MILLIMAN WHITE PAPER

Cluster modeling: A powerful way to reduce cloud costs

Case studies for variable annuities and universal life insurance

Zohair Motiwalla, FSA, MAAA
Aatman Dattani, ASA, MAAA

Milliman

Introduction

The insurance industry is currently experiencing a period of disruption, with the ongoing legacy of the COVID-19 pandemic, rapidly rising interest rates and inflationary pressures, artificial intelligence (AI), and data proliferation all having clear impacts on business-as-usual operations.

In the face of these challenges, it is more important than ever for actuaries to produce accurate and useful information in a timely fashion for all end-users and stakeholders. Seriatim actuarial models may be accurate, but they are often too costly and time-consuming to be useful for an effective and nimble response in a rapidly changing world. Accordingly, actuaries typically employ some form of proxy modeling in order to manage run-time in the face of complicated insurance products, rapidly changing regulatory frameworks, and an increasing number of "what if" requests from senior management.

A more recent but no less compelling reason for leveraging such techniques has been the gradual industry movement from a fixed-cost on-premise grid to a scalable cloud solution for distributive processing. Moving to the cloud will reduce fixed costs and increase variable costs relative to an on-premise solution, with the increase in variable costs depending on a number of factors, not least whether users are adhering to a model governance framework that determines when and how jobs should be submitted to the cloud. Depending on the behavior of these users at a company, the compute costs for such jobs have the potential to dramatically accumulate, resulting in material and often unexpected financial outlays if users are not careful about what is run.

One proxy modeling technique that may help reduce these compute costs is cluster modeling, which allows users to effectively submit "light" versions of their "heavy" production models to the cloud which run faster than the heavy models. A light model is in effect a version of the heavy model that has been scaled down in some way but still materially captures the salient features and therefore the underlying economics. Before we move on to a discussion of clustering, it is important to recognize that, prior to considering any proxy modeling technique, a review and streamlining of the model code for efficiency should first be carried out. This paper assumes that such an exercise has been completed, but that the company is struggling with the compute costs associated with running on the cloud.[1]

Cluster modeling, as its name suggests, draws on the well-developed theory of clustering in a statistical context. The technique intelligently maps similar "points"—potentially liability model points, economic scenarios, or asset CUSIP[2] numbers—together in an automated fashion, allowing for a user-defined level of compression, while preserving the fit to within some reasonable level of tolerance to the heavy model. Mapping these points together to clusters that are likely to behave similarly using key financial output allows for more accuracy in results and higher levels of compression than with more traditional model compression techniques that group in less robust ways (by inputs such as gender or issue age in the context of traditional liability compression, for example). An introduction to clustering as it pertains to actuarial modeling is not in the scope of this paper, but we refer the interested reader to a paper[3] on the topic written by Avi Freedman and Craig Reynolds. A more recent technical paper[4] on variations in clustering methodology by Michail Athanasiadis, Peter Boekel, Karol Maciejewski, and Dominik Sznajder may also be of interest to those already familiar with the basics of the clustering process.

[1] This paper does not attempt to address the model governance framework considerations that dictate how users should submit jobs to the cloud.

[2] Committee on Uniform Securities Identification Procedures

[3] Freedman, A. & Reynolds, C. (August 2008). Cluster Analysis: A Spatial Approach to Actuarial Modeling. Milliman Research Report. Retrieved March 5, 2023, from https://www.my-milliman.com/-/media/milliman/importedfiles/uploadedfiles/insight/research/life-rr/clusteranalysisaspatialrr080108pdf.ashx.

[4] Athanasiadis, M., Sznajder, D., Maciejewski, K., & Boekel, P. (January 27, 2023). Applied Unsupervised Machine Learning in Life Insurance Data. Milliman Briefing Note. Retrieved March 5, 2023, from https://www.milliman.com/en/insight/applied-unsupervised-machine-learning-in-life-insurance-data.

Clustering and key performance indicators under U.S. reporting frameworks

A fair value reserve is generally calculated as the arithmetic mean result over a set of (risk-neutral) stochastic economic scenarios, effectively giving every scenario in this set an equal weight in the final result.[5] No prescribed regulatory floor is applied to this result, which could theoretically be zero or negative. From a clustering perspective, the absence of a floor (or a cap) on the mean metric used for fair value implies a certain symmetry that, all else being equal, indicates there are fewer "artificial constraints" on the cluster calibration process. Note that the same would hold true for a stochastic set of real-world scenarios over which a mean result is calculated—in other words, there is nothing particularly special about the use of risk-neutral scenarios in this context. Rather, it is the reporting metric or key performance indicator (KPI) and therefore the underlying framework specifying that metric that is key. Other proxy modeling techniques can also work well when the reporting KPI is the arithmetic mean—for example, varying the economic scenarios by liability model point, which for a large number of model points will generally require fewer economic scenarios to achieve convergence of the mean than a more conventional approach.

In contrast, the principle-based statutory reserves and capital framework in the United States uses a conditional tail expectation (CTE) result, defined as the arithmetic mean over a small subset of the overall (real-world) stochastic economic scenarios. In a variable annuity context, a CTE 70 is calculated[6] for principle-based reserving (PBR) under VM-21 and a CTE 98 is calculated[7] for the capital total asset requirement under C-3 Phase II. The economic scenarios in question are generated each quarter using the currently prescribed American Academy of Actuaries Interest Rate Generator (AIRG). Under the statutory reporting framework for variable annuities, the heavy model effectively is a seriatim in-force file run over the full set of 10,000 economic scenarios[8] from the AIRG. At least theoretically, multiple variations of this heavy model may also need to be run for production purposes. For example, companies may have to run the model with and without dynamic hedging reflected,[9] with a varying reinvestment strategy (company actual versus the prescribed Alternative Investment Strategy) and with varying assumptions (prescribed assumptions under the Standard Projection in addition to the company prudent estimate). Clearly this can translate into a material amount of run-time and commensurate cloud computing costs.

Scenario and liability clustering

Most companies that sell variable annuities employ scenario, liability, and/or asset compression to produce a light model that is used for each of the variations described above. For scenarios, companies commonly use a subset of scenarios from the universe of 10,000 economic scenarios that are generated using the AIRG. The AIRG provides a scenario picking tool for this purpose, although it only leverages interest rates as a metric (or location variable[10]) to choose the scenarios, not both equities and interest rates. In contrast, clustering allows the use of any number of location parameters to choose scenarios thereby permitting companies to capture both the equity and interest rate movements. (We refer to this approach as "Type I Scenario Clustering.") This is a significant advantage for companies that have equity indexed or variable products, where both these types of capital market movements are important. However, for companies with complex product designs—e.g., guaranteed minimum death benefit (GMDB) riders, variable annuity guaranteed living benefits (VAGLBs), or buffers and floors—or that have hedges in place, a more robust implementation of scenario clustering can be performed by first applying liability clustering and then using the light model output as calibration points for scenario clustering to develop a model that is lighter still.

[5] Assuming no scenario reduction technique (scenario clustering or otherwise) has been implemented.

[6] The result averaged over the worst 30% of scenarios.

[7] The result averaged over the worst 2% of scenarios.

[8] Some companies (arguably erroneously) consider the seriatim in-force run over 1,000 economic scenarios to be the heavy model, neglecting the larger universe of 10,000 economic scenarios that are generated by the AIRG.

[9] If the company has a Clearly Defined Hedging Strategy.

[10] This is a reporting item that you want the light model to closely reproduce. (In this particular case, the 20-year Treasury rate.) Location variables are user-defined and can vary by the line of business clustered.

In a statutory context, liability clustering for variable annuities involves running the heavy model over a small set of three to five economic scenarios that represent the types of scenarios that would be modeled. Rather than focusing on the inputs, the output from the heavy model is captured over these calibration scenarios and used to develop the liability cluster. Importantly, this process ensures that the light model automatically captures the relevant policyholder behavior impact from the dynamic functions used to model such behavior in the heavy model.

More robust scenario clustering can then be performed by creating a (much more) heavily clustered liability in-force file, e.g., a 0.1% (1,000-to-1) liability cluster or even a 0.01% (10,000-to-1) liability cluster. Running a light model that used such a heavily clustered liability in-force file would in all likelihood produce a less-than-desirable fit to the heavy model. However, the purpose of this heavily clustered in-force file is not to produce results for downstream consumption by end-users, but rather to calibrate the scenario clustering more effectively. (We refer to this approach as "Type II Scenario Clustering.") For example, companies could run their model over all 10,000 scenarios using this heavily clustered liability file, and the output data that is so captured, while somewhat inaccurate for reporting purposes per se, is much more informative than an approach that simply focuses on the inputs as it captures the impact of the scenarios on the key drivers of the business. This extra information can be used to better calibrate the scenario cluster and automatically choose a subset of the 10,000 scenarios to use.[11]

Practical consideration and limitations

Using clustering in this way can help increase the likelihood that the CTE metric of interest that is produced with the light model matches to within a reasonable level of tolerance to what would have been obtained had the company run the heavy model (with all 10,000 scenarios).[12] That said, clustering, as with all proxy modeling techniques, often involves balancing what is most important to the actuary (and by extension, the company) with what might be considered less important. As an example, if statutory CTE 98 capital preservation is what really matters for a company, then the clustering can be calibrated so that the light model attempts to fit that KPI closely, while recognizing and accepting that by virtue of that increased focus there is the potential for slightly more noise in (say) the CTE 70 reserve. In general, matching to different KPIs is managed by calibrating the cluster to different scenarios and designing location parameters that capture the variability. There is a fine line between increasing the number of KPIs while simultaneously preserving a close fit on all these KPIs. Attempting to force the cluster to calibrate closely to many distinct KPIs will likely result in a worse fit overall.

In practice, it is also possible for asymmetries and/or nonlinearities in CTE results to occur in certain circumstances. Examples of this might include when the cash value is binding in tail scenarios,[13] as it tends to be for variable annuity books that are out-of-the-money, or where the Additional Standard Projection Amount (an additive amount to the calculated PBR requirement) is nonzero. These asymmetries can potentially distort any proxy modeling technique that is used, including clustering. Achieving a reasonable degree of fit is not an insurmountable difficulty in such cases, but a thorough analysis needs to be carried out because the implementation of clustering in practice can tend to be more of an art than a science, which often requires an experienced practitioner. This implementation is largely a onetime cost - once the calibration is set, we recommend only an annual review unless the underlying business fundamentally changes.

In certain circumstances, liability clustering may not be permitted by the reporting framework in question. For example, under the Long Duration Targeted Improvements promulgated by U.S. GAAP, policies with market risk benefits need to be modeled on a seriatim basis. Using scenario reduction techniques (such as scenario clustering) in this context is a useful alternative that can help reduce the run-time on the cloud that is required.

[11] The scenario clustering calibration process will produce a subset of the 10,000 scenarios that may be unequally weighted. For example, if clustered to 1,000 scenarios, not all these scenarios will have a weight of 1 / 1,000—some weights may be more, and some may be less. Companies will need to weight scenario results accordingly in the back-end reporting process.

[12] Other scenario reduction techniques can of course work too, but some techniques may not apply under a statutory framework. For example, it is not possible to varying economic scenarios by liability model point (for fair value calculations, as described earlier) because under PBR each scenario reserve is by definition calculated for only that scenario across all policies. Varying scenarios by contract would not allow for aggregation across policies.

[13] The time zero aggregate cash value is a prescribed floor on the calculated PBR requirement.

Defining segments for clustering is also an important practical consideration. The concept of a segment is perhaps easiest understood from the perspective of liabilities, where it refers to intentionally creating partitions (or segments) where the clustering algorithm will not map model points from one segment to another. This might naturally arise in a number of situations—for example, U.S. GAAP cohorts, considering business issued in New York versus business issued outside of New York (where the former is subject to regulation defined by the New York Department of Financial Services), business that corresponds to specific legal entities, product differences (for example, a variable annuity and a registered index-linked annuity, both of which are subject to VM-21 and C-3 Phase II and thus might be modeled collectively). However, it is important to note that, as a general rule, the more segments that are created, the more constraints will exist on the clustering algorithm and the harder it will be for the light model to achieve a high level of compression with good fit. Absent a reporting or analytical granularity requirement, segmentation usually will produce worse fit, and serves no purpose.

We also note that, in general, the compression rate that one can achieve with clustering is directly proportional to the number of policies or liability model points in the seriatim (or uncompressed) in-force. The fewer the number of policies, the lower the compression that can typically be achieved at any given quality level.

As with all proxy modeling techniques, there are certain practical considerations to keep in mind when considering clustering. For example, it is important to define the KPIs that end-users are interested in, so that the clustering algorithm can be calibrated to these metrics. These KPIs might be fair value reserves (an arithmetic mean result), VM-21 reserves (CTE 70), C-3 Phase II capital (CTE 98), and so forth. Importantly, when liability clustering, we do not need to create different clustered liability in-force files for separate metrics but rather ensure that we calibrate the cluster to be robust enough across all such KPIs. Accomplishing this requires expert judgment in many cases, and is where art is needed, not just the science. Another consideration is the model fit versus the seriatim run. Even a basic technique such as mapping “like” policies creates some noise versus a pure seriatim approach, and with a very large reduction in in-force record count management needs to establish a noise threshold it is comfortable with. For optimal liability cluster results, it is also best to calibrate to the seriatim model and not a traditionally compressed model, or else it is possible to go through substantial work to reproduce a bad answer.[14]

Lastly, it is worthwhile to point out that clustering itself can encompass a variety of methods. Choosing between methods is not always obvious. A discussion of this observation is outside the scope of this paper, but we refer the reader to the technical paper referenced earlier.[15]

Clustering case studies

The following case studies demonstrate the power of using clustering on an in-force variable annuity block and a universal life block.

CASE STUDY 1: IMPLEMENTING LIABILITY CLUSTERING FOR A VARIABLE ANNUITY BLOCK

The variable annuity block in this case study includes products that provide a range of guaranteed minimum death benefits (GMDBs) and guaranteed minimum living benefits—guaranteed minimum annuity benefits (GMAB), guaranteed minimum income benefits (GMIB), and guaranteed lifetime withdrawal benefits (GLWB)—collectively VAGLBs. We designed and implemented a 5% liability cluster on the block (a 20-to-1 or 95% reduction in liability in-force record count). The ratio of seriatim to 5% cluster liability results are shown in Figure 1, for both statutory and fair value frameworks. For the statutory framework, we present CTE results both in excess of cash surrender value (CSV) and including the cash surrender value. The seriatim run took about half a day to run across all scenarios on the cloud, whereas the 5% liability cluster run only took about 1.5 hours, a run-time savings of almost 90%.

[14] Clustering by starting with a traditionally compressed liability file rather than the seriatim file runs the risk of the cluster inheriting any noise that might be embedded in the process that led to the creation of the former.

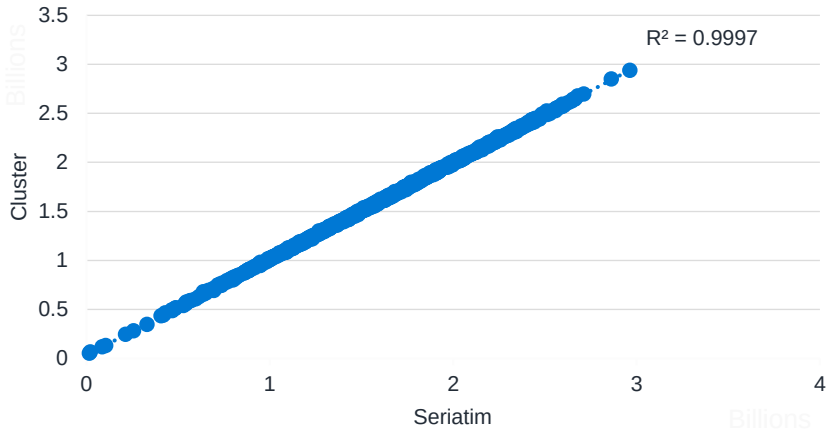
[15] Athanasiadis, M., Sznajder, D., Maciejewski, K., & Boekel, P. (January 27, 2023), op cit.

FIGURE 1: RATIO OF SERIATIM TO 5% CLUSTER LIABILITY RESULTS

KPI	RATIO OF SERIATIM TO 5% CLUSTER RESULTS (STATUTORY RESULTS ARE THE EXCESS OVER CSV)	RATIO OF SERIATIM TO 5% CLUSTER RESULTS (STATUTORY RESULTS ARE INCLUSIVE OF CSV)
Fair Value Reserves (mean over 1,000 scenarios)	99.9%	
Statutory Reserves (VM-21): CTE 70	102.4%	100.2%
Statutory Capital (C-3 Phase II): CTE 98	101.8%	100.2%
	Run-time savings of almost 90%	

The graph in Figure 2 shows the fair value reserve using the seriatim (horizontal axis) versus cluster (vertical axis) compared across all 1,000 economic scenarios. The calculated R-squared measure for this data is almost exactly 100%. Not only does the mean result match well (a 99.9% fit) but scenario-specific fair value reserves also match very well across all scenarios.

FIGURE 2: FAIR VALUE RESULT IN EACH OF 1,000 SCENARIOS (IN BILLIONS)



It is worthwhile pointing out that we could have clustered the liability in-force file further, to say a 2% (50-to-1) or 1% (100-to-1) cluster model and achieved a similar fair value result fit. However, the desire to fit closely to the heavy model for multiple KPIs, namely the fair value (mean metric), VM-21 reserve (CTE 70 metric), and C-3 Phase II capital (CTE 98 metric), led to a decision to use a model with a slightly more modest level of compression.

CASE STUDY 2: IMPLEMENTING LIABILITY CLUSTERING FOR A UNIVERSAL LIFE BLOCK

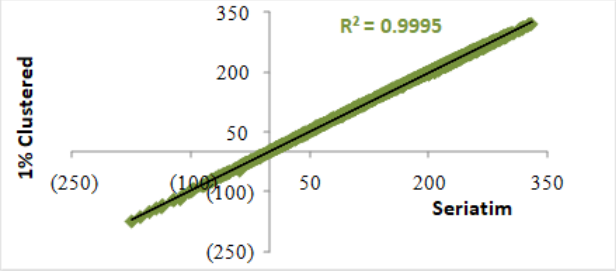
The following case study reflects a 1% (or 100-to-1) liability cluster on a universal life block that resulted in a run-time savings of approximately 90%[16] for a stochastic projection that used 1,000 economic scenarios, enabling the actuary to quickly run "what if" scenarios for senior management. The ratio of seriatim to 1% liability cluster results are shown in the table in Figure 3.

[16] Given that the clustered in-force file represented a 1% cluster, one might expect the run-time savings to be approximately 99%. However, the model initialization, asset calculations, and noncellular calculations (such as reinvestment, disinvestment, taxes, etc.) are unaffected by the liability clustering and so the actual run-time savings of 90% is slightly less than what might be expected.

FIGURE 3: RATIO OF SERIATIM TO 1% LIABILITY CLUSTER RESULTS ACROSS 1000 SCENARIOS (FAIR VALUE)

KPI	RATIO OF SERIATIM TO 1% CLUSTER RESULTS
PV of Pretax Profits	100.7%
Reserves at Time 0	98.8%
Reserves at Time 10	99.5%
Reserves at Time 20	99.8%
PV of Surrender Benefits	99.6%
PV of Death Benefits	98.7%
PV of Partial Withdrawals	99.3%
Run-time Savings	90%+

(in millions)



CASE STUDY 3: IMPLEMENTING SCENARIO CLUSTERING FOR A VARIABLE ANNUITY BLOCK

A variable annuity writer repeatedly sampled subsets of 1,000 scenarios from the universe of 10,000 AIRG scenarios using the scenario picking tool but did not observe convergence in results—that is, there was a larger-than-expected variance in results when the random seed used to choose the subset of 1,000 scenarios was changed. Effectively, the variability of the indices underlying the business was not being adequately captured by the scenario picking tool.

Using Type I Scenario Clustering on the 10,000 scenarios by calibrating the cluster to all the equity indices being modeled, the company was able to run 50% fewer scenarios and still obtain convergence in results. For validation purposes, the seriatim in-force was run through the full 10,000 scenarios and through the clustered set of scenarios. The resulting CTE 70 and CTE 98 metrics required for statutory purposes had a 103.0% and 98.9% fit, respectively.

Note that a tighter fit would very likely be possible under a Type II Scenario Clustering approach.

Other clustering use cases

Clustering can be a powerful tool for performing attribution of results between successive quarterly reporting cycles, especially where the underlying reporting is stochastic in nature. Such attribution usually requires running multiple steps in order to isolate the impacts of specific changes, and so run-time is an important concern.

Under VM-21 for example, steps in the attribution might include the impact from the following:

- Changes in equity markets
- Changes in interest rates
- Changes in volatility
- Changes from in-force business (for example changes in fund allocations, terminations, and partial withdrawals)
- Changes in assumptions
- Change in value of hedge assets
- New business
- Reinsurance

Clustering either the liability in-force file and/or scenarios can significantly reduce the run-time associated with these steps without unacceptable loss of accuracy. Another similar use case would be for running sensitivities or what-if stresses for enterprise risk management purposes. In this case, the actuary might be less concerned with whether the light model reasonably approximates the heavy model but rather that the impact (or delta) of the sensitivity with the light model is close to the corresponding delta of the sensitivity with the heavy model.

In nested deterministic and/or nested stochastic contexts, such as pricing or business forecasting, using a highly clustered liability in-force file for the inner loop reserve and capital projections can also be an effective way to reduce run-time. Typically, this might be achieved by clustering the seriatim in-force at the model start date, aging the resulting clustered in-force file, and using these aged clustered in-force files for the inner loop calculations at future points in time (sometimes referred to as “nodes”).

Irrespective of use case, it is critical that actuaries test that the light model that is developed (and that reflects the clustering) fits reasonably well against the heavy model, ideally back-tested over recent reporting cycles and under a variety of sensitivities and stresses. Such validation is generally performed when the implementation is first carried out in order to get all stakeholders comfortable with the idea of using a proxy model, but it is important to update such testing periodically. Certain financial reporting frameworks, such as PBR, require that the company attest that the light model is fit for purpose on an annual basis. Clustering is a robust process, but carrying out such testing, especially between reporting periods when staff may have downtime, can provide additional comfort.

Conclusion

Clustering has the potential to help companies reduce run-time and therefore cloud computing costs in a wide variety of use cases. As with all proxy modeling techniques, understanding the purpose, business context, and downstream use of the light model is critical to calibrate the process properly and effectively.

Milliman

Milliman is among the world’s largest providers of actuarial, risk management, and technology solutions. Our consulting and advanced analytics capabilities encompass healthcare, property & casualty insurance, life insurance and financial services, and employee benefits. Founded in 1947, Milliman is an independent firm with offices in major cities around the globe.

milliman.com

CONTACT

Zohair Motiwalla
zohair.motiwalla@milliman.com

Aatman Dattani
aatman.dattani@milliman.com