Article from:

The Financial Reporter

June 2010 – Issue 81

Model Compression and Stochastic Modeling

by Craig W. Reynolds

Craig W. Reynolds, FSA, MAAA, is principal and consulting actuary for Milliman, Inc. He can be contacted at craig.reynolds@milliman.com

For the last several years, the U.S. life insurance industry has been moving gradually towards a principle-based approach to statutory valuation. While the ultimate destination or arrival date for this journey is still far from clear, it seems likely that principle-based valuation will involve stochastic modeling in many cases. For variable annuities, stochastic principle-based statutory valuation became a reality at the end of 2009 when Actuarial Guideline 43 (AG 43) became effective. Furthermore, stochastic modeling is becoming more common for other applications as well, such as embedded value, enterprise risk management, economic capital, and fair value valuation. And of course, stochastic modeling has already been part of statutory exercises already in areas such as C-3 Phase 2 and, in some cases, cash flow testing.

While principle-based valuation has been a reality for some time in some markets (e.g., Canada) and for US GAAP (e.g., FAS 157), AG 43 is the first application of stochastic principle-based valuation in U.S. statutory accounting. As such, we should take an opportunity to look at what we learned in this exercise that we can apply to any future stochastic reserve or capital calculations that might arise in the move to a principle-based approach (PBA). Some key issues that distinguish stochastic PBA from traditional valuation approaches include:

- Valuation may be done using projection systems rather than traditional valuation systems.
- Assumptions may be largely up to the judgment of the individual actuary rather than prescribed.
- Assumptions and economic conditions may change dramatically from period to period.
- Calculations will be aggregate in nature, and then allocated to policies, rather than the reverse.
- Reinsurance will be reflected in a fundamentally different way.
- Assets need to be modeled.
- Hedging may need to be reflected.

Each of these issues combine together to create complications with respect to:

- controls,
- auditing,
- movement analysis, and
- model runtime.

Each of these complications is significant and will cause most companies to fundamentally overhaul their valuation processes. For purposes of this article, I will focus on the last issue: runtime. Runtime is significant for stochastic valuation applications because of the large number of cells, the large number of scenarios, and the need to perform principle-based forecasts rather than prescribed closed-form calculations. In contrast, for most companies using traditional valuation processes, machine runtime is not a material factor in periodic financial reporting exercises.

RUNTIME REDUCTION OPTIONS

AG 43 calculations for most companies require calculations in excess of 100,000 policies across 1,000 or more scenarios. For companies modeling dynamic hedging, each policy might need to be projected thousands of times for each scenario in order to calculate required liability "Greeks." Clearly, this can result in an extraordinarily lengthy runtime. Some of the options available for reducing runtime for such models include:

- reduce liability cell count,
- reduce asset cell count,
- reduce scenario count,
- reduce path count for hedging,
- simplify actuarial calculations,
- utilize faster or more hardware, and
- utilize faster software.

The American Academy of Actuaries has a Model Efficiency Working Group (MEWG), of which I am a member, that is charged with exploring these and related options. In this article I am speaking for myself, and not for the MEWG.

The MEWG has attempted, with some success, to identify specific actions that companies are currently taking to manage runtime efficiently. But it is clear that companies can do more to reduce runtime and most would like to do so. While more and faster hardware and software are always desirable, I believe that reduced cell or scenario counts offer the most hope for runtime improvements in the orders of magnitude that might be desired. This article summarizes a case study of one application of cell reduction.

REDUCING SCENARIOS

For AG 43, the Academy has published a set of 10,000 economic scenarios. Most companies that I work with have used 1,000 of these scenarios to drive their AG 43 work. In fact, it seems many valuation actuaries almost view this as a "safe harbor." So perhaps before we think about reducing below 1,000 scenarios, we need to consider this fundamental question: Is 1,000 enough? Unfortunately, I know of no way to resolve this issue without running 10,000 scenarios, and seeing how the answer changes as we reduce the scenario count gradually down to 1,000. This is perhaps an exercise that could be done well in advance of year-end.

Much research has been done on the topic of scenario reduction. Unfortunately, most techniques for analyzing the error in scenario reduction appear to be designed assuming that scenarios are chosen randomly from an adequately large universe of scenarios. In contrast, many reduction techniques rely on a distance measure to stratify and map scenarios, so the statistical tools used to analyze the appropriateness of the reduction may be of limited value. In practice, the best way to measure the appropriateness of the reduction might be to run a test model through a larger number of scenarios to see if the answer changes materially. As discussed below, cell compression techniques might be one useful means of creating a model that is sufficiently representative for such testing, while small enough to run in a viable amount of time.

In my experience, it is not likely that scenario count could be reduced much below 100, if we still want results that sufficiently capture the distributions illustrated by a run of 1,000 scenarios that we might start with. Thus, for truly revolutionary reduction in runtime, we need to consider liability and asset cell reduction.

LIABILITY AND ASSET CELL REDUCTION

Most actuaries have at some time in their career utilized traditional techniques to reduce cell count for projection purposes, but such compression is atypical for statutory valuation in the United States. When used, such techniques have often included strategies such as:

- mapping issues ages into quinquennial or decennial issue age bands,
- mapping similar plans together,
- mapping issue dates into central issue points within a year or a quarter of a year, and
- mapping all cells as male cells, perhaps with an age setback or a blending of mortality rates.

These techniques have their place, but for some lines of business (LOBs), including variable annuities, they have their limitations. Among other things, these sorts of mappings tend to mask factors such as "in-the-moneyness" or fund distribution, which can have a material impact on model results. After allowing for this, it is challenging to compress models by more than a factor of 10 or so.

At Milliman we have developed a technique that we call "cluster modeling" that can greatly improve model compression ratios, or improve model fit for a given level of compression.[1] Rather than presenting the details here, this article focuses on results of a single case study. Case studies such as these can serve an important purpose, in that they can give modelers and regulators increased comfort with compression techniques. This is particularly important now, as using liability or asset cell compression has not historically been common in statutory valuation exercises.

Cluster modeling is clearly not the only available option for cell compression, but we illustrate it here as a particularly effective technique that can be used for stochastic valuation calculations, such as those required by AG 43.

CASE STUDY FOR LIABILITY CELL COMPRESSION

In our case study we consider a variable annuity block with more than 100,000 policies in-force.[2] The block

FOOTNOTES:

[1] Freedman, A. & Reynolds, C. (August 2008). Cluster analysis: A spatial approach to actuarial modeling. Milliman Research Report. Retrieved March 16, 2010, from *http://www.milliman.com/expertise/life-financial/publications/rr/pdfs/cluster-analysis-a-spatial-rr08-01-08.pdf*. The technique is also described in some detail in the July 2009 issue of CompAct *(http://www.soa.org/library/newsletters/compact/2009/july/com-2009-iss32.pdf)*.

[2] While the results presented here are based on a real valuation AG 43 model, modest changes have been made to the model to ensure confidentiality. As such, calculated reserves reported here will not tie to the reserves reported by the company.

CONTINUED ON **PAGE 24**

includes account values of approximately $9.5 billion, a cash surrender value of approximately $9.0 billion, and a variety of Guaranteed Minimum Benefits (GMB), including GMAB, GMDB, GMIB, and GMWB. We will focus here on the stochastically calculated greatest present value of accumulated deficiencies, because the standard scenario reserve is straightforward to calculate on a seriatim basis, and in fact must be calculated that way. The 70 CTE value on a seriatim basis for this block is $143.6 million. Of course, this amount must be added to the starting asset amount and compared to the standard scenario reserve to get the final AG 43 reserve.

… we are running the seriatim model only five times, and we will run the compressed model at least 1,000 times. …

For convenience and ease of presentation we have ignored the impacts of reinsurance.

Using clustering, we modeled the liability cells into successively smaller models, using model criteria designed to closely reproduce values of the following key metrics across five representative scenarios. We refer to these variables as "location variables":

- initial GMB face amount for each benefit type and guarantee type,
- initial account value in-force by fund,
- present value of net revenue,
- present value of commission income,
- present value of revenue sharing,
- present value of maintenance expenses,
- present value of M&E fee income, and
- present value of net benefit costs for each GMB type (benefits paid less associated charges).

For each location variable that requires present values, we ran our model seriatim across five scenarios to get calibration data to drive our mapping process. The five scenarios were chosen to represent the 2.5 percent, 20 percent, 50 percent, 80 percent, and 97.5 percent level of the aggregate average wealth ratios across the complete set of 1,000 scenarios. The model reflected seven different equity indices and a fixed account. We used the five-year U.S. Treasury rate as the representative interest rate to be indicative of the level of interest rates for bond funds, and we weighted each of the indices by the associated initial fund allocation in order to drive average wealth ratios.

While at first it might seem counterproductive to run the model seriatim in order to get data to produce a model, remember that we are running the seriatim model only five times, and we will run the compressed model at least 1,000 times—potentially many more times than this if we conduct sensitivity testing. Thus, the investment in five seriatim runs to get data to allow us to run thousands of other runs in a time that is orders of magnitude faster is clearly worthwhile.

The table in Figure 1 summarizes the fit of selected model location variables as of the valuation date using various levels of model compression. In the compressed models, the original in-force, with more than 100,000 policies in-force, is compressed to models ranging in size from 5,000 cells to 50 cells. In these tables, the "Variable Weight" is an indicator of the priority we assigned to replicating that variable's value. As we would expect, in general, we get a better fit for higher-weighted variables. As with selecting the location variables themselves, selecting the weights requires some judgment.

Figure 1
Analysis of Fit Variables as of Valuation Date
($ millions)

			Ratio to Seriatim for Differing Cell Counts				
	Weights	Seriatim	5,000	2,500	1,000	250	50
Inforce GMB Face Amounts							
GMDB Ratchet	1	$7,733	99.8%	99.8%	99.2%	98.9%	93.6%
GMDB Rollup	1	$4,058	97.6%	96.3%	93.9%	92.4%	94.4%
GMDB ROP	1	$4,515	100.5%	100.9%	103.6%	106.6%	122.5%
GMIB Ratchet	1	$7,545	100.0%	100.0%	99.7%	100.6%	98.2%
GMIB Rollup	1	$8,181	100.4%	100.4%	100.4%	100.6%	99.3%
GMAB ROP	1	$281	99.7%	99.1%	100.0%	94.3%	63.9%
Inforce Account Values							
Separate Acct 1	1	$1,426	101.9%	102.9%	105.7%	106.1%	110.9%
Separate Acct 2	1	$1,070	99.7%	99.2%	99.0%	99.1%	94.7%
Separate Acct 3	1	$999	97.0%	96.0%	94.8%	95.6%	93.6%
Separate Acct 4	1	$267	102.5%	104.1%	104.9%	108.0%	104.7%
Separate Acct 5	1	$905	100.9%	101.3%	101.6%	102.6%	106.1%
Separate Acct 6	1	$1,330	96.2%	94.6%	92.4%	90.4%	89.2%
Separate Acct 7	1	$2,020	103.7%	105.4%	107.1%	111.3%	113.6%
General Acct	1	$654	99.9%	99.9%	99.6%	98.8%	88.2%

Figure 2 shows comparable data, but this time focusing on present values of selected results across various calibration scenarios. Note that we have used the same weight across scenarios. If we know, as is typically the case, that poor markets produce the results that drive the AG 43 results, we might choose to weight those scenarios more heavily for an even better fit of AG 43 results.

Not surprisingly, the fit degrades somewhat as the cell count goes down. However, even the 50-cell results show a surprisingly good fit compared to the original

CONTINUED ON **PAGE 26**

Figure 2
Analysis of Fit Variables, PVs by Scenario
($ millions)

			Ratio to Seriatim for Differing Cell Counts				
	Weights	Seriatim	5,000	2,500	1,000	250	50
Present Values-Scenario 1							
Net Revenue	4	$(202)	89.5%	89.0%	89.1%	78.8%	57.1%
Commissions	2	$317	99.3%	98.9%	98.5%	99.1%	98.0%
Revenue Sharing	2	$218	99.4%	99.3%	99.0%	98.9%	97.2%
Maintenance Expense	2	$150	87.3%	89.8%	94.5%	98.2%	103.6%
M&E Income	3	$872	99.6%	99.5%	99.1%	98.9%	98.9%
Net GMAB Cost	3	$5	98.3%	98.3%	98.4%	90.4%	64.1%
Net GMDB Cost	3	$93	101.1%	101.7%	101.5%	100.4%	102.1%
Net GMIB Cost	3	$395	100.3%	100.3%	100.3%	100.7%	101.0%
Present Values-Scenario 2							
Net Revenue	4	$(248)	90.9%	90.3%	88.7%	81.4%	65.9%
Commissions	2	$295	99.1%	98.6%	98.3%	98.7%	97.9%
Revenue Sharing	2	$210	99.2%	99.1%	98.9%	98.3%	96.6%
Maintenance Expense	2	$150	87.1%	89.6%	94.4%	98.2%	103.4%
M&E Income	3	$836	99.4%	99.2%	98.9%	98.5%	98.5%
Net GMAB Cost	3	$5	98.6%	98.6%	98.9%	89.6%	67.7%
Net GMDB Cost	3	$64	102.1%	102.1%	103.9%	99.1%	106.6%
Net GMIB Cost	3	$398	100.6%	100.8%	101.2%	102.1%	102.1%
Present Values-Scenario 3							
Net Revenue	4	$(787)	96.7%	96.4%	95.8%	92.8%	88.5%
Commissions	2	$176	99.7%	99.3%	98.9%	99.4%	96.8%
Revenue Sharing	2	$127	99.8%	99.9%	99.9%	99.9%	98.2%
Maintenance Expense	2	$132	85.6%	88.6%	94.0%	98.5%	104.2%
M&E Income	3	$507	100.1%	100.1%	99.9%	99.9%	99.9%
Net GMAB Cost	3	$2	102.8%	101.4%	96.9%	84.3%	90.7%
Net GMDB Cost	3	$(20)	91.4%	94.2%	86.2%	93.8%	111.3%
Net GMIB Cost	3	$44	109.3%	114.2%	123.2%	124.6%	118.2%
Present Values-Scenario 4							
Net Revenue	4	$(871)	97.4%	97.5%	97.0%	94.4%	89.1%
Commissions	2	$176	99.6%	99.2%	98.8%	98.9%	97.9%
Revenue Sharing	2	$132	99.7%	99.8%	99.8%	99.6%	98.8%
Maintenance Expense	2	$132	85.8%	88.7%	94.1%	98.3%	104.4%
M&E Income	3	$525	100.0%	100.0%	99.9%	99.6%	100.7%
Net GMAB Cost	3	$1	106.4%	103.4%	110.8%	121.9%	175.7%
Net GMDB Cost	3	$(8)	89.5%	96.2%	78.8%	84.4%	106.7%
Net GMIB Cost	3	$(167)	97.3%	98.6%	97.7%	97.4%	98.1%
Present Values-Scenario 5							
Net Revenue	4	$(1,249)	97.7%	97.7%	97.3%	95.4%	91.0%
Commissions	2	$150	99.7%	99.4%	99.1%	99.6%	98.3%
Revenue Sharing	2	$110	100.0%	100.1%	100.1%	100.2%	99.2%
Maintenance Expense	2	$130	85.1%	88.2%	93.9%	98.4%	104.4%
M&E Income	3	$437	100.2%	100.2%	100.2%	100.1%	100.8%
Net GMAB Cost	3	$(4)	96.4%	95.7%	98.1%	102.8%	44.1%
Net GMDB Cost	3	$(63)	95.9%	95.8%	92.0%	93.5%	106.2%
Net GMIB Cost	3	$(455)	98.3%	98.4%	98.2%	98.5%	96.8%

seriatim run, with a compression ratio of over 2,000 to 1, and a commensurate reduction in runtime.

One advantage of the cluster model process is that we can choose the variables that we wish to closely replicate and dial up or down the model granularity or adjust the weights as needed in order to achieve the desired level of fit for those variables. Even traditional modeling processes introduce some noise, but the cluster process allows us to measure the noise across any dimension and tweak the modeling to reduce that noise. Furthermore, we can easily analyze the implications of the modeling decisions and adjust the granularity to achieve fit objectives.

Of course the real question is: what is the impact of the modeling on the AG 43 stochastic calculation results? The table in Figure 3 provides the answer.

Figure 3
Impact of Modeling on AG 43 Results
($ millions)

Liability Cell Count	Stochastic Reserve	Ratio to Seriatim
Seriatim	$143.6	100.0%
5,000	$144.2	100.4%
2,500	$143.9	100.2%
1,000	$141.6	98.6%
250	$140.6	97.9%
50	$136.7	95.2%

While some actuaries might be troubled by even the modest levels of noise shown above, it is important to keep this in perspective. In this block, for example, a 1 percent addition to the lapse rate would change reserves by approximately $37 million. Thus, the $7 million in modeling error introduced by even the 50-cell model pales in comparison to the imprecision that we accept because of modest uncertainty in lapse assumptions.

Furthermore, the relative materiality of the difference between the seriatim stochastic reserve and the compressed model value should really be judged relative to the total reserve, which is the sum of the cash surrender value of around $9 billion and the stochastic reserve amount shown. Thus, this noise is only approximately 0.08 percent for a 50-cell model.

Of course, 50 cells might be more compression than we would feel comfortable with, but any of the intermediate values above give an even better fit, for a runtime that is far more palatable than that of the original model.

Furthermore, though we might still choose to run the valuation on a seriatim basis, the compressed model might be exceptionally useful for sensitivity testing, or for testing to see how many scenarios are necessary to run. Now such tests can be run in mere minutes on one machine, rather than in hours or days across many machines.

DO WE NEED TO DO IT?

Is liability model compression really critical? Perhaps not, to the extent that AG 43 is the only stochastic valuation application. However, there are several important reasons why a good compression technique should be considered:

- Many companies are doing traditional modeling already. A more sophisticated technique such as cluster modeling offers more robust alternatives for model validation, as well as higher compression ratios and/or better model fit.
- As stochastic calculation becomes required for the valuation of other LOBs, runtime will become more and more critical.
- Reducing runtime leaves more time for validation, sensitivity testing, and analysis.
- Similarly, while runtime considerations might make running 10,000 scenarios impractical in real time, highly compressed models can be used to run 10,000 scenarios and to analyze the impact of using lower scenario counts, predicting what those impacts might be on the seriatim model. As noted above, this is perhaps the most effective technique for validating any sort of scenario reduction.
- Nested stochastic analysis might be required to project future reserves. While seriatim valuations can be made practical for most companies with

CONTINUED ON **PAGE 28**

adequate hardware and software, nested seriatim stochastic applications are almost certainly impractical for all but the smallest blocks of business.

CONCLUSION

Model efficiency is just one issue to consider in the long list of practical issues as we move to principle-based methods for reserves and capital. But if this issue is not adequately addressed, PBA will not be practical. Early evaluation and validation of scenario reduction and model compression techniques will be an important key to success. If you would like to help in this process, feel free to contact the author at craig.reynolds@milliman.com, or Tony Dardis, chair of the MEWG, at tony.dardis@barrhibb.com. ■