Japanese Journal of Statistics and Data Science (2024) 7:1155–1170
https://doi.org/10.1007/s42081-024-00260-3

ORIGINAL PAPER

Risk and Statistics in Actuarial Science

Applications of Population Sampling to Insurance Ratemaking and Reserving

Sebastián Calcetero Vanegas[1] · **X. Sheldon Lin**[1]

Received: 24 December 2023 / Revised: 18 May 2024 / Accepted: 14 June 2024 /
Published online: 7 August 2024

Abstract
This paper explores the underutilized application of population sampling in the realm of actuarial science, a field where these statistical methodologies have been traditionally overlooked. Focusing on two distinct applications within insurance ratemaking and reserving, we unveil innovative approaches to address challenges in actuarial contexts and provide valuable insights into advancing methodologies in the field. The first application introduces population sampling as a solution to the computational complexities inherent in credibility premium calculation, particularly under Bayesian regression models. By combining population sampling with surrogate modeling, we present a method to manage computation challenges effectively. The second application delves into incurred but not reported reserves, challenging the conventional Chain–Ladder method and individual reserving models by incorporating population sampling. Proposing a reserve estimator based on inverse probability weighting techniques, we demonstrate a statistically robust, distribution-free method for IBNR reserving, emphasizing the integration of granular policyholder information

Keywords Population sampling · Surrogate modeling · Credibility premiums · IBNR reserve

1 Introduction

Population/survey sampling, as an important branch in statistics, has been widely used in censuses and election predictions and has found applications in other disciplines. However, its application has not been fully explored in insurance. In this paper, we will showcase some recent developments on its applications in actuarial science. In particular, we will present two applications to illustrate how population sampling can be used in insurance ratemaking and reserving.

✉ X. Sheldon Lin
sheldon.lin@utoronto.ca

[1] Department of Statistical Sciences, University of Toronto, Toronto, ON, Canada

The first application considers credibility premium calculation for a large and highly inhomogeneous non-life insurance (such as auto insurance) portfolio. The goal of credibility or experience ratemaking in insurance is to determine premiums that account for both the policyholders' attributes and their claim history. The underlying model for such a purpose is usually a Bayesian regression model. When we choose a data-driven model, the credibility premiums based on a premium principle (e.g., the expectation or a quantitle of predicted future losses) must be obtained via numerical methods e.g. simulation via Markov Chain Monte Carlo. Methods of this kind are clearly computationally expensive for a large portfolio as they must be applied at the policy level. Population sampling combined with surrogate modelling allows us to address the computation challenges in a manageable manner.

The second application deals with the calculation of incurred but not reported (IBNR) reserves that are commonly determined using a Chain–Ladder method in practice. The Chain–Ladder method constructs a claim triangle heuristically so that expected future claims are simply computed using development factors but it neglects the heterogeneity of policyholders. With the help of population sampling, we are able to incorporate granular information individually into the Chain–Ladder method. In particular, we view the claim reserving problem as a population sampling problem and propose a reserve estimator based on inverse probability weighting techniques, with weights driven by policyholders' attributes. The framework provides a statistically sound method for IBNR reserving in a frequency and severity distribution-free way, while also incorporating the capability to utilize granular information via a regression-type framework.

This paper is organized as follows. In the next section, we provide a brief overview of population sampling and general procedure on how population sampling is applied to actuarial problems. In Sect. 3, we apply population sampling to credibility premium calculation under any Bayesian regression model and when the credibility premium has no closed form expressions. Section 4 focuses on the application of population sampling to IBNR reserving. The last section concludes the paper.

2 Brief Overview of Population Sampling and General Procedure for Actuarial Applications

In this section, we present some known results in population sampling and propose a general procedure on how to apply them to actuarial applications. For comprehensive coverage of this topic readers may be referred to Tillé (2011).

Population sampling provides a methodological framework to sample from a population with the goal to make inferences related to the entire population. Consider a population of size N and let $L_i, i = 1, \ldots, N$, denote a quantity of interest associated with individual i of the population. In the insurance context, N is the total number of insurance policies and L_i may the liability or the premium of policy i. Suppose now we want to select a small sample of $n \ll N$: $L_1^*, \ldots, L_n^*$, from the population such that one can make a good inference to the entire population. Those selected individuals $L_1^*, \ldots, L_n^*$ must be 'representative'. A sampling design is to assign a Bernoulli random variable $I_i, i = 1, \ldots, N$, to each individual with the probability of success π_i.

I_i is known as the membership indicator, indicating whether the individual i belongs or not to the sample, respectively, and π_i as the inclusion probabilities and is defined according to the sample design.

In order for the sample to be representative, some criteria are imposed on the sampling design. One is to ensure the total liabilities/premiums $\sum_{i=1}^{N} L_i$ are satisfactorily estimated. A natural choice is the linear estimator $\sum_{j=1}^{n} w_j L_j^*$. It is shown that $\sum_{j=1}^{n} w_j L_j^*$ is an unbiased estimator estimator of $\sum_{i=1}^{N} L_i$, if $\sum_{i=1}^{N} p_i = n$ and $w_j = 1/\pi_j$. See Thompson (2012) or Särndal et al. (2003). The estimator $\sum_{j=1}^{n} \frac{1}{\pi_j} L_j^*$ is called the Horvitz–Thompson (HT) estimator. In addition, the HT estimator is very attractive due to its simplicity, and it does not rely on assumptions about the underlying distribution of the population, making it suitable for both finite and infinite population settings. It is also been called the IPW estimator where IPW stands for Inverse Probability Weight.

Another important aspect of having a representative sample is that we want the sample to be balanced. As mentioned earlier an insurance portfolio is highly heterogeneous and each policy has its unique risk profile/attributes. For example, the losses from an auto insurance policy are highly relayed to the car type, driver's age and the years of licensing, etc. The sample must take the composition of the attributes of all the policies into account, which can be described with the balanceness of the sample.

Let $\boldsymbol{x}_i$ be the attributes of individual i. A sample is said to be balanced if its Horvitz–Thompson estimator satisfies

$$\sum_{j=1}^{n} \frac{1}{\pi_j^*} \boldsymbol{x}_j^* = \sum_{i=1}^{N} \boldsymbol{x}_i.$$

It can be shown that if we assume that L_i follows a linear model:

$$L_i = (\boldsymbol{x}_i)^T \boldsymbol{\beta} + \epsilon_i, i = 1, \ldots, N,$$

where the residuals ϵ_i's are iid with zero means and standard deviation σ_i, and the sample is balanced, then the variance of the Horvitz–Thompson estimator has the minimal variance. That is, The Horvitz–Thompson estimator is a minimum-variance unbiased estimator (MVUE) estimator of $\sum_{i=1}^{N} L_i$.

In Deville and Tillé (2004), they proposed an iterative sampling algorithm termed the cube method to select a balanced sample. Suppose that each unit is equipped with r attributes. The algorithm translates the given inclusion probabilities to a vector of at least $(N - r)$ zeros (not selected) or ones (selected). Since the balanced condition

$$\sum_{j=1}^{n} \frac{1}{\pi_j^*} \boldsymbol{x}_j^* = \sum_{i=1}^{N} \boldsymbol{x}_i.$$

may be written in a matrix form:

$$\mathbf{A}\, \boldsymbol{I} = \mathbf{A}\, \boldsymbol{\pi}, \tag{2.1}$$

where $\mathbf{A} = (\boldsymbol{x}_1/\pi_1, \ldots, \boldsymbol{x}_N/\pi_N)$ is a $r \times N$ matrix, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)^T$ and $\boldsymbol{I} = (I_1, \ldots, I_N)^T$ a random column vectors, where $I_i = 1$ or 0 indicating the inclusion of individual i. Equation (2.1) implies that $\boldsymbol{I}$ can be written as $\boldsymbol{\pi} + \boldsymbol{u}$ where $\boldsymbol{u}$ is in the kernel of matrix $\mathbf{A}$, i.e. $\mathbf{A}\boldsymbol{u} = \mathbf{0}$. Using this fact. one may adjust the inclusion probabilities $\boldsymbol{\pi}$ as a vector randomly iteratively inside the kernel of $\mathbf{A}$ until it reaches to a point that is close to a vertex of the N dimensional hypercube in the following manner:

First set $\boldsymbol{\pi}(1) = \boldsymbol{\pi}$. For Iteration k,

- randomly generate a vector $\boldsymbol{u}(k)$ in the kernel of matrix $\mathbf{A}$. Set $\boldsymbol{u}_i(k) = 0$ if $\boldsymbol{\pi}_i(k) = 0$ or 1.
- compute $\lambda_1^*(k)$ and $\lambda_2^*(k)$, the largest values among $\lambda_1(k)$ and $\lambda_2(k)$ such that:

$$0 \le \boldsymbol{\pi}(k) + \lambda_1(k)\boldsymbol{u}(k) \le 1;$$
$$0 \le \boldsymbol{\pi}(k) - \lambda_2(k)\boldsymbol{u}(k) \le 1.$$

- compute $\boldsymbol{\pi}(k+1)$ as

$$\boldsymbol{\pi}(k+1) = \boldsymbol{\pi}(k) + \lambda_1^*(k)\boldsymbol{u}(k) \quad \text{with probability} \quad \frac{\lambda_2^*(k)}{\lambda_1^*(k) + \lambda_2^*(k)};$$
$$\boldsymbol{\pi}(k+1) = \boldsymbol{\pi}(k) - \lambda_2^*(k)\boldsymbol{u}(k) \quad \text{with probability} \quad \frac{\lambda_1^*(k)}{\lambda_1^*(k) + \lambda_2^*(k)}.$$

The iterations continue until $\boldsymbol{\pi}(k)$ stops changing. Finally convert all non-integer values in the final iteration to zero or one via linear programming.

A well selected balanced sample plays a pivotal role in reducing the computing time for calculating quantities of interest for a large and highly heterogeneous insurance portfolio. Instead of running through all the policies, we can now just compute the quantities on the selected policies and extrapolate them to the entire portfolio. In an application, we may employ the following procedure:

1. Select a small number of policies (1–5%) from the portfolio using the cube method. As discussed, such policies will represent the portfolio well. We call them representative policies.
2. Identify a policy-specific variable or summary statistic that is capable of extracting information from each policy including the policy attributes and claim history. The variable/summary statistic is application specific.
3. Use a flexible (surrogate) model/function such as a spline to link the statistic to the quantity of interest to be computed (premium, predicted liability for reserve calculation, etc.).
4. Perform simulation on each of the representative policies.
5. Estimate the parameters of the surrogate function using the representative policies.
6. Make use of the estimated surrogate model to compute the quantity of interest across the entire portfolio (extrapolation) and apply the result for specific applications (predictions, reserves, profit and loss analysis, etc.).

In the next two sections, we use two applications to illustrate the aforementioned approach.

3 Application to Credibility Premium Calculation

3.1 Overview on Experience Rating and the Surrogate Model Approach

In insurance, a Bayesian model is employed to upgrade premiums by integrating a policyholder's existing risk behavior understanding ("a priori" information) with their actual claims experience ("a posteriori" information). See for e.g Bühlmann and Gisler (2005). This Bayesian approach allows for a more nuanced calculation of upgraded premiums compared to traditional methods. However, the practical implementation of Bayesian models faces significant computational challenges. Insurance data is characterized for its complexities, including but not limited to big data concerns, non-neglible heterogeneity, and the inclusion of diverse sources of information e.g. telematics Pechon et al. (2018), Chan et al. (2023). As such, many of the models that provide a realistic fit to insurance data are mathematically complex and analytically intractable. Deriving Bayesian premiums often requires computationally intensive numerical approximations, such as simulations using Markov Chain Monte Carlo methods. See for e.g Xacur and Garrido (2018), Zhang et al. (2018), Ahn et al. (2021). Handling large and diverse insurance portfolios exacerbates these challenges, necessitating a considerable number of simulations for each policyholder. Moreover, the resulting premiums are obtained through numerical approximations, leading to a "black-box" scenario that lacks practical interpretability, posing a barrier to widespread adoption among practitioners who seek transparent and explainable results for clients and regulators.

Here we illustrate the surrogate modeling approach from the previous section to overcome computational challenges and the absence of analytical expressions in Bayesian credibility models. We particularly focus on efficient and transparent experience rating on complex insurance data sets and large portfolios. This discussion follows directly from our work in Calcetero-Vanegas et al. (2024), and refer the reader to it for the technical details or further information.

To introduce the main idea, note that any Bayesian pricing formula for the risk in the next period Y_{n+1} given the claim history $\mathbf{Y}_n = (Y_1, \ldots, Y_n)$ under any premium principle (see for e.g (Kaas et al., 2008, p. 115)) can be generally expressed as:

$$\Pi(Y_{n+1}|\mathbf{Y}_n) = G_\Pi(\mathbf{Y}_n, n, \mathcal{O}) \tag{3.1}$$

where $G_\Pi(\cdot)$ is the theoretical or true functional form that links the claim history of the policyholder $\mathbf{Y}_n$ and the set of model parameters and policyholder attributes $\mathcal{O}$ with the Bayesian premium, under the premium principle Π. The functional form of $G_\Pi(\cdot)$ entirely depends on the premium principle and the underlying Bayesian model, and as we mentioned earlier, it likely lacks an analytical expression. This function is the ultimate target of the surrogate model.

To streamline the estimation of the surrogate function and effectively link the policyholder with their corresponding premium, we employ an experience-based summary statistic, denoted as $T(\mathbf{Y}_n)$. This statistic encompasses both the policyholder attributes and their associated claim history. We select this statistic to be approximately sufficient, making it suitable for representing these two factors with minimal information

loss. This summary statistic serves as the primary input for the surrogate model. As demonstrated by Calcetero-Vanegas et al. (2024), one choice for such experience-based summary statistic is derived from the conditional log-likelihood function of a policyholder's claim history:

$$T(\mathbf{Y}_n) = \sum_{j=1}^{n} \log f(Y_j|\Theta = \tilde{\theta}, \mathcal{O}) \tag{3.2}$$

where $f(Y_j|\Theta = \tilde{\theta}, \mathcal{O})$ is the underlying model distribution of the Y_j, Θ is the latent random variable of the Bayesian model, and $\tilde{\theta}$ acts as tunning parameter that is policyholder specific.

That said, the aim is to find an approximation of the Bayesian premiums via a surrogate such that:

$$\Pi(Y_{n+1}|\mathbf{Y}_n) \approx \hat{G}_\Pi(T(\mathbf{Y}_n), n, \mathcal{O}). \tag{3.3}$$

In this vein, the procedure for estimating Bayesian premiums for the portfolio of policyholders is straightforward. We first choose a sample of representative policies, typically around 1% to 5%, and then proceed to fit the Bayesian model. Using the standard Markov Chain Monte Carlo (MCMC) procedure, we estimate the associated Bayesian premiums of the representative policies. Finally, a surrogate model is then derived from these estimates. The algorithm provided below delineates the general procedure to define the surrogate function (i.e. step 3 in the general framework form the previous section) through a least squares estimation. The specific structure of the surrogate is user-dependent, and while Gaussian processes are commonly utilized, other methodologies, such as B-Splines, can be equally applied. We refer to Gramacy (2020) for more information on the process of fitting a surrogate function. We will illustrate this process in the next section with a numerical illustration.

Algorithm 1 Fitting of the surrogate function

$MSE \leftarrow \text{Tol} + 1$

$\tilde{\theta}_i \leftarrow$ Random number $\forall i = 1, \ldots, M$(Number of policyholders) ▷ Start with random values for $\tilde{\theta}_i$

while $MSE \geq \text{Tol}$ **do**

$T(\mathbf{Y}_{i,n}) \leftarrow \sum_{j=1}^{n} \log f(Y_{i,j}|\tilde{\theta}_i, \mathcal{O})\ \forall i = 1, \ldots, M$ ▷ Compute the experience-based statistic

$\hat{G}(\cdot) \leftarrow \arg\min_g \sum_{i=1}^{M} \left(\hat{\Pi}_i^p - \hat{G}_\Pi(T(\mathbf{Y}_{i,n}), n_i, \mathcal{O})\right)^2$ ▷ Update the function $\hat{G}(\cdot)$ via LS

$\tilde{\theta}_i \leftarrow \arg\min_{\tilde{\theta}_i} \left(\hat{\Pi}_i^p - \hat{G}_\Pi(T(\mathbf{Y}_{i,n}), n_i, \mathcal{O})\right)^2\ \forall i = 1, \ldots, M$ ▷ Update values $\tilde{\theta}_i$

$MSE \leftarrow \sum_{i=1}^{M} \left(\hat{\Pi}_i^p - \hat{G}_\Pi(T(\mathbf{Y}_{i,n}), n_i, \mathcal{O})\right)^2$ ▷ Upgrade current interpolation error

end while

After obtaining the well-fitted surrogate function $\hat{G}_\Pi(\cdot)$, evaluating a new policyholder's experience rating involves a direct evaluation of this function. This simplifies the calculation of Bayesian premiums for extensive portfolios. The surrogate function provides an analytical link between the policyholder's attributes and claims history,

enhancing transparency in the ratemaking process. Utilizing $\hat{G}_{\Pi}(\cdot)$ for sensitivity analysis allows for interpreting premium adjustments and quantifying the impact of claim history and attributes.

3.2 Numerical Illustration with Real Data

We will now demonstrate this framework using actual data obtained from a European auto insurance company. The dataset covers policyholder contract details from January 2007 to December 2017, encompassing claim frequencies for Third Party Liability insurance (TPL) and Physical Damage (PD). The focus is on policies with both TPL and PD coverages, where the number of claims in each line may be interrelated, often stemming from the same car accident. A substantial number of policyholders renew their contracts, prompting the insurance company to conduct an experience rating analysis during contract renewal, evaluating the claims history of policyholders at that juncture.

The Bayesian model at the core of this application is a bivariate mixed negative-binomial regression model. Here, $Y_j^{(d)}$ represents the number of claims from a specific policyholder in year j, linked to the dth line of business, where $d = 1$ corresponds to PD and $d = 2$ to TPL. The associated covariates are denoted as x, and the vector of regression coefficients for the d-th coverage is represented by $\beta^{(d)}$. Additionally, $\omega^{(d)}$ is used to signify the time exposure of the contract for each coverage. This hierarchical model considered in our analysis is below, which corresponds to a bivariate negative binomial mixed model as proposed by Tzougas and di Cerchiara (2021). The fitting of this model can be easily accomplished using implementations of mixed model in `R` such as `glmer.nb`. We refer the reader to such literature for further details on estimation of the model.

$$\begin{aligned} Y_j &= \begin{pmatrix} Y_j^{(1)} \\ Y_j^{(2)} \end{pmatrix} \sim_{iid} f(y|\Theta, \langle \boldsymbol{x}, \boldsymbol{\beta} \rangle) \\ &= \text{NegBinom}(y^{(1)}; \mu^{(1)}\Theta, r^{(1)}) * \text{NegBinom}(y^{(2)}; \mu^{(2)}\Theta, r^{(2)}) \end{aligned}$$

where, for $d = 1, 2$

$$\begin{aligned} \log \mu^{(d)} = \log \omega^{(d)} &+ \beta_0^{(d)} + \beta_1^{(d)}\text{CarWeight} + \beta_2^{(d)}\text{EngineDisplace} \\ &+ \beta_3^{(d)}\text{CarAge} + \beta_4^{(d)}\text{Age} + \beta_5^{(d)}\text{EnginePower} + \beta_6^{(d)}\text{Fuel} \end{aligned}$$

and $\Theta \sim P(\theta) = \text{InvGauss}(1, \sigma^2)$. The notation $\text{NegBinom}(y; \mu, r)$ is used to signify the probability mass function of a negative binomial distribution with mean μ and dispersion r. Similarly, $\text{InvGauss}(1, \sigma^2)$ represents an Inverse-Gaussian distribution with mean 1 and variance σ^2. Note that this model could be seen as particular case of the model proposed by Pechon et al. (2018), and literature therein. It is essential to note that the model lacks analytical expressions for both the posterior and the predictive distribution. Consequently, numerical methods are indispensable for obtaining any desired quantity of interest.

Table 1 Comparison of the CPU time (in seconds) required for the calculation of premiums

Process	Total portfolio	Representative policyholders
Selecting a sample	–	32.94
Simulation for the premium	919,008.00	50,976.00
Fitting surrogate function	–	2480.22
Extrapolation	–	3.39
Total time	919,008.00 ($\approx$ 255 h)	53,492.55 ($\approx$ 15 h)

Here, we examine the implementation of the *Exponential* premium principle as an illustration, incorporating a 5% surcharge. This method is utilized by the insurance company to determine actuarial premiums.

$$\Pi(Y_{n+1}|\mathbf{Y_n}) = \frac{1}{0.05}\log(E\left(\exp(0.05 * (Y_{n+1}^{(1)} + Y_{n+1}^{(2)}))|\mathbf{Y_n}\right)).$$

To fit the surrogate model, we employ the *cube method* using the `samplecube()` function in `R` to extract a sub-portfolio of representative policies of around 5% of the total portfolio. To ensure the sample is representative of the heterogeneity of the attributes of the policies, we guarantee the balance property with respect to the average number of claims for PD and TPL, capturing claim history, as well as the average fitted values of $\mu^{(1)}$ and $\mu^{(2)}$, reflecting policyholder attributes. The computational cost of this process is minimal, as shown in Table 1.

The estimation of the premiums of the representative policies is performed via a Monte Carlo simulation scheme. It's important to note that this step is only required for the 5% sub-portfolio. However, we conduct the computationally intensive simulation for the entire portfolio to enable a comparison of premiums obtained from the surrogate model. Regarding simulation time, a single replication on the entire portfolio takes approximately 18 times longer than on the representative policyholder, as detailed in Table 1. This closely aligns with the empirical ratio of 20 associated with the proportion of 5% vs. 100% of the representative portfolio.

The surrogate model is constructed by employing a multidimensional B-Splines representation for the function $\hat{\tilde{G}}_{\Pi}$. This is fitted via least squares using the `gam()` function in `R`. See for e.g Wood (2017) for more details. The features for this model include the manual premium, the experience-based statistic, and the number of known periods in each policy. In this case, the experience-based statistic is given by the log of the probability mass function of a Negative binomial distribution, as follows

$$T(\mathbf{Y}_n) = \sum_{j=1}^{n} \log\left(\text{NegBinom}(Y_j^{(1)}; \mu^{(1)}\tilde{\theta}, r^{(1)}) * \text{NegBinom}(Y_j^{(2)}; \mu^{(2)}\tilde{\theta}, r^{(2)})\right)$$

The performance accuracy of the surrogate model is illustrated in Fig. 1 and detailed in Table 2. The outcomes presented in such table reveal that the surrogate model

Table 2 Error metrics of the surrogate model with unstructured surrogate function

Sub-portfolio	At individual level			At aggregate level		
	ME	MAE	MAPE	R^2	Error	MPE
Out of sample	− 0.0005	0.0036	0.018	0.99	71.78	0.0041
In sample	− 0.0006	0.0036	0.017	0.99	42.66	0.0024

Individual level means the metric is calculated at the policyholder level and then averaged. At the portfolio level, the metric is calculated on the aggregate premiums directly
ME mean error, *MAE* mean absolute error, *MAPE* mean absolute percentage error, R^2 coefficient of determination, *Error* True − Predicted, *MPE* mean percentage error

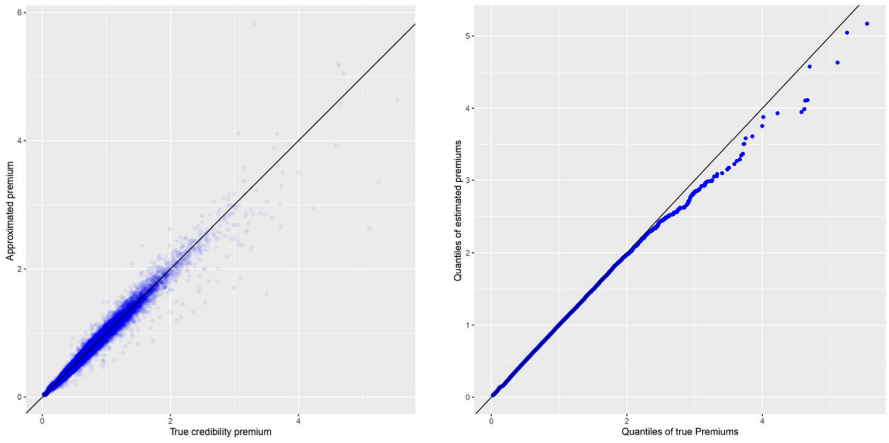

Fig. 1 Comparison of approximated premiums vs true premiums for the unstructured model. Left-hand side dispersion plot and Right-hand side QQ-plot

achieves a favorable fit with low error metrics and a high coefficient of determination of 99%, indicating almost perfect interpolation. Similarly, the graph on the left side of Fig. 1 illustrates that the fitted premiums closely align with the true premiums, displaying minimal fluctuation around the 45-degree line. Likewise, the QQplot on the right side of Fig. 1 indicates that the distribution of the fitted premiums closely resembles the pattern of the true premiums. Moreover, the results at the aggregate level in Table 2 show nearly insignificant differences with the total premiums. Lastly, it is noteworthy that the overall methodology of surrogate modeling yields favorable results in a cost-effective manner.

4 Application to IBNR Reserving

4.1 The Reserving Problem and the Sampling Approach

Consider an insurance company conducting an analysis of its total liabilities attributed to claims with accident times falling between $t = 0$ and $t = \tau$, where τ denotes the valuation time defined by an actuary. In the realm of general insurance, accidents

are often not immediately reported to the insurance company for various reasons, introducing a significant delay between the occurrence of a claimable accident and the moment the insurance company is notified. Consequently, at the given valuation time τ, the insurance company possesses information solely on the claims reported by τ and remains unaware of the unreported claims. In light of this, the insurance company seeks to estimate the total claim amount for these incurred but not reported (IBNR) claims to construct the reserve.

Let us describe the payment process as follows:

- Let $N(\tau)$ represent the total number of claims whose accident times precede the valuation time τ.
- Let Y_i, where $i = 1, \ldots, N(\tau)$, denote the sequence of the total claim amounts, or the so-called case estimates per claim if the former is not available. We assume that the indices i are arranged based on the accident time for simplicity.
- Let T_i, where $i = 1, \ldots, N(\tau)$, denote the sequence of accident times linked to the claims, and let R_i, where $i = 1, \ldots, N(\tau)$, denote the sequence of the associated reporting times.
- Let $U_i = R_i - T_i$, where $i = 1, \ldots, N(\tau)$, represent the sequence of reporting delay times associated with each claim.
- Let $\mathbf{x}_i$, where $i = 1, \ldots, N(\tau)$, be the sequence of relevant information/attributes linked to the accident, claim type, policyholder attributes, or characteristics of the claims themselves.
- Let $N^R(\tau)$ represent the number of claims paid by the valuation time τ out of the total $N(\tau)$, i.e., the number of claims reported by τ.

In a similar vein, the total liability of the insurance company linked to accidents occurring before the valuation time τ, denoted as $L(\tau)$, is expressed as:

$$L(\tau) = \sum_{i=1}^{N(\tau)} Y_i.$$

Likewise, the segment of liability that is known to the insurance company (i.e., the paid amount) by the valuation time τ, denoted as $L^R(\tau)$, is:

$$L^R(\tau) = \sum_{i=1}^{N^R(\tau)} Y_i.$$

Finally, the actuary aims on estimating the remaining liability, associated to incurred but not reported claims. This quantity, denoted as $L^{IBNR}(\tau)$, is simply given by the difference below. This quantity defines the reserve of interest.

$$L^{IBNR}(\tau) = L(\tau) - L^R(\tau).$$

The methods that are used to estimate such reserve can be broadly categorized into two frameworks: the aggregate approach and the individual approach. On one hand,

the aggregate approach involves estimating reserves using all aggregate claims data without utilizing granular information. These methods typically do not necessitate extensive statistical modeling for their formulation. The Chain–Ladder method serves as a prime example of this framework, see for e.g Wüthrich and Merz (2008). On the other hand, the individual approach entails modeling each policy using granular information, thereby accounting for heterogeneity. These methods are more sophisticated from a statistical standpoint and often utilize theories from point process predictive modeling, among others. See for e.g Pechon et al. (2018).

Our approach in this paper can be seen as a hybrid method that lies between these two frameworks. It is built upon a simple yet innovative idea of viewing the reserving problem as a *population sampling* problem as in Calcetero-Vanegas et al. (2023). Indeed, we can conceptualize all $N(\tau)$ claims as the population under consideration, while the current $N^R(\tau)$ reported claims by the valuation date serve as the selected sample for understanding this population. It is crucial to note that the sampling design and the actual sampling process are not determined or conducted by the investigator but are solely driven by the randomness associated with whether a claim is reported by the valuation date. Thus, the sample is given rather than being selected by the actuary. This distinction sets our setup apart from typical survey sampling situations.

In this analogy, the *inclusion probabilities* $\pi_i(\tau)$ can be interpreted as the likelihood of a claim Y_i belonging to the sample or, equivalently, being reported by the valuation time τ. These probabilities depend on the valuation time and are likely to vary across claims due to the different attributes $\mathbf{x}_i$ associated with each claim. Along these lines, the inclusion probabilities are given by:

$$\pi_i(\tau) = P(U_i \leq \tau - T_i | \mathbf{x}_i) \tag{4.1}$$

These probabilities are unknown to the actuary and must be estimated. To do so we need an estimation of the cumulative distribution function of the reporting delay times. By recognizing that the reporting delay time is a time-to-event random variable, we can use existing approaches from survival analysis to obtain the desired estimations without the need to develop new models. For instance, Cox regression models are by far the most popular techniques used in survival analysis (e.g., George et al. (2014)). We will further illustrate this in the numerical study of the next section.

As motivated by the population sampling literature, the *Horvitz–Thompson* (HT) estimator of the aggregate claims is described as follows

$$\hat{L}(\tau) = \sum_{i=1}^{N^R(\tau)} \frac{Y_i}{\pi_i(\tau)}, \tag{4.2}$$

Consequently, an unbiased estimator of the outstanding claims is the difference between the estimated total and the currently paid amount.

$$\hat{L}^{IBNR}(\tau) = \hat{L}(\tau) - L^R(\tau) = \sum_{i=1}^{N^R(\tau)} \frac{Y_i}{\pi_i(\tau)} - \sum_{i=1}^{N^R(\tau)} Y_i = \sum_{i=1}^{N^R(\tau)} \frac{1-\pi_i(\tau)}{\pi_i(\tau)} Y_i. \quad (4.3)$$

Confidence intervals for the reserve can be constructed relying on the sampling distribution of the HT estimator, as noted by Thompson (2012, p. 70). In brief, under minimal regularity conditions, the HT estimator follows approximately a normal distribution for large populations. We note that both the point estimator and its inference do not rely on a model for the frequency or the severity, and so the IPW provides a distribution-free approach for reserving.

Lastly, we highlight the resemblance between our estimator and the Chain–Ladder method. The inverse of the inclusion probabilities serve as development factors for the claims, akin to the Chain–Ladder. However, our approach differs by applying such development factors at the individual level, using claim-specific factors driven by claim attributes, instead of a single one for all claims as in the Chain–Ladder. Essentially, we can interpret the IPW method as an extension of the Chain–Ladder that incorporates the use of granular information.

4.2 Numerical Illustration

We demonstrate the application of the IPW estimator using a real dataset obtained from a Canadian automobile insurance company. The dataset includes information on Physical damage (PD) claims from January 2014 to December 2016. The time window from 2014 to 2015 will be utilized for training, while the last year of 2016 will be used for testing.

The only modeling required for the IPW is associated with the distribution for the reporting delay times, U_i. For this, we fit a Cox regression in which the associated hazard function $\lambda_{U|X}(u)$ depends on the attributes of the policyholder as follows:

$$\begin{aligned}\log(\lambda_{U|X}(u)) = \log(\lambda_0(u)) &+ \beta_1 \text{Car\_Age} + \beta_2 \text{Claim\_Count} \\ &+ \beta_3 \text{Horse\_Power} + \beta_4 \text{Car\_Weight} \\ &+ \beta_5 \text{Car\_Price} + \beta_6 \text{Gender} + \beta_7 \text{Driver\_Age} + S_8(\text{Accident\_day}) \\ &+ S_9(\text{Claim\_Amount})\end{aligned}$$

We estimate the log of the baseline hazard function, $\log(\lambda_0(u))$, using a B-Spline representation. Additionally, we introduce non-linear effects for the covariate "Accident_day" and "Claim_Amount" with the terms S_8(Accident-day) and S_9(Claim_Amount) respectively, also estimated through a B-Spline representation. Our approach involves a generalized additive model implementation via the piecewise exponential modeling method, accessible in `R` packages like `flexsurvreg`, `pammtools`, or `GJRM`.

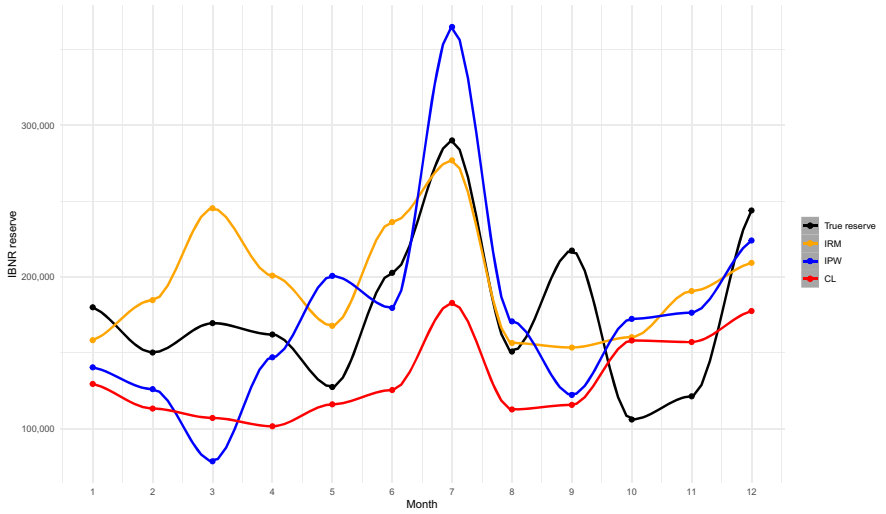

Fig. 2 Estimation for the IBNR reserve per month

Consequently, the desired inclusion probabilities are obtained using the relationship:

$$\pi_i(\tau) = Pr(U_i \leq \tau - T_i|\mathbf{x}_i) = F_{U|\mathbf{x}_i}(\tau - T_i) = 1 - \exp\left(-\int_0^{\tau - T_i} \lambda_{U|\mathbf{x}_i}(u)du\right). \tag{4.4}$$

For the sake of comparison, we also showcase two alternative modeling approaches. We include the aggregate reserving method given by the Chain–Ladder (CL), and a simplistic individual reserving method (IRM), inspired by the collective reserving model with individual data as in Delong et al. (2022). Briefly, such IRM model is constructed using a Zero-Inflated Negative Binomial regression for the frequency, a lognormal regression for the severity, and the same Cox regression illustrated above for the reporting delay time.

Along those lines, Fig. 2 presents estimations for outstanding claims compared to the true reserve value for all 12 months in the testing period, and for all models in consideration. Additionally, Table 3 provides error metrics to evaluate the disparities between the estimations across all dates.

Figure 2 depicts a close similarity between predictions generated by the IPW estimator in expression (4.3) and the actual reserve values across the majority of observed periods. It's important to note that the IPW estimator displays a more variable behavior when compared to the CL method and the IRM. However, there is no discernible pattern of over or underestimation of reserves. Despite this variability, the IPW estimator consistently outperforms the traditional CL approach and remains competitive with the IRM. This observation is further supported by the findings presented in Table 3, where the error metrics for the IPW over the 12-month period exceed those of the CL and closely approach those of the IRM. This ranking among methods is expected,

Table 3 Error metrics for the total of the reserves over the testing period

Method	ME	RMSE	MAE	MAPE (%)
IPW	17,426	57,395	49,829	23.9
CL	55,557	64,125	58,413	33.2
IRM	− 34,024	45,595	40,497	23.5

ME mean error, *RMSE* Root mean square error, *MAE* mean absolute error, *MAPE* mean absolute percentage error

given that the IPW leverages more granular information than the CL, though not to the extent of the IRM method. Hence, the predictive efficacy of the IPW, combined with the utilization of individual claims information, falls between that of aggregate and individual methods.

5 Conclusions

Sampling, a fundamental pillar of statistical methodologies, serves as a linchpin in unraveling patterns within extensive datasets, a practice of paramount importance in both general statistics and the specialized field of actuarial science. While widely utilized in censuses and election predictions, the full extent of its potential remains underexplored within actuarial science. This paper sheds light on recent developments, specifically focusing on its applications in insurance valuation and reserving. The judicious use of population sampling emerges as a powerful and efficient approach, addressing computational challenges associated with large and highly inhomogeneous non-life insurance portfolios.

On one side, the applications of population sampling extend to pricing assessments, particularly in the context of Bayesian credibility models. The intricate interplay between policyholder attributes and claim history necessitates a nuanced approach, and population sampling proves instrumental in refining the calculation of credibility premiums. By adopting a data-driven model, such as Bayesian regression, the challenge lies in obtaining credibility premiums via computationally expensive methods like simulation through Markov Chain Monte Carlo. Population sampling, coupled with surrogate modeling, provides a manageable solution, allowing for the incorporation of granular information into the premium calculation process. This not only enhances the efficiency of the computation but also ensures a statistically robust foundation for determining premiums. The integration of population sampling within Bayesian frameworks marks a significant stride in advancing the state-of-the-art in actuarial science, particularly in the domains of pricing and credibility assessments.

In the arena of reserving, particularly in the calculation of Incurred But Not Reported (IBNR) reserves, population sampling emerges as a transformative approach. Traditional methods, like the Chain–Ladder method, often overlook the heterogeneity of policyholders. Here, population sampling reframes the claim reserving problem, providing a statistically robust method for IBNR reserving. By applying inverse probability weighting techniques, this paper showcases how population sampling can seamlessly integrate granular information into the Chain–Ladder method, significantly improving the precision of IBNR reserve estimations. This represents a significant step

forward, as it enables frequency and severity distribution-free calculations, elevating the reliability of actuarial predictions.

The applications of population sampling in actuarial science extend beyond the explored realms of pricing and reserving. Indeed, Lin and Yang (2020a) and Lin and Yang (2020b) also show an application of the population framework in the context of valuation and risk management of variable annuities. Looking forward, sampling techniques hold promise in reinsurance strategy assessments, optimizing risk mitigation plans, and enhancing predictive modeling for underwriting and claims forecasting. As the landscape of actuarial science continues to evolve, the judicious use of population sampling emerges as a linchpin, guiding practitioners through the intricacies of insurance and risk management with greater precision and efficiency.

Acknowledgements This work was partly supported by Natural Sciences and Engineering Research Council of Canada [RGPIN-2023-04326]. The authors want to thank Prof. Andrei Badescu and an anonymous referee whose insightful feedback greatly enriched the final version of this work.

Data Availability The data supporting this study is proprietary data from an insurance company that we are unable to share.

Declarations

Conflict of interest The authors declare no Conflict of interest or competing interests in this paper, with no financial or personal affiliations that could compromise the objectivity or integrity of the presented work.

References

Ahn, J. Y., Jeong, H., & Lu, Y. (2021). On the ordering of credibility factors. *Insurance: Mathematics and Economics, 101*, 626–638.

Bühlmann, H., & Gisler, A. (2005). *A course in credibility theory and its applications*. Springer.

Calcetero-Vanegas, S., Badescu, A. L., & Lin, X. S. (2023). Claim reserving via inverse probability weighting: A micro-level chain-ladder method. Available at SSRN 4499355.

Calcetero-Vanegas, S., Badescu, A. L., & Lin, X. S. (2024). Effective experience rating for large insurance portfolios via surrogate modeling. *Insurance: Mathematics and Economics, 118*, 25–43.

Chan, I. W., Tseung, S. C., Badescu, A. L., & Lin, X. S. (2023). Data mining of telematics data: Unveiling the hidden patterns in driving behaviour. arXiv preprint arXiv:2304.10591

Delong, Ł, Lindholm, M., & Wüthrich, M. V. (2022). Collective reserving using individual claims data. *Scandinavian Actuarial Journal, 2022*(1), 1–28.

Deville, J.-C., & Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika, 91*(4), 893–912.

George, B., Seals, S., & Aban, I. (2014). Survival analysis and regression models. *Journal of Nuclear Cardiology, 21*(4), 686–694.

Gramacy, R. B. (2020). *Surrogates: Gaussian process modeling, design, and optimization for the applied sciences*. Boca Raton: Chapman and Hall/CRC.

Kaas, R., Goovaerts, M., Dhaene, J., & Denuit, M. (2008). *Modern actuarial risk theory: Using R*. Springer.
Lin, X. S., & Yang, S. (2020). Efficient dynamic hedging for large variable annuity portfolios with multiple underlying assets. *ASTIN Bulletin: The Journal of the IAA, 50*(3), 913–957.
Lin, X. S., & Yang, S. (2020). Fast and efficient nested simulation for large variable annuity portfolios: A surrogate modeling approach. *Insurance: Mathematics and Economics, 91*, 85–103.
Pechon, F., Trufin, J., & Denuit, M. (2018). Multivariate modelling of household claim frequencies in motor third-party liability insurance. *ASTIN Bulletin: The Journal of the IAA, 48*(3), 969–993.
Särndal, C.-E., Swensson, B., & Wretman, J. (2003). *Model assisted survey sampling*. Springer.
Thompson, S. K. (2012). *Sampling*. Oxford: Wiley.
Tillé, Y. (2011). Ten years of balanced sampling with the cube method: An appraisal. *Survey Methodology, 37*(2), 215–226.
Tzougas, G., & di Cerchiara, A. P. (2021). The multivariate mixed negative binomial regression model with an application to insurance a posteriori ratemaking. *Insurance: Mathematics and Economics, 101*, 602–625.
Wood, S. N. (2017). *Generalized additive models: An introduction with R*. Boca Raton: Chapman and Hall/CRC.
Wüthrich, M. V., & Merz, M. (2008). *Stochastic claims reserving methods in insurance*. Oxford: Wiley.
Xacur, O. A. Q., & Garrido, J. (2018). Bayesian credibility for glms. *Insurance: Mathematics and Economics, 83*, 180–189.
Zhang, J., Qiu, C., & Wu, X. (2018). Bayesian ratemaking with common effects modeled by mixture of polya tree processes. *Insurance: Mathematics and Economics, 82*, 87–94.