Backtesting reserve risk: a compound lognormal framework

31 August 2026



Neil Gedalla FIA
Principal

+44 (0)20 7432 7780
neil.gedalla@lcp.com



Jade Lagrue FIA
Consultant

+44 (0)20 7432 6720
jade.lagrue@lcp.com



Lara Palmer FIA
Senior Consultant

+44 (0)20 7432 7766
lara.palmer@lcp.com

1. Abstract

Reserve risk in non-life insurance capital models is commonly modelled using an over-dispersed Poisson (ODP) bootstrap to estimate volatility, often followed by fitting a lognormal distribution using the resulting coefficient of variation (CoV). However, these approaches typically treat the selected best estimate and estimated volatility as known with certainty. This may understate the full uncertainty associated with reserve estimates, including uncertainty arising because the selected reserving model and its assumptions are imperfect representations of the underlying claims process.

We apply a backtesting framework to historical US Schedule P paid claims data, building on the broad approach of Leong, Wang and Chen (2012). We assess the calibration of raw ODP bootstrap output, a bootstrap-parameterised lognormal distribution, and a lognormal distribution with an uplifted CoV against observed one-year movements in ultimate claims estimates. We also introduce a compound lognormal framework in which the volatility parameter of the lognormal distribution is itself uncertain.

Backtesting results suggest that both the raw ODP bootstrap and the bootstrap-parameterised lognormal distribution understate the frequency of more extreme historical reserve movements. Applying a uniform uplift to the bootstrap CoV improves tail calibration, but widens the distribution as a whole and relies on a judgemental choice of uplift. By contrast, the compound lognormal framework produces a more pronounced central peak and heavier tails, and provides a materially improved fit to the observed distribution of reserve movements. We suggest using a lognormal mixing distribution with a CoV of 0.8 to capture uncertainty around the estimate for σ produced by the bootstrap.

Contents

1. Abstract....2
2. Introduction....4
3. Background....6
4. Data....8
5. Models....12
6. Backtesting results....16
7. Conclusions and next steps....26
Acknowledgements, data and references....29
Appendices....30

2. Introduction

2.1. Purpose

Reserve risk was one of the earlier areas of capital models to be developed, and involves many assumptions that have become almost axiomatic over time. In particular:

- **Bootstrapping** – a method commonly used to parameterise reserve risk.
- **The lognormal distribution** – typically used to model ultimate reserve movements in capital models.

A commonly-used approach for reserve risk modelling is to use bootstrapping to calculate or otherwise help parameterise a CoV, and then to fit a lognormal distribution using this CoV to generate a model of ultimate reserve movements.

Market practice for reserve risk modelling has remained largely unchanged for decades, despite existing research highlighting some key limitations with common market approaches.

The purpose of this paper is to challenge existing approaches for reserve risk modelling and to propose an alternative method which addresses some of the key limitations.

2.2. Motivation

The existing literature, backed up by our analysis in this paper, shows that the bootstrap and the lognormal distribution together are typically not a good fit to the actual reserve volatility. Even after adjusting bootstrap CoVs to increase volatility, the lognormal distribution continues to be a poor fit for the actual volatility of reserves.

The limitations of bootstrapping are well-known and many capital modellers allow for this by adjusting bootstrapping parameters upwards. However, there is a lack of clarity and consistency in the market on appropriate adjustments to make to bootstrapping results, and as a result these adjustments are typically reliant on expert judgement.

One of the key limitations of bootstrapping is that it does not allow for model error, which we define in this paper as the risk that the model used to set the best estimate is not the "true" underlying model for how reserves should be estimated. This might explain why the backtesting analysis shows that bootstrapping does not typically produce a good fit to actual historical reserve volatility.

We have investigated new methods to allow for model error within reserve risk modelling, in order to address one of the key known limitations of existing market approaches and reduce the amount of expert judgement required. To do this, we have attempted to model the uncertainty in the volatility estimated by the bootstrap. This acknowledges that the ODP model commonly used for bootstrapping might not capture all sources of uncertainty in the reserves.

The main objective of this paper is therefore to propose a new approach to reserve risk modelling which addresses the shortcomings of existing market approaches by capturing an additional source of uncertainty in the reserves.

2.3. Approach

We have investigated how successful existing reserve risk approaches are at capturing historical reserve movements. The paper is structured as follows:

- **Section 3** summarises relevant existing literature, including the use of bootstrap and lognormal approaches in reserve risk modelling, and some key limitations – in particular, the extent to which conventional approaches allow for process and parameter uncertainty but do not explicitly allow for model error.
- **Section 4** details the historical data we have considered.
- **Section 5** sets out the three key models we have considered (the raw bootstrap output; a lognormal distribution fitted to the bootstrap output by method of moments; and a compound lognormal distribution).
- **Section 6** backtests each of the three models against observed historical one-year movements in ultimate claims estimates. We also consider a number of adjustments to these models, and discuss some potential impacts of using a compound lognormal compared with a simple lognormal.
- **Section 7** sets out our conclusions from these analyses, some key limitations and some potential avenues for further research.

2.4. Scope

Time horizon: The analysis focuses on **one-year reserve movements**. For each entity and cohort, the modelled distribution is assessed against the movement between the ultimate estimate at the earlier valuation date and the corresponding ultimate estimate one year later. This provides a practical basis for backtesting, since fully developed ultimate outcomes are not available for many cohorts, particularly in long-tailed classes. However, the subsequent ultimate estimate is itself an actuarial estimate rather than an observed final settlement value, and may therefore reflect judgement and information emerging after the earlier valuation date. The results should consequently be interpreted as an assessment of how well the model captures one-year movements in selected ultimate estimates, rather than as direct validation of an ultimate reserve risk distribution.

Gross vs net: The analysis is performed on a **gross of reinsurance** basis. This avoids introducing differences in firms' non-proportional reinsurance structures, retentions and recoveries into the comparison of underlying claims volatility. The findings do not directly address the appropriateness of the proposed approach for net reserve risk, where the impact of non-proportional reinsurance may materially affect the shape of the distribution.

Geography: The analysis uses **US property and casualty insurer data** reported through the NAIC Schedule P returns. The dataset covers the 2001, 2011 and 2022–2025 filing years, selected because complete and consistent data extracts for those years were available for this research. The results are therefore based on a large and diverse US dataset, but they have not been independently validated using UK, Lloyd's or other international data. Differences in claims environments, reserving practices, reporting requirements and business mix may affect the extent to which the calibration is transferable across territories.

Methodology: The analysis focuses on an **ODP bootstrap** applied to paid claims triangles, a **lognormal distribution** parameterised from the resulting bootstrap CoV, and the **proposed compound lognormal enhancement**. While we consider other models where relevant, the analysis does not seek to assess every possible reserving method, stochastic reserving model or treatment of model uncertainty. In particular, it does not test portfolio-specific calibration of the mixing parameter, interactions between origin periods or with future inflation, or the treatment of reserve risk within a full internal model.

3. Background

Bootstrapping and the lognormal distribution are widely used by practitioners in the non-life insurance industry as key assumptions in, respectively, the parameterisation and modelling of reserve risk.

The lognormal distribution has been a mainstream modelling assumption for non-life premium and reserve risk capital since before the inception of Solvency II. Indeed, the original Solvency II standard formula calibration for non-life premium and reserve risk explicitly assumed a lognormal distribution for the underlying risk (CEIOPS, 2010), although current implementation has moved away from this, instead requiring three times the standard deviation (PRA rulebook). However, the lognormal distribution remains widely used within Internal Models, eg Fröhlich and Weng (2018) note it as being "widely used as the process distribution for the reserve risk".

England and Verrall (1999) develop a bootstrapping approach which can be used to estimate the prediction error in a reserve calculated using a chain ladder model. In 2002, England extends the bootstrap chain ladder framework by combining bootstrapping to capture estimation error with simulation to capture process error, thereby producing a full predictive distribution of reserves. LCP's 2026 London Market capital modelling benchmarking survey (published during July 2026 based on survey data collected between March and May 2026 from 38 firms, predominantly from within the Lloyd's market) discovered that 59% of participants used bootstrapping of historical data to parameterise reserve risk.

Previous research has identified some important limitations in this approach.

Chan and Ramyar (2016) highlight a number of challenges in quantifying reserve risk using stochastic techniques. One key issue is the inconsistency between the bootstrap model and the basis on which reserves are selected. Bootstrapping is often performed using paid claims triangle data and chain-ladder assumptions, whereas the reserving best estimate may reflect incurred data, multiple reserving methods, more granular analysis and actuarial adjustments. This can result in a scaling issue, where the simulated distribution is adjusted to have a mean equal to the selected best estimate, without necessarily resolving the underlying differences between the bootstrap model and the reserving process.

Another important limitation is that conventional bootstrap models generally allow for process and parameter error within a selected model, but do not explicitly allow for model error. Model error arises because any model is a simplified representation of the real world and may not adequately reflect the underlying claims process. This is distinct from parameter risk, which arises because model parameters are estimated from limited data, and process risk, which reflects the inherent randomness of future claims outcomes even if parameters were known. Following Mackey and Otto (2015), let X denote the eventual claims outcome and let $\hat{X}$ denote an actuarial prediction. The uncertainty in the prediction can be measured by the mean squared error:

$$MSE(\hat{X}) = E\left[\left(X - \hat{X}\right)^2\right]$$

Under the assumption that X and $\hat{X}$ are independent, this can be decomposed as:

$$MSE(\hat{X}) = Var(X) + Var(\hat{X}) + \left(E[X] - E[\hat{X}]\right)^2$$

Conventional bootstrap approaches estimate this uncertainty conditional on a single selected model being representative of the underlying claims process. Model error arises where this assumption does not hold, that is, where the selected model is an imperfect representation of the process generating X. More explicitly, if X_M is the outcome distribution implied by model M, model error can be represented as the discrepancy ε_M between the true outcome and that implied distribution, such that $X = X_M + \varepsilon_M$. A conventional bootstrap effectively assumes that:

$$\varepsilon_M = 0$$

Allowing for model error instead recognises that this discrepancy may be non-zero and uncertain.

The importance of these limitations was demonstrated by Leong, Wang and Chen (2012), who backtested the ODP bootstrap of the paid chain-ladder method against US Schedule P data. They found that actual reserve outcomes occurred in the tails of the modelled distributions more frequently than expected, indicating that the bootstrap understated observed reserve volatility. The authors attributed this to systemic sources of uncertainty that are not captured by the bootstrap, that is the risk that future trends in the claims environment may deviate from what was seen in the past. They proposed a number of adjustments to allow for this additional risk.

This paper follows the same broad backtesting principle. We assess one-year movements in ultimate claims estimates across multiple filing years and reserving classes, with each model fitted using only information available at the earlier valuation date. We apply the method to a separate, expanded Schedule P dataset using independently selected data cleaning, cohort definition and implementation choices.

We then extend the analysis beyond the raw ODP bootstrap. Alongside the bootstrap results, we test a lognormal distribution parameterised using the bootstrap coefficient of variation (CoV), reflecting a common practical implementation in capital models. We also test the effect of applying a uniform uplift to the bootstrap CoV, representing a pragmatic allowance often made for uncertainty not captured by the base model.

We then consider a compound lognormal framework. This retains the bootstrap CoV as the principal portfolio-specific measure of volatility, but introduces a mixing distribution to allow for uncertainty around the volatility parameter used to construct the lognormal distribution. The intention is not to suggest that this fully captures every source of model error, nor that a single distribution can describe all reserve risk outcomes. Rather, it provides a transparent and empirically testable way of introducing an allowance for uncertainty that is not captured when the bootstrap-derived CoV is treated as known with certainty.

The approach is therefore best viewed as a practical enhancement to the conventional bootstrap-plus-lognormal framework. It aims to reduce reliance on purely judgemental CoV uplifts, while retaining a direct link between the simulated reserve risk distribution and the underlying claims data.

It is important to note that the analysis of historical reserve movements we present is all of one-year movements. It is common for capital models to assume that ultimate reserve movements are lognormally distributed, and then apply one-year recognition to derive the respective one-year movement. Recognition is commonly applied via linear transformation, ie

$$X_{1yr} = E(X) + \rho[X_{ult} - E(X)]$$

where X_{1yr} and X_{ult} are the one-year and ultimate reserve risk distributions for a given portfolio respectively, E(X) is the mean of the ultimate distribution, which is assumed to be preserved by the transformation, and ρ is an emergence factor, typically between 0 and 1.

Therefore, although many capital models explicitly assume that ultimate reserve movements *are* lognormally distributed, they typically implicitly assume that one-year reserve movements are *not*. This arises because a linearly transformed lognormal distribution is, in general, not itself lognormally distributed.

We note this primarily because, if implementing this approach within a generic capital model, it is likely that the approach would involve modelling ultimate reserve movements as compound lognormal, despite these having *not* been shown to behave in line with the compound lognormal, while one-year reserve movements – which we *have* shown behave more in line with the compound lognormal – would be modelled as an altogether different distribution.

4. Data

This section provides a brief overview of the main dataset we have relied upon for the analyses performed. Some further details, in particular a description of any data cleaning that we carried out, are provided alongside the descriptions of the relevant analyses.

4.1. Schedule P

We performed our analysis on the Schedule P dataset from the National Association of Insurance Commissioners (NAIC). This data includes claims information for major personal and commercial lines for all property and casualty insurers that write business in the US. This dataset was used because it is large enough to provide credible results, and easy to analyse as the data is provided in a consistent format.

The data contains paid claims triangles and historical ultimate loss estimates across multiple lines of business for 10 years of account (cohorts) and 10 years of development, providing a broad cross-section of market experience across different calendar periods and economic conditions.

Following some high-level cleaning, detailed in section 4.2, the dataset used in the backtesting analysis has total reserves of around $3.1tn. It is structured as follows:

- **Filing years**: The data spans the six filing years which were immediately and readily available to us: 2001, 2011, 2022, 2023, 2024 and 2025.
- **Reserving classes**: Within each filing year, claims are split across up to 23 classes.
- **Companies**: The dataset includes 2,381 individual companies.
- **Entities**: We define an entity as a unique company/class/filing year combination.
- **Triangles**: For a given filing year, an entity's claims history forms one paid claims triangle on which the bootstrap is run, and associated ultimate claim estimates. There are 38,714 entities in total, each corresponding to one paid triangle.
- **Cohorts and backtesting rows**: Each triangle spans 10 cohorts, and the bootstrap produces a separate result for each cohort within a triangle. A row in the backtesting dataset therefore corresponds to a unique entity/cohort combination, of which there are 177,564 in total.

The charts below show, for each reserving class (Figure 1) and filing year (Figure 2), total reserves as a bar (measured against the top/left axis, in $bn) and the number of entities as a dot (measured against the bottom/right axis).

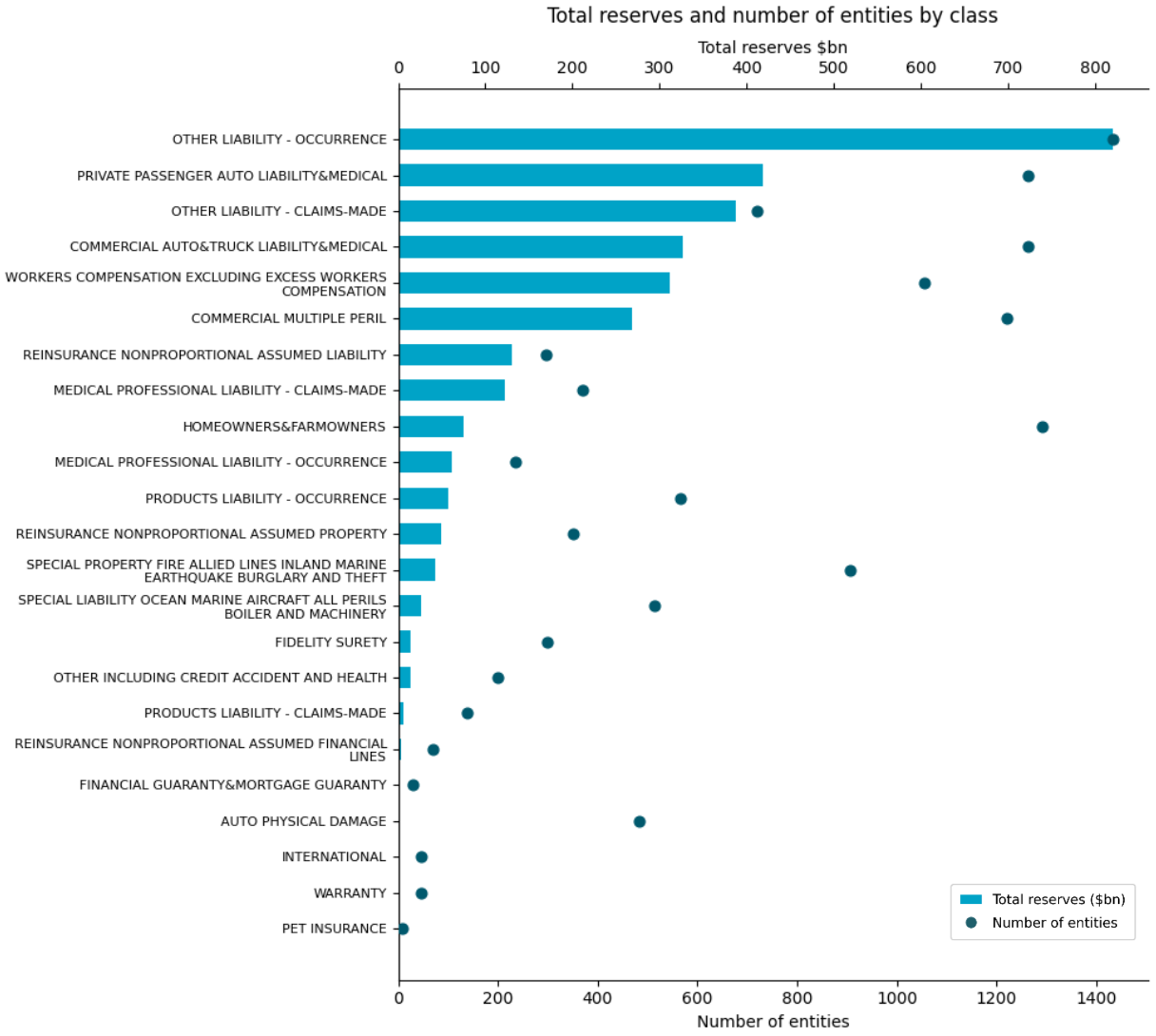


Figure 1: Schedule P reserves and entities by class

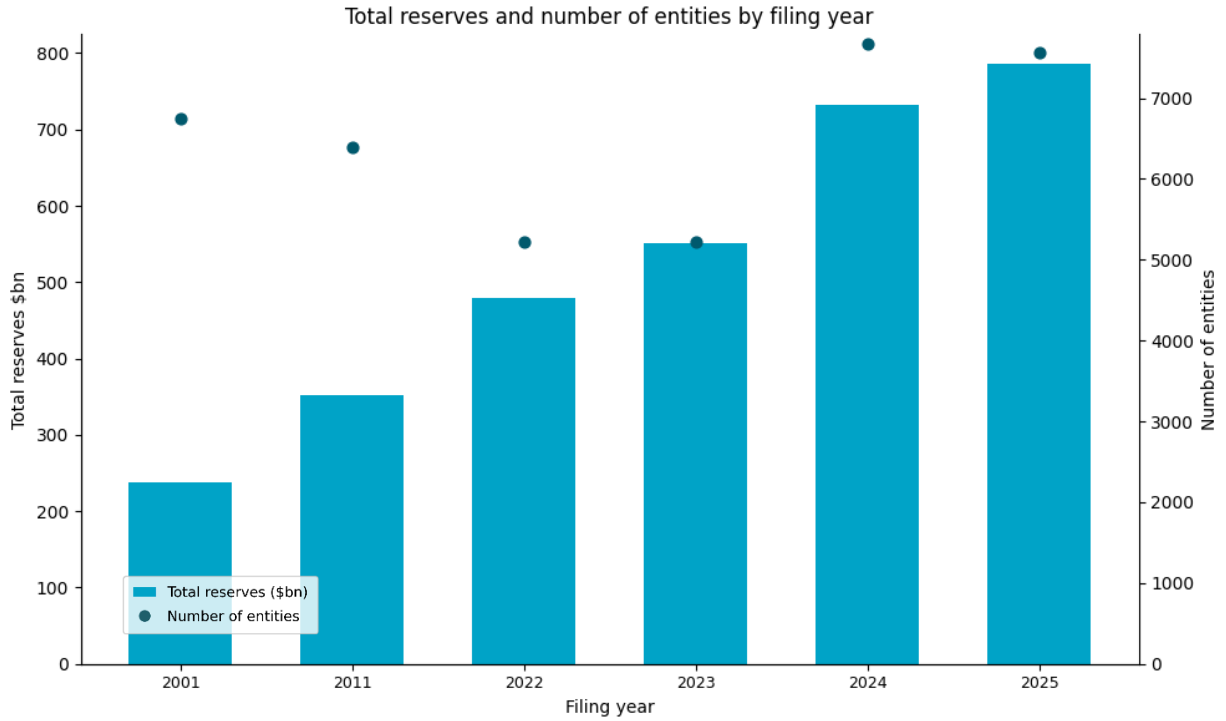


Figure 2: Schedule P reserves and entities by filing year

4.2. Data cleaning

We applied a series of cleaning and filtering steps to the data before carrying out the analysis. These steps were intended to ensure that the bootstrap could be applied consistently across the large dataset, and that the resulting distributions were credible.

The cleaning process comprised two stages. First, we applied pre-processing filters before running the bootstrap. This included retaining only entities with both paid claims and ultimate loss estimate data, requiring complete paid claims triangles, and excluding very small companies with total reserves below $1m. These filters removed incomplete or unstable triangles while preserving above 99.9% of the total reserve volume of almost $7tn in the dataset. The subsequent backtesting analysis considers a narrower view of the data, as the latest cohort for each filing year contains only one development point and therefore cannot be backtested. As a result, the reserve volume considered in the backtesting analysis is reduced to $4.2tn.

Second, we applied post-bootstrap filters to remove observations where the bootstrap output was not suitable for backtesting. This included excluding cases where the bootstrap failed to run, where prior or latest ultimate estimates were non-positive, where the simulated distribution had no variation, where the bootstrap CoV was implausibly low or high, and where prior and latest ultimate estimates were exactly identical. These checks were intended to remove cases where the comparison between the simulated reserve distribution and the observed one-year movement would not be meaningful.

The post-bootstrap filtering was more material than the pre-processing filtering. In particular, a substantial number of entity/cohort/filing-year combinations were removed because the automated bootstrap produced zero, undefined or implausible CoVs. We view this as an acknowledged limitation of applying a single automated bootstrap methodology across a very broad market dataset, rather than as a flaw specific to any individual class or filing year. The retained population remains substantial, comprising 177,564 entity/cohort/filing-year combinations, with total reserves of around $3.1tn.

Overall, the cleaning process removes observations where the bootstrap is unlikely to provide a reliable basis for backtesting, while retaining a large and diverse dataset across filing years, classes of business and companies. Further detail on the filters applied, including rationale, the number of observations removed at each stage and the impact by class and filing year, is provided in Appendix A.

5. Models

5.1. Bootstrap

Methodology

Bootstrapping is a resampling technique used to estimate the distribution of reserves from which a CoV can be derived. We use an Over-dispersed Poisson (ODP) bootstrap framework. See Appendix B for more detail on this distribution. This is equivalent to fitting a chain ladder model and is one of the most widely used implementations of bootstrapping.

This process can be summarised in five steps:

1. Fit a generalised linear model (GLM) with a log-link function and ODP error distribution to the incremental claims data. This is identical to fitting a chain ladder to the triangle.
2. Calculate the residuals from the fitted model.
3. Resample (with replacement) the residuals to create a pseudo triangle.
4. Refit the GLM using the pseudo-dataset to produce a new estimate of the reserves.
5. Repeat steps 3 and 4 n times to derive a forecast output for each pseudo-dataset. This gives a distribution of reserve estimates.

We selected $n = 10{,}000$. We therefore had, for every cohort in every reserving triangle, 10,000 reserve estimates.

See Appendix C for further detail on the bootstrapping methodology.

Benefits, uncertainty and limitations

The ODP bootstrap has a number of advantages. It is data-driven and does not rely on ultimate claims estimates, which are by nature subjective and prone to biases. It is well-established in actuarial practice and links naturally to lognormal parameterisation, making it a convenient and interpretable starting point for reserve risk modelling.

The ODP bootstrap is a well-established method, but it carries a number of known limitations that shaped how we applied it.

The first is computational cost. Generating bootstrap distributions across a dataset of this size is demanding, and we used LCP InsurSight's reserve risk tool to manage the data volumes and perform the calculations at scale.

The second concerns the stability of the resulting estimates. The bootstrap resamples from a finite set of residuals, so where a triangle contains few observations the resampled distribution is itself subject to considerable estimation error. The method is also sensitive to outliers: a single unusual development period, whether genuine or the result of a data artefact, can materially inflate the estimated CoV. Both effects are most material for smaller portfolios, where there is less data over which anomalies can average out. We therefore restricted our analysis to companies with reserve sizes greater than $1m, on the basis that the resulting CoV estimates are more stable and less likely to be dominated by individual data points.

The third is a constraint on the data to which the method can be applied. The ODP bootstrap requires non-negative incremental development at each step, which in practice means it is often only viable on paid claims data. Incurred data frequently exhibits negative increments, for instance where case reserves are released as claims settle below the initial estimate, or where subrogation and salvage form part of the claims process. A consequence is that, for example, bootstrapped results derived from paid data may not align with ultimates projected on an incurred basis, as discussed by Chan and Ramyar (2016).

Some further limitations arise specifically from the nature of this exercise. Because the calculations were performed at scale across a large number of triangles, we did not apply the actuarial judgements that would ordinarily form part of a bootstrapping exercise on an individual portfolio. In particular:

- No smoothing. Development factors were taken as calculated, without adjustment for irregularities in the observed pattern.
- No exclusion of outlier link ratios. Individual link ratios were retained regardless of how far they departed from the surrounding pattern.
- No tail factors. Development beyond the observed triangle was not extrapolated.
- No deep-dive reviews of individual triangles. Some triangles exhibit unusual development patterns that would ordinarily prompt investigation or adjustment.

The filters applied during data cleaning, described in section 4.2, partially mitigate these limitations.

5.2. Lognormal

Methodology

The lognormal distribution is one of the most commonly used distributions to model reserve risk. A positive random variable X is lognormally distributed – ie, $X \sim \text{lognormal}(\mu, \sigma^2)$ – if the natural logarithm of X is normally distributed with mean μ and variance σ^2:

$$\log X \sim \mathcal{N}(\mu, \sigma^2)$$

Figure 3 illustrates the shape of the lognormal density for a range of values of σ, with μ set equal to zero. As σ increases, the distribution becomes more dispersed and more heavily right-skewed.

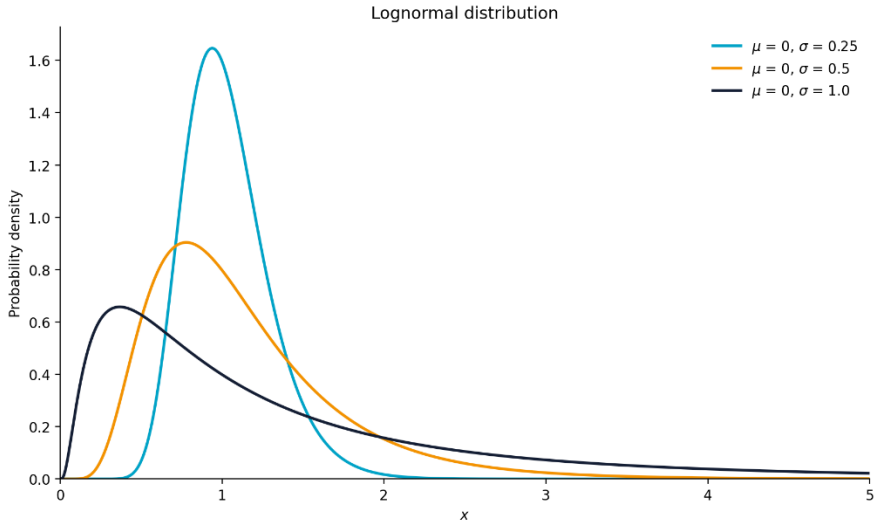


Figure 3: Some lognormal distributions

When modelling reserve risk, this approach assumes that the ultimate claims outcome, X, follows a lognormal distribution parameterised by a parameters μ and σ, where σ is typically derived from the coefficient of variation (CoV) estimated via a bootstrap. Specifically, for a given best estimate m_0 and bootstrap CoV, the lognormal parameter σ is given by:

$$\sigma_0 = \sqrt{\log(1 + \text{CoV}^2_{bootstrap})}$$

and the parameter μ is set to preserve the mean:

$$\mu = \log(m_0) - \frac{1}{2}\sigma_0^2$$

This parameterisation ensures that $\mathbb{E}[X] = m_0$ for any value of σ.

Benefits, uncertainty and limitations

The lognormal distribution is a convenient and tractable choice for reserve risk modelling. It is non-negative and right-skewed, consistent with the typical behaviour of reserve movements, and requires a single parameter (CoV), for which parameterisation via bootstrapping provides a direct and interpretable link between the empirical data and the distributional model.

However, the lognormal is, like all models, a simplification of reality, and has limitations. First, it imposes a fixed distributional shape, which may not adequately capture the full range of reserve outcomes. When using a CoV derived from a bootstrapping exercise, the lognormal distribution may understate (or overstate) tail risk, where the true distribution of outcomes may be more (or less) heavily skewed or fat-tailed than modelled.

Second, parameterisation using bootstrapped CoVs treats both the best estimate and the CoV as known with certainty. In practice, neither is truly deterministic: the bootstrap CoV is itself an estimate, subject to sampling and model uncertainty, and treating it as fixed may understate the true uncertainty in the reserve outcome. A common market response to this is to uplift the bootstrap CoV to compensate.

43% of firms in LCP's 2026 London Market capital modelling benchmark survey uplifted reserve risk CoVs for social inflation, while 62% uplift for other factors. However, simply scaling the CoV widens the entire distribution simultaneously, affecting both the body and the tails. While this may improve the fit in the right tail, in particular around the 1 in 200 return period which is of most interest for capital modelling purposes, it does so at the cost of making the body of the distribution less reliable for decisions at lower return periods.

5.3. Compound lognormal

Methodology

The lognormal parameterisation set out in section 5.2 treats both the best estimate m_0 and the volatility parameter σ as known with certainty, once the bootstrap has produced a CoV.

In practice, either (or both) of these quantities could instead be treated as uncertain: one could introduce a distribution around the best estimate itself, around the volatility parameter, or around both simultaneously, each capturing a different source of model or parameter uncertainty.

In this paper, we focus specifically on uncertainty around the volatility estimate. We refer to this as the compound lognormal framework. We relax the assumption that σ is known with certainty; instead, σ is treated as a random variable. Specifically, we consider a framework in which σ follows a lognormal distribution:

$$\sigma \sim \text{lognormal}\left(\mu_{\sigma_0}, \tau\right)$$

where μ_{σ_0} and τ are the log-scale mean and standard deviation of σ respectively, derived from a mean equal to σ_0 and a coefficient of variation CoV_σ:

$$\tau = \sqrt{\log(1 + \text{CoV}_\sigma^2)}$$

$$\mu_{\sigma_0} = \log(\sigma_0) - \frac{1}{2}\tau^2$$

The marginal distribution of X is then obtained by integrating over the distribution of σ:

$$f_X(x) = \mathbb{E}_\sigma[f_{X|\sigma}(x)]$$

This mixture does not have a closed-form expression, and therefore needs to be evaluated numerically via Monte Carlo simulation. For each simulation, a value of σ is drawn from its lognormal distribution, μ is recomputed to preserve the mean at m_0, and X is drawn from the resulting lognormal distribution. This process is repeated n_{MC} times to generate the full marginal distribution of X.

A key property of this parameterisation is that the mean of X is preserved at m_0 regardless of the realised value of σ, since:

$$\mathbb{E}[\,X \mid \sigma\,] = \exp\left(\mu(\sigma) + \frac{1}{2}\sigma^2\right) = \exp(\log(m_0)) = m_0$$

and therefore $\mathbb{E}[X] = \mathbb{E}_\sigma\big[\mathbb{E}[\,X \mid \sigma\,]\big] = m_0$. The compound lognormal therefore separates uncertainty about the spread of outcomes from uncertainty about the central estimate, with the best estimate remaining the anchor of the distribution throughout.

The compound lognormal has three inputs:

1. the best estimate m_0
2. the bootstrap $\text{CoV}_{bootstrap}$
3. CoV_σ

The first two are portfolio-specific and can generally be derived from the claims data. The third, CoV_σ, could (in principle) also be estimated on a portfolio-specific basis; for example, from the historical stability of bootstrap CoV estimates for a given class of business or entity over time. However, in this paper, we estimate a single value of CoV_σ uniformly across the whole dataset, selected on the basis of the backtesting results set out in section 6, rather than attempting to estimate it separately for each portfolio. The reason for this is that, in our view, it is likely that any portfolio-specific estimate of CoV_σ would be subject to considerable uncertainty, resulting in difficulty in applying the compound lognormal framework in practice.

Benefits, uncertainty and limitations

The compound lognormal offers several advantages over the standard lognormal. It produces a distribution with heavier tails and a narrower body, which is theoretically consistent with what we might expect from a model that explicitly accounts for parameter uncertainty. The parameterisation separates the portfolio-specific volatility estimate, captured by the bootstrap CoV, from the uncertainty around that estimate, captured by CoV_σ. The latter can be calibrated to industry data rather than set by expert judgement, giving a basis for the assumption that is empirically grounded and can be audited.

The main additional limitation is the introduction of an additional parameter CoV_σ, which requires calibration and may not be directly observable from portfolio data. We address this by calibrating CoV_σ as a universal parameter across the full dataset, and by demonstrating stability of results across a range of values. A further consideration is the computational cost of the Monte Carlo simulation, though in practice the additional run time is modest.

Otherwise, the compound lognormal retains many of the benefits and limitations of the lognormal, as discussed in section 5.2.

6. Backtesting results

6.1. Actual historical deteriorations

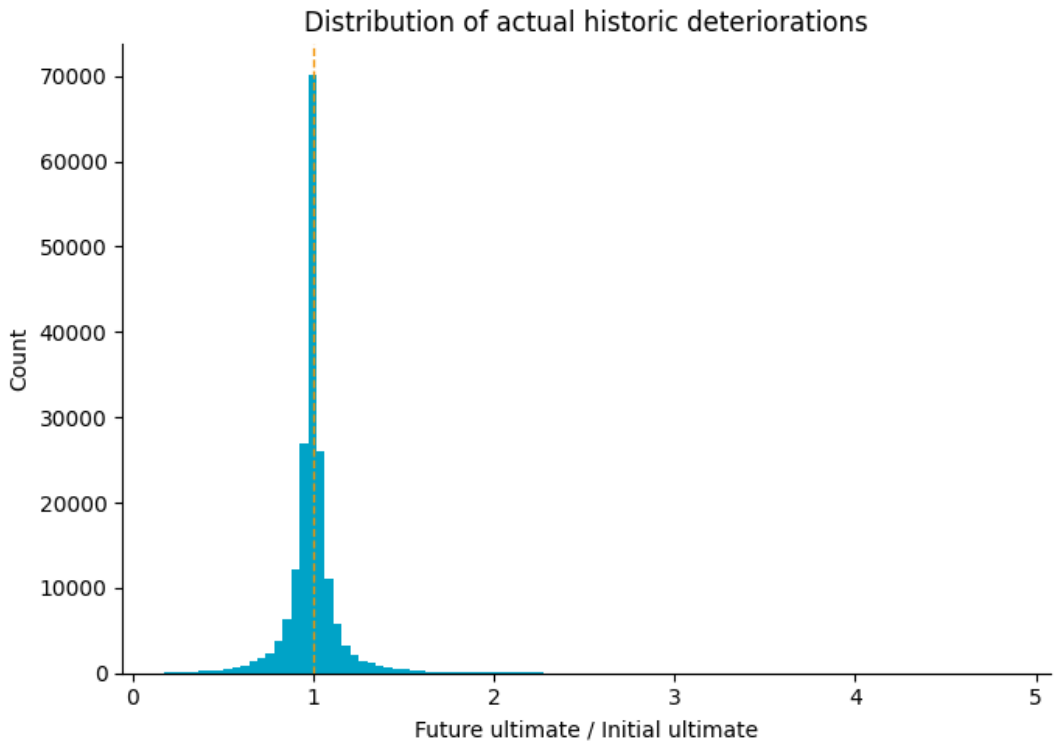


Figure 4: Summary of backtesting results

Figure 4, above, shows the distribution of actual one-year historical deteriorations across the cleaned dataset, measured as the ratio of the future ultimate to the initial ultimate. Each data point represents the deterioration for a cohort of a particular entity. The distribution is tightly centred around 1.0 (ie zero change in the ultimate estimate over the year), with a median of 0.998 and an interquartile range of 0.96 to 1.03, indicating ultimate estimates generally being stable from one evaluation to the next. The distribution does exhibit a longer right tail than left tail, demonstrating that when deteriorations do occur, adverse movements (increases in the ultimate) tend to be larger in magnitude than favourable ones. This is expected, given that there is a natural "floor" on ultimate movements at zero, but no equivalent natural "ceiling". To aid with clarity, the chart above is truncated to the 0.1st–99.9th quantile range.

6.2. Bootstrap

Bootstrap data and methodology

We used the bootstrap methodology described in section 5.1. Each filing year provides ten years of data, ie the triangles each have ten origin periods and ten development periods. We performed the bootstrap after removing the leading diagonal from each entity's paid claims triangle, ie we bootstrapped triangles with nine origin periods and nine development periods. For every cohort in every triangle, we therefore had a distribution of ultimate claims, made up of 10,000 estimates of the ultimate position as at prior year view (time 9) from the bootstrap methodology.

Figure 5 shows the 10 x 10 triangles we are using. The bootstrap is performed on a 9 x 9 triangle (ie with the blue entries on the latest diagonal removed) and the results are then compared with the ultimate position on the 10 x 10 triangle (with the blue entries restored).

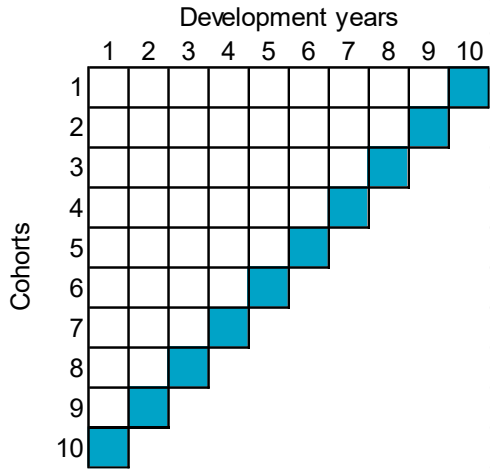


Figure 5: Schematic of triangles used in backtesting

Scaling the bootstrap

We scaled this distribution such that the mean was equal to the "actual" ultimate claims selected by the actuary as at prior year view (time 9) from the Schedule P data. This gave us a distribution of ultimate claims for each cohort in each triangle, centred around the selected ultimate.

We used multiplicative scaling, ie we multiplied each of the 10,000 estimates of ultimate claims by a factor α, where:

$$Adjusted\ ultimate_j = Ultimate_j * \alpha$$

such that:

$$\alpha = \frac{A\ priori\ (time\ 9)\ "actual"\ ultimate\ claims\ from\ Schedule\ P\ data}{A\ priori\ (time\ 9)\ mean\ of\ unadjusted\ distribution\ of\ ultimate\ claims\ from\ bootstrapping}$$

We then performed backtesting to assess the fit of the bootstrapped distribution. We took the "actual" ultimate claims as at the updated view, with the latest diagonal restored (time 10) from the Schedule P data. We then compared these updated ultimate claims estimates to the distribution obtained via bootstrapping. For each cohort, we calculated the percentile of the actual ultimates, using the distribution of adjusted ultimates.

If the bootstrap model was an accurate indication of actual reserve movements, the resulting empirical distribution of the calculated percentiles across all cohorts should be uniform. For example, 10% of entities and cohorts would have an updated ultimate that falls above the 90th percentile, and 10% should fall below the 10th percentile.

Figure 6 shows the results of backtesting. The chart plots the empirical distribution of the observed percentiles from all classes and cohorts. There are fifty buckets, each showing two percentiles of the overall distribution.

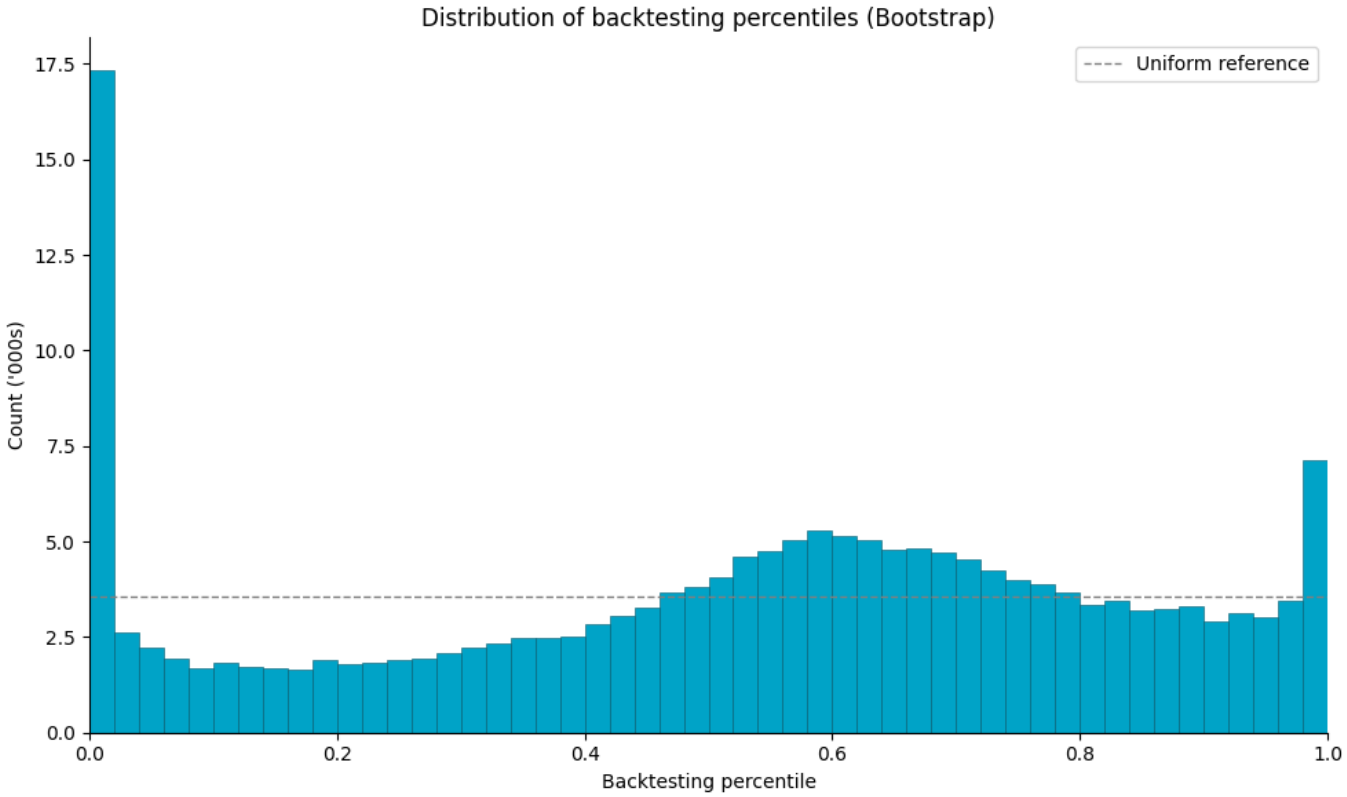


Figure 6: Backtesting the bootstrap

There is a peak in the first bucket, which indicates there are more cases falling into the lowest percentiles than would be expected if the bootstrapped distribution was a good fit. These are cases of ultimates at t = 10 falling below the entirety of the simulated distribution, ie out of 10,000 simulations, the bootstrap does not generate a single possible outcome where ultimate claims are as low as then occurred in reality. There is a similar (although smaller) peak in the last bucket, indicating that there are more cases than expected of updated ultimate positions being greater than the entirety of the simulated distribution. Finally, there is also a peak around the 60th percentile. Overall, the bootstrapped distribution is not a good model for actual one-year reserve deteriorations.

Note that we have not quoted goodness of fit metrics associated with this result or with many of the results that follow. This is because visual inspection alone makes clear that the fit is materially unsatisfactory; consequently, such metrics would add little to the assessment.

6.3. Lognormal

We performed a similar analysis on the fitted lognormal distribution, using the methodology described in section 5.2. For each cohort, we calculated the percentile of the actual t=10 ultimates, using a lognormal distribution fitted by method of moments.

The aggregated distribution of calculated percentiles is shown in Figure 7, again tested against an assumption of uniformity against historical observations.

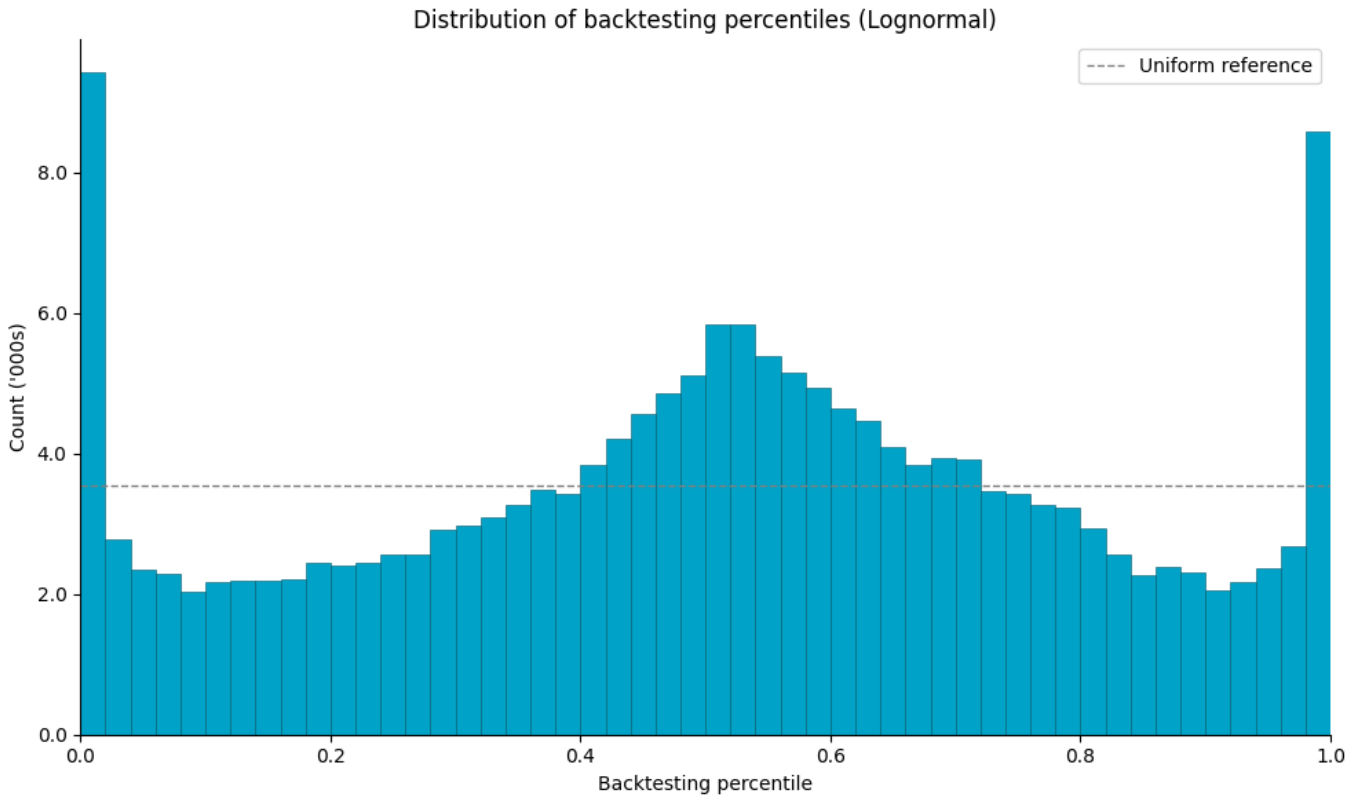


Figure 7: Backtesting the lognormal fit to the bootstrap

There are similar peaks in the first and last buckets, and a more pronounced peak around the median. This shows that a lognormal distribution fitted to the bootstrap via method of moments is not a good fit to actual one-year reserve movements.

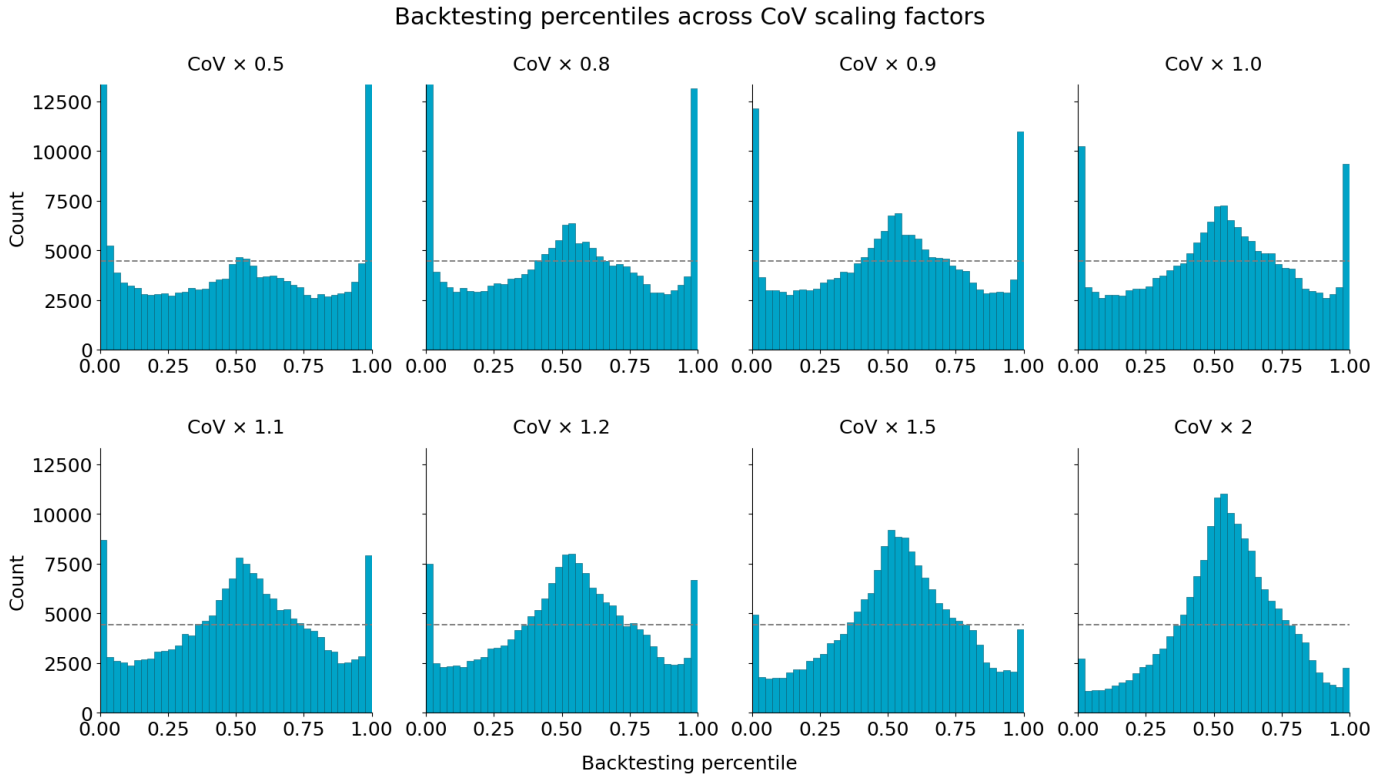
This result is not surprising, since the lognormal is fitted directly to the bootstrapped distribution via method of moments, which is a common approach for selecting reserve risk CoVs for use in capital models, and adds no new information. It is generally assumed that this provides a reasonable representation of the raw bootstrapped distribution – the similarity of the results here and the results for the raw bootstrap in section 6.2 provide some indirect support for this assumption. We note that the results here are slightly better than for the raw bootstrapped data (in particular, the peaks at either end are more balanced). This feature may be of interest to practitioners trying to better understand the implications of fitting lognormal CoVs based on bootstrapped distributions, but does not help with our problem of whether this is a sound approach in the first place.

See Appendix D for further analysis which demonstrates that the result is generally consistent when considering splits by class, filing year, and reserve volume, and that results are not materially driven by segmentation approach.

6.4. Scaled lognormal

In our experience, practitioners are typically keenly aware that the bootstrap may understate reserve volatility. A commonly-made adjustment, in particular to ensure tail risk is captured appropriately for capital setting purposes, is to judgementally increase the CoV from the bootstrap. We therefore attempt to mimic this by scaling the CoVs

obtained from bootstrapping in order to produce a better fit to the historical data. The results of this analysis are shown in Figure 8.



Note: bars exceeding the y-axis limit are clipped for readability

Figure 8: Backtesting scaled lognormals fit to the bootstrap

Reducing the CoV flattens the peak in the middle, by narrowing the body of the distribution, but at the cost of increasing the peaks in the tails as larger reserve movements are pushed into the tails of the distribution. Conversely, increasing the CoV stretches the distribution and has the desired effect of reducing the peaks in the tails, ie the first and last buckets. However, it forces more observations into the centre.

None of these results produces a satisfactory fit to historical observations. The outcome is perhaps closest to a uniform distribution around a scaling factor of 1.2, ie a 20% uplift to bootstrapped CoVs. Very anecdotally, this is not far out of line with market practice, ie a small to moderate uplift to bootstrapped CoVs is fairly typical.

6.5. Backtesting the compound lognormal

We then assessed the results from backtesting using compound lognormal distributions, as described in section 5.3. We assessed a range of different values for CoV_σ. As shown in Figure 9, $CoV_\sigma \in [0.75,1.0]$ produces much more uniform backtesting results than the previous models. This indicates the model is a better fit to observed historical reserve deteriorations than a simple lognormal, or to raw bootstrapped output.

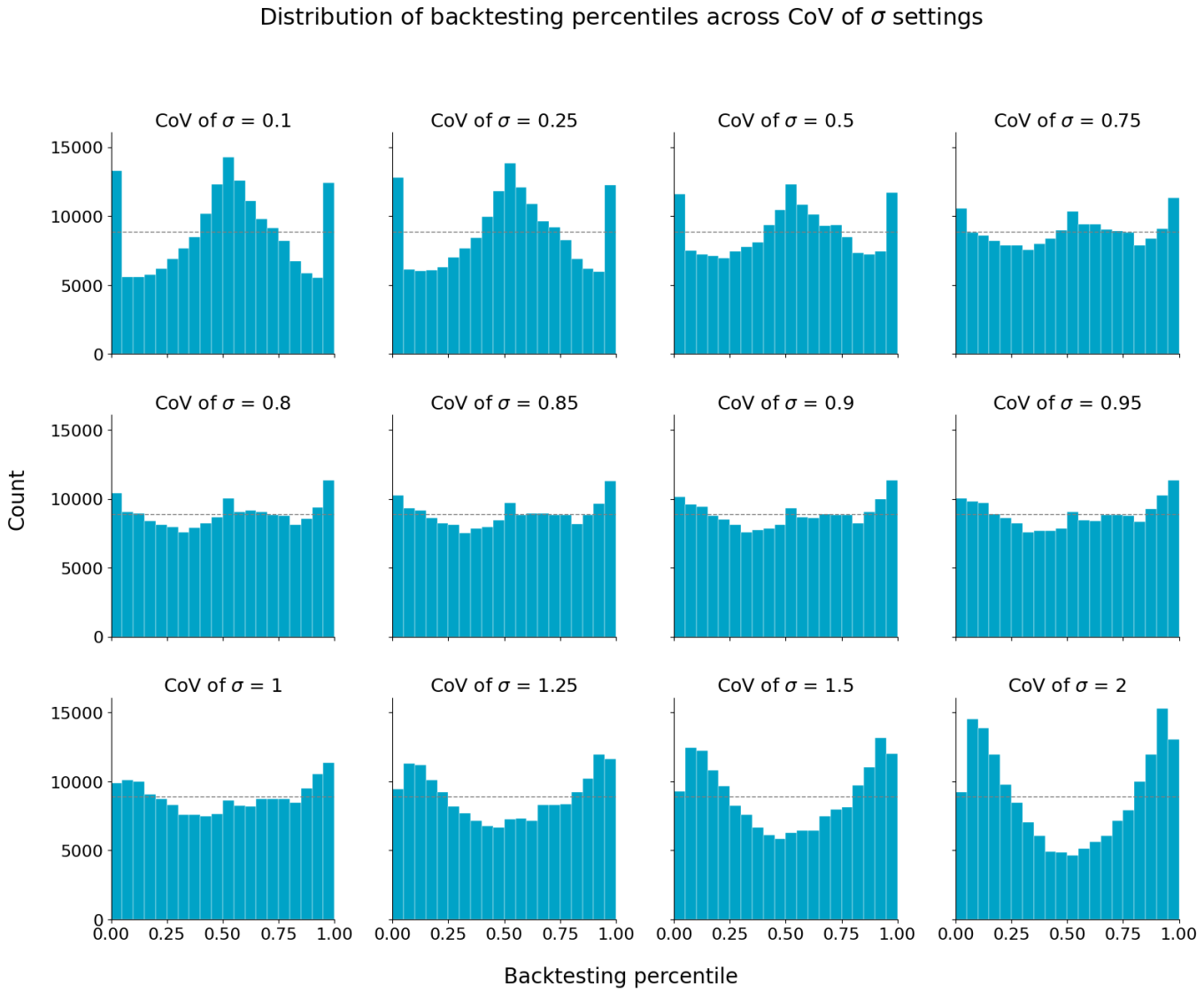


Figure 9: Backtesting compound lognormals fit to the bootstrap

For each CoV of σ, we tested their agreement with the theoretical uniform(0,1) distribution using three tests: a chi-squared goodness-of-fit test on 20 equal-width bins, the Cramér-von Mises (CvM) test, and the Kolmogorov-Smirnov (KS) test. Table 1 shows goodness of fit statistics for different values of CoV_σ. Because sample sizes were large, we have treated the resulting statistics as relative rankings across CoV values rather than as formal hypothesis tests, since even minor deviations from uniformity become statistically significant at large n.

CoV of σ	X^2	CvM	KS
0.75	2,050	25.6	0.023
0.80	1,754	25.0	0.021
0.85	1,668	27.7	0.019
0.90	1,756	33.3	0.021
0.95	2,009	41.7	0.023
1.00	2,416	53.0	0.027

Table 1: Potential values for CoV_σ

These results suggest that a CoV of σ = 0.8 or 0.85 fits best. However, it is important to note that none of these results are statistically significant; ie, none result in a backtesting distribution which is sufficiently close to uniform at any conventional significance level, and any value for $CoV_\sigma \in [0.75,1.0]$ produces reasonable results that visually improves upon a traditional lognormal approach.

Note that we have fitted for the volatility around σ. It is also possible to parameterise for the volatility around the CoV directly. We have elected to focus on CoV_σ rather than CoV_{CoV} because our analysis indicates that it tends to produce better fits. We note that the two potential approaches tend to imply similar levels of volatility. Models for CoV_{CoV} are also discussed in section 6.7.

See Appendix E for further analysis, which demonstrates that the result is generally consistent when considering splits by class, filing year, and reserve volume, and that this result is not materially driven by segmentation approach.

6.6. Impacts of the compound lognormal

We generated pseudo-distributions to assess the impact of this compound lognormal distribution compared with a traditional simple lognormal approach. Figure 10 shows, for $CoV_{bootstrap} = 30\%$ and $CoV_{\sigma} = 80\%$, that the compound places more probability in the higher peak and heavier tails (shown with green shading). The higher the value of CoV_{σ}, the more pronounced this effect. This is consistent with the results shown in the previous section, which shows that the distribution of historical reserve deteriorations is characterised by a higher central peak than can be achieved by a simple lognormal distribution. Further examples are shown in Appendix F.

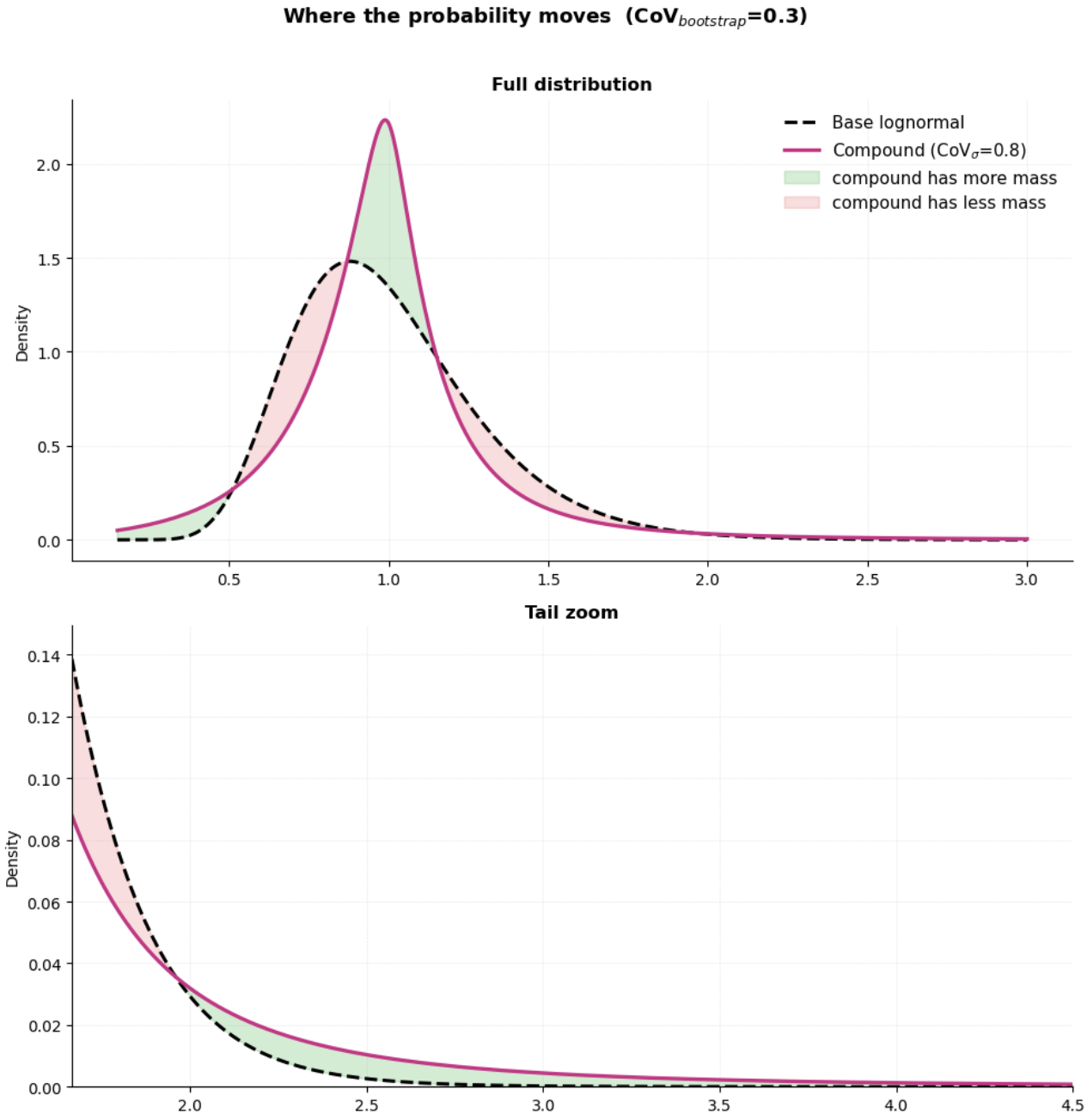


Figure 10: Simple vs compound lognormals

We then analysed the drivers of overall volatility of the distribution. Figure 11 and Figure 12 show the 99.5th percentile of the compound distribution using different combinations of $CoV_{bootstrap}$ and CoV_{σ}.

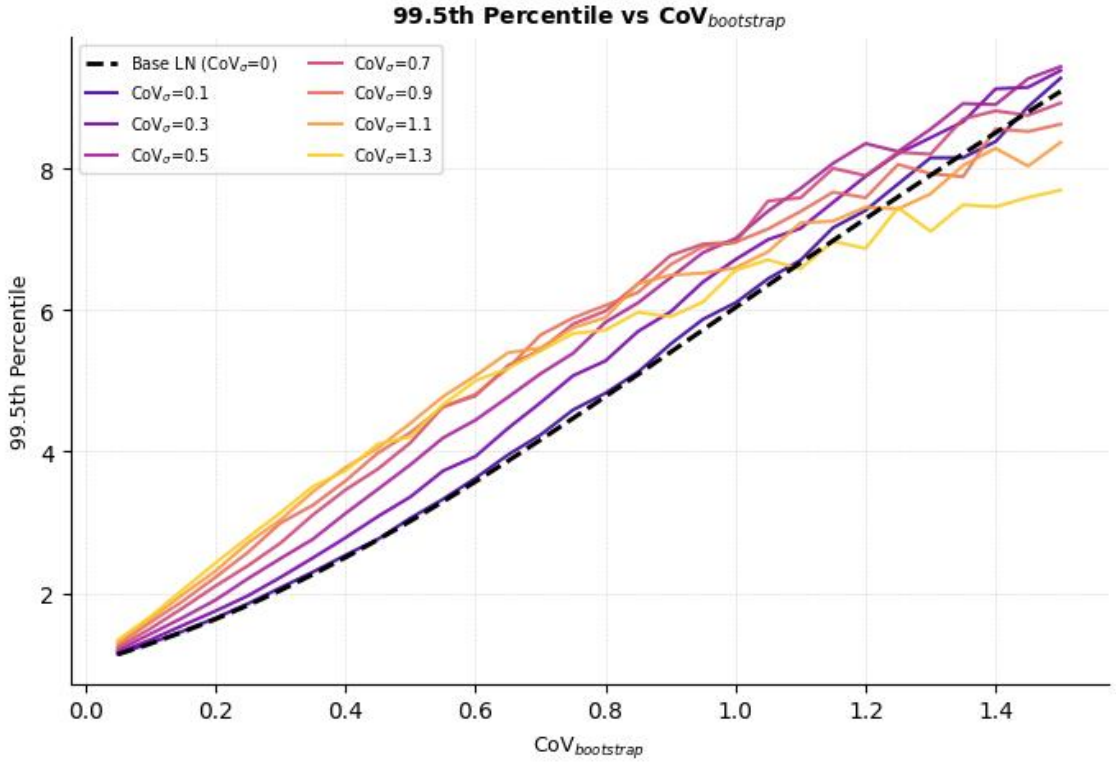


Figure 11: Bootstrap CoV vs Sigma CoV uncertainty (A)

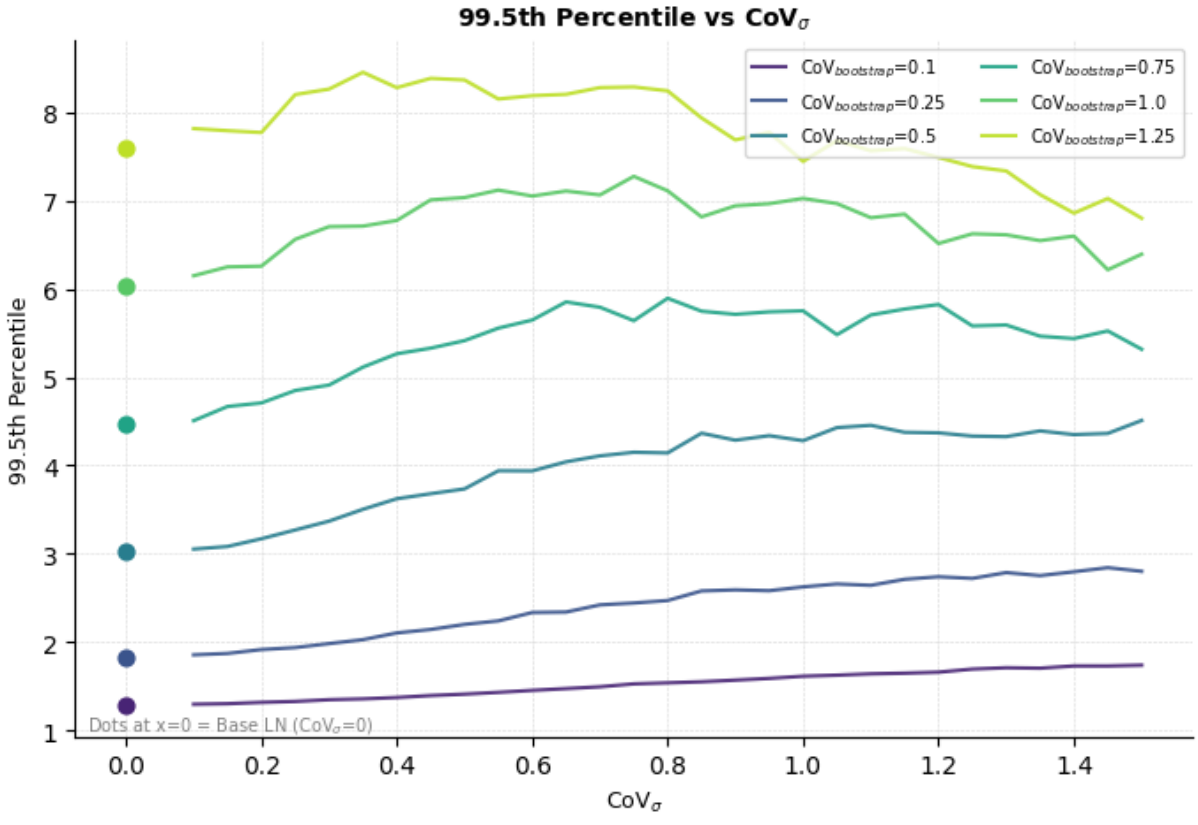


Figure 12: Bootstrap CoV vs Sigma CoV uncertainty (B)

Figure 11 shows that, for a fixed value of CoV_σ, the 99.5th percentile of the compound distribution increases approximately linearly with $CoV_{bootstrap}$, tracking the simple lognormal reference (dashed line) across most of the range. Introducing uncertainty to CoV_σ produces a modest shift relative to this baseline, but the curves for different CoV_σ values remain clustered, indicating that bootstrap uncertainty is the dominant driver of spread. In other words, the 99.5th percentile of the compound distribution is much more heavily influenced by $CoV_{bootstrap}$ than by CoV_σ.

Figure 12 isolates the effect of CoV_σ directly. Here, the compound distribution's 99.5th percentiles increase sharply with $CoV_{bootstrap}$, visible as a clear vertical separation between curves, but are comparatively insensitive to CoV_σ itself, with each curve remaining largely flat as CoV_σ increases from 0.1 to 1.4. Notably, none of the lines intersect. This again indicates that, for some of the values most likely to be selected for reserve risk modelling, $CoV_{bootstrap}$ contributes substantially more to the overall volatility in the compound lognormal distribution than CoV_σ, which plays a secondary role.

This provides comfort around the selection of the CoV_σ, as it is a less material driver of key metrics of the compound distribution, such as the 99.5th percentile. It is still the bootstrap CoV, which is calculated directly from the triangle, that drives overall volatility, but introducing CoV_σ provides a markedly better fit to historical observations throughout the body of the distribution.

6.7. Alternative models

The framework presented above introduces uncertainty in σ. This was not the only option considered: we also tested models that treated the central estimate as uncertain (both μ and the mean directly), models that treated the coefficient of variation as uncertain, and models that treated μ and σ as uncertain together, under a range of distributions for that uncertainty. Appendix G sets out the alternative structures and mixing distributions tested.

Of particular note is the finding that the choice of mixing distribution is of second-order importance. We have tested a number of different distributions, which all produce similar estimates for CoV_σ, and similar goodness of fit statistics.

The data suggest that the gamma distribution produces a similarly good fit (with a “best fit” CoV_σ of 0.7). We have suggested the lognormal distribution as an appropriate mixing distribution for practical use on the basis of familiarity and because our analysis indicates that the range of appropriate gamma parameters is smaller than the equivalent range for lognormal parameters, implying that lognormal is a more robust choice.

7. Conclusions and next steps

7.1. Conclusions

The backtesting results indicate that raw ODP bootstrap output does not provide a satisfactory representation of the observed distribution of historical reserve movements. In particular, actual outcomes occur outside, or close to the edges of, the simulated bootstrap distributions more frequently than would be expected if the model were well calibrated. A lognormal distribution parameterised using the bootstrap CoV produces a similar result. This is unsurprising, as fitting a lognormal distribution to the bootstrap output does not introduce additional information about the sources of uncertainty omitted from the underlying bootstrap model.

Applying an uplift to the bootstrap CoV can improve the fit in the tails of the distribution. However, it does so by widening the entire distribution, which can create excessive concentration around the centre of the backtesting percentile distribution and reduce the quality of fit across other parts of the distribution. In addition, the appropriate size of any uplift is difficult to determine objectively and is therefore often dependent on expert judgement.

The compound lognormal framework provides a materially improved fit to the observed historical reserve movements. It does this by treating the volatility parameter of the lognormal distribution as uncertain, rather than simply increasing the volatility of a conventional lognormal distribution. This creates a distribution with a more pronounced central peak and heavier tails than a simple lognormal distribution with the same selected best estimate and bootstrap-derived CoV.

The results suggest that the compound lognormal is a promising enhancement to conventional reserve risk modelling. It retains the useful portfolio-specific information provided by the bootstrap CoV, while introducing an explicit and auditable allowance for uncertainty around that estimate. **In our view, this is preferable to relying on judgemental uplifts to bootstrap CoVs.**

The model should not be interpreted as a complete or direct quantification of all sources of model error. Rather, it is a practical proxy for additional uncertainty not captured when the best estimate and bootstrap-derived volatility are each treated as known with certainty. The framework is supported by the backtesting results presented in this paper, but its practical use should remain subject to appropriate validation, expert judgement and consideration of portfolio-specific circumstances.

7.2. Limitations and implementation challenges

Since $\mathbb{E}[X \mid \sigma] = m_0$, the unconditional second moment is

$$\mathbb{E}[X^2] = m_0^2 \, \mathbb{E}\left[e^{\sigma^2}\right].$$

Because σ is lognormal, σ^2 is lognormal, and the lognormal distribution possesses no finite moment generating function at any positive argument. Hence:

$$\mathbb{E}\left[e^{\sigma^2}\right] = \infty \qquad \text{and} \qquad \mathrm{Var}(X) = \infty$$
$$\text{for every } c_\sigma > 0.$$

The compound lognormal with stochastic σ always has infinite variance. This is not a numerical artefact and cannot be removed by increasing simulation size; it is a property of the model. Three consequences follow, and we regard the first as the most important practical finding in this section:

- **Variance-based aggregation is unavailable**. Frameworks that combine risks through a correlation matrix applied to standard deviations cannot consume this model directly. In practice, this means that aggregating reserve risk classes would likely operate by correlating $CoV_{bootstrap}$ parameters by class, and treating CoV_σ as independent.

We have not attempted to specify how draws for CoV_σ should be correlated between classes – these could be treated as independent, or potentially fully correlated across classes, depending on how model error is prescribed.

- **Simulated CoVs are not well defined**. A Monte Carlo estimate of $\mathrm{Var}(X)$ does not converge; it drifts upward without bound as the number of simulations increases, and its value is determined by whichever extreme draws happen to occur. Reported CoVs from such simulations should be disregarded, and any stabilisation by truncation is a judgement that materially determines the answer.
- **The mean and all tail measures remain finite**. Since $\mathbb{E}[X] = m_0 < \infty$, every quantile and every tail expectation $\mathbb{E}[X \mid X > x]$ is finite and well defined.

The compound lognormal requires an additional parameter over and above the simple lognormal which must be estimated for capital modelling purposes. We have demonstrated that this parameter is necessary to produce a reasonable fit to historical reserve risk observations; however, this additional parameter still introduces limitations to capital modelling processes. In particular, the selection(s) will need to be documented and justified, and the uncertainty introduced to model results will need to be quantified and assessed via model testing. There may also be implications on model run time.

7.3. Next steps

The analysis in this paper provides a broad empirical assessment across a large and diverse dataset. Further work could develop the approach in the following areas:

- **Class-level analysis**: The results presented in this paper are based primarily on pooled data across reserving classes. A natural next step would be to assess whether the proposed calibration varies materially by class of business. In particular, long-tailed and short-tailed classes may exhibit different levels and forms of uncertainty, and a single calibration may not be appropriate in all cases. This work would need to balance improved segmentation against the reduced volume of data available within individual classes. Appendix D shows some indicative results by class of business, which indicates that the class effect may not be material, but this is not conclusive.
- **One-year versus ultimate reserve risk**: This paper backtests one-year movements in ultimate reserve estimates. Further work is required to consider how the resulting distribution should be interpreted and applied within a one-year capital framework, particularly where the underlying model is primarily parameterised based on uncertainty in ultimate claims development. This should include consideration of the relationship between ultimate deterioration, one-year emergence of reserve movements and the recognition of information over time.
- **Geographical applicability**: The analysis uses US NAIC Schedule P data. While the scale and diversity of that dataset provide a useful basis for testing the approach, further validation using data from other markets would be valuable. In particular, testing against UK and Lloyd's market data could help assess the extent to which the proposed calibration is transferable across different reserving practices, claims environments, lines of business and regulatory regimes.
- **Implementation within a capital model.** The compound lognormal framework needs further development for use within a full internal model. In particular, the model's heavy-tailed form – and the potential for infinite variance under some specifications – means that conventional correlation matrix-based aggregation may not be appropriate. Further work should therefore consider how reserve risk should be aggregated across classes of business and how different parameters (including both $CoV_{bootstrap}$ and CoV_σ) should be treated within the dependency structure of the model.
- **Broader treatment of model error**: The compound lognormal framework represents one practical way of allowing for uncertainty in a volatility estimate. Appendix H sets out one possible framework for incorporating model error, which we considered as part of our wider research. Further work could investigate whether other sources of model error should also be reflected explicitly – for example, uncertainty arising from the choice of

reserving method, data limitations, tail assumptions, parameterisation choices and expert judgement. This may help develop a more complete and structured framework for allowing for model uncertainty in reserve-risk modelling.

This list is not intended to be exhaustive. The authors would welcome correspondence with readers who have other avenues of proposed research, or who wish to discuss the methodology or implementation into a capital modelling framework.

Acknowledgements, data and references

Acknowledgements

The authors would like to thank their LCP colleagues for their support with the analysis underpinning this paper and for their time reviewing the draft:

- Cat Drummond
- Jo Thornett
- Krinal Dhiru
- Louis March
- Oli Grossman

Data source

This research uses statutory annual-statement data reported by US property and casualty insurers to the National Association of Insurance Commissioners (NAIC), including Schedule P: *Analysis of Losses and Loss Expenses*. Each Schedule P filing year provides ten years of premiums earned, unpaid losses and loss expenses, and claims reported and outstanding, analysed by line of business. The data were extracted and processed by LCP for the purposes of this research. NAIC has not endorsed the results presented in this paper. (content.naic.org)

References

- Chan, K. and Ramyar, M. (2016) 'Practical challenges in reserve risk', Institute and Faculty of Actuaries, October.
- Committee of European Insurance and Occupational Pensions Supervisors (CEIOPS) (2010) CEIOPS' Advice for Level 2 Implementing Measures on Solvency II: SCR Standard Formula – Calibration of Non-Life Underwriting Risk, CEIOPS-DOC-67/10, 8 April.
- Durkin, T. and Stone, C. (2019) Enhancing traditional reserving using data science techniques. Presentation at the Technical Issues in General Insurance (TIGI) Conference 2019, Institute and Faculty of Actuaries, 26 April 2019.
- England, P.D. (2002) 'Addendum to "Analytic and bootstrap estimates of prediction errors in claims reserving"', Insurance: Mathematics and Economics, 31(3), pp. 461–466. doi: 10.1016/S0167-6687(02)00161-0.
- England, P.D. and Verrall, R.J. (1999) 'Analytic and bootstrap estimates of prediction errors in claims reserving', Insurance: Mathematics and Economics, 25(3), pp. 281–293. doi: 10.1016/S0167-6687(99)00016-5.
- England, P.D. and Verrall, R.J. (2002) 'Stochastic claims reserving in general insurance', British Actuarial Journal, 8(3), pp. 443–518. doi: 10.1017/S1357321700003809.
- Fröhlich, A. and Weng, A. (2018) 'Parameter uncertainty and reserve risk under Solvency II', Insurance: Mathematics and Economics, 81, pp. 130–141. doi: 10.1016/j.insmatheco.2017.10.004.
- Leong, J. (Weng Kah), Wang, S. and Chen, H. (2012) 'Back-testing the ODP bootstrap of the paid chain-ladder model with actual historical claims data', Casualty Actuarial Society E-Forum, Summer 2012, pp. 1–34.
- LCP (2026) Capital Benchmarking Survey 2026. London: LCP. Survey conducted March–May 2026; published July 2026.
- Mackay, J. and Otto, D. (2015) 'Incorporating model error into the actuary's estimate of uncertainty', Casualty Actuarial Society E-Forum, Summer 2015, pp. 1–44.
- Prudential Regulation Authority (2026) 'Solvency Capital Requirement – Standard Formula', PRA Rulebook, Chapter 3A1: Non-Life Premium and Reserve Risk Sub-Module. Available at: PRA Rulebook.
- Renshaw, A.E. and Verrall, R.J. (1998) 'A stochastic model underlying the chain-ladder technique', British Actuarial Journal, 4(4), pp. 903–923. doi: 10.1017/S1357321700000222.
- Shapland, M.R. (2016) Using the ODP Bootstrap Model: A Practitioner's Guide. CAS Monograph Series, No. 4. Casualty Actuarial Society, pp. 1–126.

Appendices

Appendix A: Data cleaning

Pre-processing

The raw NAIC data contains paid claims and historical ultimate claim estimates at company, class, and cohort level. The following cleaning and filtering steps were applied to each filing year dataset independently prior to bootstrapping.

Separation of paid and ultimate data: The data was split into paid claims and ultimate loss estimates. Only entities present in both series were retained. Table 2 summarises the entity counts at this stage across each filing year.

Filing year	Total entities	Paid entities	Ultimate entities	Common entities
2001	20,046	17,702	18,435	17,684
2011	18,296	15,961	16,772	15,950
2022	15,252	9,403	13,583	9,392
2023	15,276	9,509	13,678	9,500
2024	15,897	13,599	14,273	13,562
2025	15,922	13,679	14,336	13,659

Table 2: Schedule P data by filing year

Complete triangle filter: Only entities with a non-zero total paid claims amount for every cohort triangle (ie the cohort range shown in Table 3) for each filing year were retained, that is, no cohort could be entirely missing or sum to zero. Table 3 shows the number of entities removed at this stage and those retained.

Dataset year	Cohort range	Entities before	Entities removed	Entities retained
2001	1992–2001	17,684	10,244	7,440
2011	2002–2011	15,950	9,074	6,876
2022	2013–2022	9,392	3,829	5,563
2023	2014–2023	9,500	3,988	5,512
2024	2015–2024	13,562	4,942	8,620
2025	2016–2025	13,659	5,121	8,538

Table 3: Schedule P data by dataset year

Minimum reserve filter: To ensure statistical credibility and remove very small entities where bootstrap estimates would be unstable, only companies with total reserves exceeding $1m were retained. As shown in Table 4, this threshold sits around the 6th to 10th percentile of the reserve size distribution across filing years, meaning the vast majority of the market by reserve volume remains included.

Filing year	Proportion of companies with less $1m reserves (%)
2001	10.1
2011	9.2
2022	6.9
2023	6.2
2024	7.7
2025	7.0

Table 4: Proportion of companies with less than $1m total reserves

The impact of filtering on companies with larger reserve sizes is shown in the charts below. Figure 13 shows total reserves (bars) and the number of triangles (dots) by reserving class, comparing the dataset before and after the minimum reserve filter was applied; Figure 14 shows the same comparison by filing year.

Across both cuts, total reserves are materially unaffected by the filter — the blue (original) and orange (filtered) bars are nearly indistinguishable in almost every case, confirming that the $1m threshold is removing small companies without meaningfully reducing the reserve volume of the data. The impact is more visible in the number of triangles. Table 5 shows the number of entities and corresponding reserve volume removed.

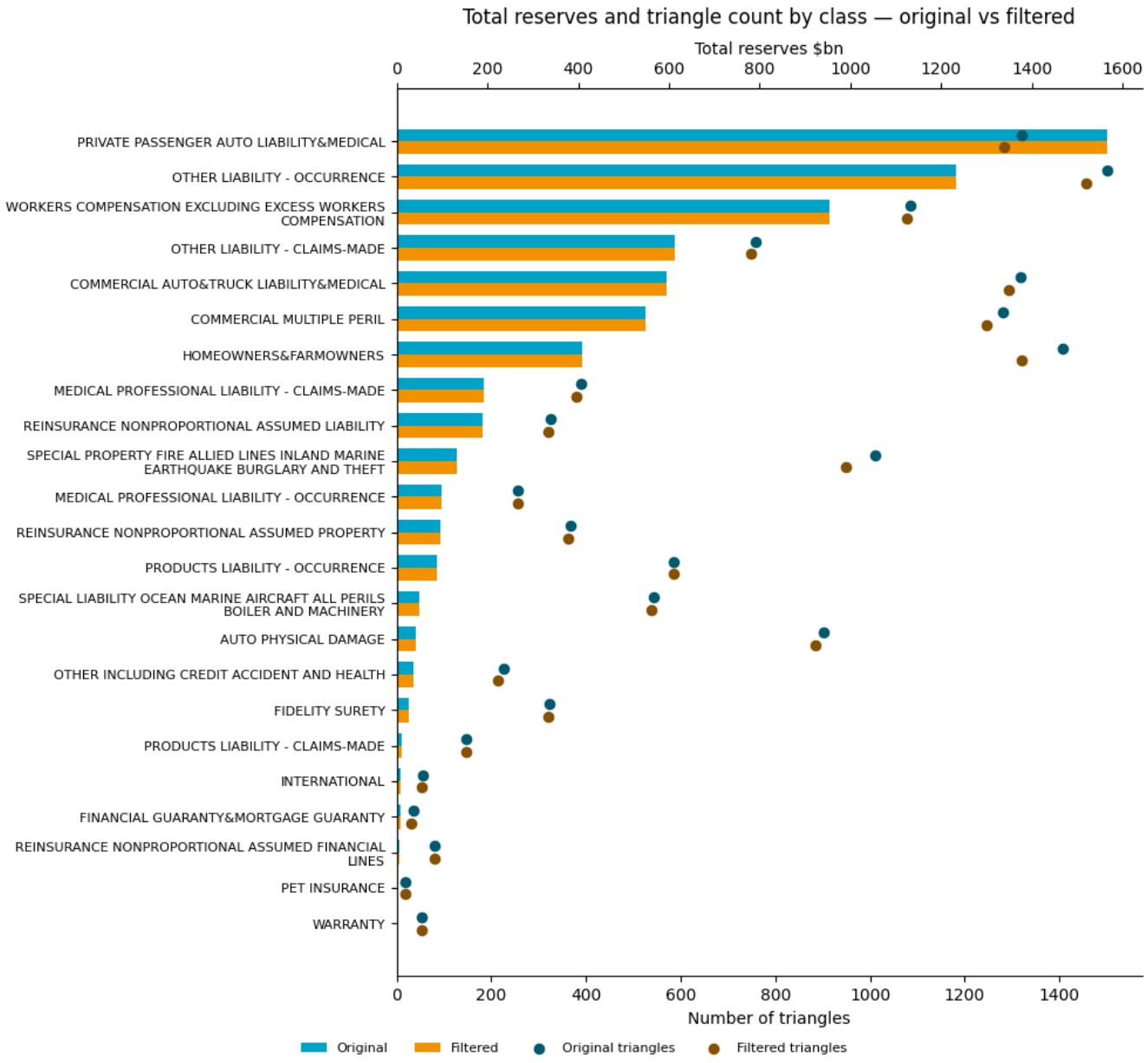


Figure 13: Impact of filtering by class (pre-processing)

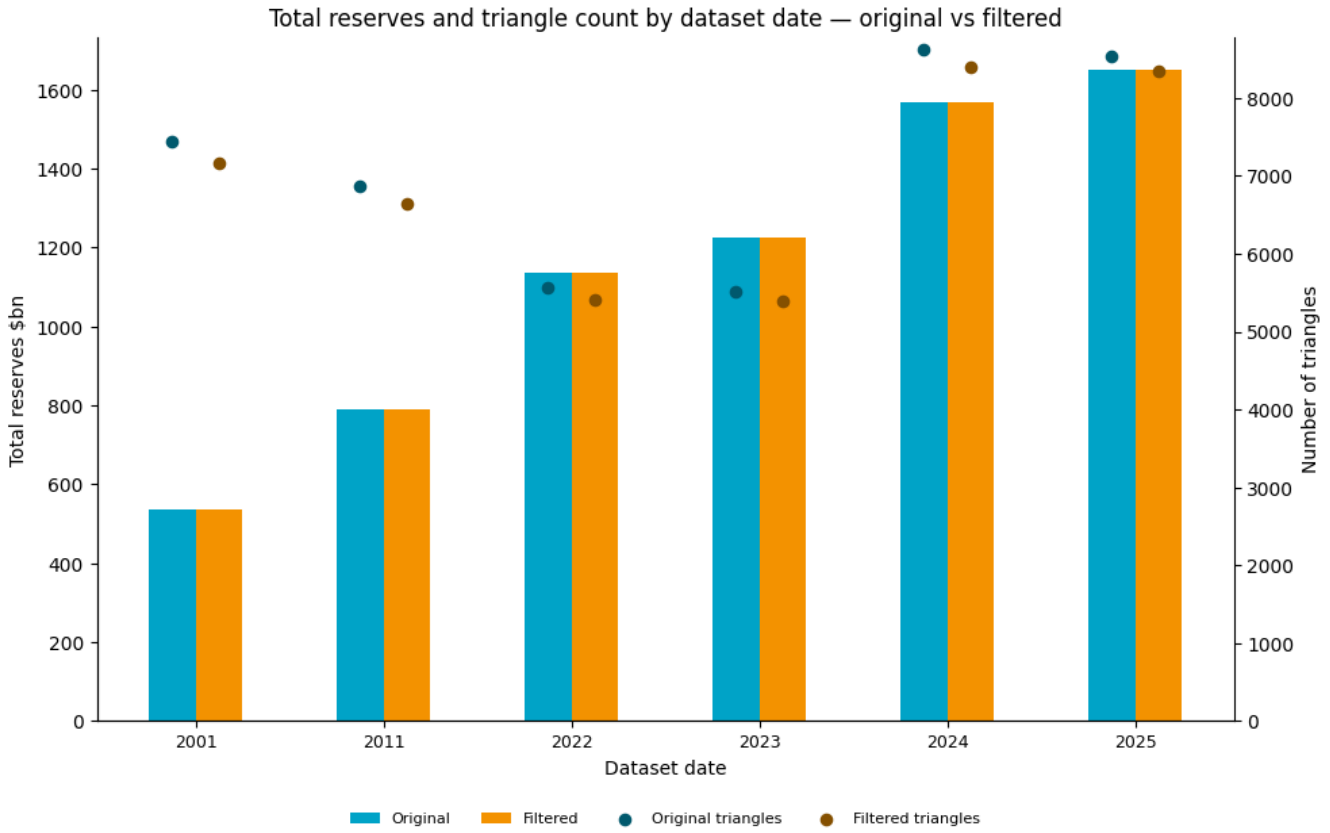


Figure 14: Impact of filtering by filing year (pre-processing)

Filing year	Reserves (original) $bn	Reserves (filtered) $bn	No. triangles (original)	No. triangles (filtered)	Reserves dropped $bn	No. triangles dropped	Reserves dropped %	Triangles dropped %
2001	535.14	535.09	7,440	7,172	0.06	268	0.01%	3.6%
2011	789.35	789.41	6,876	6,636	-0.06	240	-0.01%	3.5%
2022	1,135.73	1,135.71	5,563	5,414	0.02	149	0.00%	2.7%
2023	1,226.68	1,226.66	5,512	5,390	0.02	122	0.00%	2.2%
2024	1,567.03	1,566.99	8,620	8,405	0.04	215	0.00%	2.5%
2025	1,650.51	1,650.48	8,538	8,352	0.04	186	0.00%	2.2%

Table 5: Impact of filtering by filing year

The profile of the dataset is relatively unchanged. Total reserves are materially unaffected by this filter. Across every filing year, the change in reserves is under 0.1%, even though a meaningful number of triangles are removed (2.2%–3.6% depending on the year). This is expected: the filter is specifically designed to remove small companies (by reserve size), so removing them necessarily has a small effect on total reserves relative to their share of the triangle count. No individual class of business or filing year is disproportionately affected by this filter. The reduction in triangle count and the (negligible) change in reserves are both spread broadly across the dataset, rather than being concentrated in any particular class or year. In this sense, the filter does not meaningfully reshape the composition of the dataset; it removes a small, evenly-distributed set of small companies.

Following this filtering, we had 41,369 triangles to bootstrap, with total reserves of almost $7tn across the different filing years.

The analysis compares movements in ultimate claims estimates over successive development periods. For each filing year, the latest cohort contains only one observed development point and therefore cannot be backtested. On this basis, the total reserves are $4.2tn, compared with $6.9tn when the latest cohorts are included. Figure 15 shows total reserves (bars) and the number of triangles (dots) by reserving class, comparing the dataset before and after excluding the latest cohort; Figure 16 shows the same comparison by filing year.

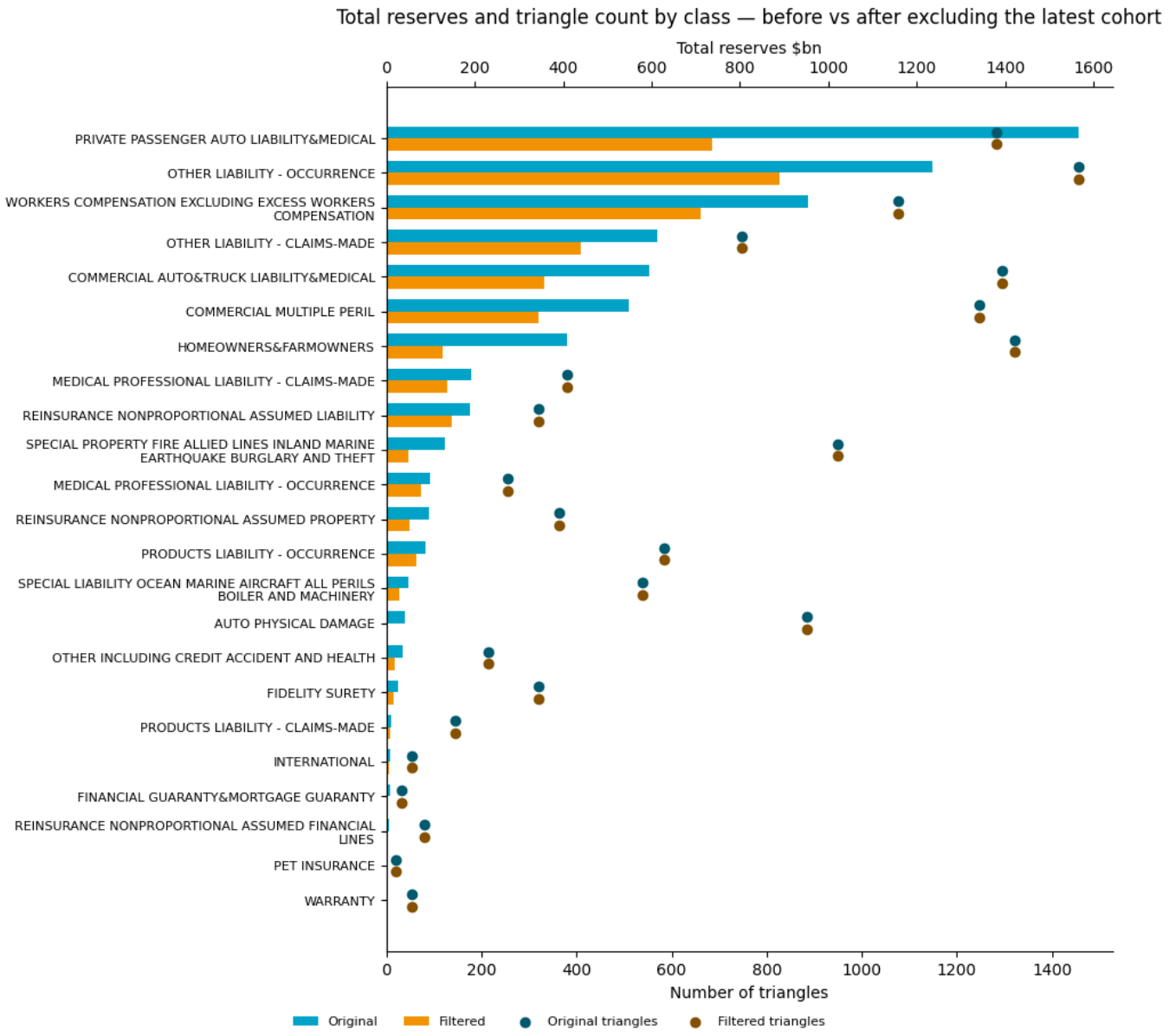


Figure 15: Impact of removing latest cohort by class

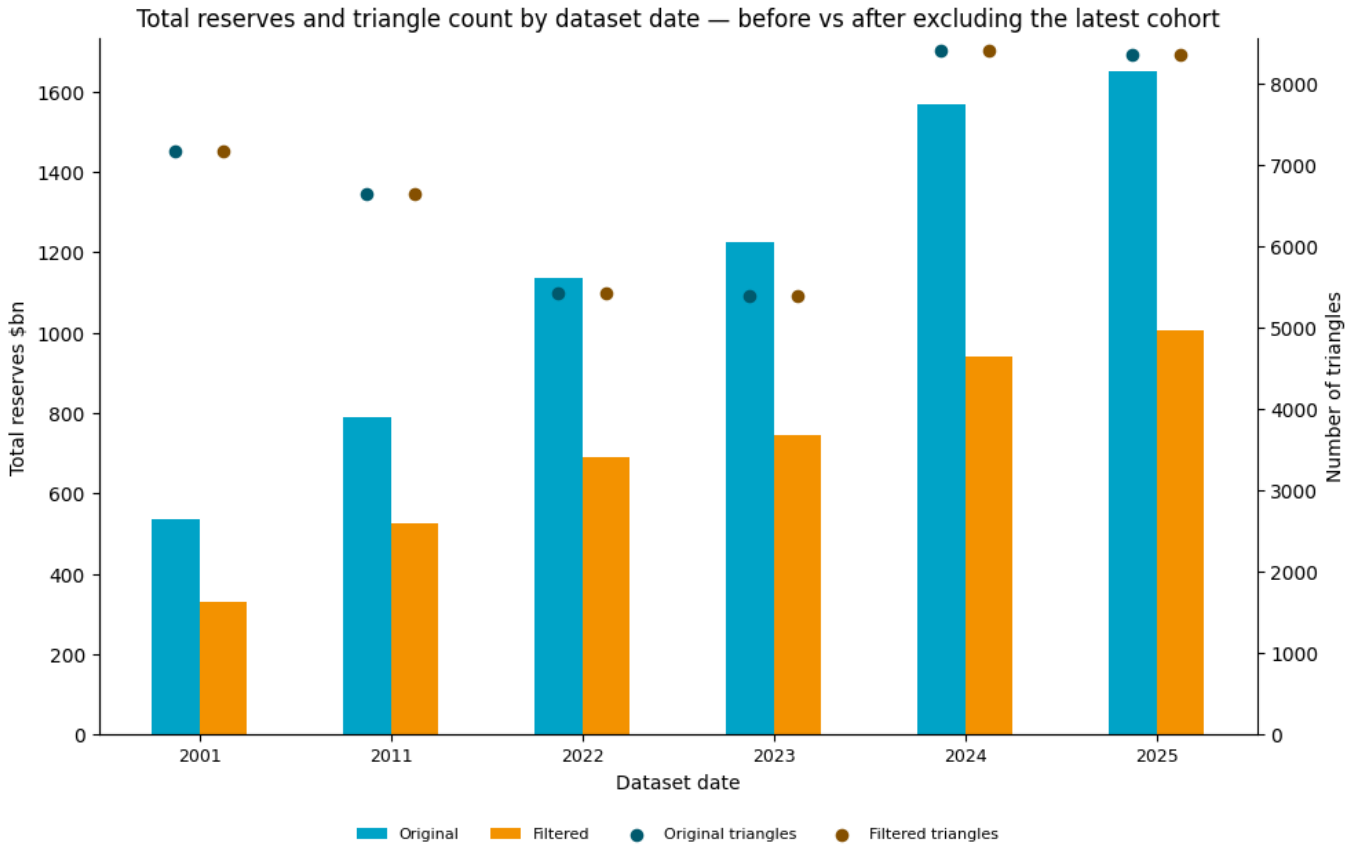


Figure 16 Impact of removing latest cohort by filing year

Post-bootstrap filtering

Following bootstrapping, we performed further cleaning. While we focus on triangles as an input to the bootstrap, the output is considered in entity/cohort combinations.

Bootstrap errors: For a small number of entity/filing year combinations, the bootstrap process itself failed to run, most commonly due to insufficient variation in the underlying paid claims triangle for the resampling procedure to draw from. These entities were excluded entirely, since no simulated distribution was available to compare against. Table 6 summarises the impact of this step.

	Entity/cohort/filing year combinations
Before bootstrap	372,321
Failed to bootstrap	1,386 (154 entity/filing year combinations)
Combinations retained	370,935
Proportion retained	99.6%

Table 6: Bootstrap errors

This represents a loss of 0.37% of the pre-bootstrap population, concentrated in a small number of entities where the bootstrap could not be run.

Unlikely deteriorations: Two further checks were applied to identify cohorts where the comparison between the prior and latest ultimate loss estimate was not considered representative. First, cohorts where either the prior or the latest ultimate loss estimate was zero or negative were removed. Second, cohorts where the simulated distribution

showed no variation at all (a coefficient of variation of exactly zero, or undefined) were removed. Table 7 summarises the impact of these two steps.

	Entity/cohort/filing year combinations
Combinations before deterioration checks	370,935
Removed: prior or latest ultimate ≤ 0	3,306
Removed: no variation in simulated distribution (CoV undefined or zero)	71,434
Combinations retained	296,195
Proportion retained	79.9%

Table 7: Unlikely deteriorations

Implausible CoVs: Cohorts with a simulated coefficient of variation outside the range of 2% to 500% were removed as implausible: a CoV below 2% implies an unrealistic degree of certainty in the reserve estimate, while a CoV above 500% suggests the bootstrap failed to produce a stable or meaningful distribution. Based on the authors' experience and with reference to market benchmarks, in particular the range of CoVs observed in LCP's 2026 London Market capital modelling benchmarking survey, we are satisfied that this provides a suitably wide range of CoVs that does not remove relevant data. Table 8 summarises the impact of this step.

	Entity/cohort/filing year combinations
Combinations before CoV threshold	296,195
Removed: CoV < 2%	99,495
Removed: CoV > 500%	11,928
Combinations retained	184,772
Proportion retained	62.4%

Table 8: Implausible CoVs

Identical prior and latest ultimates: Finally, cohorts where the prior and latest ultimate loss estimates were exactly identical were removed. We do not consider it likely that a company's true ultimate estimate remains completely unchanged between development periods, in cohorts that are still developing. Since the backtesting analysis is intended to assess how well the model captures genuine year-on-year movement, we do not consider these cohorts to represent reliable observations of that movement, and therefore we have removed them. 7,208 rows (entity/cohort/filing-year combinations), or 3.9% of the remaining population at this stage, were removed as a result. This represents a 0.2% reduction in the total reserves, involving mainly older cohorts, which provides reassurance that this filter has not materially reduced the relevance or coverage of the dataset. Table 9 summarises the impact of this step.

	Entity/cohort/filing year combinations
Combinations before unchanged reserves	184,772
Removed: Prior = latest ultimate	7,208
Combinations retained	177,564
Proportion retained	96.1%

Table 9: Unchanged ultimates

In total, post-bootstrap cleaning removed 193,371 combinations (52.1% of the 370,935 that were in the bootstrap), leaving 177,564 entity/cohort/filing year combinations for the final analysis. The majority of this reduction, 111,423

of the 193,371 removed combinations, comes from the implausible CoV filter, and reflects the practical limits of applying a bootstrap method. Rather than indicating a flaw specific to any one class or cohort, we view this largely as evidence that a single bootstrap methodology, applied without further tailoring, does not produce a stable, plausible distribution for every cohort in a dataset this broad and that a non-trivial share of cohorts needs to be set aside as a result.

We consider this an acknowledged limitation of the approach, rather than a sign that the retained population of 177,564 combinations is unrepresentative. As shown in Figure 17 and Figure 18 below, the extent of post-bootstrap cleaning varies somewhat by filing year and by class of business. The post-bootstrap removals, while material in aggregate, are concentrated in the sources of instability described above (near-zero or extreme simulated CoVs, non-positive ultimates, and identical prior/latest ultimates) rather than reflecting a systematic bias in the data used for the analysis.

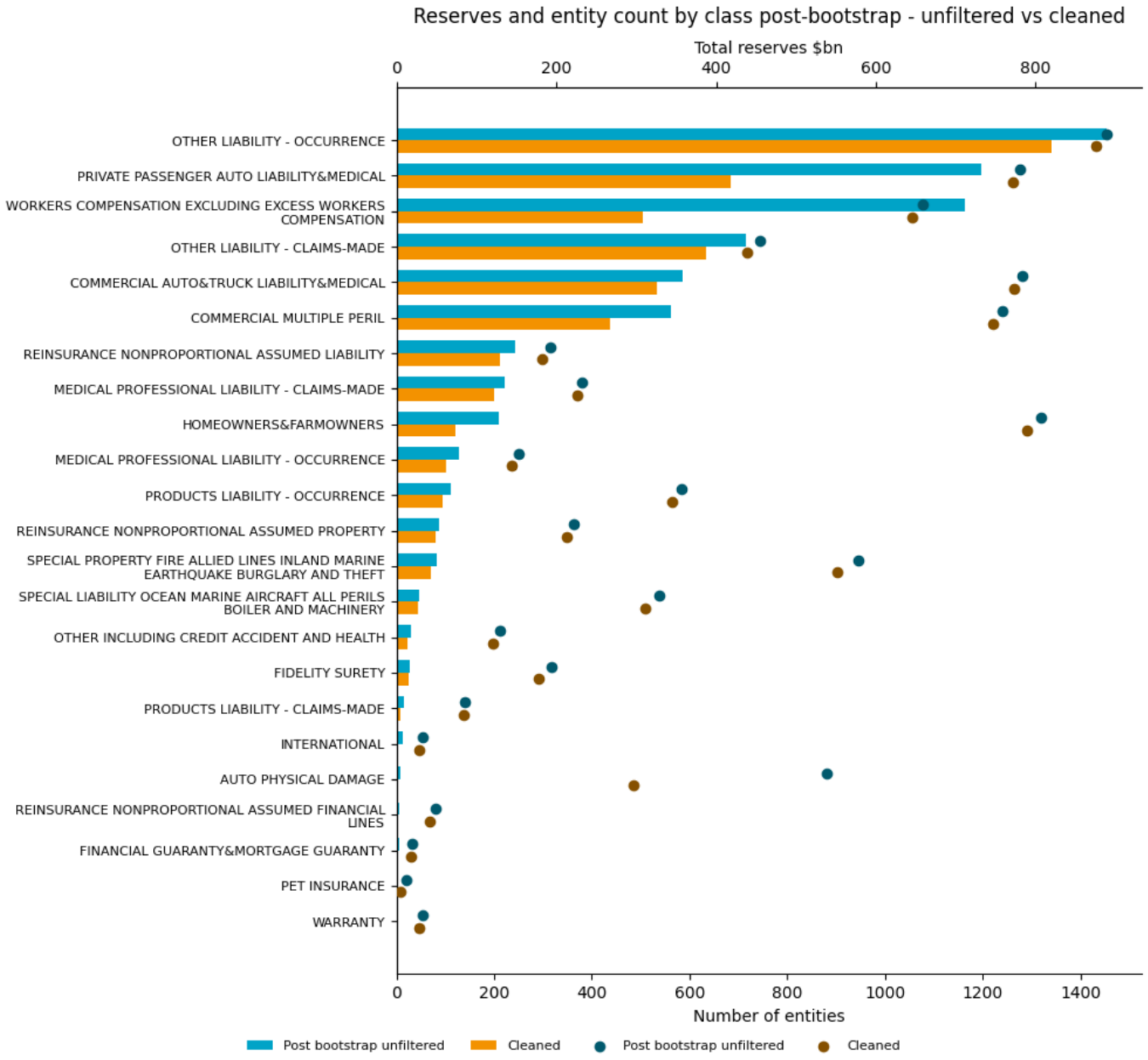


Figure 17: Impact of filtering by class (post-processing)

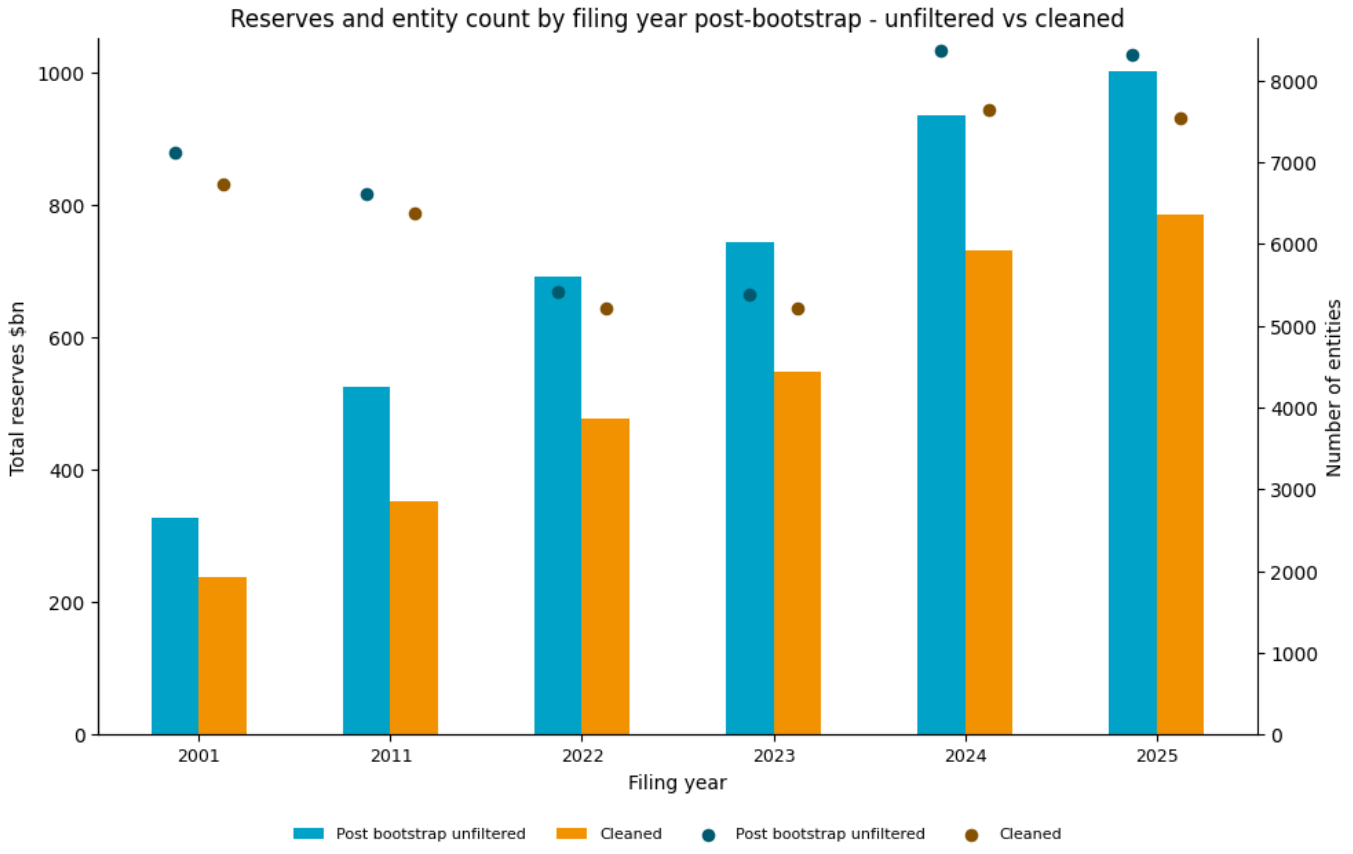


Figure 18: Impact of filtering by filing year (post-processing)

Appendix B: Over-dispersed Poisson distribution framework

Unlike the Poisson distribution, where the variance equals the mean, the over-dispersed Poisson distribution has a variance proportional to the mean. In claims reserving, this model assumes that the incremental claims C_{ij} are independent over-dispersed Poisson random variables with mean and variance:

$$E[C_{ij}] = \mu_{ij} = x_i y_j \text{ and } Var[C_{ij}] = \phi\, x_i y_j$$

where x_i is the expected ultimate claims and y_j is the proportion of ultimate claims emerging in each development period. The ϕ parameter introduces the over-dispersion which is estimated from the data. Further details are provided by Renshaw and Verrall (1998) and England and Verrall (2002).

The main limitation of the bootstrap is that the sum of the incremental claims for each development period must be positive. This is due to the y_j in the variance parameter. Some negative increments can occur as long as the sum of any column is not negative.

The log link function in the GLM is introduced to re-parameterise the mean structure so that it has a linear form such that:

$$log(\mu_{ij}) = c + \alpha_i + \beta_j$$

Since there is a parameter for each row i and column j, the structure is still a chain ladder type.

Appendix C: Bootstrapping methodology

Fitting the GLM

Given an incremental claims triangle, we can calculate an expected triangle, from which Pearson residuals can also be calculated. We define η_{ij} to be the loglink of μ_{ij}, where the loglink function is defined as:

- $\ln(\mu_{ij})$ for $C_{ij} > 0$
- 0 for $\mu_{ij} = 0$
- $-\ln(-\mu_{ij})$ for $C_{ij} < 0$

We then further define $\eta_{ij} = \alpha_i + \beta_j$, where $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, n$, and define a system of equations as $Y = XA$ to calculate η_{ij}, with:

$$Y = \begin{pmatrix} \ln(C_{11}) \\ \ln(C_{21}) \\ \ln(C_{31}) \\ \vdots \\ \ln(C_{n1}) \end{pmatrix}$$

$$X = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & 0 & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

$$A = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_n \end{pmatrix}$$

We can solve this using a least squares approach for the chosen distribution, and the calculate the fitted matrix, $\hat{Y} = XA$, which will be of the form:

$$\hat{Y} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_1 + \beta_2 + \cdots + \beta_n \end{pmatrix} = \begin{pmatrix} \eta_{11} \\ \eta_{21} \\ \eta_{31} \\ \vdots \\ \eta_{n1} \end{pmatrix}$$

This can then be exponentiated using the inverse of the loglink function to find u_{ij}, the expected values for each entry of the triangle according to the GLM as:

- $\exp(\eta_{ij})$ for $C_{ij} > 0$
- 0 for $C_{ij} = 0$
- $-\exp(\eta_{ij})$ for $C_{ij} < 0$

A more comprehensive explanation of the ODP bootstrap methodology is provided by Shapland (2016).

Residuals

After calculating all values of μ_{ij}, we can now calculate residuals for all entries in the incremental triangle, from which sampling can begin. Firstly, the unscaled Pearson residuals can be calculated as:

$$r_{ij} = \frac{C_{ij} - \mu_{ij}}{\sqrt{|\mu_{ij}^z|}} * f^{DoF}$$

f^{DoF} is an adjustment made to effectively allow for over-dispersion of the residuals in the sampling process and to add process variance to approximate a distribution of possible outcomes. It is defined as $f^{DoF} = \frac{N}{N-p}$, where N and p are defined below.

The scale parameter is calculated as either of the following two formulae, depending on whether it is set to be constant by the user, or changing for each development period j:

$$\phi = \sqrt{\frac{\sum_{ij} r_{ij}^2}{N-p}} \quad \text{or} \quad \phi_j = \sqrt{\frac{\sum_i r_{ij}^2}{o_j}},$$

where N is the number of observations or incremental data cells in the triangle, p is the number of parameters and o_j is the number of entries in the incremental triangle for that development period. Typically, $N = \frac{n(n+1)}{2}, p = 2(n-1)$ and $o_j = n + 1 - j$ for an $n \times n$ triangle.

Sampling residuals

The residuals are then randomly sampled from with replacement and used to create a pseudo triangle, using the following formula:

$$C_{ij}^* = r_{ij}^* * \phi_j * \sqrt{|\mu_{ij}^z|} + \mu_{ij}$$

where C_{ij}^* is our resampled residual, and r_{ij}^* is the new incremental value for the pseudo triangle and origin period i and development period j.

The user can decide if negative incremental values are allowed in the bootstrapped triangle. If they are not allowed, all negative incremental values will be set to 0. The user could also decide to prevent negative cumulative values in the bootstrapped triangle, even if negative incremental values are allowed.

Refitting the GLM

The pseudo triangle is then projected across all future development periods using the basic chain ladder approach in order to calculate a completed triangle. The link ratios are calculated across all origin years in the pseudo triangle and this is used to derive an estimate for the ultimate claims in the pseudo triangle and a reserve for each origin year can be calculated by subtracting the latest paid in the pseudo triangle from the calculated ultimate claims.

Ultimate

From the set of reserves calculated using the bootstrap method the standard deviation, mean, and CoV can be calculated.

Appendix D: Lognormal backtesting results

We show the backtesting results for different segmentations of the data to understand the generalisability of results and identify any material drivers.

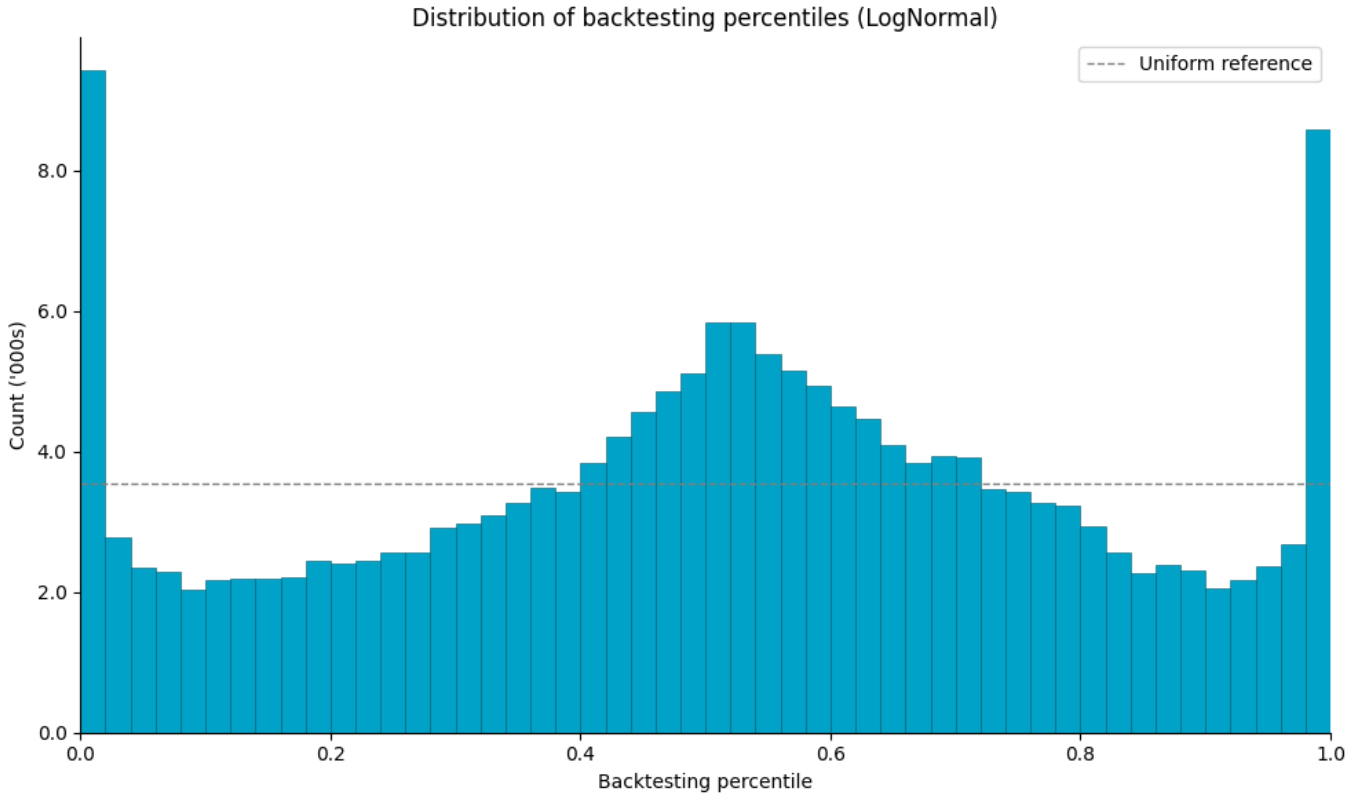


Figure 19: Overall lognormal backtesting results

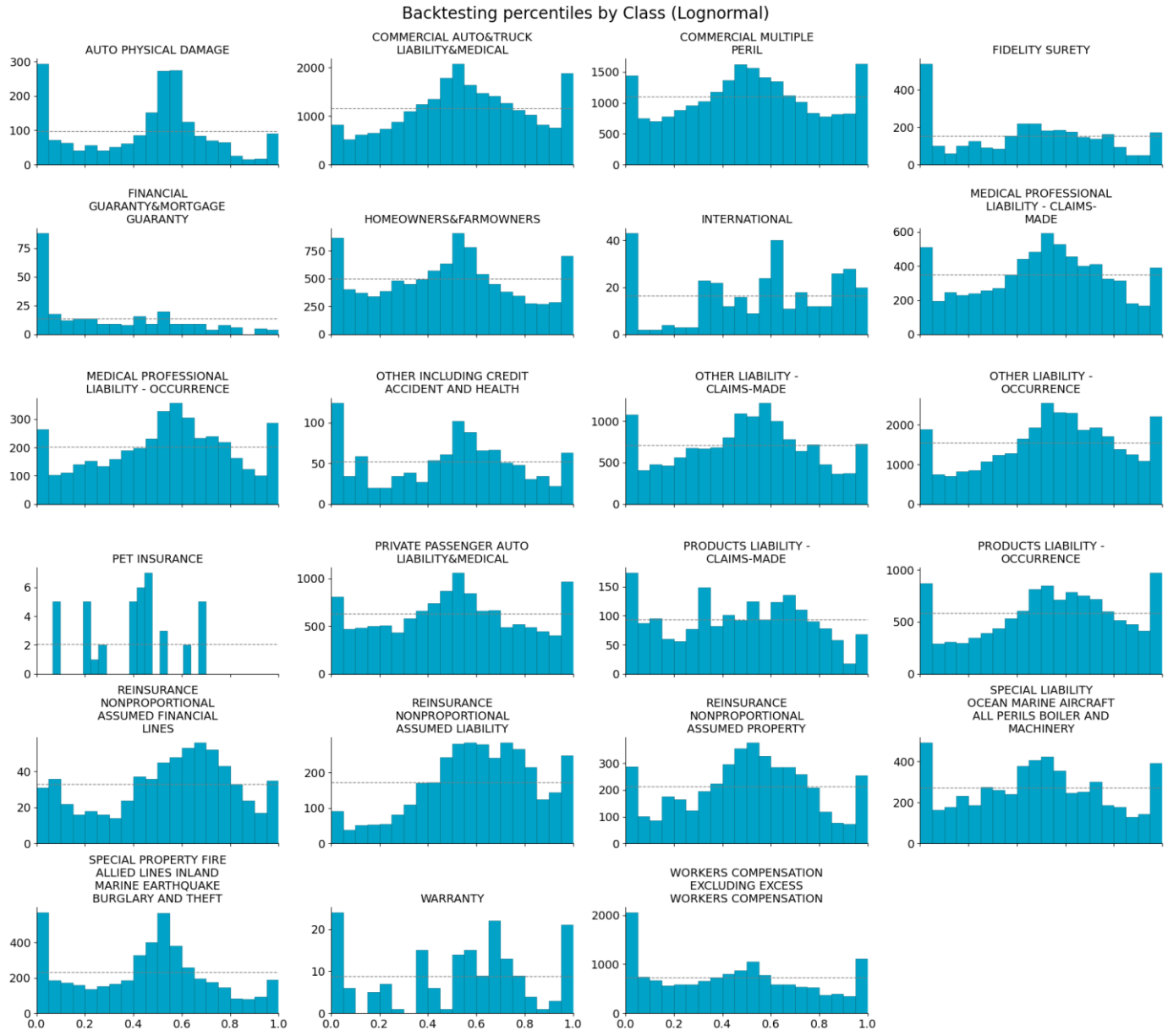


Figure 20: Lognormal backtesting results by class of business

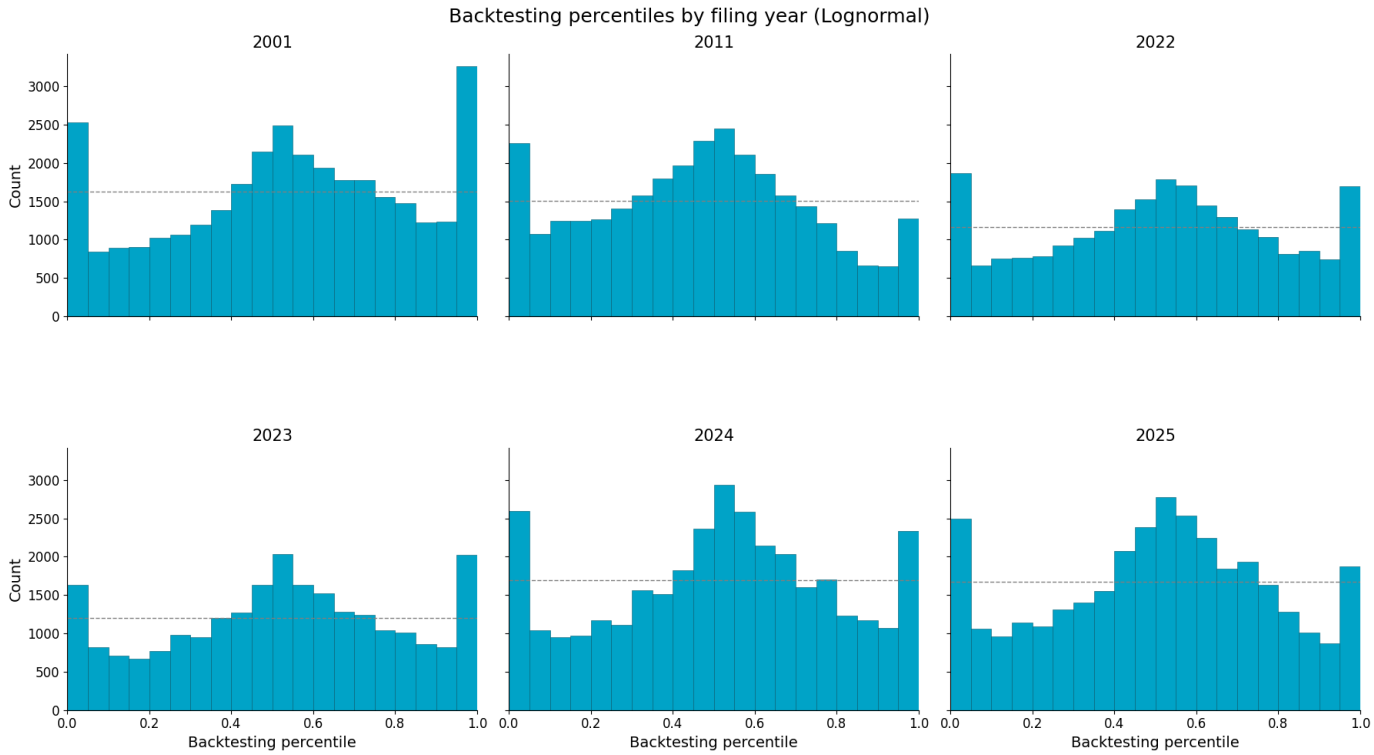


Figure 21: Lognormal backtesting results by filing year

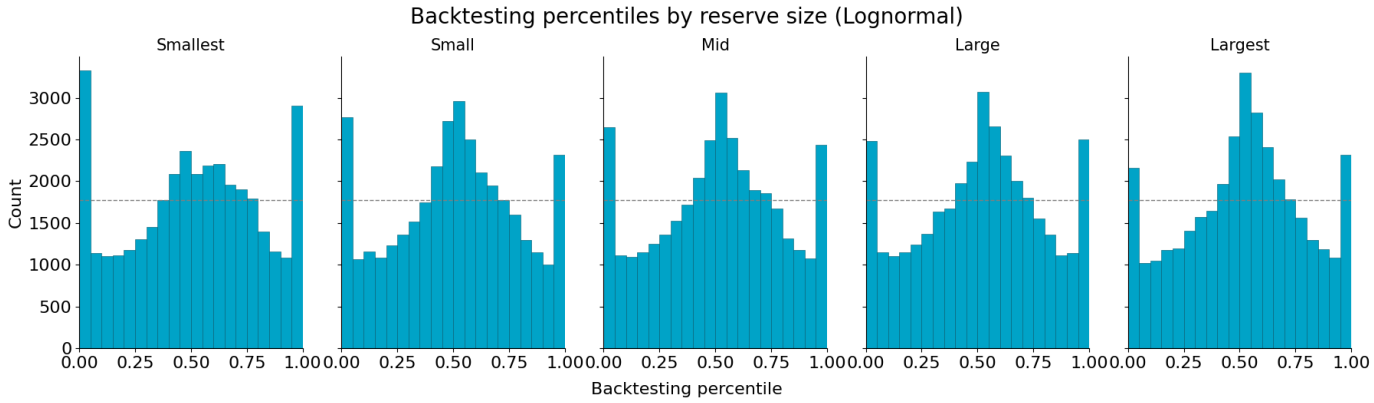


Figure 22: Lognormal backtesting results by reserve size

We conclude:

- **Classes of business** generally behave similarly, especially classes with large data volumes available. In particular, the lowest (and, usually, highest) buckets are generally over-represented when backtesting. There are no classes that appear to produce an adequate fit to the historical data without adjustment.
- Neither **filing year** nor **reserve size** seems to have a material effect on whether backtesting produces a reasonable fit.

Appendix E: Compound lognormal backtesting results

We show the backtesting results for different segmentations of the data to understand the generalisability of results and identify any material drivers.

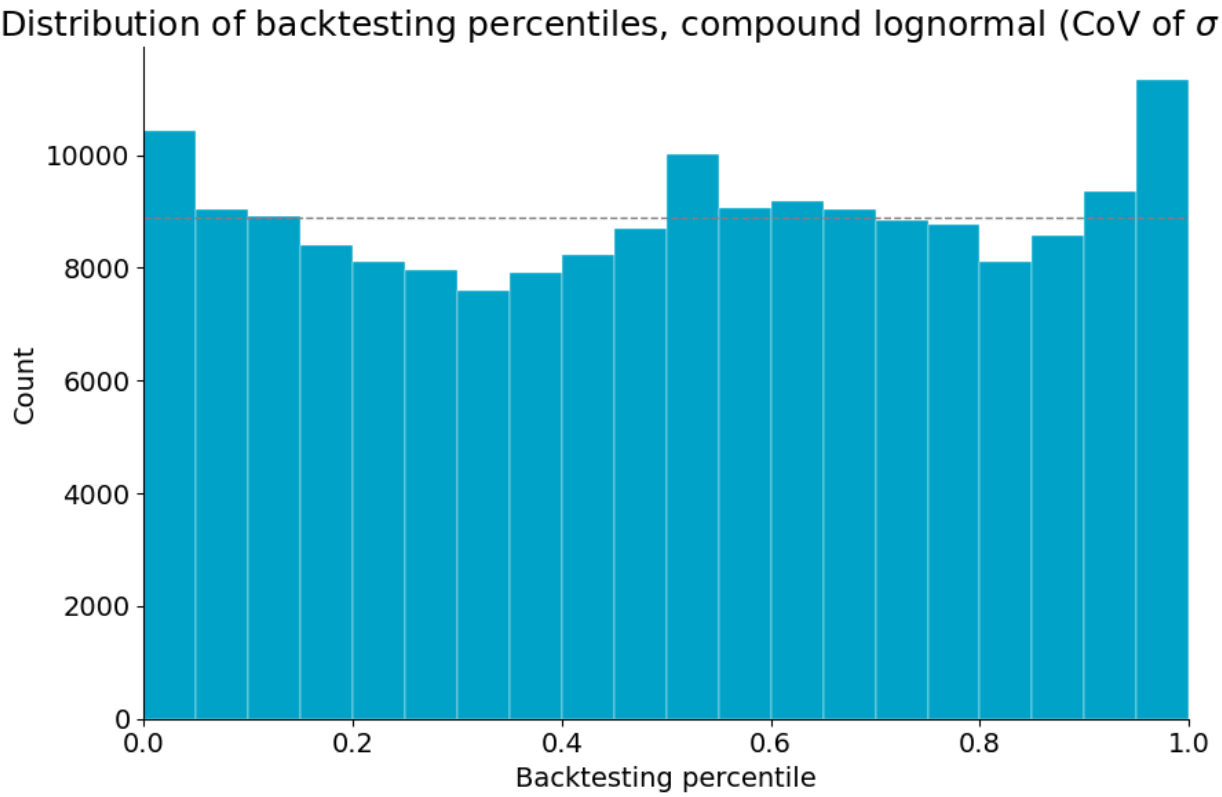


Figure 23: Overall compound lognormal backtesting results

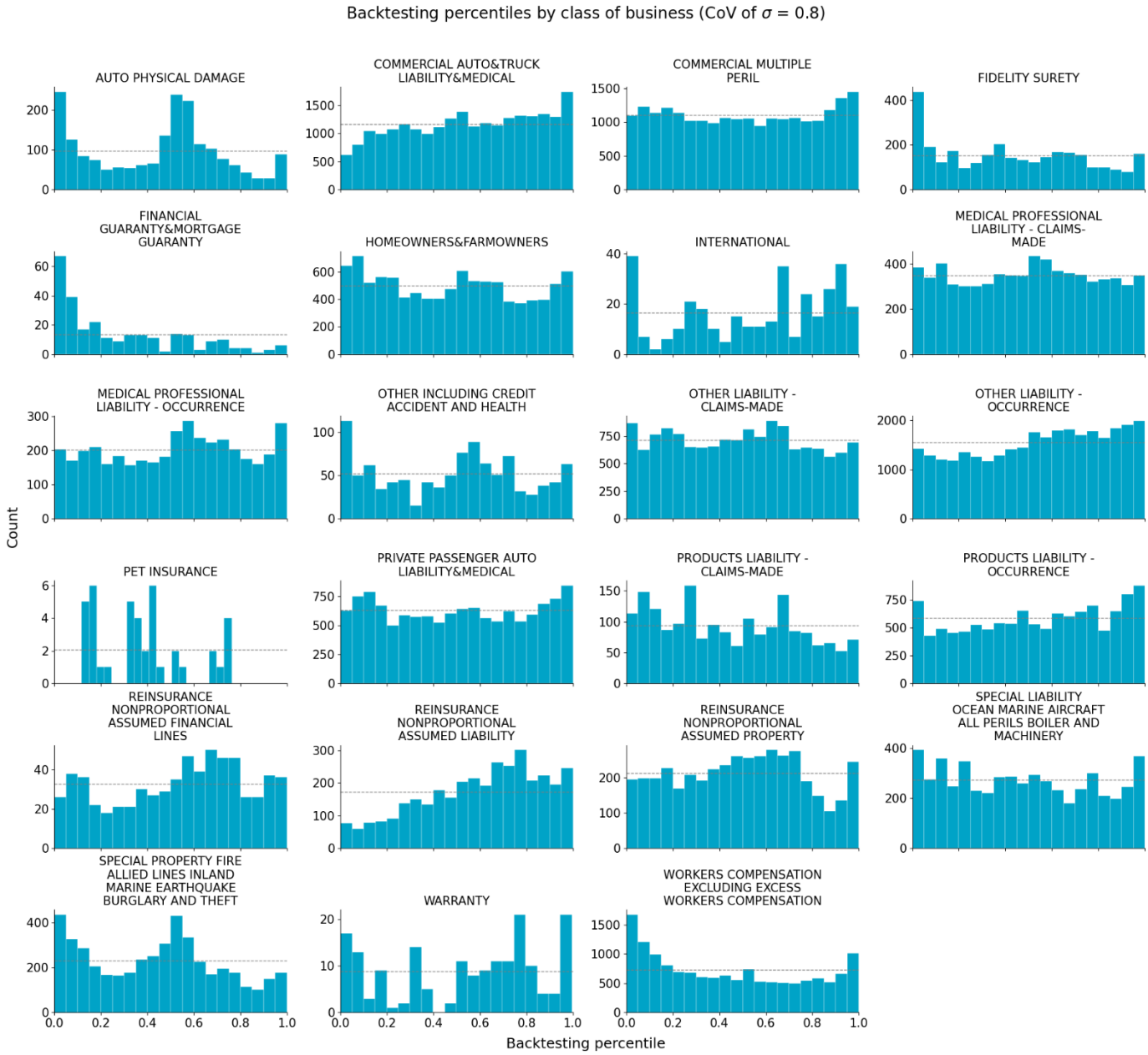


Figure 24: Compound lognormal backtesting results by class of business

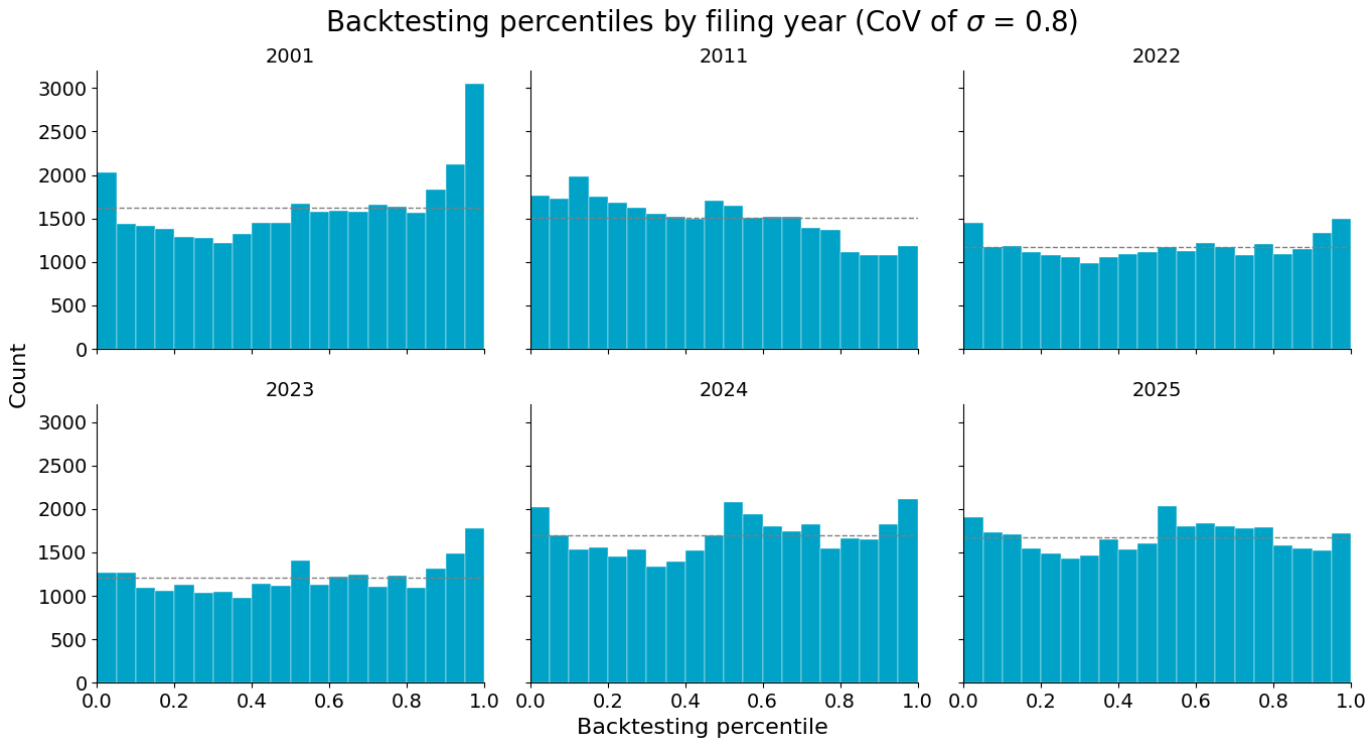


Figure 25: Compound lognormal backtesting results by filing year

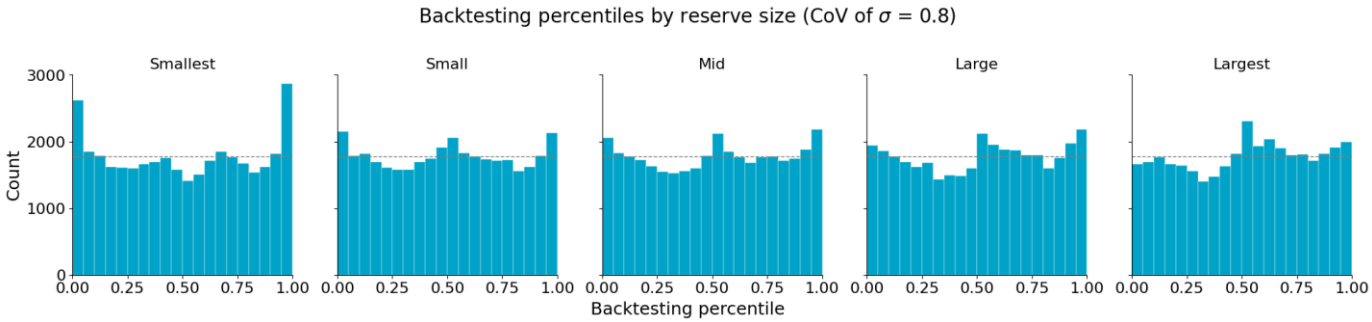


Figure 26: Compound lognormal backtesting results by reserve size

We conclude:

- **Class of business** likely has a small amount of influence on CoV_σ. However, the larger classes all produce reasonable fits for 0.8 (and for a range of values close to this). Our main conclusion from this is that large data volumes are needed to estimate CoV_σ with a meaningful degree of precision.
- **Filing year** likely does not have any influence on CoV_σ, and the anomaly for 2001 is likely due to poorer data quality than in other years.
- CoV_σ likely decreases as **reserve size** increases. This makes intuitive sense, since larger reserves will typically be associated with more stable triangles for which the bootstrap will tend to produce better estimates. However, the ranges of values we have focused on produce reasonable fits for all but the very smallest reserve volumes, which will generally not be material to capital.

Appendix F: Compound lognormal vs simple lognormal

This appendix expands on the analysis of the compound vs simple lognormal presented in section 6.6.

Figure 27 shows the change in probability mass for different values of $CoV_{bootstrap}$ and CoV_{σ}. The right-hand side panel zooms into the tail to see the impacts. Higher CoV_{σ}s produce higher peaks and heavier tails.

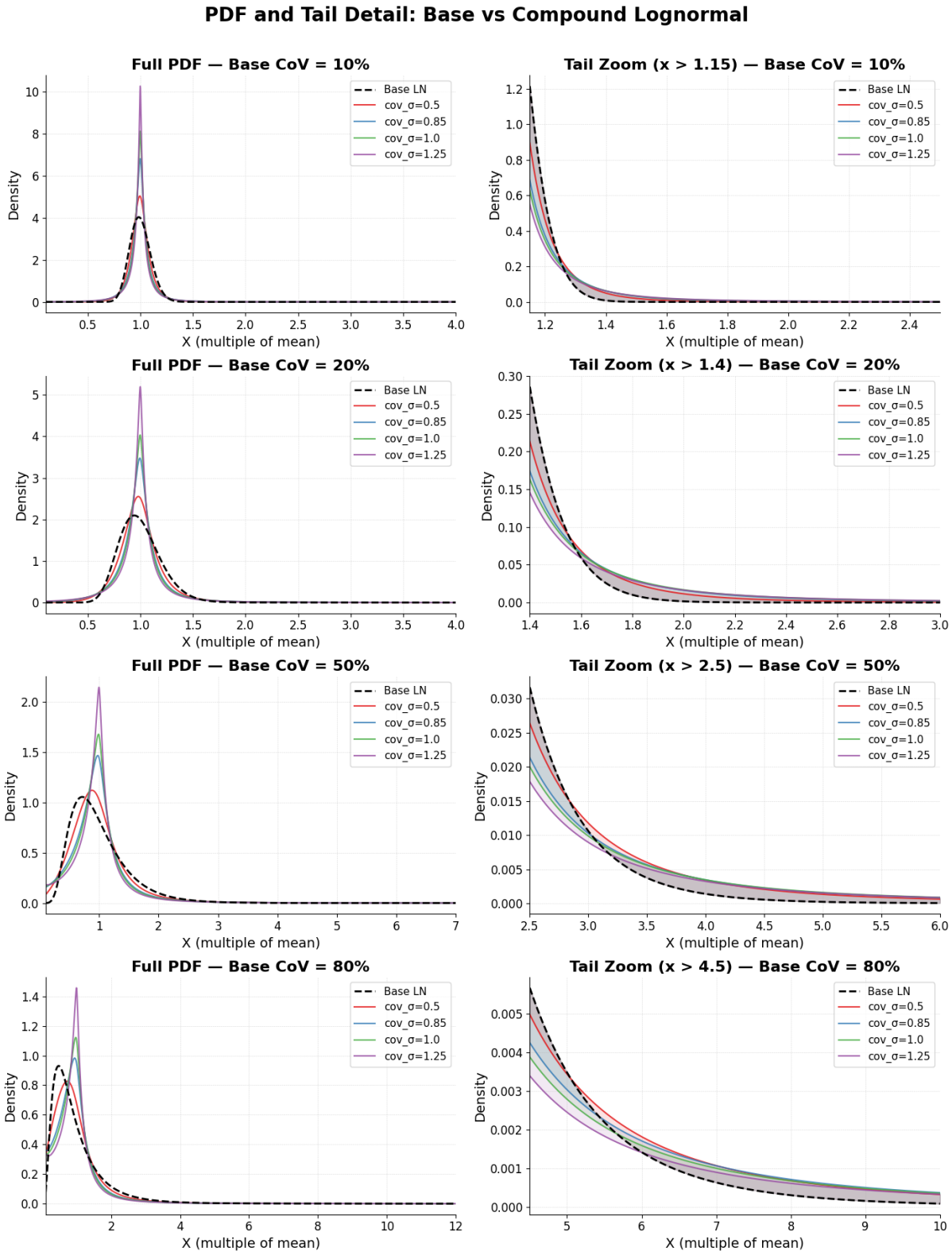


Figure 27: Compound lognormal vs simple lognormal for different CoVs

Figure 28 plots the ratio of compound to base across every percentile. It provides a complementary view to the density comparison shown in Figure 10: rather than showing where probability sits, it shows the consequence for the quantiles themselves for a $CoV_\sigma = 0.8$ over a range of values for $CoV_{bootstrap}$. A ratio above 1 means the compound distribution produces a higher outcome at that percentile than the simple lognormal and a ratio below 1 means it gives a lower outcome. The left panel shows this for Value at Risk (VaR), and the right for Tail Var at Risk (TVaR). Each curve in the left panel sits below 1 in the extreme lower tail, rises above it through the lower-middle of the distribution, falls back below between roughly the 70th and 95th percentiles, and rises steeply in the far-right tail. The compound distribution therefore produces quantiles that are alternately smaller and larger than those of the lognormal fitted to the same $\mathrm{CoV}_{bootstrap}$, depending on in which part of the distribution they are evaluated.

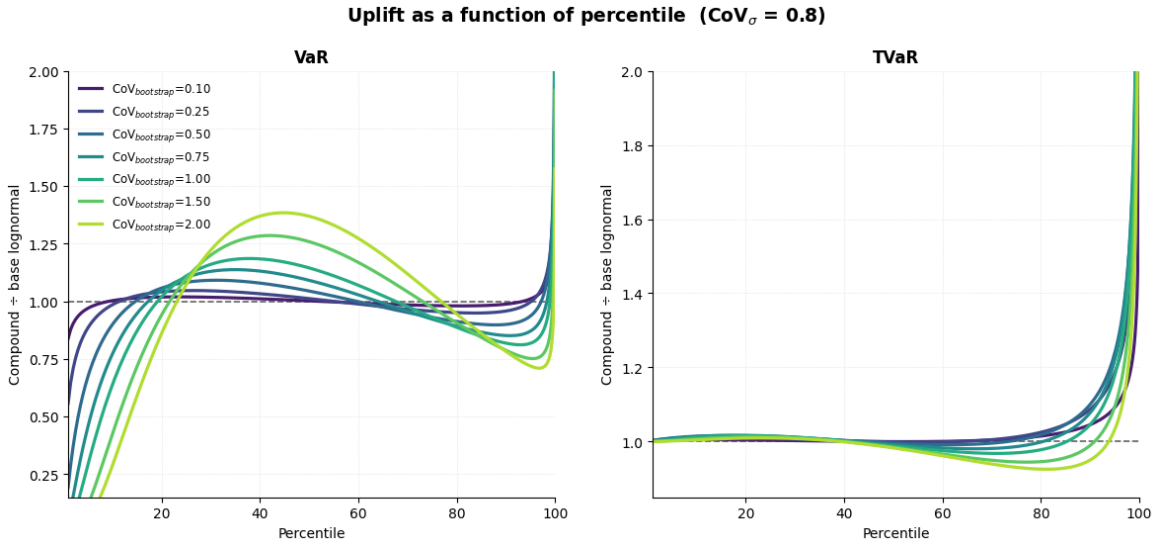


Figure 28: Simple vs compound lognormals (uplifts)

This shape follows from the mean-preserving parameterisation. Both distributions have mean m_0 by construction. The compound distribution places materially more weight on extreme adverse outcomes, and because those outcomes are large in magnitude, they contribute substantially to the mean. This contribution then gets offset in the "shoulders" of the distribution, roughly between the 5th and 30th percentiles, and between the 70th and 95th percentiles. The compound distribution does not inflate the lognormal uniformly; it redistributes probability, and the heavier tail is "funded" from the shoulders.

The right panel has a quite different shape: close to 1.0 across the body of the distribution, rising only as the threshold approaches the tail. The difference between the panels reflects what each measure looks at. VaR_q depends only on the value of the distribution at a single point, and so tracks whichever of the regions above that percentile falls in. TVaR_q averages over the whole region beyond the threshold, and therefore always incorporates the extreme tail regardless of the selected quantile q; the redistribution within the body largely cancels in that average, leaving the tail to dominate.

We then examine the ratio of the compound lognormal to the simple lognormal at different percentiles. Figure 29 and Figure 30 present the same underlying calculation from two directions. Figure 29 fixes CoV_σ and varies $\mathrm{CoV}_{bootstrap}$ along the horizontal axis; Figure 30 does the reverse. Each figure shows six return periods down the page, with VaR in the left column and TVaR in the right. In Figure 30 every curve begins at exactly unity, since $\mathrm{CoV}_\sigma = 0$ recovers the simple lognormal by construction. Note that the horizontal axes are identical in all exhibits (hence only the bottom pairs are labelled); the vertical axes are not (but are labelled individually).

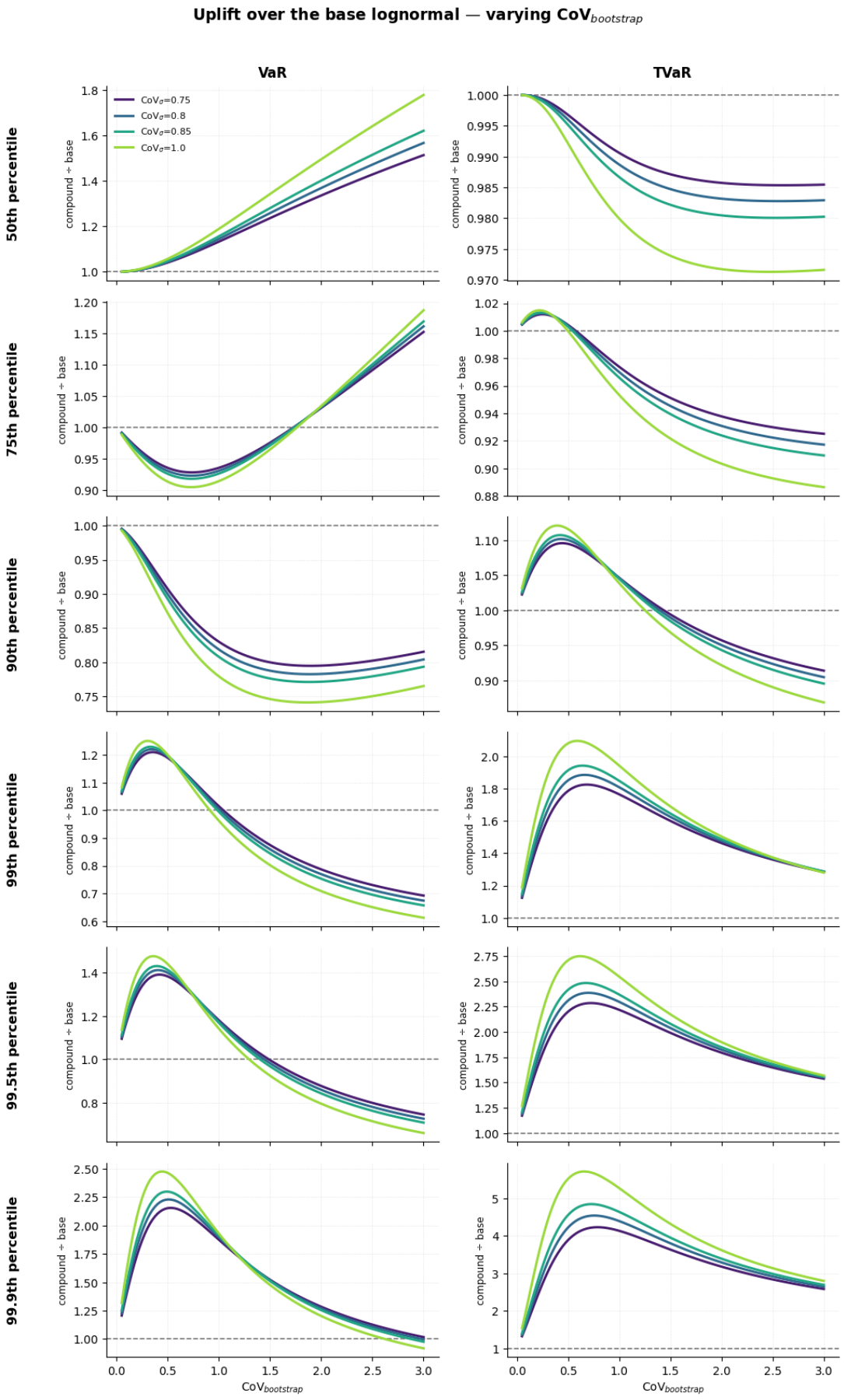


Figure 29: Uplift over base – varying $CoV_{bootstrap}$

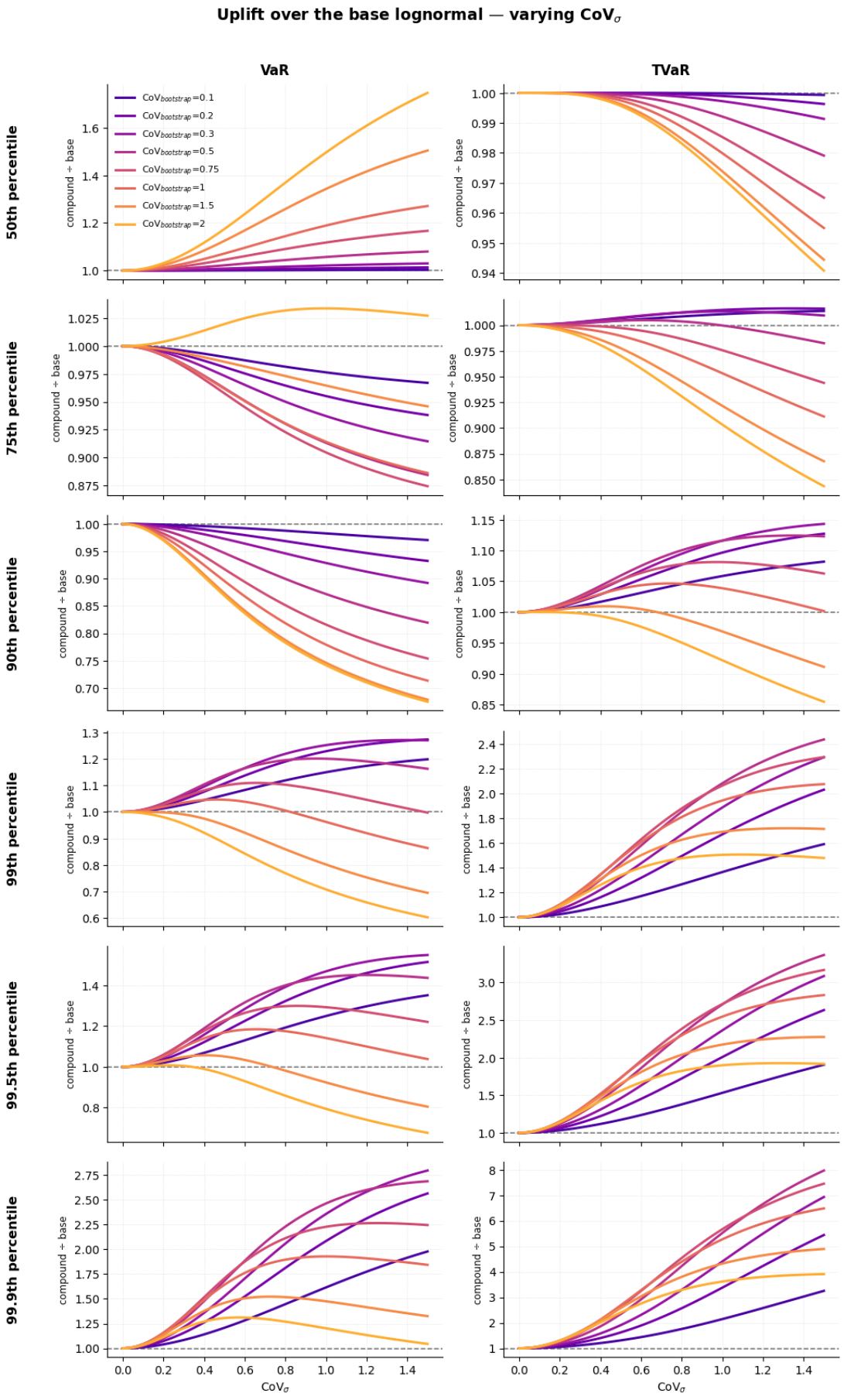


Figure 30: Uplift over base - varying CoV_σ

Appendix G: Alternative models considered

This appendix summarises alternative models considered during the development of the compound lognormal framework. The intention is not to provide a full treatment of each alternative, but to document the main variants tested and explain why the stochastic σ model detailed in the main analysis was selected.

Each model was calibrated using the same backtesting framework as the main analysis, on the full dataset (n = 177,564). For every alternative considered, a wide range of parameterisations was tested over the model's additional parameter(s), in order to identify the configuration under which that model performs best. The statistics and figures reported below are the parameterisation achieving the lowest Cramér–von Mises (CvM) statistic against the Uniform(0,1) distribution so that each alternative is represented at its most favourable setting. The remaining parameterisations are not shown, for conciseness. Lower CvM, KS and χ^2 values indicate better calibration; these statistics are used throughout as relative diagnostics rather than as formal hypothesis tests, and because the reported parameters were selected to minimise CvM, they should not be interpreted as out-of-sample measures of fit. Table 10 summarises the results for the main models considered.

Two reference points are used throughout. The first is the simple lognormal model with no compounding, representing the calibration achieved by taking the bootstrap output at face value. The second is the main analysis of this paper, the compound lognormal with stochastic σ, $\sigma \sim LN(\sigma_0, \sigma_\sigma)$, $\text{CoV}_\sigma = 0.80$.

Model	Parameterisation	Suggested parameter(s)	CvM	KS	χ^2
Simple lognormal	No compounding	—	234.0	0.067	19,405
Stochastic σ (main analysis)	$\sigma \sim \text{lognormal}(\sigma_0, \sigma_\sigma)$	$\text{CoV}_\sigma = 0.80$	**25.3**	**0.021**	**1,792**
Stochastic μ	$\mu \sim \text{lognormal}(\mu_0, \sigma_\mu)$	$\text{CoV}_\mu = 0.0005$	248.8	0.068	19,360
Stochastic μ	$\mu \sim \text{normal}(\mu_0, \sigma_\mu)$	$\text{CoV}_\mu = 0.0005$	248.5	0.068	19,256
Stochastic mean	$\text{mean} \sim \text{lognormal}(\mu_{BE}, \sigma_{BE})$	$\text{CoV} = 0.002$	234.5	0.067	19,360
Stochastic mean	$\text{mean} \sim \text{normal}(BE, \sigma_{BE})$	$\text{CoV} = 0.002$	234.0	0.066	19,288
Stochastic μ and σ	$\mu \sim \text{lognormal}$, $\sigma \sim \text{lognormal}$	$\text{CoV}_\mu = 0.001$, $\text{CoV}_\sigma = 0.95$	21.4	0.024	1,675
Log Student's t	ν degrees of freedom	$\nu = 4.8$	112.1	0.047	11,605
Stochastic σ	$\sigma \sim$ gamma	$\text{CoV}_\sigma = 0.70$	19.0	0.017	862
Stochastic σ	$\sigma \sim$ Weibull	$\text{CoV}_\sigma = 0.60$	23.0	0.020	1,417
Stochastic σ	$\sigma \sim$ beta	$\text{CoV}_\sigma = 0.70$	19.7	0.020	750
Stochastic CoV	$\text{CoV} \sim \text{lognormal}$	$\text{CoV}_{CoV} = 0.75$	24.5	0.022	2,190
Stochastic CoV	CoV ~ gamma	$\text{CoV}_{CoV} = 0.70$	16.8	0.017	931
Stochastic CoV	CoV ~ Weibull	$\text{CoV}_{CoV} = 0.60$	21.2	0.020	1,523
Stochastic CoV	CoV ~ beta	$\text{CoV}_{CoV} = 0.70$	16.7	0.019	852

Table 10: Alternative models considered

Compound lognormal with stochastic μ

In both cases σ is held fixed at the value implied by the bootstrap CoV, and $\mu_0 = \log(BE) - \sigma^2/2$ retains the centring of the underlying lognormal model. The dispersion parameter σ_μ is expressed relative to μ_0 so that a single grid may be searched across cohorts of very different size.

Neither specification improves upon the simple lognormal model. Goodness of fit degrades monotonically as CoV_μ increases from zero, so that the grid-search optimum lies at the lower boundary of the search range in both cases,

and is best interpreted as indicating $\text{CoV}_\mu \to 0$, i.e. no compounding at all. Even at the smallest values tested, both models are marginally worse than the uncompounded baseline (CvM = 248.8 and 248.5 versus 234.0).

The choice between the lognormal and normal specifications is immaterial. At the small values of σ_μ where these models are least poorly calibrated, the two distributions are near-indistinguishable, and their fit statistics differ only in the third significant figure.

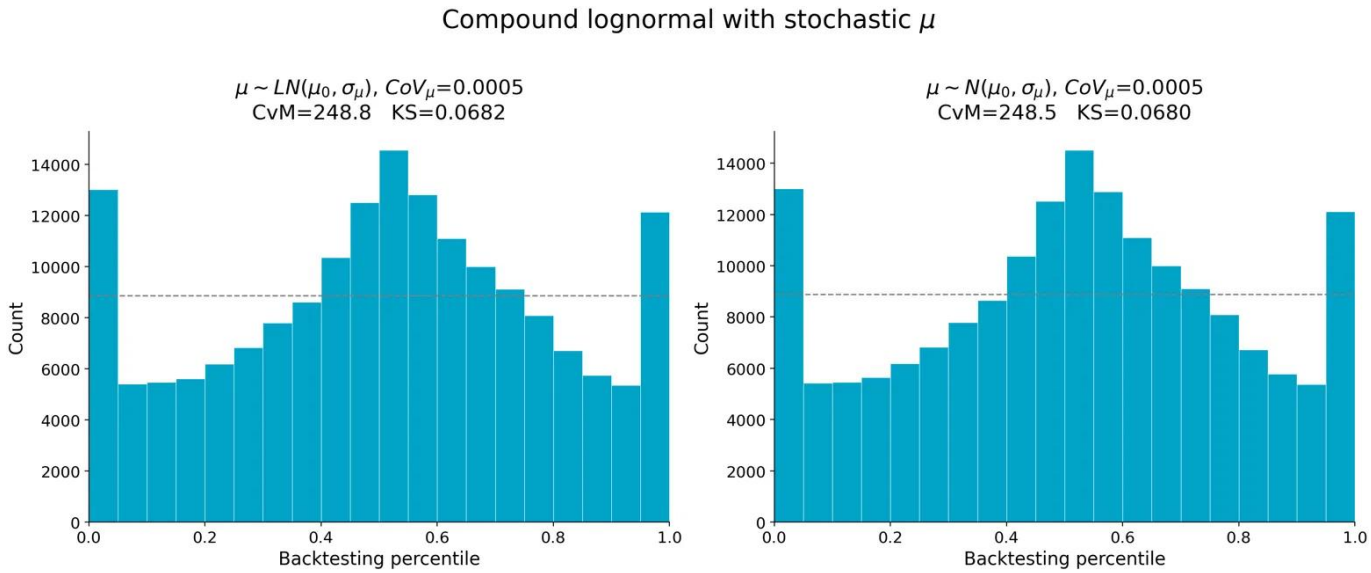


Figure 31: Compound lognormal with stochastic μ

Compound lognormal with stochastic mean

We next considered placing the uncertainty on the mean (the best estimate) rather than on μ directly:

$$\text{mean} \sim LN(\mu_{BE}, \sigma_{BE}), \text{mean} \sim N(BE, \sigma_{BE})$$

with μ re-derived from each drawn mean as $\log(\text{mean}) - \sigma^2/2$ and σ again held fixed. This is a more natural parameterisation, in that the perturbation is applied to a quantity on the scale of the reserve itself rather than to its logarithm, and the resulting displacement in log space is correspondingly better scaled.

These models are, at their optimum, indistinguishable from the uncompounded baseline (CvM = 234.5 and 234.0 respectively, against 234.0). The grid search identifies a shallow optimum near $\text{CoV} = 0.002$, but the improvement over no compounding is not material, and fit degrades steadily for larger values. As with the previous model, the lognormal and normal specifications perform equivalently.

We note one practical distinction between the two. The normal specification is symmetric in level space and can therefore in principle draw a negative mean, which is not typically meaningful for a reserve; such draws were floored at a small positive value. At the CoV values relevant here such draws are rare and do not affect the results, but the lognormal specification avoids the issue by construction.

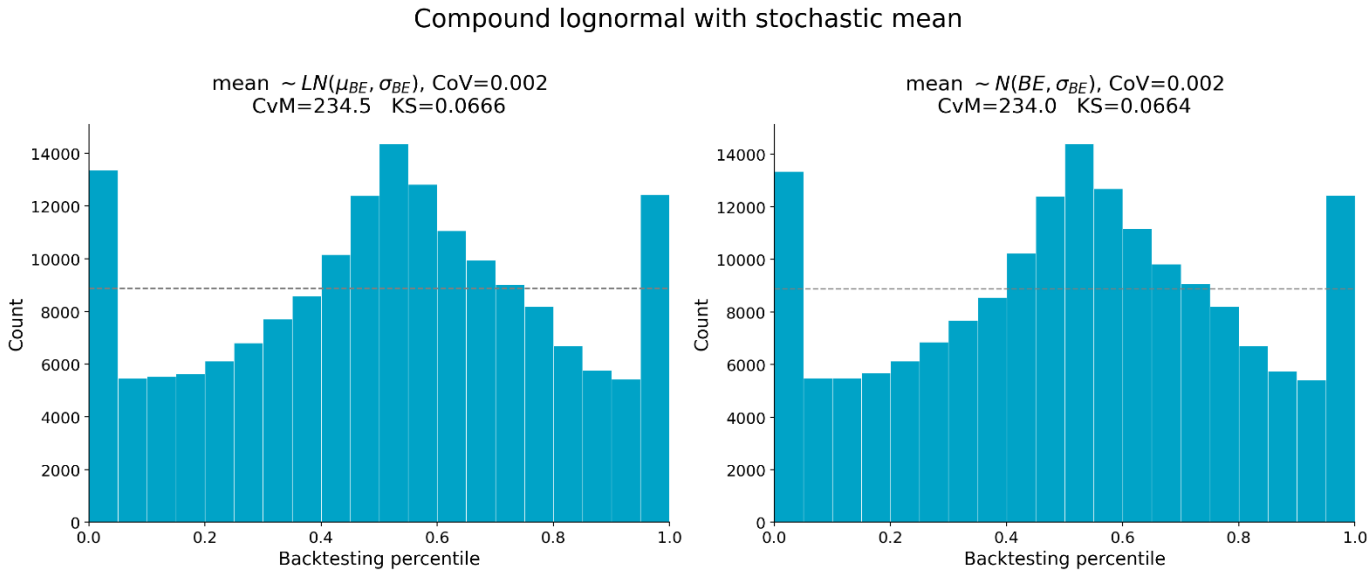


Figure 32: Compound lognormal with stochastic mean

Compound lognormal distribution with stochastic μ and σ parameters

We then considered compounding both parameters simultaneously, with independent lognormal mixing distributions:

$$\mu \sim LN(\mu_0, \sigma_\mu), \sigma \sim LN(\sigma_0, \sigma_\sigma)$$

The grid search was conducted jointly over both dispersion parameters, and identifies $\text{CoV}_\mu = 0.001$ and $\text{CoV}_\sigma = 0.95$, giving CvM = 21.4 against the main analysis's 25.3. The improvement is modest, and is accompanied by a slight deterioration in the KS statistic (0.0241 against 0.0207), so the two models are not unambiguously ranked. As previously noted, the contribution of the μ dimension is confined to very small values of CoV_μ, for the scale reasons discussed; essentially all of the improvement over the uncompounded baseline continues to come from the σ dimension.

We record one implementation point: for large draws of σ, the quantity $\mu_0 = \log(BE) - \sigma^2/2$ may become negative, in which case a lognormal mixing distribution centred on μ_0 is not defined. We therefore implement the μ compounding as a mean-preserving multiplicative lognormal shock, which is well defined for either sign of μ_0 and coincides with the intended specification wherever $\mu_0 > 0$.

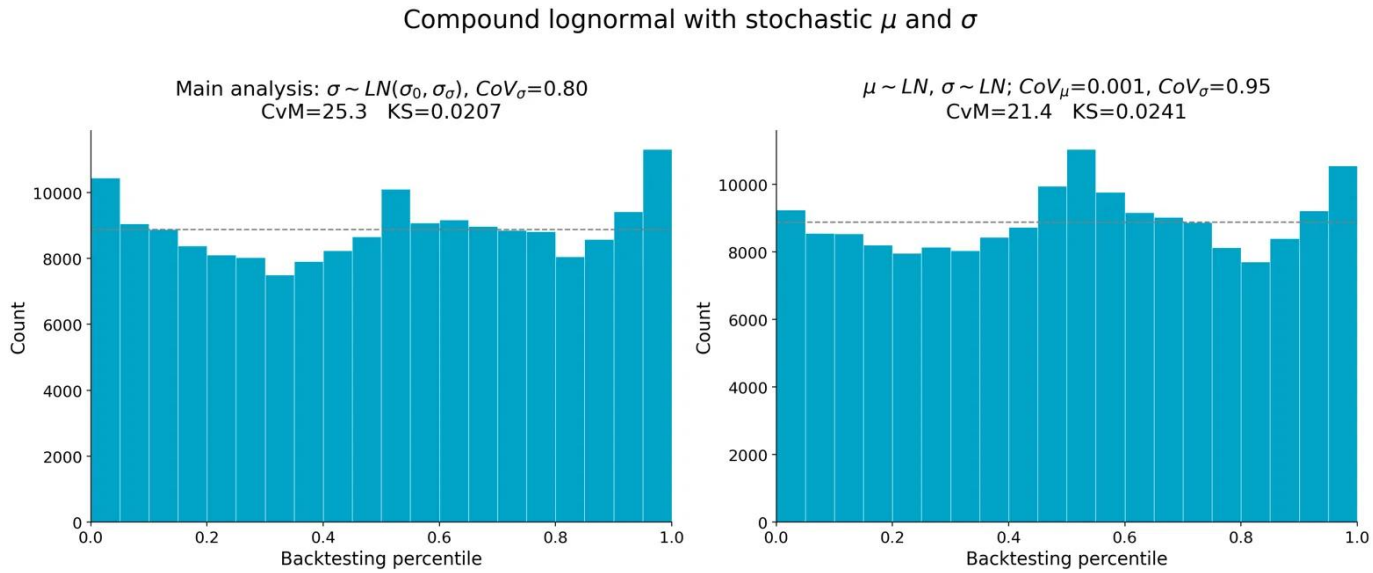


Figure 33: Compound lognormal with stochastic μ and σ

Log Student's t distribution

As an alternative to compounding, we considered replacing the lognormal with a log Student's t: that is, taking $\log(\text{ultimate})$ to follow a location-scale Student's t distribution with ν degrees of freedom, with location μ_0 as before and scale chosen as $\sigma\sqrt{(\nu-2)/\nu}$ so that the total variance in log space matches that of the corresponding lognormal fit. This has the practical advantage of a closed-form distribution function, requiring no simulation.

The degrees of freedom parameter ν governs tail weight: smaller values produce heavier tails and a correspondingly higher probability of extreme outcomes, while larger values cause the distribution to converge to the lognormal as $\nu \to \infty$. A grid search over ν identifies $\nu = 4.8$, giving CvM = 112.1. This is an improvement on the uncompounded lognormal, approximately halving the CvM statistic, and confirms that the data exhibit materially heavier tails than a single lognormal admits. It does not, however, approach the calibration achieved by the compound models, which is to be expected: a single global ν applies the same degree of tail weight to every cohort, whereas the compound models adapt the dispersion to each cohort's own bootstrap CoV.

The residual pattern in Figure 34 is consistent with this interpretation. The N-shape, or central hump of the uncompounded model is reduced, but the characteristic W shape remains.

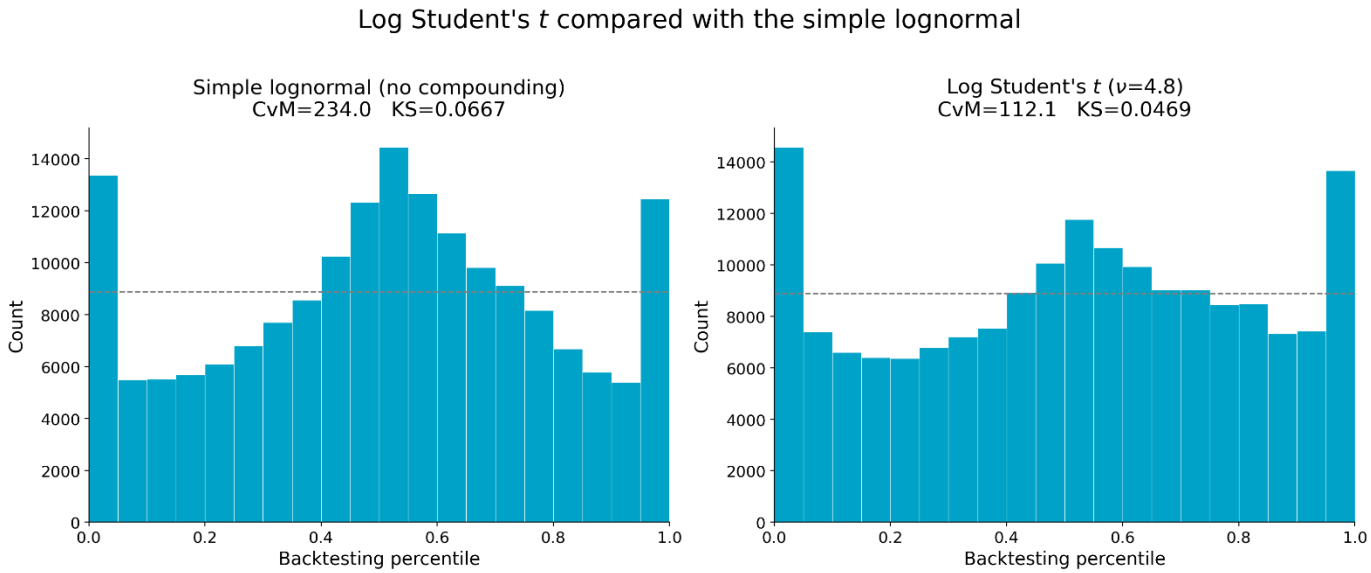


Figure 34: Log student's t

A more fundamental objection is that this log Student's t does not possess a finite mean. For any v, the polynomial tails of the t distribution decay too slowly to offset the exponential of the transformation, so E[X] is infinite. In contrast, the compound models, which preserve E[X] = BE exactly by construction. The backtesting percentiles remain well defined, since they depend only on the distribution function.

Alternative mixing distributions for σ and CoV

Finally, having established stochastic σ as the most promising structure, we tested whether the lognormal mixing distribution used in the main analysis is itself the best choice, by substituting gamma, Weibull and beta mixing distributions in its place. Each was parameterised by moment-matching so that the mixing distribution has mean equal to the cohort's bootstrap-implied σ (or CoV) and coefficient of variation equal to the grid-searched dispersion parameter. For the Weibull this requires solving numerically for the shape parameter; for the beta, draws were mapped onto a bounded support of $[0, 8\hat{\sigma}]$, reflecting the fact that neither σ nor CoV is plausibly unbounded above.

We tested each family in two forms: mixing σ directly, and mixing the CoV from which σ is then derived.

All six variants perform comparably to, and several modestly better than, the lognormal mixing used in the main analysis (CvM = 25.3). The best performers are the gamma and beta mixing distributions applied to CoV (CvM = 16.8 and 16.7 respectively), with the σ-mixing forms close behind (19.0 and 19.7). The Weibull is the weakest of the three families in both forms (23.0 and 21.2), though still competitive.

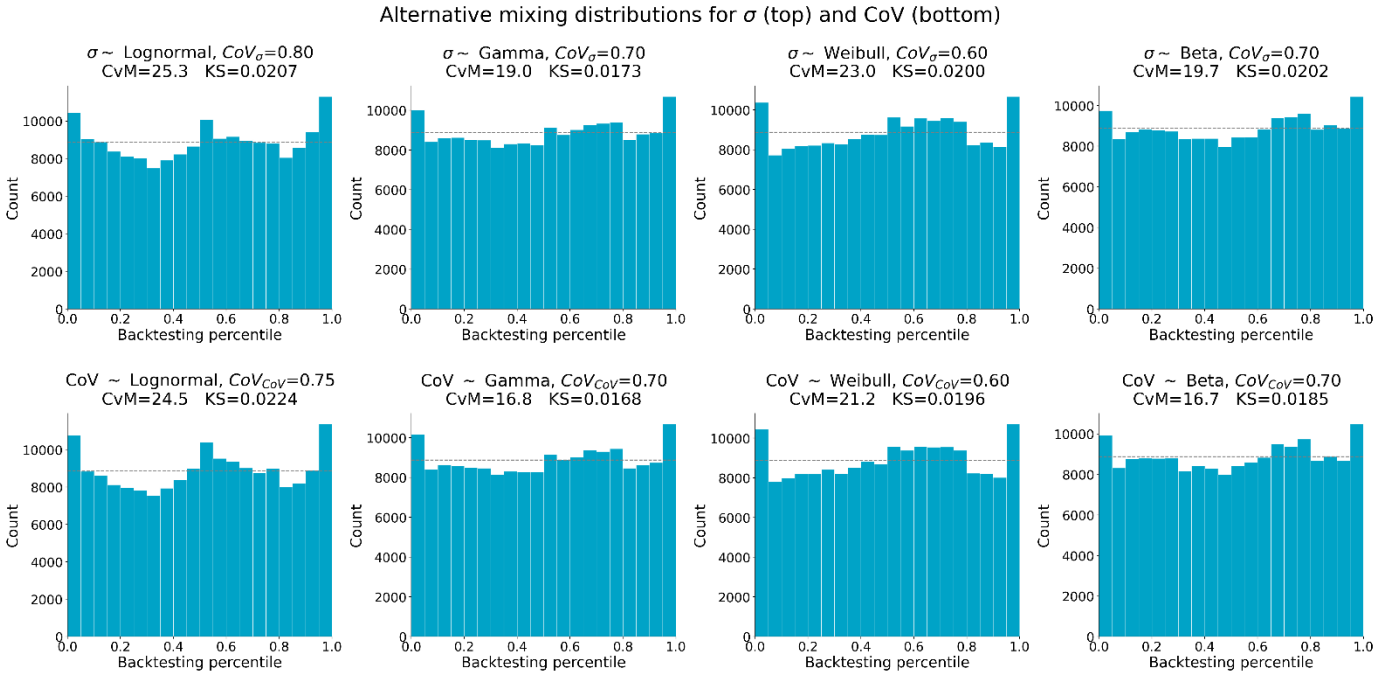


Figure 35: Mixing distributions

The substantive conclusion is that the choice of mixing distribution is of second-order importance. Once a sensible dispersion parameter is selected, all four families: lognormal, gamma, Weibull and beta, produce broadly similar calibration, and the differences between them are small relative to the improvement that compounding itself delivers over the uncompounded lognormal.

The lognormal specification retained in the main analysis has the advantage of analytical familiarity. It is also the more robust choice: our analysis indicates that the gamma's calibration deteriorates considerably more sharply than the lognormal's calibration when the dispersion parameter is set above or below its optimum.

Conclusions

Of the alternatives considered, none improves on the stochastic-σ compound lognormal adopted in the main analysis in terms of both fit and parsimony. Models placing uncertainty on the location parameter alone – whether on μ or on the mean, and whether lognormally or normally distributed – fail to improve on the uncompounded baseline. The joint μ–σ model offers a marginal improvement in CvM at the cost of an additional parameter and a slightly worse KS statistic. The log Student's t improves on the uncompounded lognormal while remaining well short of the compound models. The alternative mixing families confirm that the results of the main analysis are robust to the choice of mixing distribution rather than being an artefact of the lognormal assumption.

Appendix H: Incorporating model error

Model error, which we define in this paper as the risk that the model used to set the best estimate is not the "true" underlying model for how reserves should be estimated. Common reserve risk methodologies, specifically using the bootstrap to parameterise a lognormal distribution, do not allow for this source of uncertainty.

One way to allow for model error is to produce a model of the uncertainty in the best estimate reserves. The **Universe of Reserves®** [1] technique, explained in Appendix I, provides us with a measure of this uncertainty. The Universe of Reserves is a range of reserve estimates from many alternative reserve projections, each with a quality score assigned. The quality score is based on the volatility and bias of the projection method. The highest quality projections are those with low volatility and low bias.

We have explored approaches for using the Universe of Reserves to produce a model of the uncertainty in the best estimate reserves, ie a model that captures model error.

We have performed this analysis at class level. The logic behind this approach is that model error is likely to impact multiple entities and cohorts in similar ways.

An outline of the approach we followed is as follows:

- For each class, there are n reserve estimates, each with a projection quality.
- Normalise the projection quality to get a number between 0 and 1.
- Remove any duplicates to ensure all projections result in unique reserve estimates.
- To turn the projection quality scores into a distribution of reserve volatility, we need the probability that a reserve actuary would select each reserve estimate in practice. In reality, the relationship between projection quality and likelihood of selecting that reserve estimate is non-linear. Ie, the actuary is more than twice as likely to select an estimate with a confidence score of 0.9 than an estimate with a confidence score of 0.45.

[1] Universe of Reserves® is a registered trade mark of Lane Clark & Peacock LLP

We explored various potential transformations of projection quality scores, including exponential, power, and rank. Figure 36 shows some of these distributions for the Homeowners and Farmowners class for the 2022 filing year.

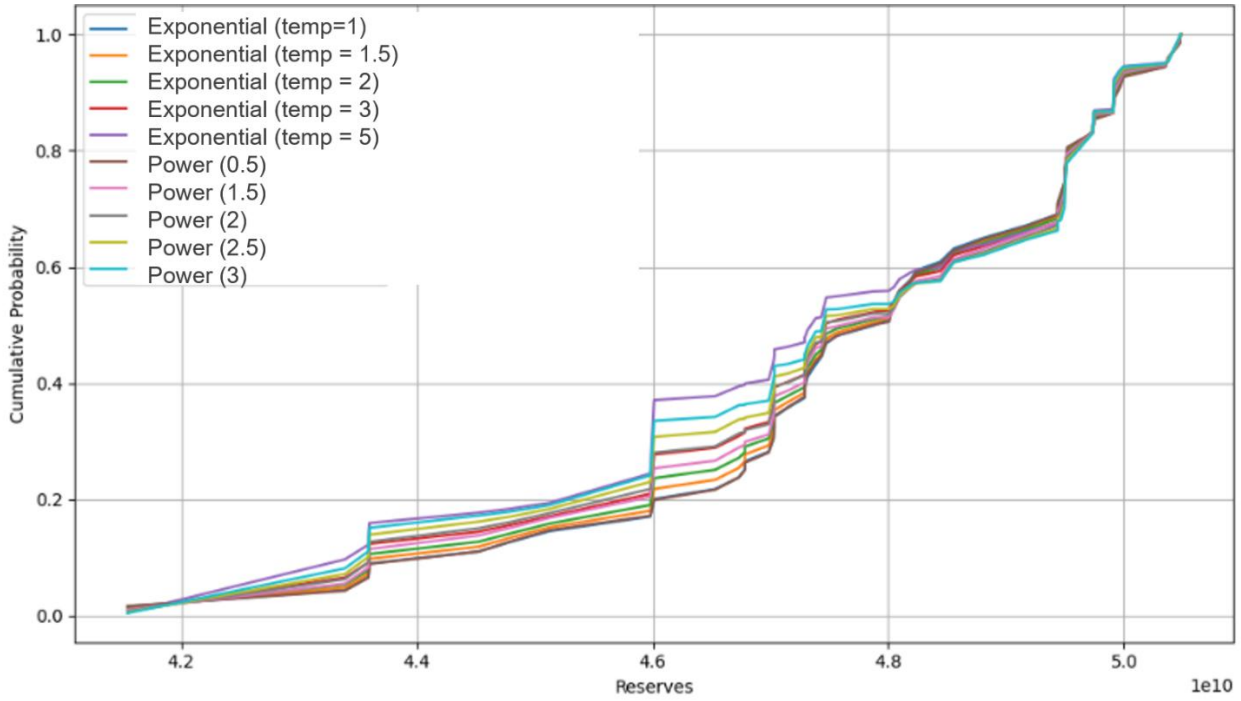


Figure 36: Potential transformations for projection quality scores

The table below shows summary statistics of CoVs of these distributions. This shows that the CoV of the best estimate reserves isn't materially sensitive to the transformation selected.

Class	Max	Min	Range	Average	Std Dev
COMMERCIAL AUTO&TRUCK LIABILITY&MEDICAL	0.0871	0.0707	0.0164	0.0816	0.0049
COMMERCIAL MULTIPLE PERIL	0.0791	0.0657	0.0135	0.0728	0.0047
HOMEOWNERS&FARMOWNERS	0.0488	0.0435	0.0053	0.0458	0.0018
MEDICAL PROFESSIONAL LIABILITY - CM	0.1679	0.1054	0.0625	0.1412	0.0223
MEDICAL PROFESSIONAL LIABILITY - OC	0.2560	0.1303	0.1257	0.2019	0.0432
OTHER LIABILITY - CLAIMS-MADE	0.1121	0.0882	0.0239	0.1006	0.0075
OTHER LIABILITY - OCCURRENCE	0.1188	0.1156	0.0032	0.1169	0.0012
PP AUTO LIABILITY&MEDICAL	0.0659	0.0554	0.0105	0.0603	0.0035
PRODUCTS LIABILITY - CLAIMS-MADE	0.2887	0.2525	0.0362	0.2705	0.0123
PRODUCTS LIABILITY - OCCURRENCE	0.1096	0.0910	0.0186	0.1017	0.0056
WC EXCL EXCESS WC	0.2127	0.1901	0.0226	0.2049	0.0077

Table 11: Distribution of CoVs from potential transformations

Our view is that a full model for reserve risk would incorporate volatility around both μ and σ (or, equivalently, around both the best estimate reserve and standard deviation as calculated by the bootstrap). We view using the CoV as calculated above as a potentially promising avenue for this parameterisation. Further work is required to consider how this estimate of model error could be identified and incorporated within this framework, including whether it is best captured through uncertainty around the mean, or as a separate explicit model error component. In either case, the question of how to translate between reserve quality projection scores and probabilities remains unsolved.

Appendix I: Universe of Reserves

A key source of uncertainty that is commonly not allowed for in current reserve risk modelling techniques is the uncertainty in the best estimate (ie model error). One way to quantify this source of uncertainty is via a Universe of Reserves technique. A potential way of doing this was outlined by Durkin and Stone (2019) in their talk at the IFoA's Technical Issues in General Insurance conference on 26 April 2019.

Methodology

The Universe of Reserves is a range of reserve estimates from many alternative reserve projections. Each reserve estimate is set using traditional actuarial techniques and is based on a different set of assumptions. For example, alternative averaging periods for development patterns and initial expected loss ratios, and using paid and incurred data.

Each projection is then assigned a "quality" score. The assessment of projection quality is based on the volatility and bias of the projection, by using a backtesting approach and removing years of experience from the claims data:

- **Bias**: based on actual versus expected experience over each of the backtesting periods, for the relevant set of assumptions. The actual claims experience in a given period is compared against the claims that would have been projected for that period by the set of assumptions used in the projection.
- **Volatility**: based on assessing the standard deviation of the ultimates from the projection method, as they are rolled-forward across the backtesting periods.

The "highest quality" projections are those with low volatility and low bias.

Figure 37 shows an illustrative example of the Universe of Reserves.

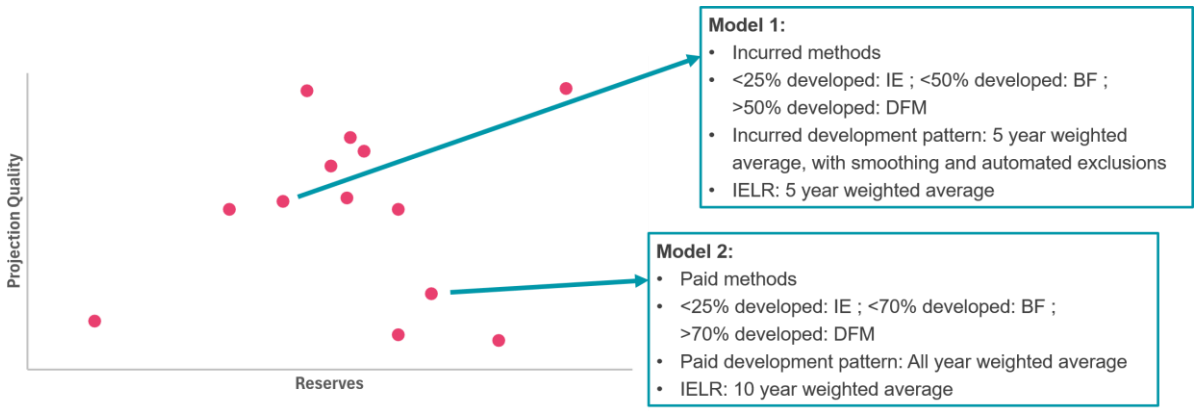


Figure 37: Universe of Reserves

Limitations

The assessment of projection quality described above is based on the assumption that past experience is a good indicator of future experience. If future experience is expected to be significantly different to past experience, the projection quality measure described above is not necessarily a good indicator of how well the particular model and set of assumptions will perform at projecting future experience.

The range of reserve estimates might not be the full range of reasonable best estimate reserves. Each projection is based on an alternative set of assumptions. In order to quantify the uncertainty in the best estimate, it is important that the range of reserve estimates adequately captures the reasonable range of alternative methods and assumptions. However, in practice, the range of alternative methods and assumptions considered is limited and will not represent the full range of possible outcomes.