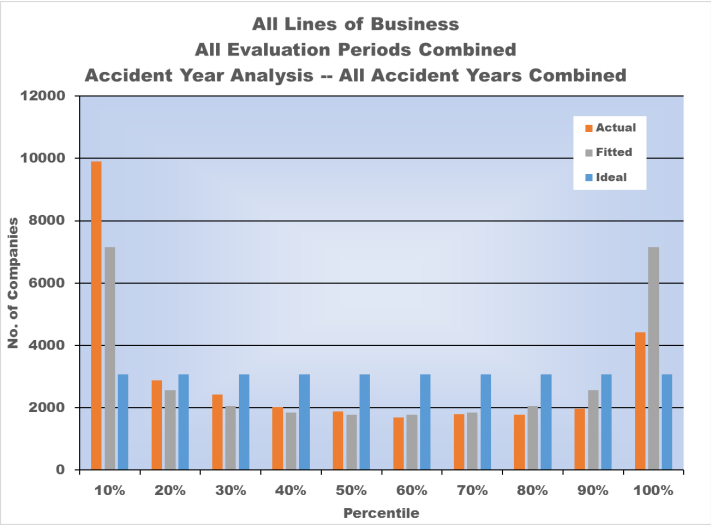


In Graph 6.2, it is not surprising that the standard deviations also increase in line with the premium, but the deviations around the trend line are more pronounced, which is likely due to the mixture of all lines of business, but at least to some degree a few of these could be considered outliers. A more important ingredient of Graph 6.2 is that the slope of the trend line is much lower, which confirms that the Coefficient of Variation is consistent with statistical principles, meaning for smaller companies the standard deviation is a larger percentage of the mean compared to larger companies.

The results shown in Graphs 6.1 and 6.2 are consistent for all other views of the data discussed in Section 5 (i.e., for each accident year, each calendar year, all years combined, etc.). In addition, similar graphs by line of business are also consistent with Graphs 6.1 and 6.2, except that they are more specific to the data for each line of business. This new insight lead to the idea of combining regression results (based on pure premiums instead of premiums) by line of business to create a benchmark algorithm for the means and related standard deviations by accident year, calendar year, etc., which at a minimum reflects the independent risks in the data.

As these regression results are based on the original simulation results, without any further adjustment the benchmarks would also reflect the biases shown in the back-testing results. In order to adjust for this bias an optimal variance correction factor was included similar to the factors proposed by Leong, Wang & Chen [11], except that the factor does not change each year during the reserving cycle. As an example, consider Graph 6.3 for all accident years and all lines of business combined.

Graph 6.3. ODP Bootstrap Bias Adjustment – All Accident Years Combined



For the fitted results in Graph 6.3 the optimal adjustment factor is 1.755, meaning the

benchmark standard deviations would be increased by 75.5%. There are variations in the optimal factor when looking at individual accident years, but they are reasonably consistent so only one factor based on the total of all years combined is used for the unpaid claims benchmarks. As noted for Graph 5.7, the results for the time zero to ultimate loss ratios are much closer to the ideal histogram so a lower adjustment factor is appropriate for the loss ratio benchmarks.

Because of the cyclical bias it is not possible to increase the factor to the point where the ideal histogram is achieved. However, this does seem to address the variance mis-estimation component of the internal systemic risk and external systemic risk to the extent that external systemic risks have influenced the outcomes in this research database. Assuming this is correct, the remaining "goal post" shape of the fitted histogram in Graph 6.3 is due to the mis-estimation of the mean during the reserving cycle that can be addressed by the actuary as part of the selection of the booked reserves.

It is possible that future research could help the actuary further understand the timing of the reserving cycles, but assuming the actuary can use caution to ensure their modeling assumptions are not being biased by the reserving cycle, the unpaid claim distribution benchmarks can be used as a guide to assess an estimated distribution from any stochastic model. For example, consider the results in Table 6.1 which compare standard results for the ODP Bootstrap paid and incurred chain ladder models with the corresponding benchmarks using commercial auto data from a randomly selected company in the research database.

Table 6.1. Comparison of ODP Bootstrap with Benchmark Unpaid

			ODP Bootstrap Paid Chain Ladder			ODP Bootstrap Incurred Chain Ladder			Unpaid Claim Benchmark		
Accident Year	Earned Premium	A priori Loss Ratio	Mean	Standard Error	CoV	Mean	Standard Error	CoV	Mean	Standard Error	CoV
2008	83,943	55.0%	125	194	154.9 %	135	216	160.7 %			
2009	94,343	55.0%	225	267	118.5 %	234	293	125.3 %	669	1,325	198.1 %
2010	115,098	55.0%	568	453	79.8 %	593	503	84.9 %	1,184	1,540	130.0 %
2011	126,714	55.0%	975	639	65.5 %	1,010	717	71.0 %	1,960	2,055	104.8 %
2012	138,148	55.0%	2,564	978	38.1 %	2,618	1,206	46.1 %	3,632	2,689	74.0 %
2013	156,046	55.0%	6,222	1,648	26.5 %	6,404	2,455	38.3 %	7,301	4,475	61.3 %
2014	173,621	55.0%	13,146	2,529	19.2 %	14,781	4,841	32.7 %	15,027	7,609	50.6 %
2015	181,416	55.0%	27,524	3,888	14.1 %	32,868	9,345	28.4 %	28,179	11,947	42.4 %
2016	184,422	55.0%	45,759	5,518	12.1 %	49,668	15,204	30.6 %	48,125	18,504	38.4 %
2017	186,444	55.0%	66,947	9,017	13.5 %	80,709	24,784	30.7 %	75,007	27,104	36.1 %
Totals	**1,440,195**		**164,055**	**12,928**	**7.9 %**	**189,019**	**31,310**	**16.6 %**	**181,085**	**38,898**	**21.5 %**

The benchmark algorithm is based on 10 years of data so the earned premium and a priori loss ratios are used to enter pure premiums by year into the algorithm. The ODP Bootstrap results in Table 6.1 do include using the optimal hetero groups and a few other model options to replicate what an actuary could easily produce as a first draft of the unpaid claim distribution. Not surprisingly, the benchmark results indicate that the CoV should be higher compared to either of the ODP Bootstrap models.

Reviewing Table 6.1, three more observations can be made. First, the benchmark does not include the 9th prior accident year (i.e., 2008). This is due to tail factors being excluded from the back-testing to date, but future back-testing can include tail factors which will allow the benchmark algorithm to be expanded to include the 9th prior accident year (and an additional calendar year). Second, even though the data used in the back-testing was from 1996-2004, the algorithm is independent of the year and based on fitting to distributions by size of exposure (e.g., pure premiums) so using the algorithm to create a benchmark for 2017 makes sense as long as the a priori loss ratios are reasonable given the reserving cycle.

The third observation from Table 6.1 is that the benchmarks are essentially based on the average loss development pattern from the industry data. Thus, it would be a reasonable critique to note that the loss development pattern for the company under review does not clearly match with a Schedule P line of business. Because this is such a common issue, the algorithm also includes an option to adjust benchmarks based on the loss development pattern assumption used by the actuary. In Table 6.2 the benchmark has been updated for the loss development pattern for the data being used in the example. Comparing Table 6.2 with Table 6.1, note that while the mean and standard deviations both decreased a bit the CoV essentially stayed the same or even increased a bit, which makes sense given the increased uncertainty. Also keep in mind that all benchmarks are only intended to serve as a guideline for the actuary and a perfect match is not a goal.

Table 6.2. Comparison of ODP Bootstrap with Benchmark Unpaid & Custom LDF Pattern

			ODP Bootstrap Paid Chain Ladder			ODP Bootstrap Incurred Chain Ladder			Unpaid Claim Benchmark		
Accident Year	Earned Premium	A priori Loss Ratio	Mean	Standard Error	CoV	Mean	Standard Error	CoV	Mean	Standard Error	CoV
2008	83,943	55.0%	125	194	154.9 %	135	216	160.7 %			
2009	94,343	55.0%	225	267	118.5 %	234	293	125.3 %	93	211	225.4 %
2010	115,098	55.0%	568	453	79.8 %	593	503	84.9 %	260	384	148.0 %
2011	126,714	55.0%	975	639	65.5 %	1,010	717	71.0 %	641	745	116.1 %
2012	138,148	55.0%	2,564	978	38.1 %	2,618	1,206	46.1 %	1,778	1,404	79.0 %
2013	156,046	55.0%	6,222	1,648	26.5 %	6,404	2,455	38.3 %	4,643	2,942	63.4 %
2014	173,621	55.0%	13,146	2,529	19.2 %	14,781	4,841	32.7 %	11,306	5,809	51.4 %
2015	181,416	55.0%	27,524	3,888	14.1 %	32,868	9,345	28.4 %	24,133	10,287	42.6 %
2016	184,422	55.0%	45,759	5,518	12.1 %	49,668	15,204	30.6 %	44,007	16,974	38.6 %
2017	186,444	55.0%	66,947	9,017	13.5 %	80,709	24,784	30.7 %	72,694	26,296	36.2 %
Totals	**1,440,195**		**164,055**	**12,928**	**7.9 %**	**189,019**	**31,310**	**16.6 %**	**159,555**	**34,460**	**21.6 %**

In order to illustrate how the benchmark algorithm responds to different input assumptions, Table 6.3 includes a comparison of the benchmarks from Table 6.2 with benchmarks based on only 10% of the original premiums (i.e., all other assumptions are the same). This shows how the benchmarks for a smaller company would compare to those for a larger company. Following statistical principles, and the regressions illustrated in Graph 6.1 and 6.2, with only 10% of the premium the mean is reduced by 90% but CoV increases to reflect the additional uncertainty.

Table 6.3. Comparison of Benchmarks by Size of Company

			Unpaid Claim Benchmarks						Unpaid Claim Benchmarks		
Accident Year	Earned Premium	A priori Loss Ratio	Mean	Standard Error	CoV	Accident Year	Earned Premium	A priori Loss Ratio	Mean	Standard Error	CoV
2008	83,943	55.0%				2008	8,394	55.0%			
2009	94,343	55.0%	95	214	224.8 %	2009	9,434	55.0%	10	47	495.6 %
2010	115,098	55.0%	262	387	147.8 %	2010	11,510	55.0%	26	91	346.5 %
2011	126,714	55.0%	616	719	116.8 %	2011	12,671	55.0%	62	166	269.1 %
2012	138,148	55.0%	1,735	1,374	79.2 %	2012	13,815	55.0%	173	289	166.8 %
2013	156,046	55.0%	4,525	2,874	63.5 %	2013	15,605	55.0%	452	523	115.6 %
2014	173,621	55.0%	11,154	5,736	51.4 %	2014	17,362	55.0%	1,115	877	78.6 %
2015	181,416	55.0%	23,905	10,194	42.6 %	2015	18,142	55.0%	2,390	1,369	57.3 %
2016	184,422	55.0%	43,759	16,882	38.6 %	2016	18,442	55.0%	4,376	2,253	51.5 %
2017	186,444	55.0%	72,465	26,216	36.2 %	2017	18,644	55.0%	7,246	3,436	47.4 %
Totals	**1,440,195**		**158,515**	**34,245**	**21.6 %**		**144,020**		**15,851**	**4,835**	**30.5 %**

As noted earlier, the benchmark algorithm includes more than the accident year unpaid claims, so Table 6.4 illustrates the cash flow and unpaid claim runoff benchmarks which would be comparable to the unpaid claim benchmarks in Table 6.2. The benchmark algorithm also includes time zero to ultimate loss ratios, but these are not illustrated in any of the Tables.

Table 6.4. Comparison of Unpaid, Cash Flow and Runoff Benchmarks

	Unpaid Claim Benchmarks				Cash Flow Benchmarks				Unpaid Claim Runoff Benchmarks		
Accident Year	Mean	Standard Error	CoV	Calendar Year	Mean	Standard Error	CoV	Calendar Year	Mean	Standard Error	CoV
2008								2017	158,515	34,245	21.6 %
2009	95	214	224.8 %	2018	61,896	16,101	26.0 %	2018	96,619	23,882	24.7 %
2010	262	387	147.8 %	2019	40,956	12,296	30.0 %	2019	55,663	16,877	30.3 %
2011	616	719	116.8 %	2020	24,529	9,031	36.8 %	2020	31,133	12,234	39.3 %
2012	1,735	1,374	79.2 %	2021	13,581	6,286	46.3 %	2021	17,552	8,841	50.4 %
2013	4,525	2,874	63.5 %	2022	7,252	4,308	59.4 %	2022	10,301	6,585	63.9 %
2014	11,154	5,736	51.4 %	2023	4,067	3,349	82.4 %	2023	6,234	5,381	86.3 %
2015	23,905	10,194	42.6 %	2024	2,437	2,827	116.0 %	2024	3,797	4,346	114.5 %
2016	43,759	16,882	38.6 %	2025	1,698	2,268	133.6 %	2025	2,099	4,164	198.3 %
2017	72,465	26,216	36.2 %	2026	2,099	4,164	198.3 %				
Totals	**158,515**	**34,245**	**21.6 %**		**158,515**	**34,245**	**21.6 %**				

6.2. Correlation Benchmarks

As noted at the end of Section 3, the data from 1,182 of the companies had at least 2 LOBs with Valid Data for at least one year. For each company (and year) with 2 or more LOBs, the correlation between the residuals was also calculated and saved, including the P-Values and the Degrees of Freedom, both before and after the hetero group factor adjustments, for both paid and incurred data. This database of 195,228 pairs of LOBs with correlation values were used to create separate benchmarks of correlation between Schedule P lines of business.

The correlation benchmarks include each year separately and all years combined, but only a sample from 1996 is illustrated in Table 6.5. In addition to calculating the sample average and standard deviations by pair, the number of pairs are also shown.

Table 6.5. Sample Correlation Benchmarks for 1996 – Paid After Hetero Adjustment – Raw Data

Mean Values	MPL-O	HO	WC	CA	PPA
MPL-O	100.0%	-0.5%	-10.9%	2.8%	-1.9%
HO	-0.5%	100.0%	4.0%	5.9%	11.8%
WC	-10.9%	4.0%	100.0%	11.9%	13.9%
CA	2.8%	5.9%	11.9%	100.0%	13.3%
PPA	-1.9%	11.8%	13.9%	13.3%	100.0%

Standard Deviations	MPL-O	HO	WC	CA	PPA
MPL-O	0.0%	11.1%	14.0%	16.1%	16.7%
HO	11.1%	0.0%	20.0%	18.9%	20.8%
WC	14.0%	20.0%	0.0%	23.5%	23.7%
CA	16.1%	18.9%	23.5%	0.0%	24.3%
PPA	16.7%	20.8%	23.7%	24.3%	0.0%

Count of Pairs	MPL-O	HO	WC	CA	PPA
MPL-O	-	57	62	59	48
HO	57	-	618	757	851
WC	62	618	-	688	570
CA	59	757	688	-	784
PPA	48	851	570	784	-

The P-Values are a measure of how significantly different from zero the correlation value is for each calculated pair. The lower the P-Value the more significantly different from zero the correlation. Thus, a second set of correlation benchmarks, using one minus the P-Value as the weights, were calculated for weighted means and weighted standard deviations. For comparison, the weighted benchmarks for the same sample are included in Table 6.6.

Table 6.6. Sample Correlation Benchmarks for 1996 – Paid After Hetero Adjustment – Weighted

Mean Values	MPL-O	HO	WC	CA	PPA
MPL-O	100.0%	0.0%	-16.2%	5.9%	-1.7%
HO	0.0%	100.0%	5.4%	9.5%	16.7%
WC	-16.2%	5.4%	100.0%	17.1%	18.9%
CA	5.9%	9.5%	17.1%	100.0%	19.3%
PPA	-1.7%	16.7%	18.9%	19.3%	100.0%

Standard Deviations	MPL-O	HO	WC	CA	PPA
MPL-O	0.0%	14.0%	14.6%	18.8%	18.6%
HO	14.0%	0.0%	23.6%	22.9%	22.9%
WC	14.6%	23.6%	0.0%	26.6%	26.0%
CA	18.8%	22.9%	26.6%	0.0%	27.1%
PPA	18.6%	22.9%	26.0%	27.1%	0.0%

Count of Pairs	MPL-O	HO	WC	CA	PPA
MPL-O	-	57	62	59	48
HO	57	-	618	757	851
WC	62	618	-	688	570
CA	59	757	688	-	784
PPA	48	851	570	784	-

While it was noted in Section 4 that aggregate simulations were not captured, and thus not available for additional benchmarks, it is quite straightforward to use the correlation benchmarks in conjunction with the unpaid benchmarks to create a customized aggregate unpaid benchmark. Finally, as noted above the Degrees of Freedom was also captured and could have been included as part of Tables 6.5 and 6.6. In practice, this would be a valuable benchmark for copulas used for aggregation as they are intended to strengthen the tail of the aggregate distribution given a selected correlation.

6.3. LDF Pattern Benchmarks

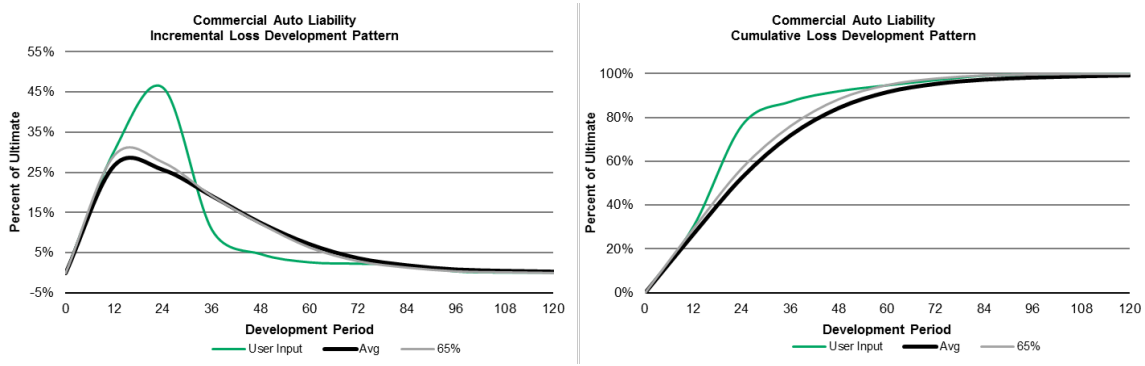
In addition to all of the simulation results, for each dataset the all year volume weighted average loss development pattern from the original paid triangle (actual), along with the implied pattern from the average of all the simulated sample paid triangles (simulated), were captured. Using all of the paid patterns by line of business, the mean and percentiles of these patterns can be used as LDF pattern benchmarks. For example, the development patterns for Commercial Auto sample used in the Tables in Section 6 are included in Table 6.7.

Table 6.7. Sample LDF Pattern Benchmarks – Commercial Auto

Development Periods:	12-24	24-36	36-48	48-60	60-72	72-84	84-96	96-108	108-120	120+
Actual LDF Pattern:	30.4%	76.4%	87.2%	91.9%	94.6%	97.0%	99.0%	99.5%	99.7%	99.9%
Average LDF Pattern:	26.9%	52.6%	71.8%	84.3%	91.5%	95.2%	97.2%	98.1%	98.7%	99.1%
65% LDF Pattern:	29.3%	56.9%	76.1%	88.3%	94.8%	97.7%	99.1%	99.6%	99.8%	99.9%

The actual LDF pattern was calculated using the all year volume weighted average LDF factors from the sample dataset. The average and 65% LDF patterns are the average and 65th percentile from all of the simulated patterns in the database, respectively. By systematic testing and a little trial and error, the 65th percentile was found to be the best fit to the actual pattern. The patterns from Table 6.7 are illustrated in Graph 6.4, since one of the uses of LDF pattern benchmarks could be to help smooth the selection of age-to-age factors.

Graph 6.4. Comparison of Actual with Benchmark LDF Patterns

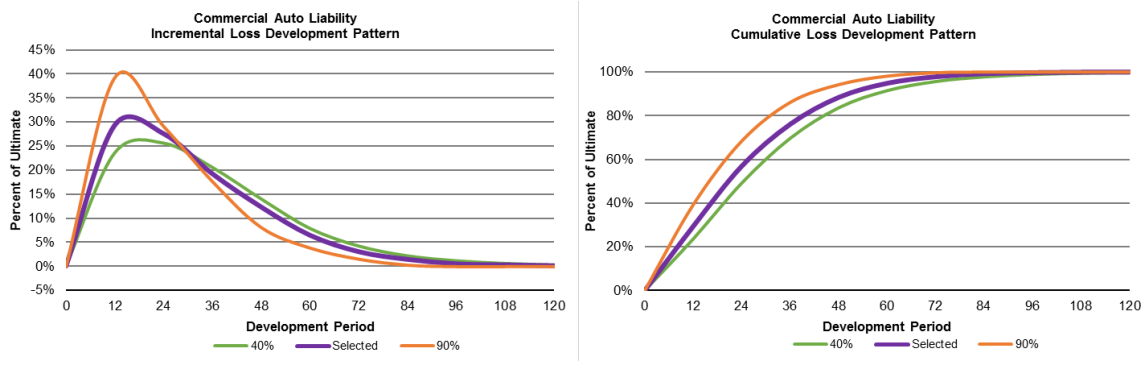


Once the actual LDF pattern has been smoothed, or a suitable percentile pattern has been selected, another use of the LDF pattern benchmarks is to help create a range of deterministic central estimates. For example, assuming the 65th percentile pattern is selected, the actuary could then base a deterministic range on the patterns which are 25 points above and below the 65th percentile as illustrated in Table 6.8 and Graph 6.5.

Table 6.8. Sample LDF Pattern Range – Commercial Auto

Development Periods:	12-24	24-36	36-48	48-60	60-72	72-84	84-96	96-108	108-120	120+
40% LDF Pattern:	23.6%	49.1%	69.6%	83.6%	91.4%	95.5%	97.7%	98.8%	99.3%	99.5%
65% LDF Pattern:	26.9%	52.6%	71.8%	84.3%	91.5%	95.2%	97.2%	98.1%	98.7%	99.1%
90% LDF Pattern:	39.3%	68.5%	86.1%	94.3%	98.2%	99.7%	100.0%	100.0%	100.0%	100.0%

Graph 6.5. Range of Benchmark LDF Patterns



7. CONCLUSIONS

Using an extensive database pulled from historical Schedule P data, the results from back-testing various ODP Bootstrap models and the Mack Bootstrap model has confirmed similar prior research on how effective these models predict the distribution of possible outcomes. For the versions of the ODP Bootstrap model not previously tested, the back-testing results are both encouraging and inconclusive. In particular, for the ODP Bootstrap incurred chain ladder model, as described in Shapland [17], using both the paid and incurred data significantly improves the results. For the ODP Bootstrap Bornhuetter-Ferguson and Cape Cod models the results were inconclusive due to the need to test more model parameters. However, even with inconclusive results for four of the six ODP Bootstrap models, testing of weighted results demonstrated that weighing multiple models, to at least partially address model risk, is a significant improvement over using a single model.

Due to the size of the database used in the back-testing, the data allows us to use benchmarking algorithms as a guide when evaluating the estimated distribution of possible outcomes from any stochastic model. These benchmarking algorithms are quite sophisticated in the sense that they address the statistical properties of real data sets (e.g., more relative variance for smaller exposures) and can be customized to more closely approximate the data being analyzed (e.g., using selected ATA factors). Additional uses from the data include correlation benchmarks and LDF pattern benchmarks.

Acknowledgment

The author gratefully acknowledges the many authors listed in the References (and others not listed) that contributed to the foundation of the ODP bootstrap and Mack models, without which this research would not have been possible. He would like to thank all the peer reviewers, and is grateful to the CAS referees, in particular Lynne Bloom and Jiao Ziyi, for their comments that greatly improved the quality of the paper.

Supplementary Material

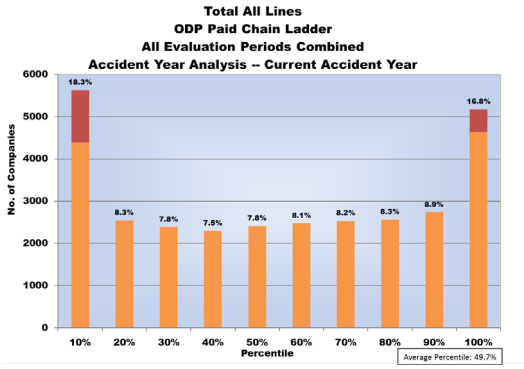
Given the vast size of the database used in this research and the proprietary nature of the results, no supplementary materials can be provided. However, the interested reader can contact the author to learn more about the proprietary benchmarks.

Appendix A – Back-Testing Results by Accident Year

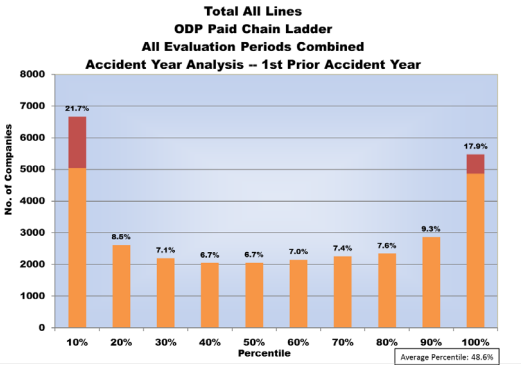
The back-testing results for the current accident year are shown in Graphs 5.3 and 5.9 for the ODP Bootstrap paid chain ladder and incurred chain ladder, respectively, and for completeness are repeated here in Graphs A.1 and A.10. All of the Graphs in Appendix A show results for the ODP Bootstrap paid chain ladder and incurred chain ladder models using the "Baseline Limits & Hetero" assumptions for all lines of business and all evaluation periods combined.

ODP Paid Chain Ladder:

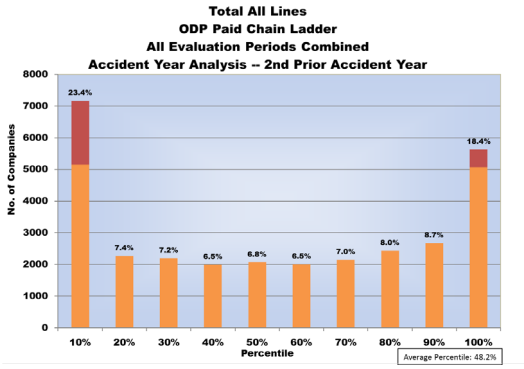
Graph A.1. Current Accident Year



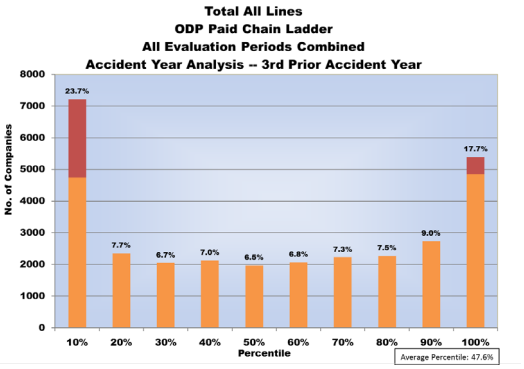
Graph A.2. 1st Prior Accident Year



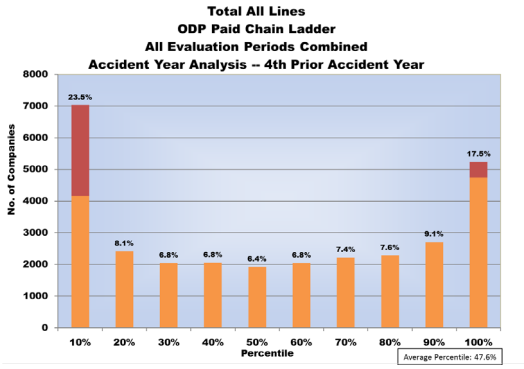
Graph A.3. 2nd Prior Accident Year



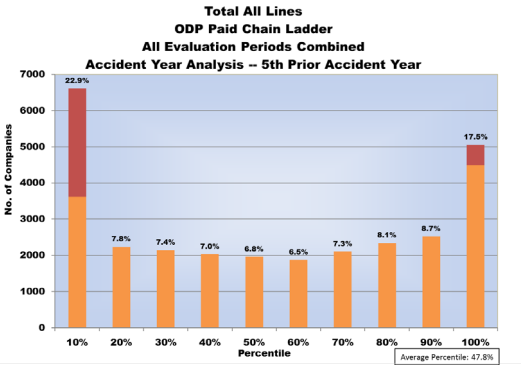
Graph A.4. 3rd Prior Accident Year



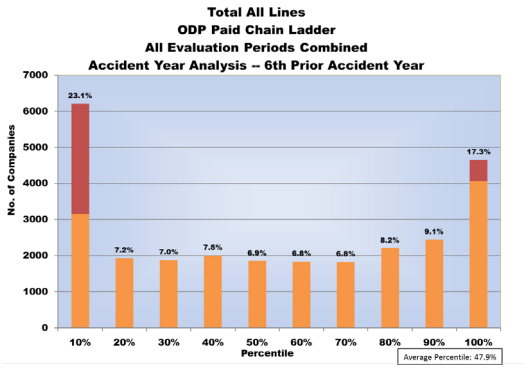
Graph A.5. 4th Prior Accident Year



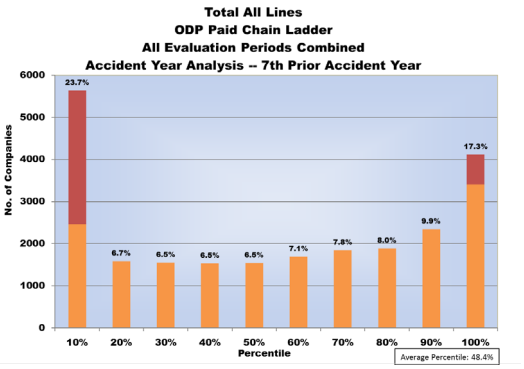
Graph A.6. 5th Prior Accident Year



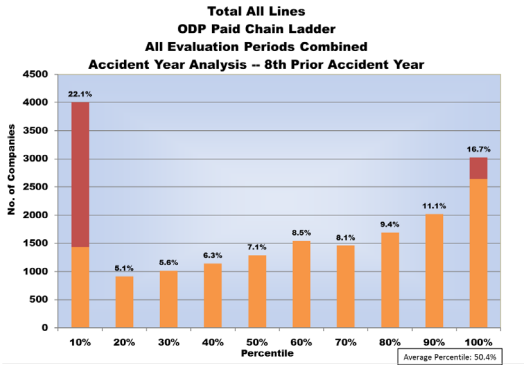
Graph A.7. 6th Prior Accident Year



Graph A.8. 7th Prior Accident Year

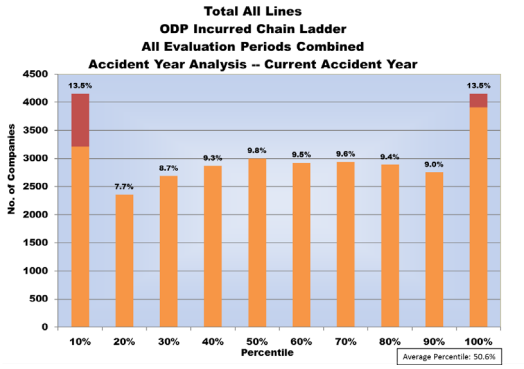


Graph A.9. 8th Prior Accident Year

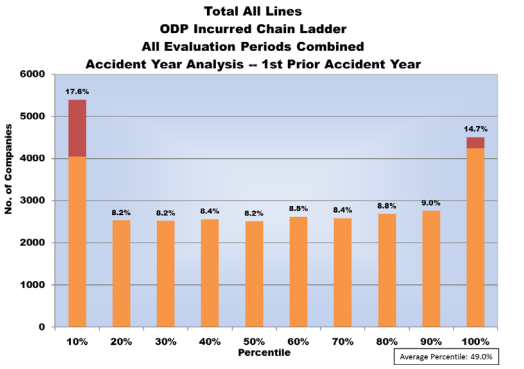


ODP Incurred Chain Ladder:

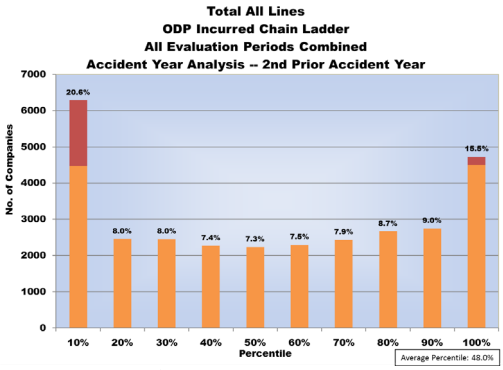
Graph A.10. Current Accident Year



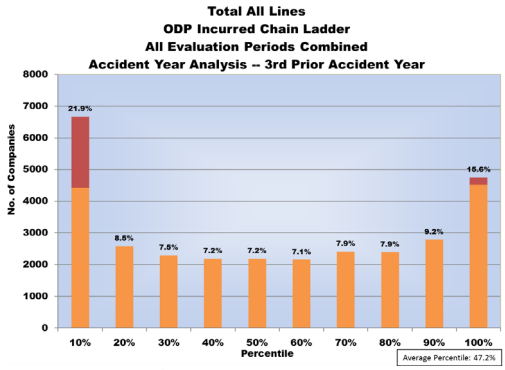
Graph A.11. 1st Prior Accident Year



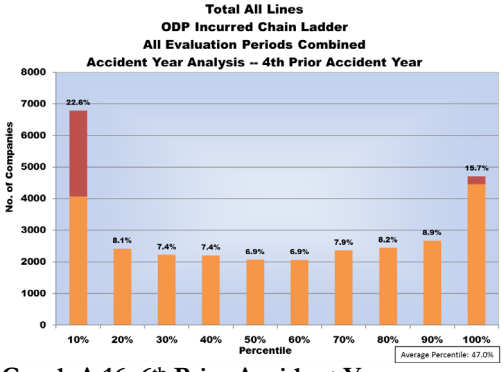
Graph A.12. 2nd Prior Accident Year



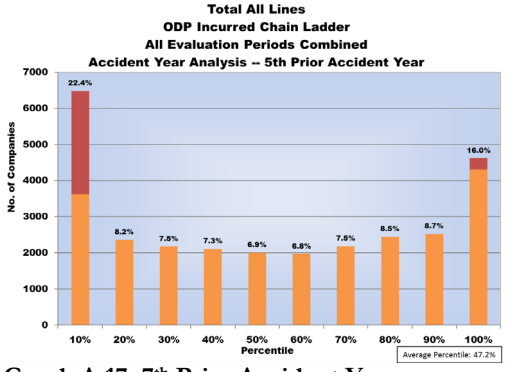
Graph A.13. 3rd Prior Accident Year



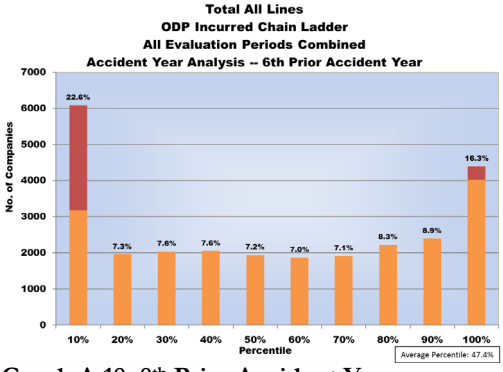
Graph A.14. 4th Prior Accident Year



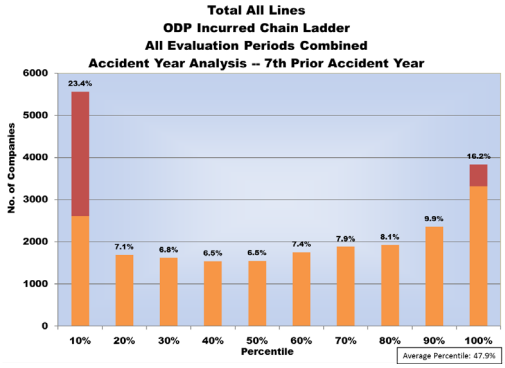
Graph A.15. 5th Prior Accident Year



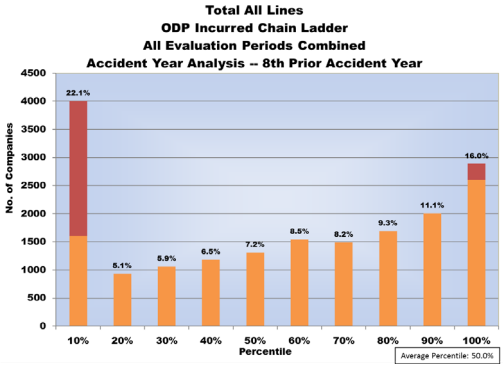
Graph A.16. 6th Prior Accident Year



Graph A.17. 7th Prior Accident Year



Graph A.18. 8th Prior Accident Year

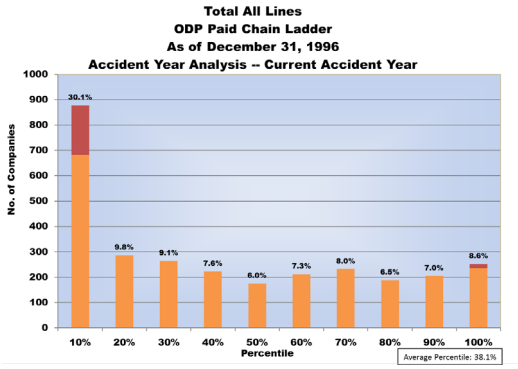


Appendix B – Back-Testing Results by Evaluation Year

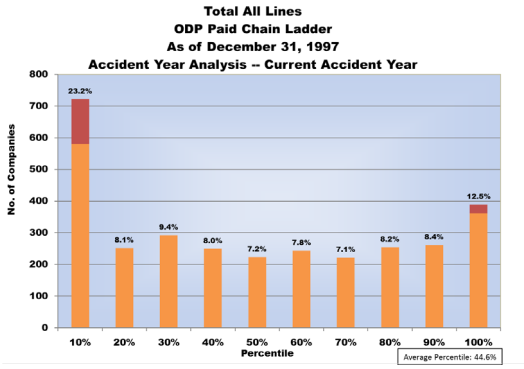
The back-testing results for the current accident year in Graphs 5.3 and 5.9, for the ODP Bootstrap paid chain ladder and incurred chain ladder, respectively, is for all evaluation years combined. All of the Graphs in Appendix B show results for the current accident year for the ODP Bootstrap paid chain ladder and incurred chain ladder models using the "Baseline Limits & Hetero" assumptions for all lines of business by evaluation periods.

ODP Paid Chain Ladder:

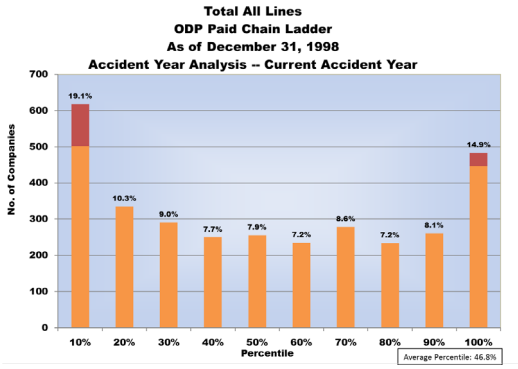
Graph B.1. Evaluation Year 1996



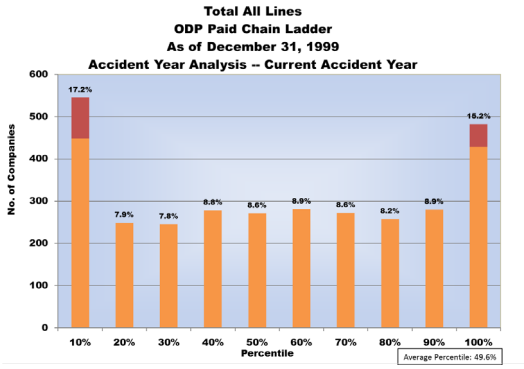
Graph B.2. Evaluation Year 1997



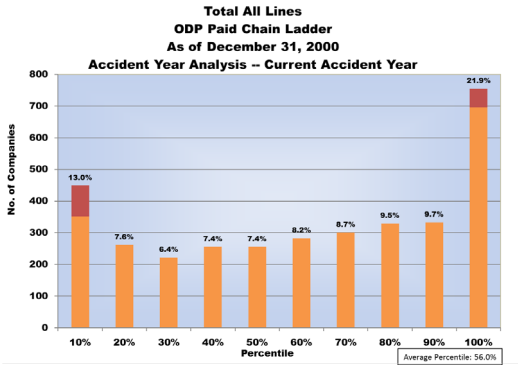
Graph B.3. Evaluation Year 1998



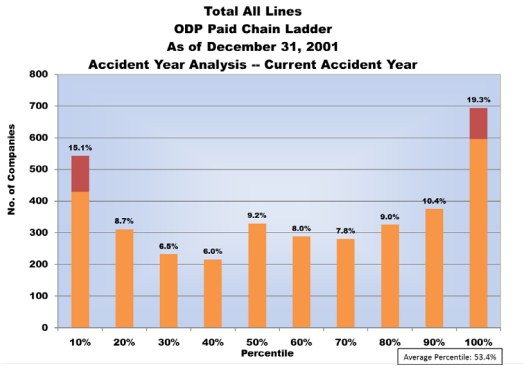
Graph B.4. Evaluation Year 1999



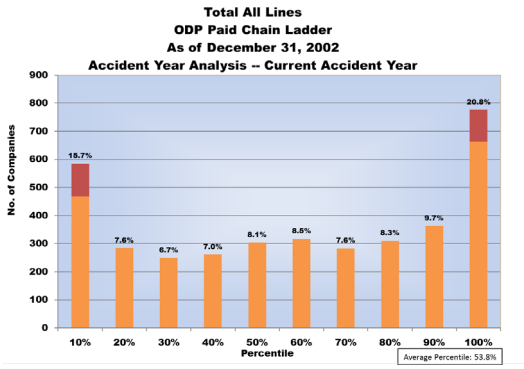
Graph B.5. Evaluation Year 2000



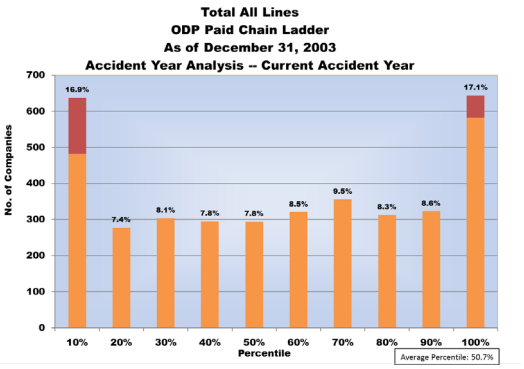
Graph B.6. Evaluation Year 2001



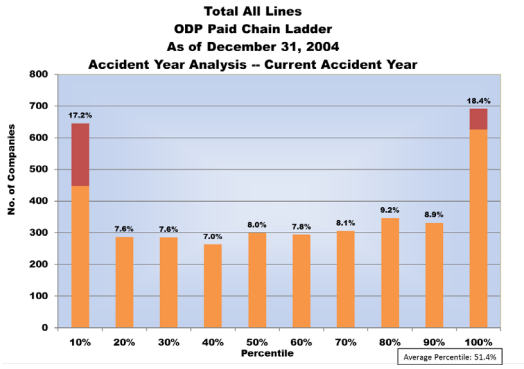
Graph B.7. Evaluation Year 2002



Graph B.8. Evaluation Year 2003

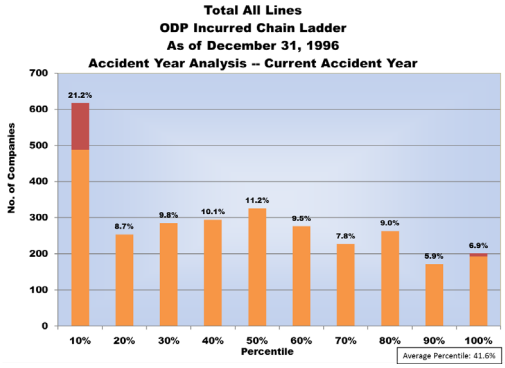


Graph B.9. Evaluation Year 2004

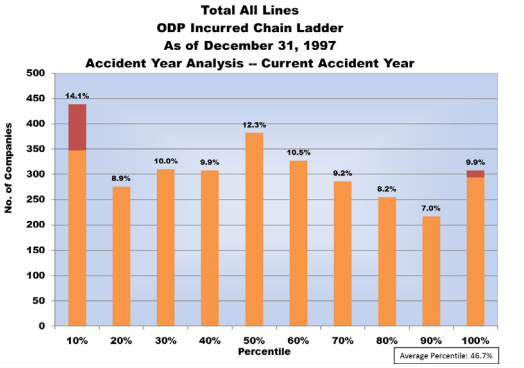


ODP Incurred Chain Ladder:

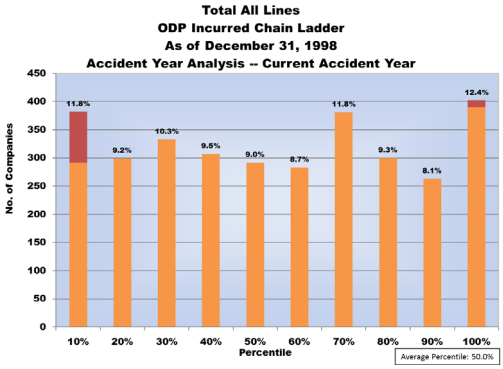
Graph B.10. Evaluation Year 1996



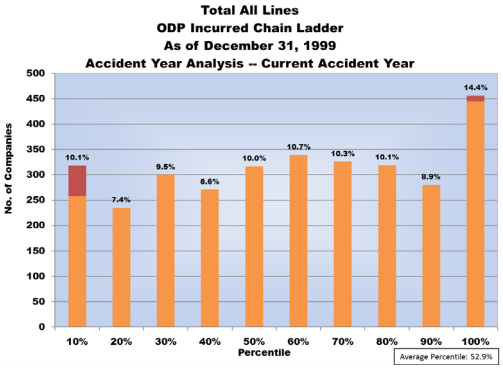
Graph B.11. Evaluation Year 1997



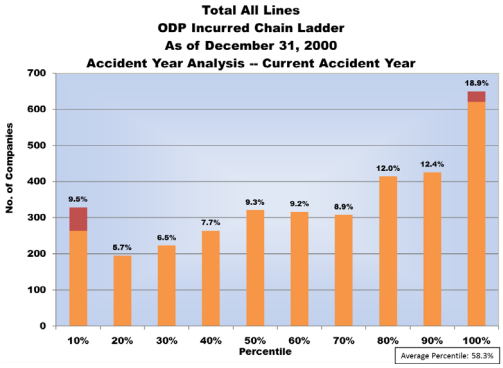
Graph B.12. Evaluation Year 1998



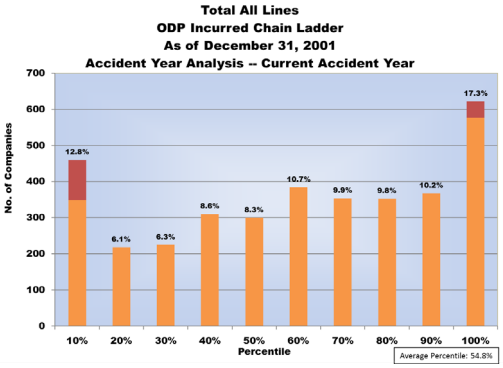
Graph B.13. Evaluation Year 1999



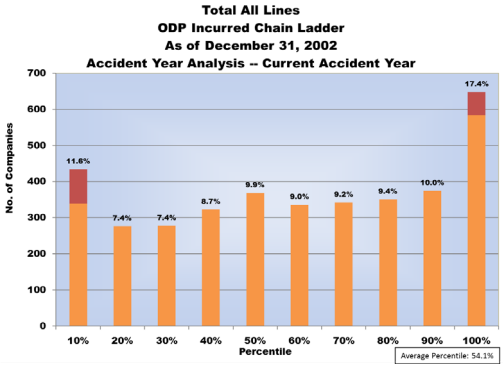
Graph B.14. Evaluation Year 2000



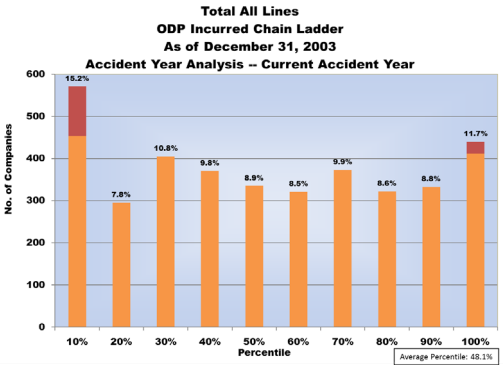
Graph B.15. Evaluation Year 2001



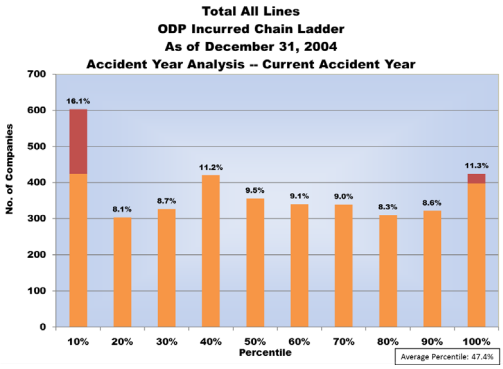
Graph B.16. Evaluation Year 2002



Graph B.17. Evaluation Year 2003



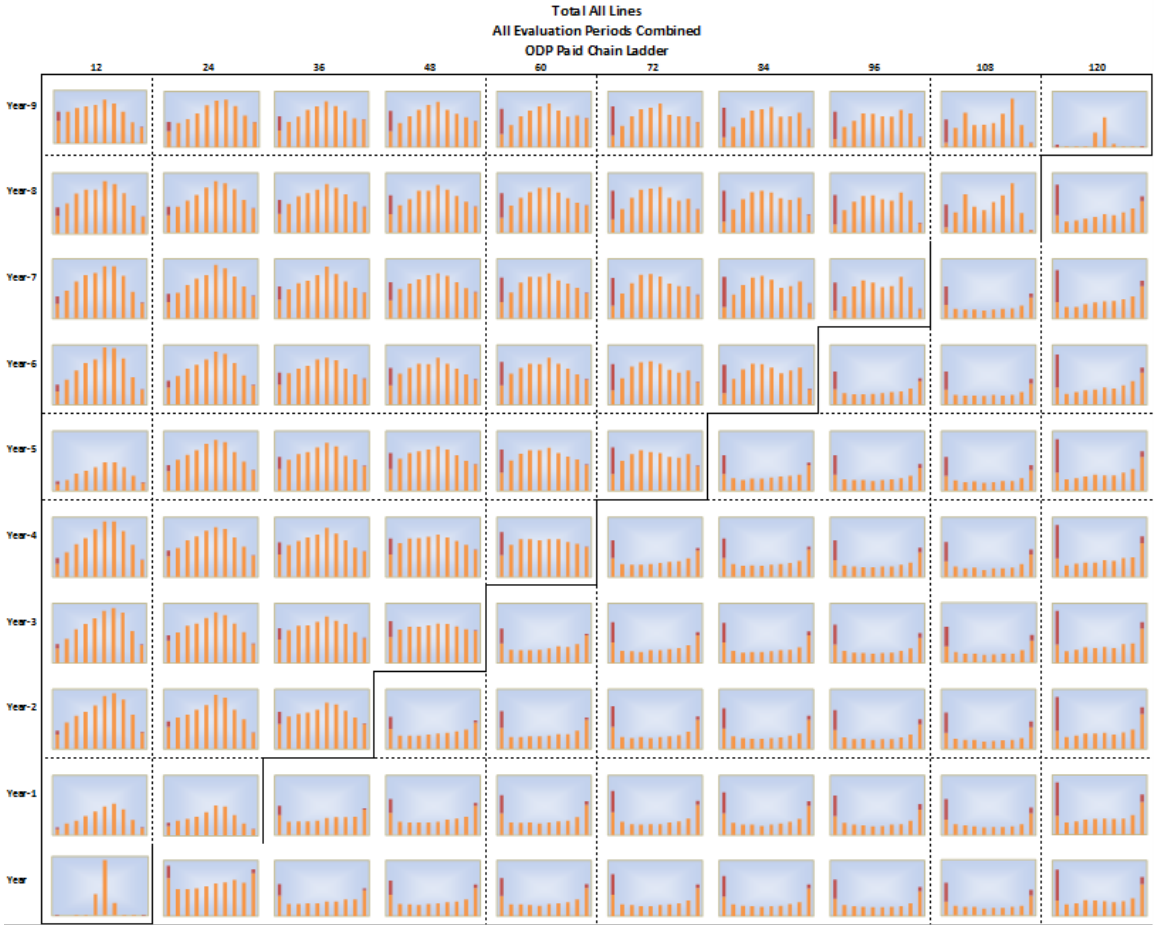
Graph B.18. Evaluation Year 2004



Appendix C – Back-Testing Results by Incremental Cell

The back-testing results in Appendix C show results for the ODP Bootstrap paid chain ladder and incurred chain ladder models using the "Baseline Limits & Hetero" assumptions for all lines of business and all evaluation periods combined.

Graph C.1. ODP Bootstrap Paid Chain Ladder by Incremental Cell



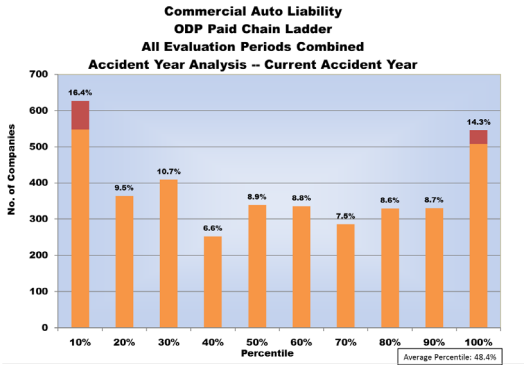
Graph C.2. ODP Bootstrap Incurred Chain Ladder by Incremental Cell

Appendix D – Back-Testing Results by Line of Business

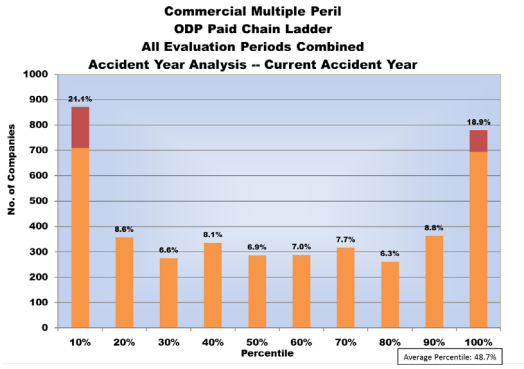
The back-testing results for the current accident year for all lines of business is included as Graphs 5.3 and 5.9 for the ODP Bootstrap paid chain ladder and incurred chain ladder, respectively. All of the Graphs in Appendix D show results for the ODP Bootstrap paid chain ladder and incurred chain ladder models using the "Baseline Limits & Hetero" assumptions for all evaluation periods combined, separately for each Schedule P line of business.

ODP Paid Chain Ladder:

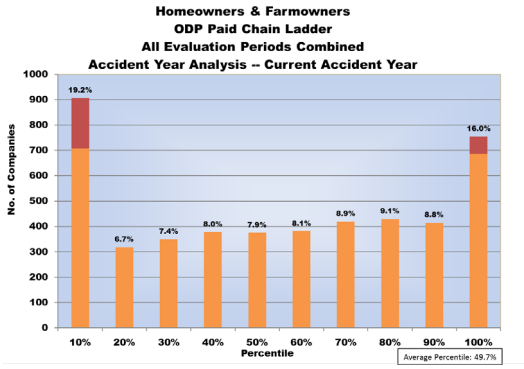
Graph D.1. Commercial Auto Liability



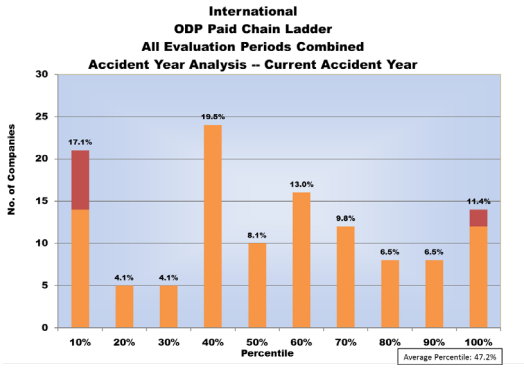
Graph D.2. Commercial Multiple Peril



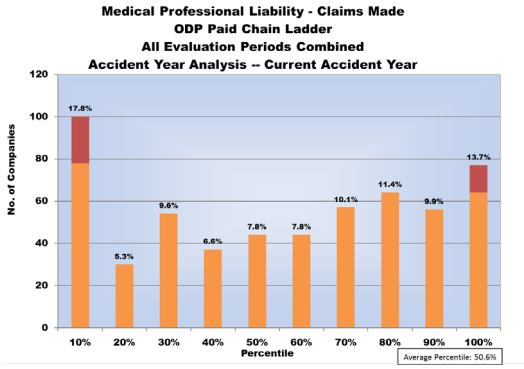
Graph D.3. Homeowners & Farmowners



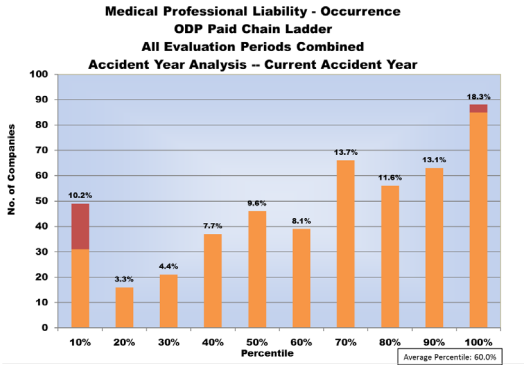
Graph D.4. International



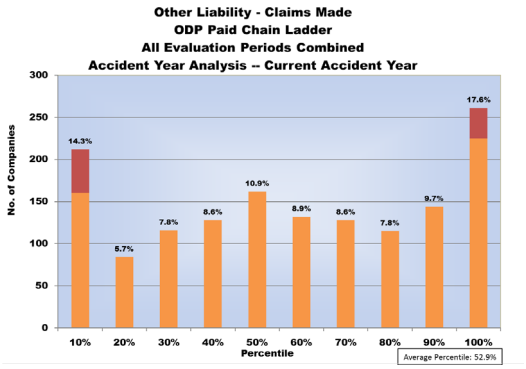
Graph D.5. Med. Prof. Liab. - Claims Made



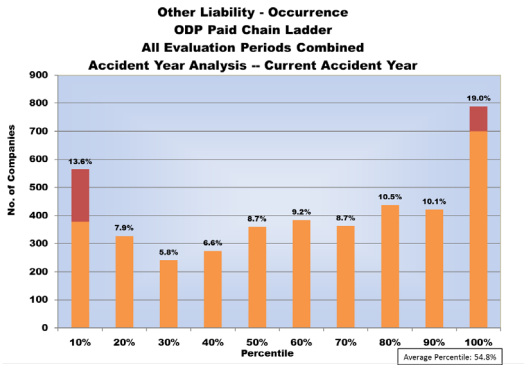
Graph D.6. Med. Prof. Liab. - Occurrence



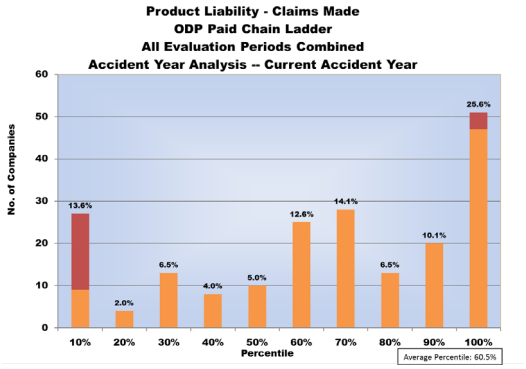
Graph D.7. Other Liability - Claims Made



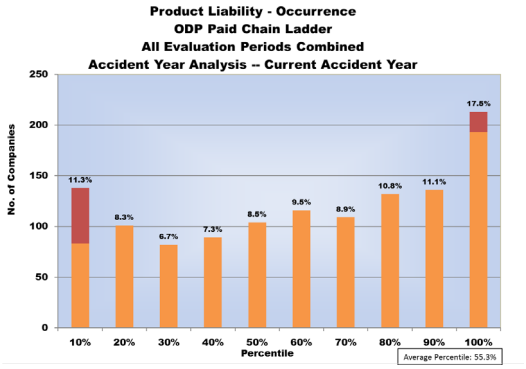
Graph D.8. Other Liability - Occurrence



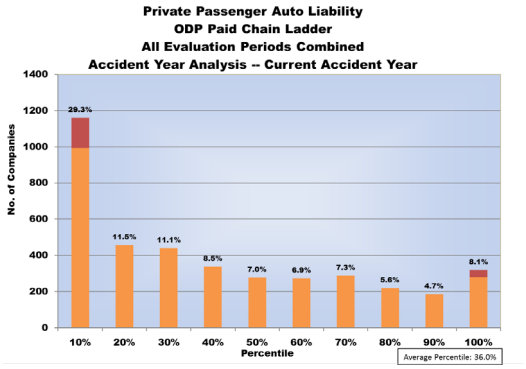
Graph D.9. Product Liability - Claims Made



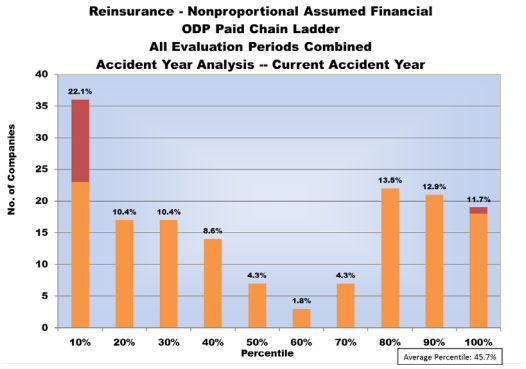
Graph D.10. Product Liability - Occurrence



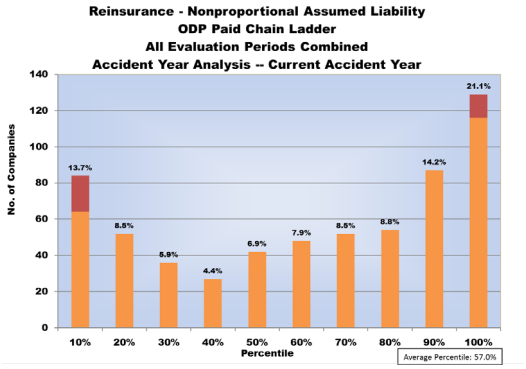
Graph D.11. Private Passenger Auto Liability



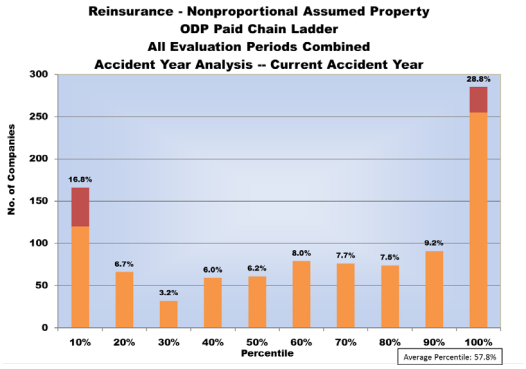
Graph D.12. Reins. – NP Assumed Financial



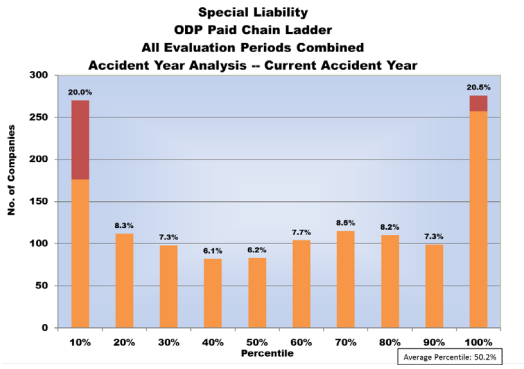
Graph D.13. Reins. – NP Assumed Liability



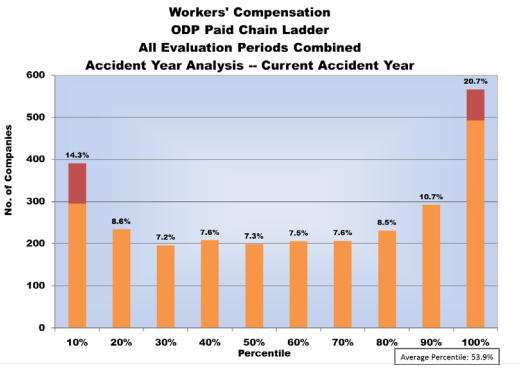
Graph D.14. Reins. – NP Assumed Property



Graph D.15. Special Liability

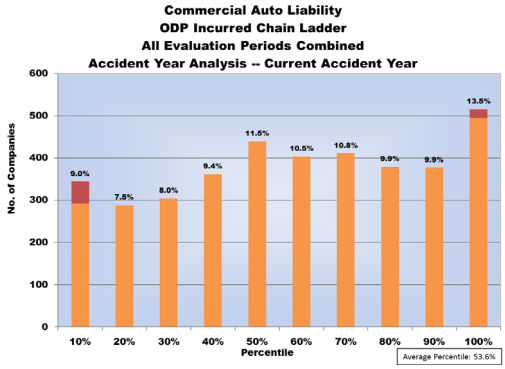


Graph D.16. Workers' Compensation

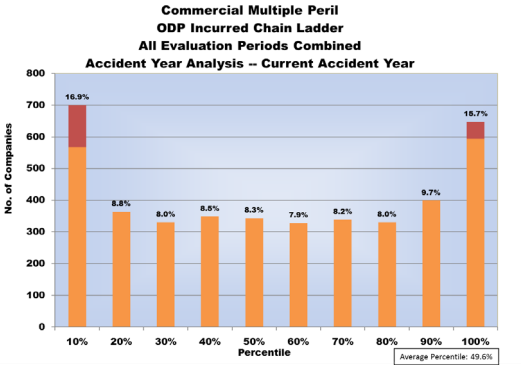


ODP Incurred Chain Ladder:

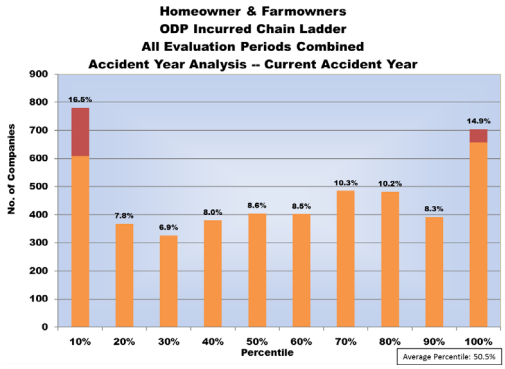
Graph D.17. Commercial Auto Liability



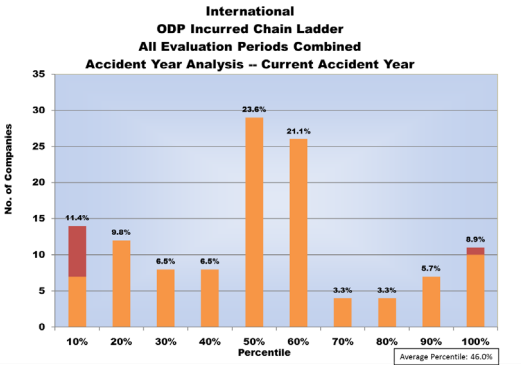
Graph D.18. Commercial Multiple Peril



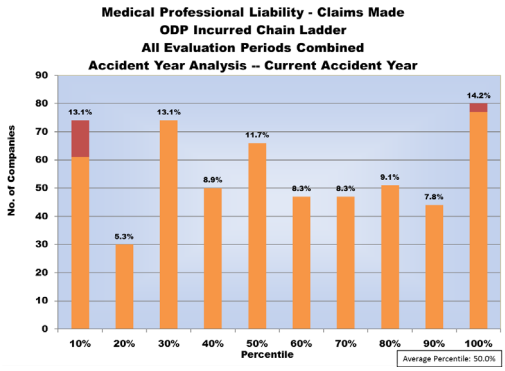
Graph D.19. Homeowners & Farmowners



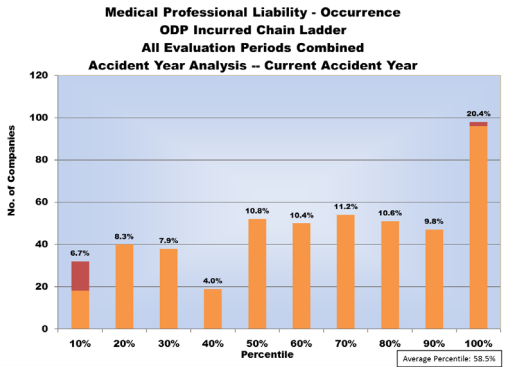
Graph D.20. International



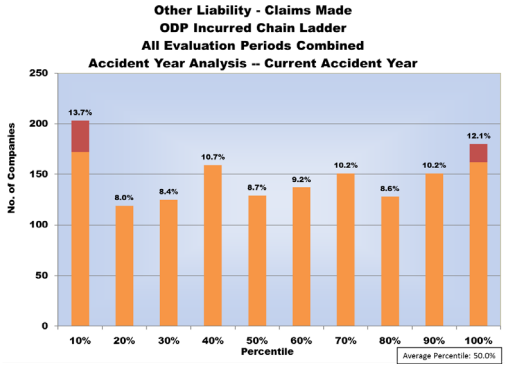
Graph D.21. Med. Prof. Liab. - Claims Made



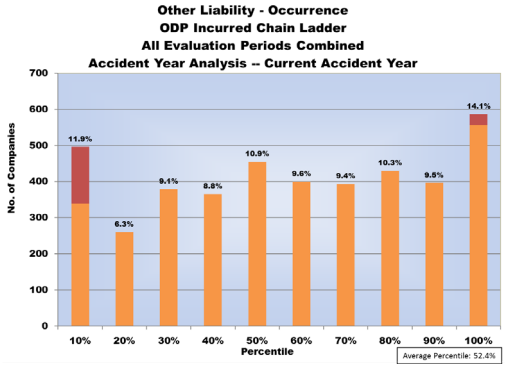
Graph D.22. Med. Prof. Liab. - Occurrence



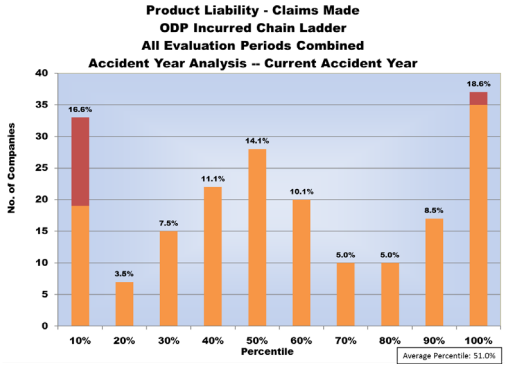
Graph D.23. Other Liability - Claims Made



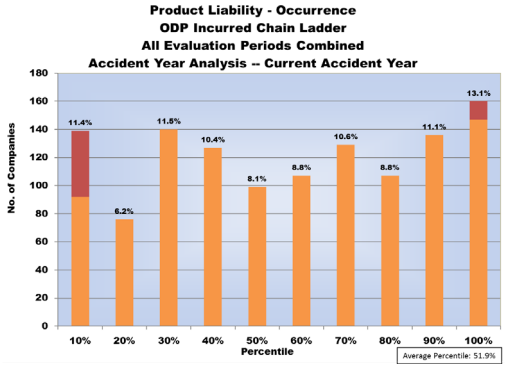
Graph D.24. Other Liability - Occurrence



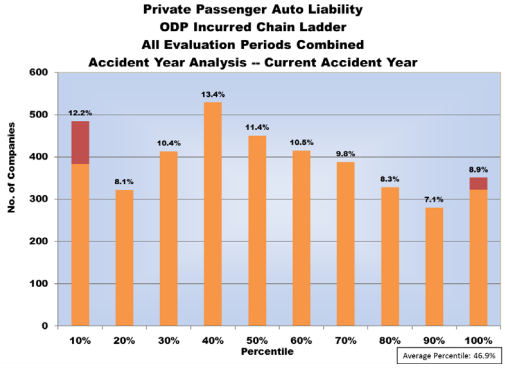
Graph D.25. Product Liability - Claims Made



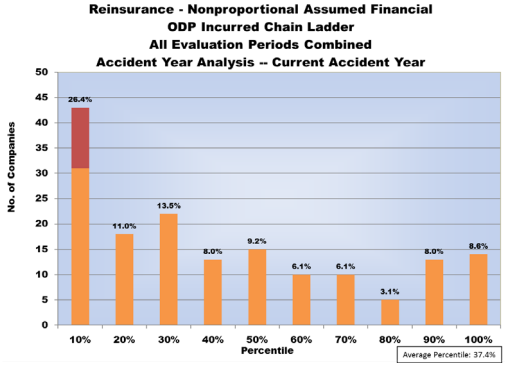
Graph D.26. Product Liability - Occurrence



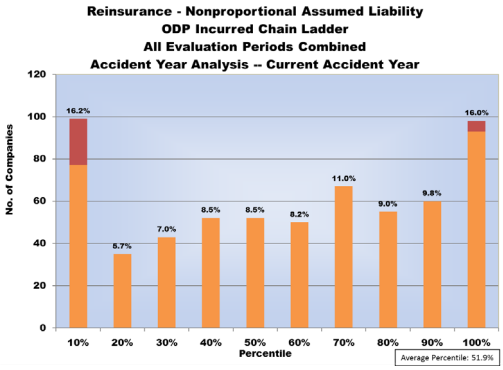
Graph D.27. Private Passenger Auto Liability



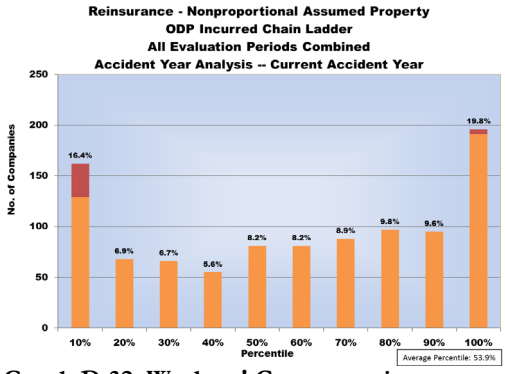
Graph D.28. Reins. – NP Assumed Financial



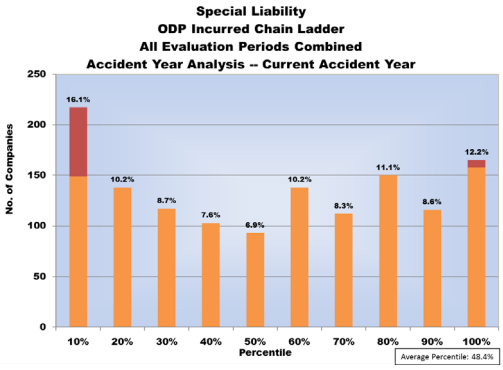
Graph D.29. Reins. – NP Assumed Liability



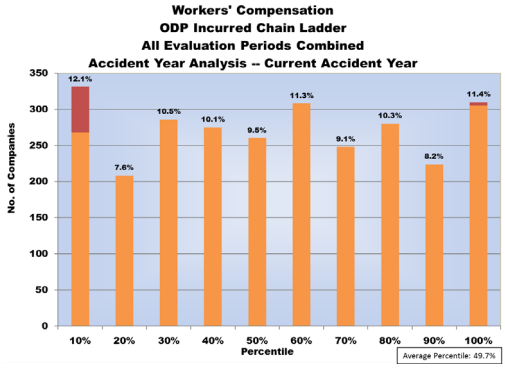
Graph D.30. Reins. – NP Assumed Property



Graph D.31. Special Liability



Graph D.32. Workers' Compensation



REFERENCES

[1] Bornhuetter, Ronald, and Ronald Ferguson. 1972. "The Actuary and IBNR." *Proceedings of the Casualty Actuarial Society* LIX: 181-195.

[2] CAS Working Party on Quantifying Variability in Reserve Estimates. 2005. "The Analysis and Estimation of Loss & ALAE Variability: A Summary Report." *CAS Forum* (Fall): 29-146.

[3] England, Peter D., and Richard J. Verrall. 1999. "Analytic and Bootstrap Estimates of Prediction Errors in Claims Reserving." Insurance: Mathematics and Economics 25: 281-293.

[4] England, Peter D., and Richard J. Verrall. 2002. "Stochastic Claims Reserving in General Insurance." British Actuarial Journal 8-3: 443-544.

[5] England, Peter D., and Richard J. Verrall. 2006. "Predictive Distributions of Outstanding Liabilities in General Insurance." Annals of Actuarial Science 1-2: 221-270.

[6] Forray, Susan J. 2012. "Looking Back to See Ahead: A Hindsight Analysis of Actuarial Reserving Methods." *CAS Forum* (Summer): 1-33.

[7] Gremillet, Marion, Pierre Miehe, and José Luis Vilar Zanón. 2014. "Back-Testing the Reversible Jump Markov Chain Monte Carlo & further extensions." Presented at 2014 ICA in Washington DC.

[8] Iman, R., and W. Conover. 1982. "A Distribution-Free Approach to Inducing Rank Correlation Among Input Variables." *Communications in Statistics--Simulation and Computation* 11(3): 311–334.

[9] Jing, Yi, Joseph R. Lebens, and Stephen P. Lowe. 2009. "Claim Reserving: Performance Testing and the Control Cycle." *Variance* 3-2: 161-193.

[10] Kirschner, Gerald S., Colin Kerley, and Belinda Isaacs. 2008. "Two Approaches to Calculating Correlated Reserve Indications Across Multiple Lines of Business." *Variance* 1: 15-38.

[11] Leong, (Weng Kah) Jessica, Shaun S. Wang and Han Chen. 2014. "Back-Testing the ODP Bootstrap of the Paid Chain-Ladder Model with Actual Historical Claims Data." *Variance* 8-2: 182-202.

[12] Meyers, G., and P. Shi. 2011. "The Retrospective Testing of Stochastic Loss Reserve Methods." CAS Forum (Summer):

[13] Mildenhall, Stephen J. 2006. "Correlation and Aggregate Loss Distributions with an Emphasis on the Iman-Conover Method." *Casualty Actuarial Society E-Forum* (Winter): 103-204.

[14] Pinheiro, Paulo J. R., João Manuel Andrade e Silva, and Maria de Lourdes Centeno. 2003. "Bootstrap Methodology in Claim Reserving." Journal of Risk and Insurance 70: 701-714.

[15] ROC/GIRO Working Party. 2007. "Best Estimates and Reserving Uncertainty." Institute of Actuaries.

[16] ROC/GIRO Working Party. 2008. "Reserving Uncertainty." Institute of Actuaries.

[17] Shapland, Mark R. 2016. "Using the ODP Bootstrap Model: A Practitioner's Guide." *CAS Monograph 4*.

[18] Skurnick, David. 1973. "A Survey of Loss Reserving Methods." *Proceedings of the Casualty Actuarial Society* LX: 16-58.

[19] Struzzieri, Paul J., and Paul R. Hussian. 1998. "Using Best Practices to Determine a Best Reserve Estimate." *Casualty Actuarial Society Forum* (Fall): 353-413.

[20] Venter, Gary G. 1998. "Testing the Assumptions of Age-to-Age Factors." *Proceedings of the Casualty Actuarial Society* LXXXV: 807-47.

Abbreviations and notations

APD, automobile physical damage
ATA, age-to-age
CL, chain ladder
DFA, dynamic financial analysis
GLM, generalized linear models
OLS, ordinary least squares
ERM, enterprise risk management

Biography of the Author

Mark R. Shapland is a Principal & Consulting Actuary in Milliman's Dubai office where he is responsible for various reserving and pricing projects for a variety of clients and was previously the lead actuary for the Property & Casualty Insurance Software (PCIS) development team at Milliman. He has a B.S. degree in Integrated Studies (Actuarial Science) from the University of Nebraska-Lincoln. He is a Fellow of the Casualty Actuarial Society, a Fellow of the Society of Actuaries, a Fellow of the Institute of Actuaries of India, and a Member of the American Academy of Actuaries. He was the leader of Section 3 of the Reserve Variability Working Party, the Chair of the CAS Committee on Reserves, co-chair of the Tail Factor Working Party, and co-chair of the Loss Simulation Model Working Party. He is also a co-developer and co-presenter of the CAS Reserve Variability Limited Attendance Seminar and has spoken frequently on this subject both within the CAS and internationally. He can be contacted at mark.shapland@milliman.com.