Actuaries
Risk is Opportunity.®

Model Efficiency Study Results

NOVEMBER 2011

SPONSORED BY

Financial Reporting Section
Product Development Section
Committee on Life Insurance Research
Society of Actuaries

PREPARED BY

Bruce B. Rosner, FSA, MAAA
Ernst & Young LLP

Acknowledgments

Bruce Rosner led Ernst & Young's project team on this assignment, with oversight from David Minches and support from Ilan Man and Cathy Fan. We would like to acknowledge and thank a number of other individuals who contributed to the success of this study:

- Ronora Stryker and Jan Schuh from the Society of Actuaries for providing leadership and coordination
- The Project Oversight Group (POG) for guidance throughout this project. The members of the POG are:
  - Mark Alberts
  - Mary Bahna-Nolan
  - Sue Deakins
  - Trevor Howes
  - Lisa Lefkowitz
  - Craig Reynolds (chair)
- Steve Strommen from Northwestern Mutual and Yvonne Chueh from Central Washington University for various discussions during the project
- The companies that volunteered to participate in the study:
  - Allstate Financial
  - Ameriprise Financial
  - Assurity
  - Aviva USA
  - Commonwealth
  - Horace Mann
  - ING
  - Lincoln
  - MetLife
  - Milliman
  - Nationwide
  - Sunlife

Table of Contents

Executive Summary

The purpose of the model efficiency study was to test certain mathematical approaches that can reduce the number of model points or scenarios required to achieve a given level of precision in stochastic actuarial modeling.

- Twelve companies volunteered to participate in the research.
- A total of six techniques were tested.
- Each company tested approximately two of the techniques.

The following summarizes each technique that was tested during the study and the key observations for each technique.

Techniques Tested

The model efficiency study focused on six techniques. The first four techniques below focus on reducing the number of scenarios required (scenario reduction) and the last two techniques focus on reducing the number of model points required to represent the inforce policies (inforce compression).

A. Transfer Scenario Order	Determine the ranking of scenarios based on the ranking from running a subset of policies
B. Representative Scenarios	Select a subset of scenarios that are "representative" of the full scenario set based on certain characteristics of the scenarios
C. Importance Sampling	Sample more scenarios in parts of the distribution that are more critical to the overall result, and subject those scenarios to reduced weights
D. Curve Fitting	Determine an underlying distribution that fits well to the measured distribution of a variable, and report using the underlying distribution

E. Cluster Modeling	Mathematically locate policies whose results are "close" and combine them to produce a reduced scaled subset of policies that will have similar characteristics to the full inforce
F. Replicating Liabilities	Use optimization to determine a reduced scaled subset of policies that will have similar characteristics to the full inforce

Summary of Observations

In practice, companies will face a selection process if they choose to go down the path of seeking model efficiency. Below is a summary of the key strengths and weaknesses of each technique as experienced by the participating companies:

	Initial Effort Required	Runtime Reduction Experienced	Limitations Experienced
A. Transfer Scenario Order	Low	Low	Only useful for tail metrics
B. Representative Scenarios	Moderate	Moderate	More useful for the mean; limited by accuracy of baseline scenarios
C. Importance Sampling	Moderate	Moderate (subject to limitations)	More useful for out-of-the-money options
D. Curve Fitting	High	Unknown	More useful for tail metrics; bias is a significant concern
E. Cluster Modeling	High	High	Bias is a concern for all compression techniques
F. Replicating Liabilities	High	High	Bias is a concern for all compression techniques

Limitations

In summarizing the results of the study, we relied on the accuracy and completeness of the underlying data and modeling performed by the participating companies. We did not audit or

independently verify such information, or confirm the accuracy of the data or the information and explanations provided by the participating companies.

It is important to note that these results are limited by the following:

- Each technique has only been tested by a small number of companies.
- The depth of the testing varied by technique, often based on the level of effort required to set up the process.
- Participating companies only produced test results, and the techniques would require additional refinement and review in order to be used in practice.
- Samples sizes were less than ideal.

As a result, the results of this study are only useful to provide general guidance and direction for actuaries.

I. Introduction

The purpose of this study is to test certain mathematical approaches that can reduce the number of model points or scenarios required to achieve a given level of precision in stochastic actuarial modeling. This has been termed "model efficiency."

The insurance industry now stands at a critical junction where:

(a) Options and guarantees in the products that insurance companies now offer demand more sophisticated measurements for pricing, financial reporting and risk management.

(b) Regulatory reporting regimes across the world are moving toward prospective, principle-based modeling approaches to valuation.

(c) Management at companies across the finance industry demands high levels of precision.

(d) Management also demands a variety of sensitivity testing, and attribution of movement often requiring additional stochastic runs.

(e) Reporting time frames are compressed.

(f) Financial organizations are striving to contain and reduce costs.

The insurance industry has not reached an upper limit on the number of computers that can be combined to perform actuarial calculations. The following table shows the number of cores in use by 14 companies spanning the life insurance industry. Results are separate for companies with more complex modeling requirements (e.g., variable annuities or economic capital) and for those with less complex modeling requirements.

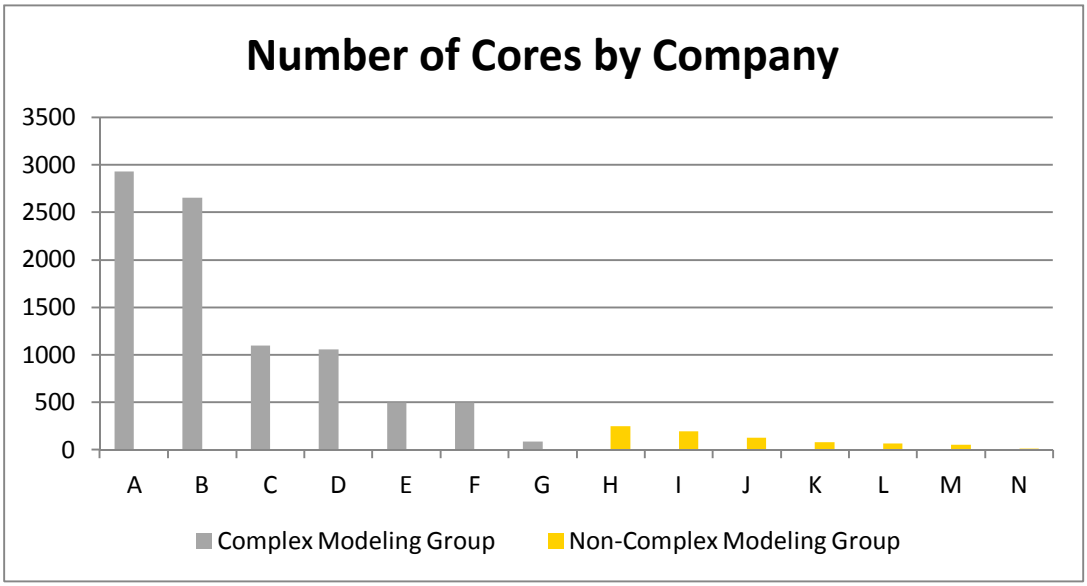

Figure I.1. Cores per company
Source: Ernst & Young 2011 Actuarial Benchmarking Survey

There has been a rising need in actuarial departments in North America and around the world to increase modeling capacity, precision and sophistication, and actuaries are now actively seeking alternatives to help them meet the demands of the business. Model efficiency is one alternative, encompassing a class of techniques that can help to address these objectives. (Note that high performance computing techniques are beyond the scope of this study.)

In setting up this study, Ernst & Young and the Society of Actuaries jointly reached out to the life insurance industry requesting cooperation and support on a volunteer basis. In return, we received a very favorable response. Twelve companies volunteered to join in the effort. More than a dozen other companies responded with interest in learning from the results of the study, but were unable to actively participate.

The majority of companies that participated in this study have stated that they plan to further develop the techniques that they tested and ultimately use them for live modeling. We hope that this report is successful at establishing another precedent that the industry and others can learn from and use to better and more efficiently manage risk.

Techniques Tested

The model efficiency study focused on the following six techniques:

1. Transfer Scenario Order
2. Representative Scenarios
3. Importance Sampling
4. Curve Fitting
5. Cluster Modeling
6. Replicating Liabilities.

The techniques were selected from a much larger group of available techniques that Ernst & Young has encountered in their research. Readers should not assume that these are the only available techniques; they were simply chosen because they met certain objectives and constraints of this study, including:

- The study is intended to investigate certain advanced cell compression and scenario reduction techniques.
- The participating companies had to be able to test the techniques with a limited budget and time, which largely restricted them to techniques that have software readily available, or otherwise require minimal setup time.
- The study is intended to capture techniques that are already emerging in the actuarial community, and techniques that are widely used in other industries but have not yet been widely applied in the actuarial community.

The remainder of this report will walk through each technique in some detail, including background information and study results.

II. Study Findings

This section examines each of the techniques and the testing that was performed. The following terminology will be used throughout the discussion of testing results:

- Error — This generally refers to the percentage difference between the results from the original run, and the results from the compressed/reduced run.
- Compression/Reduction — These terms can be used interchangeably to refer to the change in the size of the model. "Compression" is used in this report to refer to a smaller inforce, and "reduction" is used to refer to a smaller scenario set. These may not translate directly into runtime savings, but should be closely correlated.

A. Transfer Scenario Order

Transfer Scenario Order is a simple technique that companies use to avoid running unnecessary scenarios. For example, a company may be producing a CTE70 measure using 1,000 scenarios. In this situation, the actuary only wants results from the worst 300 scenarios, but still has to run all 1,000 in order to determine which scenarios are the worst 300.

The core of this technique is to find a way to determine which scenarios are the worst 300 without having to run all the policies through them. This is done by first running a small, random subset of policies through the full 1,000 scenarios to determine scenario order. Once the scenarios are ordered based on the results from the subset of policies, the full inforce likely has a similar ordering and can therefore be run only through the worst 300 scenarios.

The actuary will still want to understand

- How many policies should be in the subset used to determine scenario order, and
- How to buffer against the impact of differences in scenario order from that subset of policies to the full inforce.

The following is an example of the process that an actuary might follow in order to perform the technique:

1. Select a random 5 percent of policies.
2. Run these policies through the full set of 1,000 scenarios.
3. Order the results, and identify the worst 400 scenarios.
4. Run the full set of policies through the 400 scenarios identified.
5. Order the results and calculate the final CTE70 based on the worst 300 out of the 400 scenarios that were run (this can be viewed as a CTE25 of the new scenario set).

This sample case includes the worst 400 scenarios on the second pass instead of the worst 300 in order to buffer against differences in scenario order. For example, suppose that the 295th scenario in the full set landed as scenario 305 for the subset – if only the worst 300 scenarios from the subset are taken, this would not be caught. However, because 400 scenarios are used in this example, it makes it much less likely that this sort of thing will occur. The testing results that follow will demonstrate whether or not this was effective. In practice, the actuary may want to demonstrate periodically that the buffer is adequate to produce required accuracy.

In this sample case, the technique should produce approximately 55 percent reduction to total runtime. Hypothetically, if a computer takes 0.01 seconds to run a policy through a scenario, and there are 50,000 policies and 1,000 scenarios, the calculation is as follows:

	Policies	Scenarios	Runtime
Original	50,000	1,000	5.8 days
New Run 1	2,500	1,000	0.3 days
New Run 2	50,000	400	2.3 days

Table II.A.1. Hypothetical runtimes

$\frac{0.3+2.3}{5.8} - 1 =$ -55% impact to runtime

Testing Results

Three companies tested this technique during this study:

Company	Product(s)	Metric	Inforce	Scenarios
A	VA/GMxB	CTE70	500,000 policies	1,000 AAA interest rate scenarios and four internally generated indices
B	VA/GMIB /GMDB	CTE70	10,000 policies	10,000 AAA interest rate scenarios and one equity index
C	VA/GMWB /GMDB	CTE70	Test 1: 20,000 cells Test 2: 5,000 cells	1,000 AAA interest rate scenarios and six indices

Table II.A.2. Testing background

The primary measure used to gauge success is the percentage difference in the **greatest present value of accumulated deficiencies (GPVAD)**[1] using CTE70 (referred to as the "error"), as defined under AG43 statutory reporting requirements for variable annuities. A secondary measure that provides additional perspective is the number of scenarios from the worst 30 percent of the full run that are successfully captured in the reduced run.

The following table shows the average results obtained for each of the four blocks of business included in the study. Note that this aggregates several runs performed by each company (discussed further below), and the runtime reduction is an average of the different levels of compression utilized for all the blocks:

	Error	Scenarios Captured	Reduction
Company A	0.00%	100%	51%
Company B	0.00%	100%	51%
Company C, Block 1	-0.43%	91%	51%
Company C, Block 2	-3.80%	65%	51%

Table II.A.3. Average results by block

The Scenarios Captured column was only relevant for this technique as it directly explains part of the reason for differences in accuracy. Scenarios Captured refers to the number of scenarios from the worst 30 percent of the original unreduced run that were successfully identified using Run 1, and therefore used in Run 2.

Several observations are immediately apparent from this set of numbers above:

1. The technique provided results which had a relatively low GPVAD error for two of the participating companies, but greater error percentages for the third company. The average reduction tested was roughly 50 percent, and the error was within 0 to -4 percent.
2. Results vary depending on the block of business. Company A noted that only 10 percent of the scenarios produced a positive GPVAD, which means that considerable shifting of scenarios may occur without impacting the results at all. Company B's inforce includes a more mature and homogeneous set of policies, and it is expected that the more homogeneous the inforce, the more powerful this technique will be. The two blocks of business at Company C both have a mix of riders, but the company was unable to determine during the study what was driving the disparity in results between the two blocks.

[1] This is calculated for AG43 (VACARVM) and C3 Phase II. At a high level, the calculation is as follows: (i) The accumulated deficiency is determined for each year of the projection (liability less assets), (ii) discounted to time zero, (iii) the largest result over all the projection periods is selected. A GPVAD is determined for each scenario, and then the CTE70 or CTE90 of the GPVADs is used for the two metrics, respectively. Some companies that participated in the study floored results at zero for each scenario.

3. For Company C, Block 2, we found it interesting that despite only capturing 65 percent of the worst 300 scenarios in the baseline run, the results are only off by -3.8 percent. A fair amount of swapping must take place in the scenario order, but the distribution must be relatively flat in order to produce results similar to the baseline run.

As the results for Companies A and B matched the baseline runs perfectly, the remainder of the analysis will focus only on the sets of results from Company C.

We broke down the results further in order to determine how the compression correlates with accuracy. Let's first analyze the sensitivity to the number of policies used in the initial run:

	GPVAD Error	Scenarios Captured	Reduction
Every 100th Policy	-2.58%	76%	60%
Every 10th Policy	-1.93%	78%	51%
Every 5th Policy	-1.78%	79%	41%

Table II.A.4. Sensitivity to the number of policies

There is measured variation in accuracy coming from the number of policies used in the initial run. The number of scenarios captured does not appear to be a key driver of accuracy here.

Next, we analyze the sensitivity of the results to the number of scenarios used in the final run:

Number of Scenarios	GPVAD Error	Scenarios Captured	Reduction
30%	-2.90%	71%	60%
35%	-2.39%	76%	55%
40%	-1.78%	80%	50%
50%	-1.34%	86%	40%

Table II.A.5. Sensitivity to the number of scenarios

Note that in the above table, the "30%" in the first column corresponds to 300 scenarios out of a full set of 1,000.

The variation is greater here. Again, the tests remain in the 40 to 60 percent range of reduction in runtime, but the range of errors is wider, indicating one may achieve higher runtime reduction through running fewer policies on the initial run.

The major driver of increases in accuracy when adjusting the number of scenarios in the final run is through capturing more of the targeted scenarios (between 71 and 86 percent).

Here is a comparison of the strongest and weakest cases that were tested:

	GPVAD Error	Scenarios Captured	Reduction
Every 100, 30%	-3.48%	69%	69%
Every 5, 50%	-1.04%	87%	30%

Table II.A.6. Strongest and weakest cases tested

The full range of GPVAD differences that were experienced after aggregating across the two blocks for Company C is -1.04 to -3.48 percent.

Considerations

Positive:

A. The technique is easy to understand, easy to apply, and easy to audit. It was largely because of this that we chose to test this technique for the study, despite less potential for runtime reduction than the other techniques.

B. Participating companies estimated approximately 10 hours of ongoing effort each reporting period to perform the additional steps required for this technique.

C. This technique can be layered over existing traditional inforce compression techniques. This statement will generally apply to all scenario reduction techniques, but it is worth mentioning here because most companies today are already using traditional inforce compression techniques for some purposes, and having a scenario reduction technique that can be applied to the reduced inforce might be an easy first step for them.

D. Alternatively, an advanced cell compression algorithm can be used to select the small subset of policies for the initial run, which will likely improve performance of the technique.

E. Unlike most scenario reduction techniques, this technique should theoretically be *more* powerful under a CTE90 metric than a CTE70 metric. Since the worst 10 percent of scenarios are a subset of the worst 30 percent of scenarios, it follows mathematically that an equivalent number of scenarios will capture a higher proportion of the worst 10 percent than the worst 30 percent, and it follows conversely that a higher degree of reduction can be obtained at CTE90 with a similar degree of accuracy.

Negative:

A. This technique is clearly limited to tail metrics. Testing was based on AG43, which uses a CTE70 metric. The technique will similarly apply to other CTE and VaR-type metrics, but has no application for metrics that rely on the expected value of a distribution.

B. It is impossible to improve the accuracy of a result by altering the scenario order because, for testing purposes, "perfect" accuracy is defined as the CTE70 of the baseline run. Unfortunately, any errors introduced by altering the scenario order of a left-tail reserve metric must, by definition, reduce the reserve. Knowing this, companies will have to determine whether it is acceptable to apply this technique to a statutory reserve, as it may be preferable to see a positive bias for conservatism. Many companies may determine that as a result of the negative bias, the technique is only suitable for sensitivity testing. Others will find that they can apply this with a high enough degree of accuracy that they consider the error to be immaterial, regardless of the direction.

C. One company noted that it would be difficult for it to use this in practice, as it introduces additional steps to the valuation process, requiring multiple runs in the primary modeling platform in order to obtain a single result.

Areas for Future Research

This study primarily tested the first variation of the technique; that is, extrapolating from a subset of policies to the full inforce.

This technique can also be used to transfer scenarios by using the full inforce for a baseline run and applying the order to various sensitivities. One area for further research into effectiveness for sensitivities will be whether the error introduced is reliably consistent between the baseline run and the sensitivities, in which case the baseline error can be used as a **control variate**[2] to improve accuracy in the sensitivities.

As noted earlier, this technique should be more effective under metrics that are farther out in the tail. However, this has not been tested, and it bears further research to determine how much more effective the technique will be.

[2] A variance reduction technique that exploits error introduced into a known quantity (the Control Variate) under Monte Carlo simulation in order to reduce the error in the estimate of an unknown quantity (e.g., a liability model) in the same simulation.

B. Representative Scenarios

Representative Scenarios is a class of techniques that select a subset of scenarios that are "representative" of the full scenario set based on certain characteristics of the scenarios.

Four variations of this technique were tested by the participating companies. See the bibliography for details of the original publications describing these techniques. The following descriptions refer to interest rates, and these have been extended to incorporate index/fund returns as well where appropriate.

1. **Modified Euclidean Distance Method** — The premise behind this technique is that each scenario can be plotted in an n-dimensional space, where:
   i. n = the number of years that are used
   ii. The value of each scenario in dimension_t is the interest rate in year_t

   The actuary will apply the Euclidean distance formula to locate a subset of scenarios that are "far" from each other in the n-dimensional space. In order to allocate a higher impact to the interest rates in the early years of the projection, a weight is added to each dimension of the Euclidean distance formula equal to the present value of a dollar at time *t*.

   The final formula for the distance between scenarios A and B is as follows:

   $$\sqrt{\sum_1^n \left(i_{A,t} - i_{B,t}\right)^2 \times PV01_{A,t}}$$

   Each scenario will be weighted based on the number of scenarios mapped into each representative scenario.

   For more information on this technique, see Chueh (2002), page 92. Where applicable, this technique was extended to include equity/bond index returns following the methodology described in Longley-Cook (2003).

2. **Relative Present Value Distance Method** — This is very similar to the prior method, with the exception that the value of each scenario is the present value of a dollar ("PV01") each year, rather than the interest rate itself. Weights are not required for this technique within the distance formula. The early years of the projection inherently have a higher impact on the results than the later years, as the interest rates in the early years influence the PV01 each year in the early years and the later years, whereas the opposite is not the case.

   The final formula for the distance between scenarios A and B is as follows:

   $$\sqrt{\sum_1^n \left(PV01_{A,t} - PV01_{B,t}\right)^2}$$

For more information on this technique, see Chueh (2002), page 92. Where applicable, this technique was extended to include equity/bond index returns following the methodology described in Longley-Cook (2003).

3. **Significance Method** — This technique also utilizes the PV01 each year. However, instead of calculating the difference between each scenario, the PV01 each year is calculated, and the "significance" of the scenario is set equal to:

$$\sqrt{\sum PV01_t{}^2}$$

This can also be interpreted as the distance of the scenario from zero.

Once the significance of each scenario is obtained, the scenarios can then be ordered. The scenarios with lower significance correspond to the left tail, and the scenarios with higher significance correspond to the right tail. The actuary can select every n^{th} scenario from the entire distribution, thereby obtaining a good distribution even with a smaller scenario set.

The Significance Method is the basis of the Scenario Picker Tool published by the American Academy of Actuaries for use in selected C3 Phase II scenarios. However, the Scenario Picker Tool is limited to selection based on interest rates only at the time this report is being written.

For more information on this technique, see Chueh (2002), pages 92–93. Where applicable, this technique was extended to include equity/bond index returns following the methodology described in Longley-Cook (2003).

4. **Scenario Cluster Modeling** — This is a variation of Representative Scenarios that is included in MG-ALFA®. The algorithm is related to Modified Euclidean Distance Method as it is based on distance between scenarios, keyed on selected interest rates and equity/bond index returns. As with the Modified Euclidean Distance Method, this technique results in unequally weighted scenarios, based on the number of scenarios mapped into each representative scenario. There is one key difference—the Modified Euclidean Distance Method picks scenarios that are farthest from all others, whereas Cluster Modeling groups together similar scenarios and picks the center of each cluster.

Testing Results

Five companies tested this technique during this study:

Company	Product(s)	Metric	Inforce	Scenarios	Reduction Method
A	VA/GMIB /GMDB	CTE70	10,000 policies	10,000 AAA interest rate scenarios and one equity index	1,2,3
B	VA/GMWB /GMDB	CTE70	A: 20,000 cells B: 5,000 cells	1,000 AAA interest rate scenarios and six indices	4
C	EIA, Indexed UL/NLG	Mean	100,000 policies	1,000 RN internally generated with interest rates and one equity index	3
D	VA/GMIB /GMWB	Mean	200,000 cells	6,000 internally generated with interest rates and eight indices	3
E	VA/GMIB /GMDB	Mean	14,000 cells	1,000 RN internally generated with interest rates and six indices	3

Table II.B.1. Testing background

Based on the characteristics of the testing detailed above, it is difficult to summarize how well the test performed overall. Not only are the environmental characteristics unique for each company, but four variations of the Representative Scenarios algorithm were used, and the techniques used to measure accuracy varied as well.

Three companies tested the Significance Method only. One company tested each of the first three methods (Modified Euclidean Distance Method, Relative Present Value Distance Method and Significance Method), and one company tested results under Scenario Cluster Modeling only.

The "error," as it will be used below, refers to the absolute percentage difference between the reduced run and the original run for the metric in question. The metrics included in aggregated results are all based on Mean or CTE70. Note that since errors are both positive and negative, we converted to absolute form prior to aggregating results from each test to prevent them from averaging to near-zero.

The following table shows the overall performance of the technique for each block of business tested:

	Metric	Product	Variation	Error	Reduction
Block A	Mean	Indexed UL, NLG	Significance	0.9%	58%
Block B	CTE70	VA	Cluster Modeling	1.1%	79%
Block C	Mean	Indexed UL	Significance	1.3%	58%
Block D	Mean	EIA	Significance	2.0%	58%
Block E	Mean	VA	Significance	2.4%	91%
Block F	Mean	VA	Significance	3.3%	85%
Block G	Mean	VA	Significance	5.3%	93%
Block H	CTE70	VA	Cluster Modeling	10.7%	79%
Block I	CTE70	VA	Various (1,2,3)	11.8%	73%

Table II.B.2. Average results by block

Several observations are immediately apparent from these results:

1. The range of errors was large.
2. Two companies noted that this technique is far more successful at capturing the mean than it is at capturing the tail. We observed that the weakest two results captured were based on CTE70 metrics.
3. It is difficult to determine the strength of the relationship between the level of compression and the error from these aggregated results.

For each of the runs performed by the companies, the compression level and the error are plotted below:

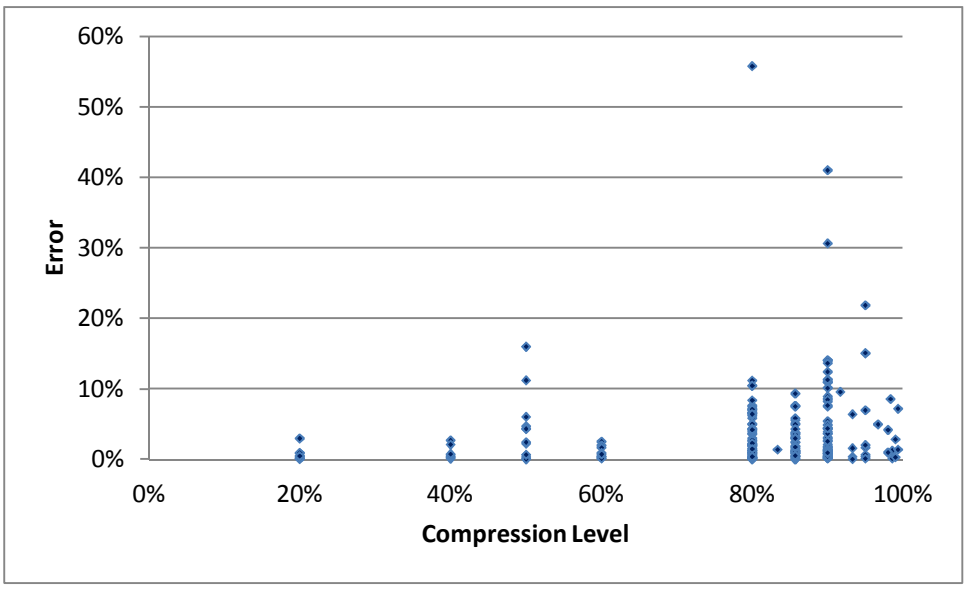

Figure II.B.3. Compression vs. error

This graph provides good insight into the range of errors at different compression levels. We can clearly see a positive correlation between compression and error, as expected, but there are instances with errors of up to 16 percent at 50 percent compression, and instances where errors were less than 5 percent at compression levels of more than 90 percent. Note that the extreme errors seen in this graph (>20 percent) are nearly all from CTE70 runs.

One company provided us with results for 33 risk neutral simulations, each based on a unique starting account value (from different real-world outer loop scenarios), and using different random **seeds**[3]. This provides insight into the range of errors that can occur at different compression levels. The following results include the actual sign on the error terms, as no aggregation takes place:

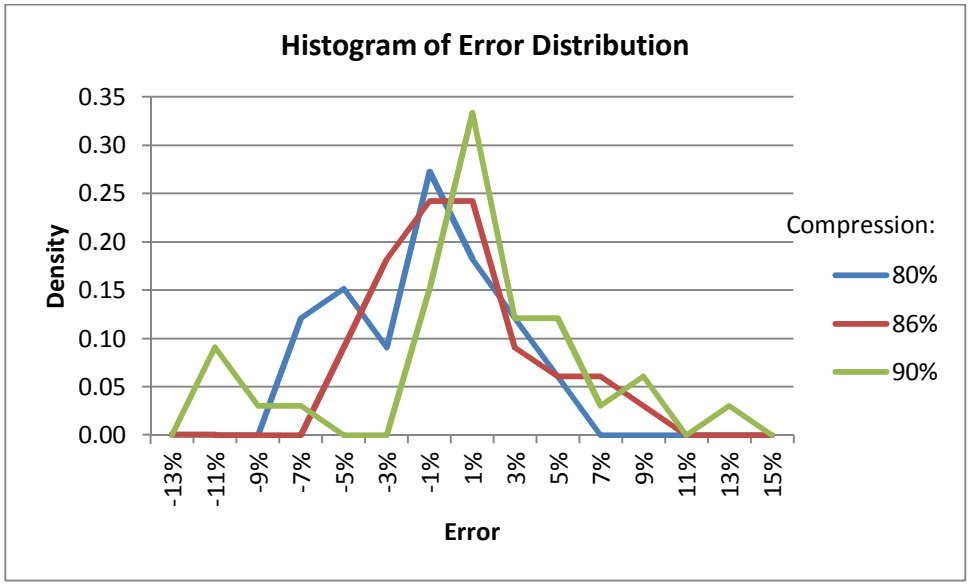

Figure II.B.4. Distribution of errors

- Even with 33 results at each compression level, it is still remarkably difficult to determine a clear relationship between the compression level and the error.
- The widest range of errors appears to be at the 90 percent compression level.
- The majority of results at the 80 percent and 86 percent compression levels have errors of +/-5 percent.
- There does not appear to be any visible bias in the results.

This graph also highlights something that is lacking in many of the test results produced by the participating companies. Most companies did not have the resources or capability to provide a series of estimates based on different random seeds. Suppose that a company

[3] A set of random numbers always begins with a seed number. This seed number can generally be entered manually or derived from the computer's internal clock. In repeating a simulation, one would begin with a different seed in order to obtain a different set of scenarios.

provides a single result showing an error of 2 percent. As illustrated on the graph above, it is entirely possible to produce an error of 2 percent without giving any indication whatsoever of the range of possible errors. This is why we have drawn upon the aggregate testing results across all of the companies wherever possible.

We investigated further to determine why the results are so poor for the tail risk metrics. There are two primary reasons why we felt this could occur:

1. A bias in the distribution — possibly underweighting or overweighting both tails
2. Increased statistical error in the tails, as results are driven by a smaller set of scenarios.

One company provided us with the present value of the liability in each scenario for the full run as well as various compressed runs using the Significance Method. The following results show these results, converted into a probability density function:

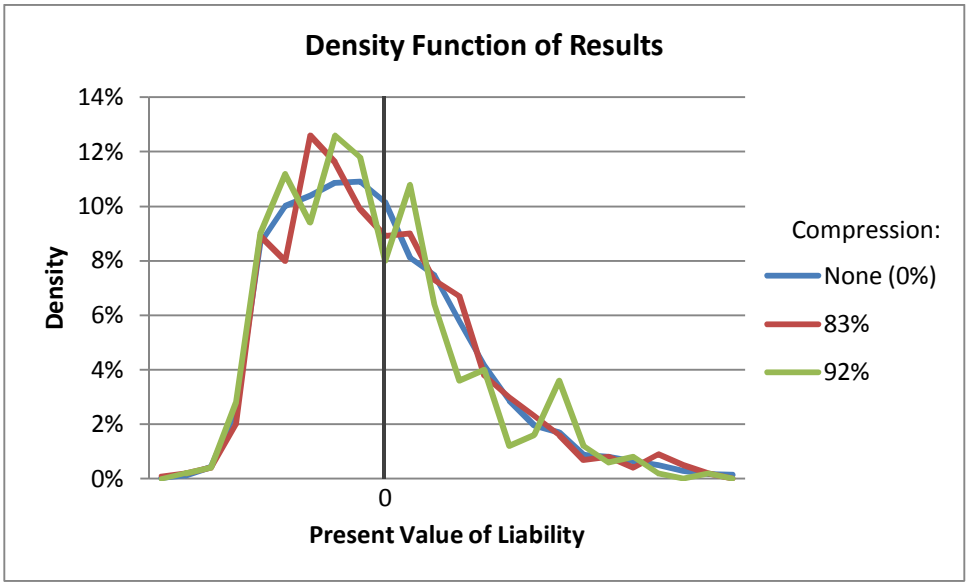

Figure II.B.5. Density of results for one company

While these only reflect the results for one company, we can tentatively make certain observations:

1. The Representative Scenarios technique does not appear to have introduced visible bias or altered the overall density function. Note that the company tested this using the Kolmogorov-Smirnov test and the two-sample Anderson-Darling test, and based on both tests it was unable to reject the null hypothesis that the compressed distributions are drawn from the same distribution as the full set.
2. The compressed runs have a far more jittery appearance, which translates into a less stable result for any one part of the distribution. For this reason, and due to the lack of

visible bias, it is likely that the increased statistical error in the tails is driving the poor performance for the CTE70 results.

One company tested the same block of business under CTE70 and CTE95 using the first three techniques described above. The following results show aggregate performance under each of the techniques:

	Error
Modified Euclidean Distance Method	19.2%
Relative Present Value Distance Method	5.6%
Significance Method	3.2%

Table II.B.6. Comparison of techniques for one company

The Significance Method produced the highest accuracy. There is not sufficient data from the study to determine whether the differences in accuracy are a result of bias or increased statistical error.

We also compared average error under CTE70 versus CTE95 from the same set of data:

	Error
CTE70	11.8%
CTE95	10.3%

Table II.B.7. Comparison of CTE levels under Significance Method

Results are comparable at the two CTE levels. It is unlikely that the difference is statistically significant. (Note that other companies did observe reduced precision farther out in the tails.)

The other company that provided results on a CTE basis used the Cluster Modeling algorithm in MG-ALFA. This company provided results for two blocks of business at CTE70 and CTE90:

	CTE70	CTE90
Block B	1.1%	8.0%
Block H	10.7%	41.7%

Table II.B.8. Comparison of CTE levels for two blocks under Cluster Modeling

As there are no other data points available for Cluster Modeling, it is not possible to determine whether the higher aggregate error was the result of using the tail metrics or the algorithm used. The fact that the CTE90 performs more poorly lends some credence to the

former. There is also insufficient information to determine the basis for the differences between the two blocks.

Limitations in the Definition of Error

Of the six techniques included in the study, Representative Scenarios was the most difficult one to organize and analyze. As noted earlier, companies tested a variety of algorithms under a variety of circumstances; they were generally constrained in the type of output and the volume of runs they were able to perform; and they were not working under laboratory conditions. This should be considered when interpreting the results.

Additionally, we would like to point out a key issue that we had to work around when organizing the data.

All error terms in the study were described as a percentage. Generally, this is equal to $\frac{B}{A} - 1$, where:

A is the result of the baseline run.

B is the result of the compressed run.

Unfortunately, this approach to calculating the error does not always work. Suppose that we have a set of results as follows:

Scenario	Base Result	Compressed Result
1	-3	-2
2	-2	-1
3	-1	0
4	0	1
5	1	2
6	2	3
7	3	4
Mean	**0**	**1**

Table II.B.9. Hypothetical error calculation

The formula $\frac{B}{A} - 1$ will result in an infinite error when measuring the mean. This is an extreme case designed to illustrate the difficulty with measurement of the error. Unfortunately, this was a reality for one of the companies in the study. Some of the results provided showed a liability close to zero, as the moneyness of the inforce was set in some cases at a point where the present value of fees was approximately equal to the present value of claims.

In order to calculate a percentage error, first consider that $\frac{B}{A} - 1$ is equal to $\frac{(B-A)}{A}$, and then simply find a suitable denominator to switch with A. We chose the standard deviation of the distribution. This was convenient because: (i) it is representative of the magnitude of the results; (ii) it is nearly constant across different seeds; and (iii) this error term now has an appealing explanation: it represents the difference between the two results as the number of standard deviations between the two sets (e.g., the compressed run produced a result that is one standard deviation away from the full run). Note that this does create an inconsistency in the measurement of error between the blocks of business, and it was not possible at the time to revise the definition of error for all participating companies. However, we believe that the information presented is still useful and relevant.

There is no possibility of this issue arising under the GPVAD design, as results in each scenario are floored at zero. However, this does beg the question, how should a company address the error? Will they be forced to use an arbitrary component in the error term?

We think that companies will generally have the cleanest analysis if errors are analyzed in absolute dollars. Rather than accepting/rejecting a technique best on a percentage difference, the actuary should go into the testing with a clear idea of what dollar error is material to the company. If the tolerance is +/- $10 million, there will be no question whether the actuary stayed within the target for a given test. We were unfortunately unable to do that for this study, as there was not a way to provide comparability across the results.

Considerations

Positive:

A. The Significance Method produced lower errors than the other three variations of the technique, with errors of 0.9 to 5.3 percent versus errors of 1.1 to 11.8 percent. However, the scope of the testing was very limited, and we encourage companies to test the other three variations further, as they may very well find that they are more effective for certain products and under different modeling arrangements.

B. The Significance Method is easy to understand and easy to use. All of the participating companies estimated less than 10 hours of ongoing effort to maintain this process.

C. In the full distribution of results provided by one company using the Significance Method, there was no bias observed.

D. This technique can be applied using a wide range of compression levels based on the need for accuracy. Higher compression can potentially be used for sensitivity testing.

E. This technique does not require the company to generate additional scenarios.

Negative:

A. Two companies noted that precision was reduced in the tails of the distribution.

B. Unfortunately, it is often the case that there are barely enough scenarios to begin with to capture high quality results in the tail. Any reduction technique like Representative Scenarios, no matter how cleverly implemented, will take away from the number of scenarios remaining in the tail, and can potentially reduce accuracy below acceptable thresholds for tail metrics.

C. Like other techniques, a considerable amount of testing is recommended before a company can implement Representative Scenarios. It would be important to test multiple samples, and it would be prudent to determine for which blocks of business and for which metrics/applications the method will produce satisfactory results.

Areas for Future Research

A. This study primarily tested the Significance Method. We would encourage companies to explore the other techniques further to determine their effectiveness, and understand where they work best.

B. Explore further regarding use for tail metrics.

C. Importance Sampling

Importance Sampling is a class of techniques that sample more scenarios in parts of the distribution that are more critical to the overall result, and subject those scenarios to reduced weights to produce a final result with no bias.

The variation of Importance Sampling tested in this study is based on the Significance Method variation of Representative Scenarios. Under the Significance Method, the scenarios are ordered from worst to best based on the interest rates and/or accumulation factors in the scenarios. Every n^{th} scenario is selected after ordering, providing a way to collect a smooth distribution of scenarios.

The Importance Sampling extension is to select more heavily from the tail of the distribution, which is generally where more liability cash flows are produced for products with interest rate or accumulation guarantees. The following table provides an example comparing the original selection process to the weighted selection process:

Scenario	Original Weight		Revised Weight	
1		Set 1	4.8%	Group 1
2	14.3%		4.8%	
3			4.8%	
4				Group 2
5	14.3%		14.3%	
6				
7				
8	14.3%		14.3%	
9		.		
10				
11	14.3%	.	14.3%	
12		.		
13				Group 3
14	14.3%			
15				
16				
17	14.3%		42.9%	
18				
19		Set 7		
20	14.3%			
21				
Total	100.0%		100.0%	

Table II.C.1. Hypothetical Importance Sampling example

- Each set of three scenarios in the column of original weights is represented by one of the scenarios (the central scenario).
- The original weighting scheme samples smoothly throughout the distribution and weights the representative scenarios equally, with a total of 100 percent.
- In the revised weights, we set up three groups. The first group is sampled heavily and receives low weights that total to the original weight. The second group is unaltered. The third group is sampled sparsely, and receives high weights that total to the original weight.

Importance Sampling does not specifically improve runtime. Instead, it may improve accuracy, indirectly enabling the actuary to use fewer scenarios to obtain the same level of accuracy.

One other important note about this technique is that it has the ability to increase or decrease accuracy. For example, if the actuary chooses to sample more from the right tail instead of the left tail, and the guarantees are generating more cash flows in the left tail, we would expect accuracy to go down. For this reason, Importance Sampling in general can never be used without careful support for the adjusted weights.

Testing Results

Two companies tested this technique during this study:

Company	Product(s)	Metric	Inforce	Scenarios
A	VA/GMIB /GMDB	Mean	14,000 cells	1,000 RN internally generated with interest rates and six indices
B	VA/GMIB /GMDB	Various CTEs	10,000 policies	1,000 AAA interest rate scenarios and one equity index

Table II.C.2. Testing background

In order to perform testing, we took a scattershot approach and asked participating companies to perform the following series of tests:

1. 20-80 to 50-50 (this is shorthand for: lowest 20 percent of distribution is allocated 50 percent of scenarios, highest 80 percent of distribution is allocated 50 percent of scenarios)
2. 10-20-70 to 40-40-20
3. 30-70 to 50-50.

The third variation is the least aggressive combination, and the others are both fairly aggressive.

Companies tested each of these using 80 percent reduction and 90 percent reduction.

Unless otherwise specified, the errors described in this section refer to absolute value of the difference between original results and compressed results.

One company performed the testing using a large number of full simulations, and compared the results to the unweighted Significance Method, as well as a random selection:

Weighting Scheme	Error
20-80 to 50-50	4.8%
10-20-70 to 40-40-20	6.7%
30-70 to 50-50	3.2%
Unweighted	3.5%
Random	4.7%

Table II.C.3. Company A results

The other company only tested a single seed, and compared results to the unweighted Significance Method:

Weighting Scheme	Error
20-80 to 50-50	4.6%
10-20-70 to 40-40-20	7.7%
30-70 to 50-50	3.8%
Unweighted	0.8%

Table II.C.4. Company B results

The following observations can be made from these results:

1. The original weighting scheme is an improvement over random selection (3.5 percent versus 4.7 percent) as expected.
2. In both cases, the 30-70 to 50-50 weighting scheme performs better than the other two variations. This indicates that cash flows in these blocks of business do not have extreme levels of asymmetry.
3. The improvement in the first company under 30-70 to 50-50 is not very large (3.2 percent versus 3.5 percent).
4. Even under 30-70 to 50-50, the second company drops in accuracy (3.8 percent versus 0.8 percent).

The second company may have lost accuracy because its block of business is mostly in-the-money, indicating that liability cash flows occur in both up and down scenarios. This technique specifically works well when the cash flows are only significant in one tail, which would occur with an out-of-the-money option.

The following table shows average results at different levels of scenario reduction:

Reduction	Error
80%	2.7%
90%	4.3%

Table II.C.5. Reduction versus error

As expected, the errors increase as the level of reduction is increased (4.3 percent versus 2.7 percent).

Considerations

The testing above can be described as a first attempt at the technique. Importance Sampling has been used much more extensively in other industries, and there are suitable applications for which the technique has been known to perform extremely well. One area where it has been used successfully is in the pricing of deep out-of-the-money options. Under circumstances where only 1 out of every 10,000 scenarios produces a payoff, this technique can reasonably reduce the number of scenarios required by 99 percent.

This technique only generated a 0.3 percent improvement to accuracy compared with the unweighted Significance Method, and in many cases decreased accuracy due to weighting schemes that were not appropriate for the business tested.

We would expect this technique to provide strong results only when the following two conditions hold true:

1. The distribution of scenario weights is designed to match the cash flow pattern of the inforce.
2. The method used to rank the scenarios is well-defined and closely correlated with the drivers of the cash flows.

Areas for Future Research

This study introduces a single variation of Importance Sampling which is relatively simple to use. More testing is required to determine at what levels of out-of-the-moneyness this technique begins to significantly improve accuracy and reduce runtime.

Much more powerful (but complex) variations of Importance Sampling are used outside of the actuarial profession. One example is as follows:

1. Instead of converting random numbers to the underlying distribution, convert them to a different distribution that will produce more scenarios with lower returns.
2. Multiply the result for each scenario by A/B, where A=the density of the scenario under the original distribution, and B=the density of the scenario under the altered distribution. This acts as a set of weights that assigns a lower weight to the parts of the distribution that have been oversampled.

The differences with the technique tested in this study are as follows:

1. Requires developer access to the scenario generator
2. Produces an entirely different set of scenarios, rather than selecting a subset
3. Allows a great deal of flexibility in the choice of altered distributions.

D. Curve Fitting

The results of a model projection across a large enough number of scenarios will generally produce a distribution that closely follows an underlying continuous distribution. If the actuary is ultimately looking to determine the mean of the distribution, there is little to gain by attempting to determine the underlying distribution. For tail metrics, however, it is possible to use the underlying distribution to improve accuracy.

If we can improve accuracy, this in turn allows us to reduce the number of scenarios required to produce a given level of accuracy.

The following is an example of how this may be used to calculate a CTE70 for a set of cash flows that follow a Normal distribution: $f(x) = N(\mu, \sigma)$:

1. Apply curve fitting to determine μ and σ (described in more detail below).
2. Determine X such that $P(x < X) = 0.7$ (Microsoft Excel's NORMINV() function can do this easily).
3. CTE70 = $\int_{.7}^{1} x \cdot f(x)$, which can be approximated as $\sum_{.7}^{1} x \cdot f(x)$.

If the data has a good fit to the statistical distribution, this can produce a result with increased accuracy and no bias. The purpose of the testing was to determine how feasible it is to achieve that goal.

There are three primary areas where some effort is required to set up a Curve Fitting algorithm:

1. Selection of appropriate statistical distribution, or appropriate mix of distributions. For example, the actuary may know that the distribution of returns is approximately Normal, in which case the remaining exercise would be to find an appropriate mean and variance.
2. Selection of an underlying optimization algorithm to fit the curve to the data. Optimization is discussed in more detail under the Replicating Liabilities inforce compression section.
3. Setting an appropriate objective function and constraints for the optimization algorithm. The objective function should represent the overall difference between the curve and the data, and optimization is used to minimize this amount.

Testing was initially based on the approach taken in the AMOOF tool produced by Dr. Yvonne Chueh of Central Washington University, as this tool is specifically designed to fit a simulation of insurance contract liabilities:

1. It attempts to parameterize 22 different statistical distributions, including combinations of distributions.
2. Optimization is performed using Microsoft Excel's Solver tool.

3. The objective function in this tool matches the first four moments of the distribution.

Testing Results

One company performed preliminary testing:

- Variable annuities, GMxB
- 900,000 policies, run seriatim
- 1,000 AAA scenarios, including interest rates and five equity indices
- Metric: AG43 stochastic reserve (CTE70).

The following graph shows the density for functions where the fit was relatively strong:

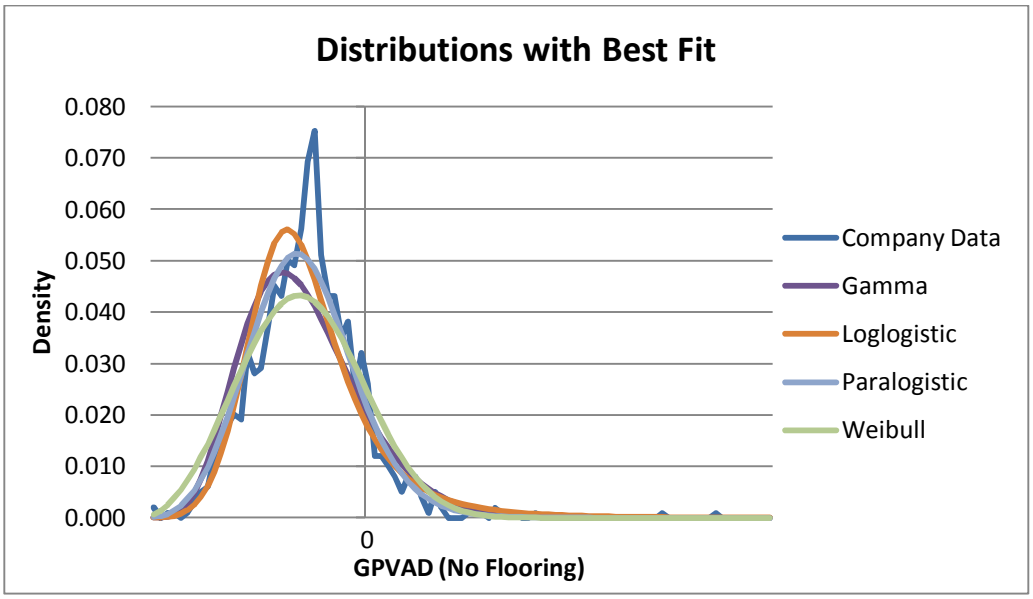

Figure II.D.1. Distributions with best fit

For contrast, the following graph shows the density for functions that did not fit very well:

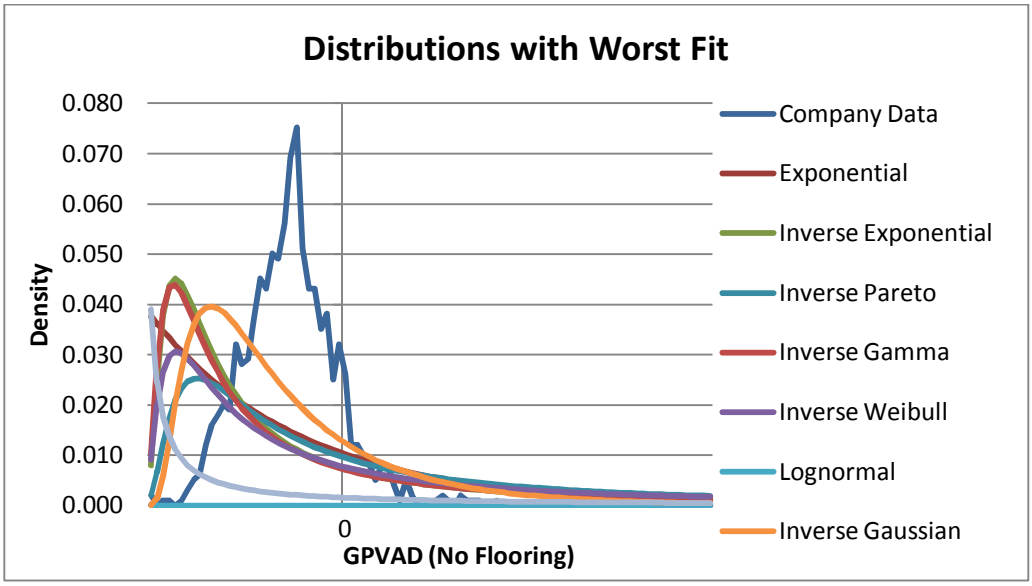

Figure II.D.2. Distributions with worst fit

The goal for this exercise was to obtain a distribution with a good fit in the tail. Returning to the distributions with the best fit, the following graph zooms in on the left tail of the distribution:

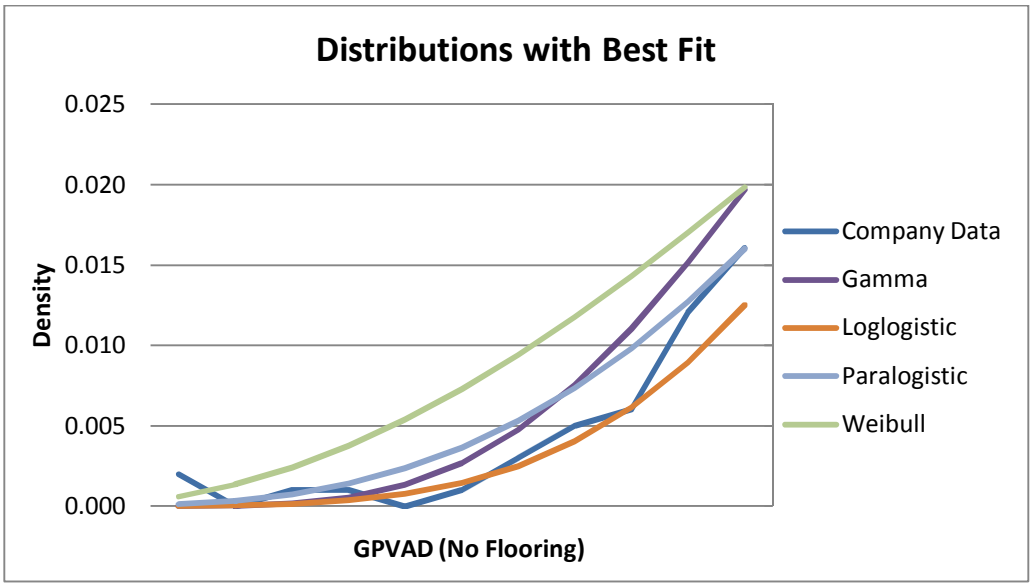

Figure II.D.3. Distributions with best fit, tail results

This shows more clearly that the Loglogistic distribution is most successful at passing through the points of the underlying data in the left tail.

The following graph shows the results for the Loglogistic curve only, covering the worst 30 percent of the distribution:

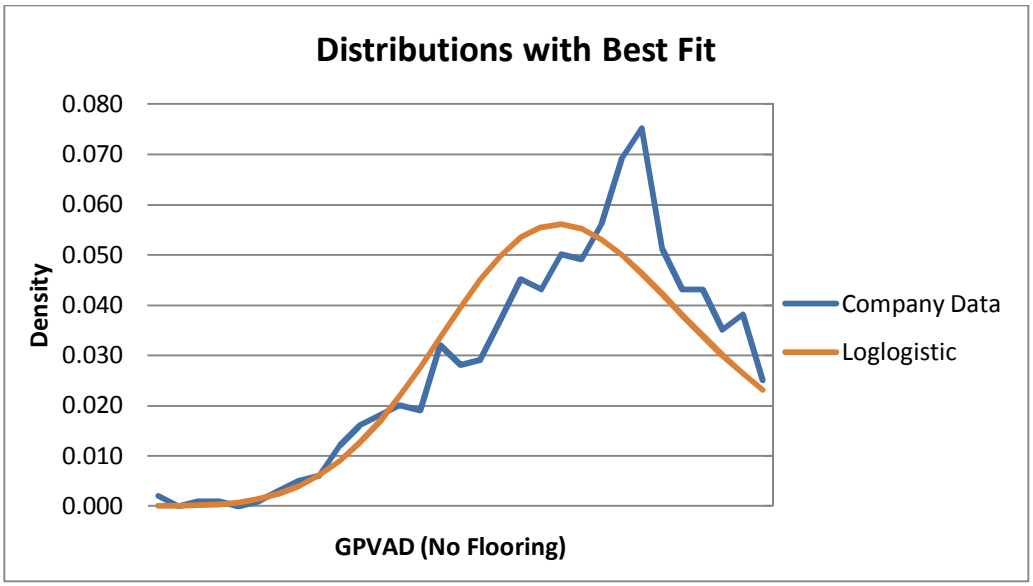

Figure II.D.4. Loglogistic distribution results

The following table compares the CTE70 and CTE95 of the data used for fitting with the modeled result from the Loglogistic distribution:

	CTE70	CTE95
Company Data	$19.9	$26.7
Loglogistic Result	$25.9	$29.6
Difference	30%	11%

Table II.D.5. Results under various CTE levels

The primary objective of this technique is to reduce or eliminate the impact of sampling error on results generated from a finite number of scenarios. As a result, it is difficult to evaluate the effectiveness of the technique by comparing its results against a sample that includes sampling error. With that caveat, observations on the initial set of results are:

1. There was no single function that was very successful at matching a wide area of the distribution. The Loglogistic has a reasonable fit overall, but even it does not match all that well.
2. As a result of the weak fit overall, the Loglogistic CTE70 results are 30 percent higher than the initial sample.
3. The CTE95 is off by 11 percent. This is a little more difficult to interpret, because visually the results appear quite strong, and it is entirely possible that the curve fitting results are more accurate than the original result based on 1,000 scenarios. Unfortunately, the company performing the testing was unable to provide results for a

larger set of scenarios to test whether the difference is really caused by error in the set of 1,000.

We noted earlier that one of the features of the AMOOF tool is that it determines the parameters for each distribution by matching the moments of the distribution. One weakness of this approach is that there is no way to target the left tail of the distribution, which certainly contributes to difficulty in matching the left tail.

As part of the study, tests were performed where the optimization parameters were swapped out and replaced with a fitting that minimizes the sum of the squared errors for the lowest 40 percent of the distribution:

	Error	
Scenarios	CTE70	CTE95
1,000	28%	11%
200 A	7%	-6%
200 B	6%	-6%
200 C	27%	12%
200 D	33%	13%
200 E	21%	6%
Average	20%	5%
Average (Absolute)	20%	9%

Table II.D.6. Results under various scenarios sets

In the table above, the left-hand column represents the number of scenarios used for fitting. The first test was performed using 1,000 scenarios, and the results are comparable to the prior test, which used the moment matching logic in AMOOF.

The 1,000 scenarios were then randomly split into five sets of 200, and the next five rows in the table represent fittings performed using those five sets of scenarios. The following observations can be made from these results:

1. The errors are still high, ranging from 6 to 33 percent for CTE70 and -6 to 13 percent for CTE95, despite fitting based on the left tail directly.
2. There is a bias in the results — approximately 20 percent for the CTE70 and 5 percent for CTE95.

3. In addition to the bias, there is a significant amount of variability around the 20 percent and 5 percent average errors, pointing to a more general lack of precision as well with reduced scenarios.
4. The CTE95 continues to have a lower error overall than CTE70 (9 percent versus 20 percent). This is reflective of the Loglogistic curve and the data, and will be unique for each combination of data/curve that is chosen.

Considerations

The testing above can be described as a first attempt at the technique. Curve Fitting in general has been used in the industry, and there are suitable applications for which the technique will perform well. This analysis does demonstrate that there will be a significant amount of effort and expertise required for the initial setup.

Areas for Future Research

There is a considerable amount of research that can be done with Curve Fitting. This study only tested the technique for use in AG43. Variations of the technique are used in the industry to fit nested stochastic cash flows, reducing the number of nested loops required to produce accurate results. Additionally, each of the three aspects of the technique discussed above would be interesting to explore:

- Selection of statistical distribution — Determine which distributions tend to perform well in modeling different metrics/products. Consider use of polynomial distributions instead of statistical distributions.
- Optimization algorithm — Study how sensitive the results are to the starting assumptions. Test different optimization algorithms. Consider use of least squares regression instead of optimization.
- Objective function — Research can be done to determine how best to match results for different metrics. Consider use of absolute errors instead of squared errors to minimize the impact of outliers.

E. Cluster Modeling

Cluster Modeling is a cell compression technique that produces a reduced subset of policies that will have similar characteristics to the full inforce.

Cluster Modeling is a technique more commonly used in other disciplines of science to group ("cluster") together points that are similar to each other. It was widely introduced into the actuarial community through inclusion in MG-ALFA, the actuarial projection software developed by Milliman. It is also currently under development in GGY Axis in order to perform similar compression functionality.

While companies do have the ability to code their own Cluster Modeling algorithm, to expedite the research our study relied on the implementation of this technique in MG-ALFA.

Cluster Modeling is more commonly referred to in other disciplines as Cluster Analysis. The algorithm used for testing is related to the k-Nearest Neighbor (KNN) algorithm, where k = 1.

The premise behind this technique is that each policy can be plotted in a n-dimensional space, where each of the n dimensions represents an aspect of the inforce record (e.g., age) or results from prior calibration scenarios (e.g., present value of cash flows in a scenario). The algorithm will apply the Euclidean distance formula to group small policies with nearby larger policies. Note that the dimensions are referred to in Milliman's literature as Location Variables.

The algorithm is iterative and will continue to group policies sequentially until it reaches a specified compression level. The remaining clusters should be representative of the larger group, and will generally be centered on a policy with a large account value or other size-related variable. Nevertheless, it is possible that some smaller policies are very distinct with respect to the specified attributes and will remain uncompressed, or compressed only with similar small policies. The policies nearest to the center of each cluster will then be grossed up so that critical values like total account value will be the same size as the original inforce.

For more details on Cluster Modeling, including the technique and more testing results, see referenced publications Freedman (2009) and Reynolds (2010).

Testing Results

Five companies tested this technique during the study:

Company	Product(s)	Metric	Inforce	Scenarios
A	VA/GMIB	Various CTEs	200,000 policies	12,600 internally generated with interest rates and six indices
B	VA/GMxB	Various CTEs	500,000 policies	1,000 C3P1 interest rate scenarios and six internally generated indices
C	FIA, DA, Indexed UL	Mean	600,000 policies	1,000 RN internally generated with interest rates and one equity index
D	VA/GMDB	CTE90	100,000 policies	1,000 AAA interest rate scenarios and five internally generated indices
E	VA/GMxB	CTE90	600,000 policies	1,000 AAA interest rate scenarios and eight internally generated indices

Table II.E.1. Testing background

Unless otherwise specified, the errors described in this section refer to absolute value of the difference between original results and compressed results.

The following table provides a summary of the average results from each metric tested by each company:

	Product	Metric	Error	Compression
Block A	VA	Various CTEs	0.5%	99.9%
Block B	NLG	Mean	1.0%	96.5%
Block C	FIA	Mean	1.1%	99.4%
Block D	VA	Various CTEs	2.3%	95.9%
Block E	VA	CTE90	4.0%	94.7%
Block F	VA	CTE90	4.3%	86.8%
Block G	DA	Mean	23.8%	99.4%

Table II.E.2. Average results by block

We can make several observations based on these results:

1. The errors introduced after compression varied from 0.5 to 4.3 percent at high compression levels, and there is one outlier with an error of 23.8 percent.
2. The company determined that the outlier of 23.8 percent occurred in a block of business with an average PV of profits close to zero, which causes percentage errors to be high even when the errors are small in dollars. Therefore, we will disregard this result for the remainder of the analysis. See the discussion "Limitations in the Definition of Error" in the Representative Scenarios section of this report.

3. Aside from the one outlier, the companies that found higher errors (~4 percent) were both testing C3 Phase II at CTE90. As we have acknowledged for other techniques, it is generally harder to capture results farther out in the tail.
4. We cannot determine a relationship between compression level and accuracy from these results. We do have more data from a series of tests performed by one company, and we will use that to explore the impact of compression further below.

One company provided results for a block of variable annuities under varying CTE levels, at an average compression level of approximately 96 percent. Note that the results in these scenarios are not floored at zero.

Metric	Error
CTE0 (Mean)	0.8%
CTE70	1.5%
CTE90	2.7%
CTE95	4.2%

Table II.E.3. Results for various CTE levels

These results provide a clear example where the tail scenarios are more difficult to capture accurately. It is possible that the fit can be improved by including appropriate calibration scenarios if a company is specifically looking for more accurate results in the tail.

As noted earlier, we were unable to determine the relationship between compression and accuracy from the overall set, likely as a result of using different blocks of business and metrics. One company provided us with results for a block of variable annuities under varying compression levels:

Compression	Error
70%	1.2%
75%	1.1%
80%	1.2%
85%	1.7%
90%	4.5%
95%	4.7%
98%	8.7%
99%	14.1%

Table II.E.4. Results for various compression levels

The above results provide a clear example where higher levels of compression result in a larger error. However, please note the following:

- Each of these results is an average over several runs, and the errors for individual runs do fluctuate.

- These results provide valuable insight into the shape of the compression/accuracy curve, but not the absolute level of the curve, as the level of accuracy is unique to this block of business.

For the two companies that provided large sets of results, we found that the errors tended to occur in a particular direction:

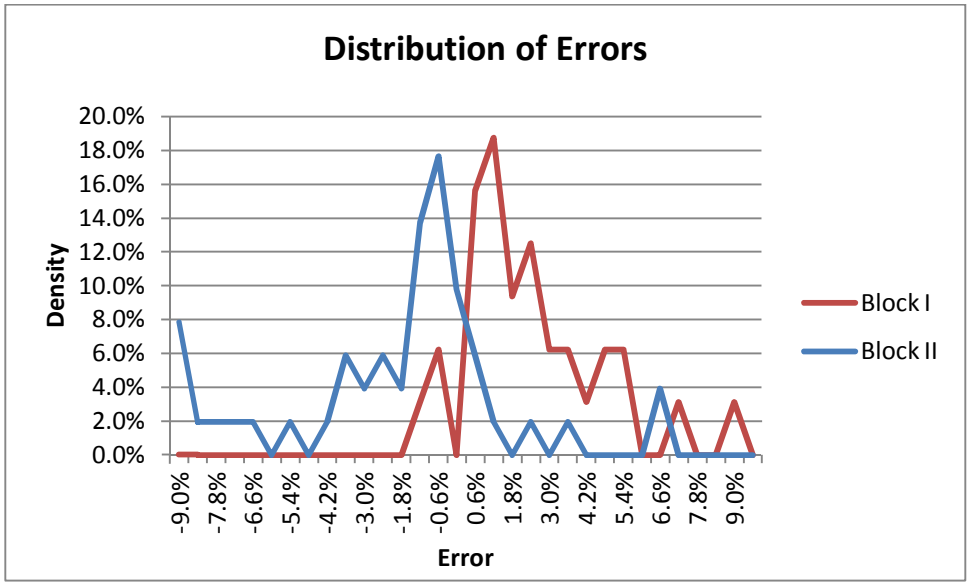

Figure II.E.5. Distribution of errors for two companies

One block tends to have positive errors, while the other block tends to have negative errors. This indicates that the errors introduced by compression are a combination of accuracy and bias. This is not necessarily a good or bad thing, but it is important to understand that bias is generally a concern with all compression techniques, and something that needs to be considered separately when looking for ways to improve the results of compression.

Companies also found that, in some cases, adding more constraints actually decreased accuracy. For example, one company first relied on both location variables including many inforce parameters and calibration scenarios. When it later removed many of the inforce parameters, the results improved significantly. The implication is that the location variables need to be tailored very carefully to the business to maximize the potential of this technique.

Considerations

Positive:

A. The technique is built into MG-ALFA, and can be coded to work with other modeling platforms as well.

B. The technique produced results with errors ranging from 0.5 to 4.3 percent at very high levels of compression. This level of compression would enable a company to perform many more sensitivities on a relatively small computing budget.

C. The results of the study show that accuracy is potentially reduced for tail metrics. However, the range of errors was lower than that produced by most of the other techniques, and companies have the opportunity to adjust the compression levels if necessary to improve accuracy.

D. Participating companies indicated that while it took a fair amount of effort initially to get the technique running, the ongoing time to maintain the process should be minimal.

Negative:

A. Secondary metrics (e.g., premium levels) can move around from period to period as a result of the compression, adding some noise to rollforwards and metrics.

B. It is necessary to be aware of the potential for differences due to bias as well as precision.

Areas for Further Research

Cluster Modeling is the most developed technique of all the techniques tested in this study. Milliman has produced articles, and many companies have already implemented the technique and refined their thought processes.

For companies using other software or considering building a clustering algorithm, there is certainly room for research into clustering techniques. For some examples of questions that can be explored, consider the following situation where the algorithm seeks to add policy X to one of the existing clusters:

1. Should the algorithm pair policy X with (i) the closest policy, (ii) the cluster with the closest center, or (iii) the cluster whose farthest member is closest?
2. Should the algorithm consider the nearest policy only, or should it determine which cluster contains the highest number of policies within the closest n-policies to policy X?

3. Can a policy belong to multiple clusters?

Actuaries are encouraged to review literature from other industries to explore clustering algorithm techniques and uses.

While not tested here, like other compression techniques clustering can be used for asset model compression as well.

F. Replicating Liabilities

Replicating Liabilities is a cell compression technique that uses optimization to produce a small subset of policies that closely reproduces specified characteristics of the full inforce.

Optimization is used widely in other industries, including various engineering applications and operations research. In the insurance industry, replication is now commonly used to find a set of assets that replicate the liabilities, which poses a separate set of challenges from the technique described here. In the wider financial industry, replication is commonly used in portfolio optimization to find a subset of assets that replicate a larger set of assets. As a result, optimization software is widely available, and it is possible to perform optimization without getting into the details of the exact algorithm used to solve the equations.

Optimization contains two main components:

1. A function to be optimized — This can be minimization or maximization of the function. In the absence of constraints, it can generally be solved using basic optimization techniques taught in introductory Calculus, and with a small number of constraints can be solved using Lagrange Multipliers. This application does not permit these simpler techniques.
2. A set of constraints — This is a set of equalities or inequalities that must be maintained in any solution. Solutions that meet the constraints are referred to as being in the feasible range. Sets of constraints that are unsolvable are considered infeasible.

Suppose that an actuary is calculating the average present value of cash flows across the inforce. There is a weight attached to each policy, and the starting weight will normally be 1/n (where n is the number of policies in the inforce). This represents an equally weighted inforce. The goal for cell compression is to adjust those weights so that a large number of policies have a weight of zero, and a small number of representative policies have much higher weights.

The actuary will then create a set of constraints that may be fed into the optimization software. Suppose that the actuary wants the account value of the compressed inforce to be equal to the account value of the original inforce. The software will be directed as follows: Minimize the number of policies with a weight greater than zero, subject to the constraint that

$$\sum_{i=1}^{n} Account\ Value_i \times W_i = \sum_{i=1}^{n} Account\ Value_i \times \frac{1}{n}$$

Where:

n is the number of policies in the inforce

W_i is the weight associated with each policy.

Notice that the only variables present are the weights. Everything else is a constant, including the entire right-hand side of the equation.

This can be expanded to cover any number of constraints. In practice, actuaries will generally constrain on key aspects of the inforce, as well as results from prior runs.

Actuaries are already accustomed to banding policies by age, gender, issue year and other pieces of data that are readily available. With this technique, it becomes very easy to expand that further and place constraints that reflect combinations or cross-sections of the various aspects of the inforce.

In order to explain how to use results from prior runs, we will work through an example:

1. Run the uncompressed inforce through two sample Scenarios A and B.
2. The two data points are added to the inforce record for each policy, much like a starting account value.
3. The two data points are set as additional constraints during the optimization process.
4. One constraint will be that the present value of cash flows for all policies under Scenario A will be close to the weighted present value of cash flows for the compressed inforce. A second constraint will provide for the same equality under Scenario B.

Effectively, the optimization using these constraints results in a compressed inforce that not only has similar characteristics to the full inforce (e.g., age distribution and account value) but also matches the present value of cash flows almost exactly for two scenarios.

Testing Results

Three companies tested this technique during the study:

Company	Product(s)	Metric	Inforce	Scenarios
A	VA/GMDB	Various CTEs	3,500 cells	SOP: 1,000 internally generated Other: 1,000 AAA scenarios
B	VA/GMxB	Various CTEs	900,000 policies	1,000 AAA interest rate scenarios and five equity indices
C	Interest Sensitive Life	Mean	40,000 policies	200 prescribed statutory scenarios

Table II.F.1. Testing background

Unless otherwise specified, the errors described in this section refer to absolute value of the difference between original results and compressed results.

The following table provides a summary of the average results from each metric tested by each company:

	Metric	Product	Error	Compression
Company A	CTE70	VA	0.8%	97%
Company A	Mean	VA	0.9%	96%
Company A	CTE90	VA	1.0%	97%
Company B	CTE70	VA	5.1%	97%
Company B	CTE90	VA	5.1%	97%
Company C	Mean	ISL	8.6%**	99%

**Note for Company C these results measure against the company's current compression technique.

Table II.F.2. Average results by block and metric

We can make several observations based on these results:

1. The results varied significantly, ranging from 0.8 to 5.1 percent (excluding results from Company C).
2. The calculated errors do not appear to vary by metric – rather, other factors must be creating differences between the different companies.
3. All the compression ratios tested for this technique were quite high, ranging from 96 to 99 percent. There appears to be a loose relationship where higher compression reduces accuracy. However, we believe that other factors drove the differences, and we will analyze this further below.

In order to understand the differences further, we analyzed the distribution of the GPVAD or PVS for each company.

Company A:

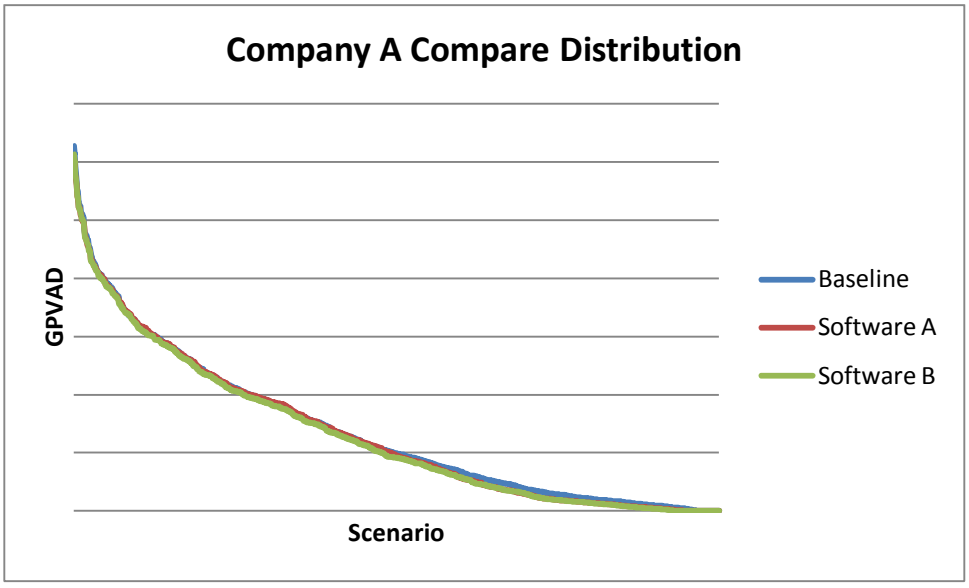

Figure II.F.3. Company A distribution

- Company A performed tests using two different optimization engines. One engine produced differences of 0.5 percent, and the other produced differences of 1.3 percent. The difference in results between the two engines is too small to definitively state that one is better than the other, and this aspect of the technique was not studied further.

Company B:

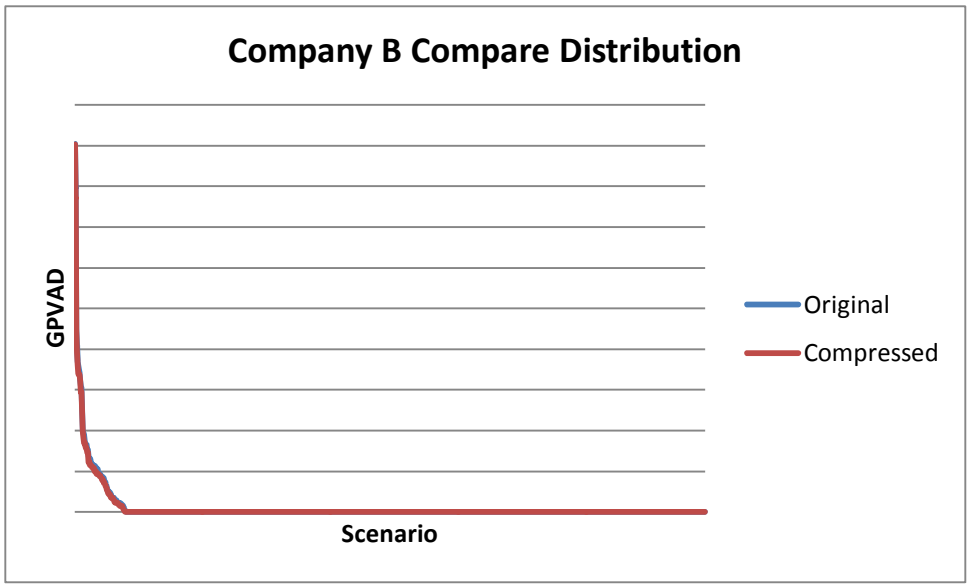

Figure II.F.4. Company B distribution

- Very few scenarios (8 percent) produce a positive GPVAD. As a result, despite fitting the original distribution well, the error is somewhat higher as a percentage of the average GPVAD for the worst CTE90, and significantly higher as a percentage of the worst CTE70. However, the company actually noted that the dollar differences created by the compression were only about 0.01percent of the starting assets, and they were quite happy with the outcome.

Company C:

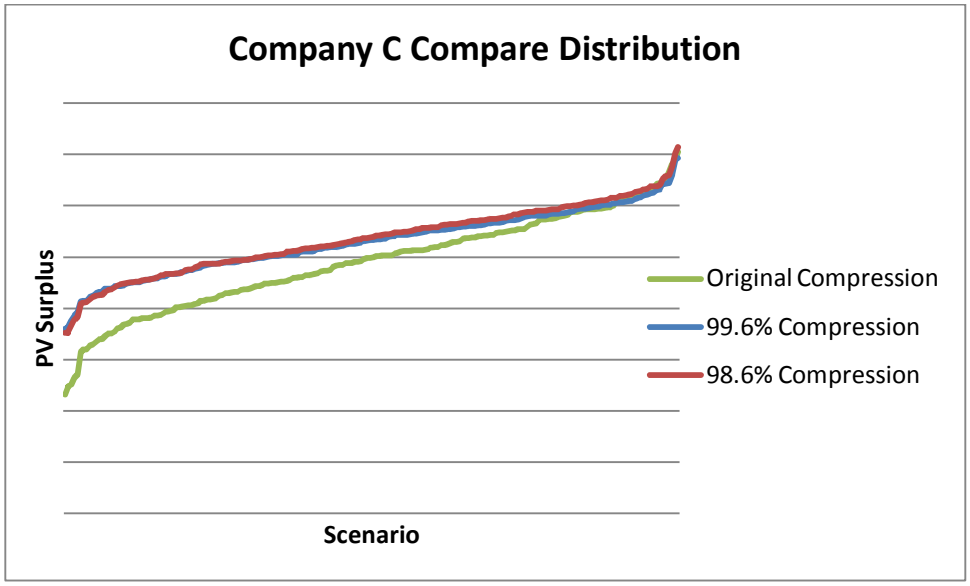

Figure II.F.5. Company C distribution

- We noted that this was the only company testing Replicating Liabilities that compared the results of the compression with the results from its current traditional compression technique. Unfortunately, it is not possible to determine whether the bias comes from the Replicating Liabilities or the current compression technique in use by the company.
- There is clearly a bias in the resulting distribution of results relative to the uncompressed run. The difference is 9 percent on average, but smaller in the right tail and larger in the left tail.
- This company tested using interest-sensitive life. However, we could not see any reason that this product would be more difficult to capture, and we do not believe that this is the reason for the difference.

Note that the compression levels tested are all very high. This technique inherently tends to produce small sets of policies. The reason is that we are essentially providing the algorithm a matrix with a set of equations to solve, where each constraint represents an equation, and a system of X equations can generally be solved with X variables. Companies in this study tended to use approximately 100 to 200 constraints, so the algorithm found about that many policies to solve the set of equations.

The practical implication is that if the actuary does not want such extreme compression, it will be necessary to create workarounds, including splitting blocks of business to be compressed separately, or iteratively compressing and removing the selected policies from the available policies to choose from.

One company tested the latter technique, increasing the number of cells from 169 to 526 through four iterations, but did not find noticeable differences in results.

It is worth noting that participating companies spent considerably more time to set up this technique than other techniques tested in this study. Most of the techniques that were selected were specifically chosen because software was readily available, or the technique was simple to implement. This technique required setting up constraints manually in spreadsheets, and learning to use a new piece of software. During the study, companies generally budgeted approximately 100 hours to test a technique, so if it took 100 hours to set it up instead of 50 hours, that would have left minimal time to refine and troubleshoot the results.

As with Cluster Modeling, companies found that, in some cases, adding more constraints can decrease accuracy. One company first set constraints so that the account value at every age would be equal to the original inforce. When they later relaxed the constraint so that it only matched in total for each 10-year age band, the results improved significantly. The implication is that the constraints need to be tailored very carefully to the business to maximize the potential of this technique.

Considerations

Positive:

A. The technique produced results with errors ranging from 1 to 5 percent at extremely high levels of compression. One company tested Replicating Liabilities against its traditional compression technique and found a 9 percent difference between the two.

B. The calculated errors did not vary by metric for companies that measured multiple metrics.

C. Participating companies indicated that while it took a fair amount of effort initially to get the technique running, the ongoing time to maintain the process should be minimal.

Negative:

A. There is a learning curve to understanding and using this technique, particularly because it is not specifically built into actuarial modeling platforms (to our knowledge) or widely recognized in the actuarial community. (These are short-term rather than long-term issues if the technique proves effective.)

B. Replicating Liabilities tends to produce extremely high levels of compression. Counterintuitively, producing lower levels of compression requires more research and more effort.

C. It is necessary to be aware of the potential for differences due to bias as well as precision.

Areas for Further Research

There is a considerable amount of research that can still be performed with this technique. The following list provides some recommendations for research:

1. While we noted that the actuary can use the results of prior runs as constraints, companies did not do that for this study. These constraints provided considerable benefit for Cluster Modeling, and are likely to produce a similar benefit for Replicating Liabilities.
2. This study did focus on the differences between different algorithms or software available to perform optimization. Optimization software differ with respect to pricing, compatibility, speed, accuracy of results, flexibility, etc. A good list of available software can be found at: http://en.wikipedia.org/wiki/Linear_programming

III. Summary

Summary of Observations

In practice, companies will face a selection process if they choose to go down the path of seeking model efficiency. Below is a summary of the key strengths and weaknesses of each technique as experienced by the participating companies:

	Initial Effort Required	Runtime Reduction Experienced	Limitations Experienced
A. Transfer Scenario Order	Low	Low	Only useful for tail metrics
B. Representative Scenarios	Moderate	Moderate	Only useful for the mean; limited by accuracy of baseline scenarios
C. Importance Sampling	Moderate	Moderate (subject to limitations)	Only useful for out-of-the-money options
D. Curve Fitting	High	Unknown	Only useful for tail metrics; bias is a significant concern
E. Cluster Modeling	High	High	Bias is a concern for all compression techniques
F. Replicating Liabilities	High	High	Bias is a concern for all compression techniques

Limitations

In summarizing the results of the study, we relied on the accuracy and completeness of the underlying data and modeling performed by the participating companies. We did not audit or independently verify such information, or confirm the accuracy of the data or the information and explanations provided by the participating companies.

It is important to note that these results are limited by the following:

- Each technique has only been tested by a small number of companies.
- The depth of the testing varied by technique, often based on the level of effort required to set up the process.
- Participating companies only produced test results, and the techniques would require additional refinement and review in order to be used in practice.
- Samples sizes were less than ideal.

As a result, the results of this study are only useful to provide general guidance and direction for actuaries.

References

AMERICAN ACADEMY OF ACTUARIES' MODELING EFFICIENCY WORK GROUP. 2010. *Modeling Efficiency Bibliography for Practicing Actuaries*. May. http://www.actuary.org/risk/pdf/bibliography.pdf

CHRISTIANSEN, SARAH L. M. 1998. "Representative Interest Rate Scenarios." *North American Actuarial Journal* 2(3): 29–59.

CHUEH, YVONNE C. M. 2002. "Efficient Stochastic Modeling for Large and Consolidated Insurance Business: Interest Rate Sampling Algorithms." *North American Actuarial Journal* 6(3): 88–103.

CHUEH, YVONNE. 2005. "Efficient Stochastic Modeling: Scenario Sampling Enhanced by Parametric Model Outcome Fitting." *Contingencies*. January–February.

FREEDMAN, AVI, and CRAIG REYNOLDS. 2009. "Cluster Modeling: A New Technique To Improve Model Efficiency." CompAct. July.

LONGLEY-COOK, ALISTAIR G. 1997. "Probabilities of 'Required 7' Scenarios (and a Few More)." *The Financial Reporter*. July.

LONGLEY-COOK, ALISTAIR G. 2003. "Efficient Stochastic Modeling Utilizing Representative Scenarios: Application to Equity Risks." Stochastic Modeling Symposium.

MORRISON, STEVEN, and MARK CATHCART. 2011. "Least-Squares Monte Carlo Simulation." *Contingencies*. March–April.

REYNOLDS, CRAIG. 2010. "Model Compression and Stochastic Modeling." *The Financial Reporter*. June.

