Malaysian Journal of Mathematical Sciences

Journal homepage: https://einspem.upm.edu.my/journal

Reserve Estimate Based on the Claims Data of Individual Customers

Ang, S. L.*[1], Pooi, A. H.[1,2], and Ng, K. H.[3]

[1]*School of Mathematical Sciences, Sunway University, Malaysia*
[2]*Centre for Mathematical Sciences, Universiti Tunku Abdul Rahman, Malaysia*
[3]*Institute of Mathematical Sciences, Faculty of Science, University of Malaya, Malaysia*

E-mail: siewlinga@sunway.edu.my
**Corresponding author*

Received: 4 August 2020
Accepted: 27 April 2021

Abstract

Suppose the claims data of individual customers consist of the delay times in reporting the claims, delay times in payments and the severities of claims. A mixture of two multivariate power-normal (MPN) distributions and a degenerate distribution is fitted to the vector of variables consisting of the sum insured, the claim and payment records until the present time and the outstanding claims liabilities (OCL). When the sum insured together with the claim and payment records of a customer until the present time are given, a conditional distribution of the OCL is derived from the fitted MPN mixture distribution. The distribution of the sum of the OCL over the customers in a company is obtained from the conditional distributions of the OCL pertaining to the individual customers. From the distribution of the sum of OCL, the provision of risk margin for adverse deviation can be calculated to provide a 75% level of capital adequacy at the company level.

Keywords: Outstanding claims liabilities; reserve estimation; multivariate power-normal distribution; individual claims data.

1 Introduction

Outstanding claims reserves in non-life insurance is established to provide the future liability for claims which are incurred but not reported or which have been reported but not settled. The actuary makes use of a variety of available methods to calculate the best reserve estimate needed for an insurer and apply a provision of risk margin for adverse deviation (PRAD).

In the literature, recently, instead of using the data aggregated in run-off triangles, some authors used the individual claims data to study loss reserving. The summary given by the run-off triangle has discarded a lot of information. Thus, it is sensible to examine instead the original individual data on claims. The individual data consist of the delay times in reporting the claims, delay times in payments and the severities of claims.

In the analysis based on individual data, a number of authors modelled the times of occurrence of claims as a Poisson process. Jewell [4] studied the formulation of the loss reserve using reporting delay in continuous time. Norberg [6, 7] using position dependent marked Poisson processes developed a mathematical framework to estimate loss reserving based on individual claims data. Haastrup and Arjas [3] used a model which was close to that of Norberg [6]. They assumed that the claims occurred in accordance with a Poisson process with intensity which depended on the calendar time and the characteristics of the insured via some piecewise constant structures. Larsen [5] illustrated the framework of marked Poisson processes with a small case study in estimating loss reserving using individual claims data. Antonio and Plat [1] revisited the works of Norberg [6, 7] with an extensive case study developed in likelihood based framework and implemented a micro-level stochastic model for individual loss reserving with paid or incurred losses. Zhao et al. [12] and Zhao and Zhou [11] proposed a semi-parametric structure for an individual claims development from survival analysis and copula methods.

Lately, because of the popularity in big data analytics and also the flexibility of the technique, some authors presented machine learning tools in estimating loss reserve. Wüthrich [9, 10] tried using regression trees to obtain the claims reserves on individual claims. Gabrielli and Wüthrich [2] used a stochastic simulation machine to study individual claims history data for individual claims reserving.

In this paper, we study the claims data of individual customers. In Section 2, we present the claim timeline. A description of the individual claims data is given in Section 3. In Section 4 we introduce the procedure of estimating the claims reserve. Section 5 provides some numerical results. Finally, Section 6 concludes the paper.

2 Claim Timeline

Assume that a premium is paid at time 0 for an insurance protection during the period $(0, T]$. Suppose an accident occurs at the time $t_a < T$ and is reported at time $t_r \geq t_a$. After time t_r, the company starts to collect information about the accident. Thus, after time t_r, the claim process will begin. The claim process will typically consist of

1. The times when the claims are reported.
2. The times when the payments are made.

3. The times when the claims amounts are adjusted.
4. The amounts involved in the above claims, payments and adjustments.

Let t_c be the time when the claims process is closed.

3 Individual Claims Data

Let N_s be the number of individual customers under consideration in the estimation of reserve. To record the individual claims data for the j-th customer $(1 \leq j \leq N_s)$, we may first take note of the number n_j of occurrence of the event given by claim (C), payment (P) or adjustment (D) until the present time. If $n_j = 0$, then we take note of the sum insured S_j and use A_j to denote the total amount paid to the insured from the time occurrence of the n_j-th event until t_c. If $n_j \geq 1$, then for the i-th event which has occurred, we take note of the time t_{ji} elapsed before the occurrence of the event after time t_r or after the time of occurrence of the $(i-1)$-th event, and the amount A_{ji} of claim/payment/adjustment involved. The following codes may be used to indicate the type of event which has occurred:

Table 1: Codes for type of event.

	e_{ji1}	e_{ji2}
Claim	0	1
Payment	1	0
Adjustment	0	0

Again, we use S_j to denote the sum insured, and use A_j to denote the total amount to be paid to the insured from the time of occurrence of the n_j-th event until t_c. Thus for the j-th customer, after knowing the value of A_j and using $\overline{A}_j$ and $\overline{A}_{ji}$ to denote A_j/S_j and A_{ji}/S_j respectively, the complete data recorded may be expressed as $\mathbf{G}_j = (S_j,\ \overline{A}_j)$ if $n_j = 0$, or

$$\mathbf{G}_j = (S_j,\ e_{j11},\ e_{j12},\ t_{j1},\ \overline{A}_{j1},\ e_{j21},\ e_{j22},\ t_{j2},\ \overline{A}_{j2},\ \ldots,\ e_{jn_j1},\ e_{jn_j2},\ t_{jn_j},\ \overline{A}_{jn_j},\ \overline{A}_j)$$

if $n_j \geq 1$.

In practice, the values of A_j are unknown at the time t_{jn_j} and we need to estimate the reserve given by

$$R = \sum_{j=1}^{N_s} A_j.$$

The value $\overline{A}_j$ may first be estimated based on the company's historical data which may be denoted as $\mathbf{G}_{j'}^{'} = (S_{j'}^{'},\ \overline{A}_{j'}^{'})$ if $n_j = 0$, or

$$\mathbf{G}_{j'}^{'} = (S_{j'}^{'},\ e_{j'11}^{'},\ e_{j'12}^{'},\ t_{j'1}^{'},\ \overline{A}_{j'1}^{'},\ e_{j'21}^{'},\ e_{j'22}^{'},\ t_{j'2}^{'},\ \overline{A}_{j'2}^{'},\ \ldots,\ e_{j'n_j1}^{'},\ e_{j'n_j2}^{'},\ t_{j'n_j}^{'},\ \overline{A}_{j'n_j}^{'},\ \overline{A}_{j'}^{'})$$

if $n_j \geq 1$.

The procedure to estimate the reserve R will be introduced in the next section.

4 Estimation of Reserve

To estimate the reserve R, we may first find a conditional distribution for A_j, $1 \leq j \leq N_s$. If $n_j = 0$, then we select n_s number of $(S_{j'}^{'},\ \overline{A}_{j'}^{'})$ where $S_{j'}^{'}$ is as close as possible to S_j. Let the n_s number of selected $\mathbf{G}_{j'}^{'} = (S_{j'}^{'},\ \overline{A}_{j'}^{'})$ be denoted as $\mathbf{G}_{j^+}^{+} = (S_{j^+}^{+},\ \overline{A}_{j^+}^{+}), 1 \leq j^+ \leq n_s$. When $n_j \geq 1$, we select n_s number of $\mathbf{G}_{j'}^{'}$ where $S_{j'}^{'}$ is as close as possible to S_j. Let the n_s number of selected $\mathbf{G}_{j'}^{'}$ be denoted as

$$\mathbf{G}_{j^+}^{+} = (S_{j^+}^{+},\ e_{j^+11}^{+},\ e_{j^+12}^{+},\ t_{j^+1}^{+},\ \overline{A}_{j^+1}^{+},\ e_{j^+21}^{+},\ e_{j^+22}^{+},\ t_{j^+2}^{+},\ \overline{A}_{j^+2}^{+},\ \dots\ ,\ e_{j^+n_j1}^{+},\ e_{j^+n_j2}^{+},\ t_{j^+n_j}^{+},\ \overline{A}_{j^+n_j}^{+}, \overline{A}_{j^+}^{+})$$

for $1 \leq j^+ \leq n_s$.

The values of $\mathbf{G}_{j^+}^{+}$ may be arranged in the form of a table of which the j^+-row contains the $4n_j + 2$ components in $\mathbf{G}_{j^+}^{+}$, for $1 \leq j^+ \leq n_s$.

We sort the values of $\overline{A}_{j^+}^{+}$ in an ascending order and partition the n_s selected values of $\mathbf{G}_{j^+}^{+}$ into three subsets $\mathbf{H}^{+(k)}$, $k = 1, 2, 3$, such that the n_1^+ values of $\overline{A}_{j^+}^{+}$ in $\mathbf{G}_{j^+}^{+}$ in the first subset $\mathbf{H}^{+(1)}$ are all smaller than the n_2^+ values of $\overline{A}_{j^+}^{+}$ in $\mathbf{G}_{j^+}^{+}$ in the second subset $\mathbf{H}^{+(2)}$ and $\mathbf{H}^{+(3)}$ contains the l-th $(1 \leq l \leq n_3^+)$ largest values of $\overline{A}_{j^+}^{+}$, and $n_1^+ + n_2^+ + n_3^+ = n_s$. The l-th $(1 \leq l \leq n_3^+)$ largest values of $\overline{A}_{j^+}^{+}$ may be chosen to be those which are at least 0.95. The values of n_3^+ would then be small.

Multivariate power-normal (MPN) distribution given in Pooi [8] will next be used to handle the data in the subset $\mathbf{H}^{+(k)}$, $k = 1, 2$. The reasons for choosing the MPN distribution are that:

1. The power-normal distribution is a very general distribution which can be used to fit data with wide ranges of skewness and kurtosis.
2. The power-normal distribution has a probability density function which can be expressed in an explicit form. This explicit form makes it convenient to compute the required conditional distributions from the MPN distribution.

For $k = 1, 2$, we fit an MPN distribution to the n_k^+ values of $\mathbf{G}_{j^+}^{+}$ in $\mathbf{H}^{+(k)}$, and find a conditional distribution for $\overline{A}_{j^+}^{+}$ when the value of $S_{j^+}^{+}$ is given by S_j in the case where $n_j = 0$ or when the value of

$$(S_{j^+}^{+},\ e_{j^+11}^{+},\ e_{j^+12}^{+},\ t_{j^+1}^{+},\ \overline{A}_{j^+1}^{+},\ e_{j^+21}^{+},\ e_{j^+22}^{+},\ t_{j^+2}^{+},\ \overline{A}_{j^+2}^{+},\ \dots\ ,\ e_{j^+n_j1}^{+},\ e_{j^+n_j2}^{+},\ t_{j^+n_j}^{+},\ \overline{A}_{j^+n_j}^{+})$$

is given by

$$(S_j,\ e_{j11},\ e_{j12},\ t_{j1},\ \overline{A}_{j1},\ e_{j21},\ e_{j22},\ t_{j2},\ \overline{A}_{j2},\ \dots\ ,\ e_{jn_j1},\ e_{jn_j2},\ t_{jn_j},\ \overline{A}_{jn_j})$$

in the case where $n_j \geq 1$.

Let the first four raw moments of the k-th $(1 \leq k \leq 2)$ conditional distribution be denoted as $m_k^{(1)}$, $m_k^{(2)}$, $m_k^{(3)}$ and $m_k^{(4)}$.

With small values of n_3^+ in the case when $k = 3$, it would be difficult to fit an MPN distribution to the n_3^+ values of $\mathbf{G}_{j^+}^+$ in $\mathbf{H}^{+(3)}$. However as large values of $\overline{A}_{j^+}^+$ (which are in the range [0.95,1]) can occur irrespective of the sum of insured, and the width of the range [0.95,1] is small, we may approximate the conditional distribution of $\overline{A}_{j^+}^+$ by a degenerate distribution which assigns a probability of one to the following average value of the l-th ($1 \le l \le n_3^+$) largest values of $\overline{A}_{j^+}^+$ in the third partitioned subset:

$$m_3^{(1)} = \frac{1}{n_3^+} \sum_{l=1}^{n_3^+} (l \text{ -th largest value of } \overline{A}_{j^+}^+).$$

We note that the second to fourth moments $m_3^{(2)}$, $m_3^{(3)}$ and $m_3^{(4)}$ of the above degenerate distribution are all equal to zero. Next let $P_k = n_k^+/n_s$, $k = 1, 2, 3$. The weighted q-th moment for A_j can be expressed as $M_j^{(q)} = \sum_{k=1}^3 P_k m_k^{(q)} S_j^{(q)}$, $1 \le q \le 4$.

The values of the above n_1^+ and n_2^+ are chosen as described below:

1. From the mixture distribution formed by MPN distribution fitted to the n_k^+ values of $\mathbf{G}_{j^+}^+$ in $\mathbf{H}^{+(k)}$, $k = 1, 2$, and the degenerate distribution fitted to the n_3^+ values of $\mathbf{G}_{j^+}^+$ in $\mathbf{H}^{+(3)}$, we find a conditional distribution for $\overline{A}_{j^+}^+$ when the value of

$$(S_{j^+}^+,\ e_{j^+11}^+,\ e_{j^+12}^+,\ t_{j^+1}^+,\ \overline{A}_{j^+1}^+,\ e_{j^+21}^+,\ e_{j^+22}^+,\ t_{j^+2}^+,\ \overline{A}_{j^+2}^+,\ \dots,\ e_{j^+n_j1}^+,\ e_{j^+n_j2}^+,\ t_{j^+n_j}^+,\ \overline{A}_{j^+n_j}^+)$$

   is given by the first $4n_j + 1$ components in the j^+-row of the table formed by $\mathbf{G}_{j^+}^+$.

2. Find the mean $\widehat{\overline{A}_{j^+}^+}$ of the conditional distribution yield from Step 1.

3. Choose the values of n_1^+ and n_2^+ such that the average $\sum_{j^+=1}^{n_s} \widehat{\overline{A}_{j^+}^+}/n_s$ of the n_s predicted values is closest to the average $\sum_{j^+=1}^{n_s} \overline{A}_{j^+}^+/n_s$ of the n_s observed values.

The q-th moment for R is then given by $M^{(q)} = \sum_{j=1}^{N_s} M_j^{(q)}$. From the raw moments $M^{(q)}$, $1 \le q \le 4$, we can compute the coefficients of skewness and kurtosis for R.

If the computed coefficient of skewness is close to 0 while that of kurtosis is close to 3, then we may fit a normal distribution to R. The difference between the 0.75-quantile and the mean of the distribution for R becomes approximately the value of the PRAD.

5 Numerical Results

The data for $\mathbf{G}_j$ ($1 \le j \le N_s$) and $\mathbf{G}_{j'}^{'}$ ($1 \le j' \le n_s$), used in this study are obtained by coding the data from an insurance company in Malaysia over the period August 2002-August 2007. With these data, we investigate the goodness of fit of the fitted MPN distributions.

From the conditional distributions for $\overline{A}_{j+}^{+}$ when the value of

$$(S_{j+}^{+},\ e_{j+11}^{+},\ e_{j+12}^{+},\ t_{j+1}^{+},\ \overline{A}_{j+1}^{+},\ e_{j+21}^{+},\ e_{j+22}^{+},\ t_{j+2}^{+},\ \overline{A}_{j+2}^{+},\ \ldots,\ e_{j+n_j1}^{+},\ e_{j+n_j2}^{+},\ t_{j+n_j}^{+},\ \overline{A}_{j+n_j}^{+})$$

is given by

$$(S_j,\ e_{j11},\ e_{j12},\ t_{j1},\ \overline{A}_{j1},\ e_{j21},\ e_{j22},\ t_{j2},\ \overline{A}_{j2},\ \ldots,\ e_{jn_j1},\ e_{jn_j2},\ t_{jn_j},\ \overline{A}_{jn_j}),$$

we find the weighted mean $M_j^{(1)}$ and the 0.95-quantile of A_j.

A nominally 95% one-sided prediction interval for A_j is then given by $(0, q_{0.95})$. A predicted value of A_j is given by the weighted mean $M_j^{(1)}$ of the conditional distributions for A_j.

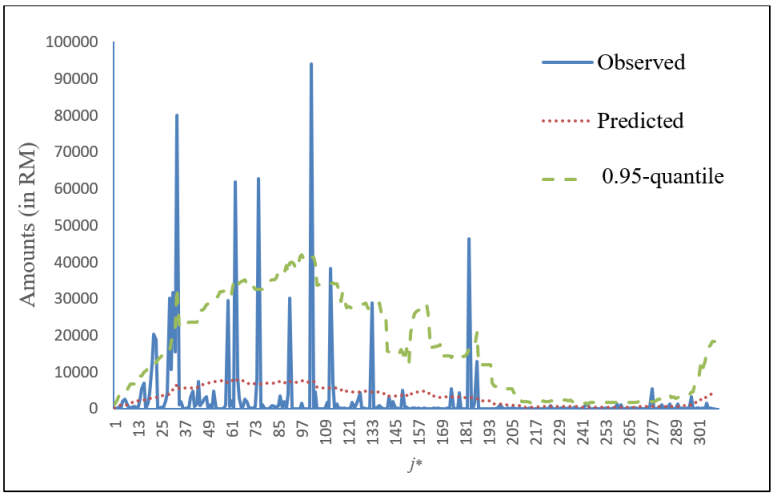

Figure 1: The j^*-th observed and predicted values of A_j together with the 0.95-quantile of the conditional distribution for A_j when $n_j = 0$ ($N_s = 500$, $n_s = 300$, estimate coverage probability = 0.9579).

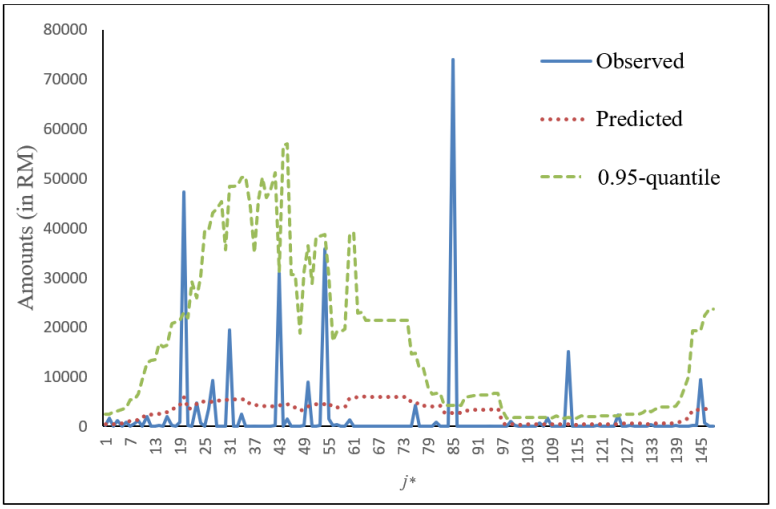

Figure 2: The j^*-th observed and predicted values of A_j together with the 0.95-quantile of the conditional distribution for A_j when $n_j = 1$ ($N_s = 500$, $n_s = 300$, estimate coverage probability = 0.9730).

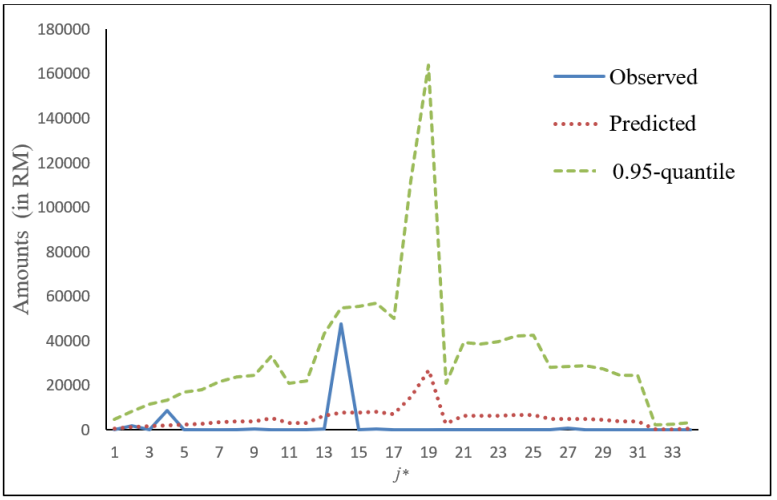

Figure 3: The j^*-th observed and predicted values of A_j together with the 0.95-quantile of the conditional distribution for A_j when $n_j = 2$ ($N_s = 500$, $n_s = 300$, estimate coverage probability = 1.0000).

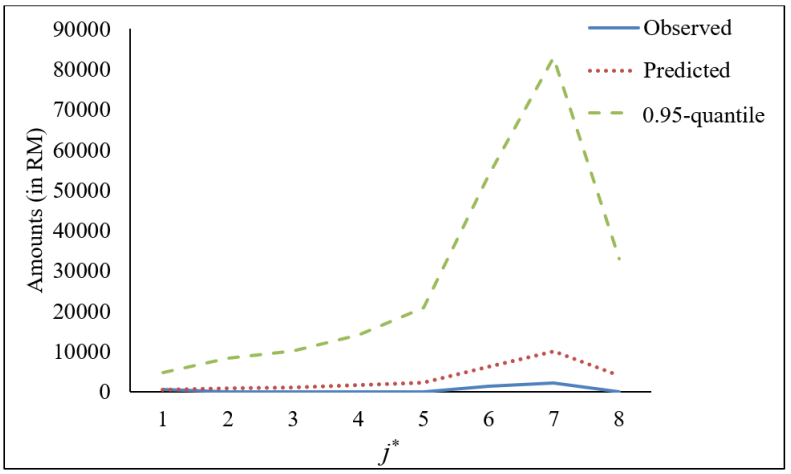

Figure 4: The j^*-th observed and predicted values of A_j together with the 0.95-quantile of the conditional distribution for A_j when $n_j = 3$ ($N_s = 500$, $n_s = 300$, estimate coverage probability = 1.0000).

Figures 1-4 exhibit the observed and predicted values of A_j together with the 0.95-quantile of the conditional distribution for A_j. The figures show that when the number of j^* for a given n_j is large, the observed coverage probabilities of the one-sided prediction intervals are fairly close to the target value 0.95. Thus, we expect that the fitted distribution of R would give a good estimate of the PRAD.

6 Conclusion

The estimation procedure in this research is carried out by partitioning the data into three parts and use a mixture of two MPN distributions and a degenerate distribution to fit the data. By controlling the sizes of the partitioned subsets of data, it is possible to reduce the influence of extremely large claims to the mean and variance of the estimate for the outstanding claims liability.

The comparison of the observed and predicted values of the outstanding claims together with the estimated coverage probability of the prediction interval for the outstanding claims are found to be useful for determining the performance of the reserves estimation.

However, the idea of partitioning the data into three parts of suitable sizes and using a mixture distribution to fit the data may be explored further. For example, we may consider partitioning the data into four parts instead. To find out the suitable number of partitioned parts, we may try to use the criterion given by the coverage probability and average length of the prediction intervals for the outstanding claims.

Another area of future research is the inclusion of the characteristics of the individual customers to achieve possible further improvement in estimating the reserves.

References

[1] K. Antonio & R. Plat (2014). Micro-level stochastic loss reserving for general insurance. *Scandinavian Actuarial Journal*, *2014*(7), 649–669. https://doi.org/10.1080/03461238.2012.755938.

[2] A. Gabrielli & M. V. Wüthrich (2018). An individual claims history simulation machine. *Risks*, *6*(2), 29. https://doi.org/10.3390/risks6020029.

[3] S. Haastrup & E. Arjas (1996). Claims reserving in continuous time: A nonparametric Bayesian approach. *ASTIN Bulletin: The Journal of the IAA*, *26*(2), 139–164. https://doi.org/10.2143/AST.26.2.563216.

[4] W. S. Jewell (1989). Predicting IBNYR events and delays: I. Continuous time. *ASTIN Bulletin: The Journal of the IAA*, *19*(1), 25–55. https://doi.org/10.2143/AST.19.1.2014914.

[5] C. R. Larsen (2007). An individual claims reserving model. *ASTIN Bulletin: The Journal of the IAA*, *37*(1), 113–132. https://doi.org/10.1017/S0515036100014768.

[6] R. Norberg (1993). Prediction of outstanding liabilities in non-life insurance. *ASTIN Bulletin: The Journal of the IAA*, *23*(1), 95–115. https://doi.org/10.2143/AST.23.1.2005103.

[7] R. Norberg (1999). Prediction of outstanding liabilities II: Model variations and extensions. *ASTIN Bulletin: The Journal of the IAA*, *29*(1), 5–25. https://doi.org/10.2143/AST.29.1.504603.

[8] A. H. Pooi (2012). A model for time series analysis. *Applied Mathematical Sciences*, *6*(115), 5735–5748.

[9] M. V. Wüthrich (2018). Machine learning in individual claims reserving. *Scandinavian Actuarial Journal*, *2018*(6), 465–480. https://doi.org/10.1080/03461238.2018.1428681.

[10] M. V. Wüthrich (2018). Neural networks applied to chain–ladder reserving. *European Actuarial Journal*, *8*(2), 407–436. https://doi.org/10.1007/s13385-018-0184-4.

[11] X. B. Zhao & X. Zhou (2010). Applying copula models to individual claim loss reserving methods. *Insurance*: *Mathematics and Economics*, *46*(2), 290–299. https://doi.org/10.1016/j.insmatheco.2009.11.001.

[12] X. B. Zhao, X. Zhou & J. L. Wang (2009). Semiparametric model for prediction of individual claim loss reserving. *Insurance*: *Mathematics and Economics*, *45*(1), 1–8. https://doi.org/10.1016/j.insmatheco.2009.02.009.