Prediction region for average claim occurrence rate and average claim size in motor insurance

Wei Yeing Pan[1*], *Huei Ching* Soo[2], and *Ah Hin* Pooi[3]

[1]Department of Mathematical and Actuarial Sciences, Lee Kong Chian Faculty of Engineering and Science, Universiti Tunku Abdul Rahman, Malaysia
[2]School of Mathematical and Computer Sciences, Heriot-Watt University Malaysia
[3]Centre for Mathematical Sciences, Universiti Tunku Abdul Rahman, Malaysia

Abstract. The third-party motor insurance data from Sweden for 1977 described by Andrews and Herzberg in 1985 contain average claim occurrence rate (P_c), average claim size (C_a) for category of vehicles specified by the kilometres travelled per year (K), geographical zone (Z), no claims bonus (B) and make of car (M). The categorical variables Z and M may first be represented respectively by the vectors $(Z_1, Z_2, \dots, Z_6)$ and $(M_1, M_2, \dots, M_8)$ of binary variables. The variable (P_c, C_a) is next modelled to be dependent on $\boldsymbol{X}^* = (K, Z_1, Z_2, \dots, Z_6, B, M_1, M_2, \dots, M_8)$ via a conditional distribution which is derived from an 18-dimensional power-normal distribution. From the conditional distribution, a prediction region for (P_c, C_a) can be obtained to provide useful information on the possible ranges of average claim occurrence rate and average claim size for a given category of vehicles.

1 Introduction

Determination of fair and accurate tariffs is an important issue in the car insurance industry. Decision-trees and combination of regression techniques have been used to analyse the car insurance data [1-4]. Jørgensen and de Souza [5] assumed that the claim number is Poisson distributed while the cost for individual claims is gamma distributed and modelled the expected cost μ of claims per insured unit as a function of the explanatory variables. Double generalized linear models of which the underlying exponential families of distributions are restricted to the Tweedie family were investigated by Jørgensen and Smyth [6] and Andersen and Bonat [7]. These authors modelled the dispersion and mean of the costs simultaneously and produced the tariffs based on the model fitted to the Swedish third-party automobile portfolio of 1977. In the double generalized linear models, the logarithm of the mean μ is very often set to be a linear function of the explanatory variables. Yang et al. [8] used a gradient tree-boosting algorithm to replace the logarithmic mean by a highly complex functional form. An extension to the lasso which imposes the grouped elastic net penalty was used by Qian et al. [9] to select the explanatory variables to be included in the final Tweedie's compound Poisson model. Another alternative to determine the motor insurance rate was proposed by Pan et al. [10]. They used the multivariate power-normal distribution to fit the

[*] Corresponding author: panwy@utar.edu.my

Swedish third-party motor insurance data for 1977 and proposed using a conditional distribution for the payment per insured to determine the motor insurance rate.

The approach based on the Tweedie family of generalized linear models implicitly assumes that the claim counts and amounts are independent. However, in practice, claim frequency and severity may be dependent. Several authors attempted to deal with the case when dependency between the claim frequency and severity exists [11-16].

In this paper, without assuming that claim counts and amounts are independent, we proceed to derive a two-dimensional conditional distribution for the variables average claim occurrence rate and average claim size from the multivariate power-normal distribution for the vector of variables given by the average claim occurrence rate, average claim size, and the those formed from the various characteristics of the insured vehicles. From the conditional distribution, we construct a prediction region for the variables average claim occurrence rate and average claim size. This region provides useful information on the possible ranges of average claim occurrence rate and average claim size for a given category of vehicles. From the conditional distribution, we next find the distribution for the variable payment per insured which equals the product of the average claim occurrence rate and the average claim size. From the distribution for the payment per insured, we construct a prediction interval for the payment per insured. The first 100 resulting prediction intervals are found to have a shorter average length, but comparable estimated coverage probability, when compared with those given in Pan et al. [10]. Thus, the conditional distribution derived in this paper would provide a good alternative method for determining the motor insurance rate.

This paper contains 5 sections. The second section outlines a numerical method for finding a two-dimensional conditional distribution from a multivariate power-normal distribution (MPN). The third section describes the construction of two-dimensional prediction region for (P_c, C_a) = (Average claim occurrence rate, Average claim size) as well as in-sample and out-of-sample prediction intervals for the claim size $C = P_c C_a$ per insured. In Section 4, we fit an MPN distribution to the Swedish third-party motor insurance data for 1977 and derive a two-dimensional conditional distribution for (P_c, C_a). A prediction region for (P_c, C_a) is next derived from the two-dimensional conditional distribution. From the Swedish data for 1977 we also find prediction intervals for the claim size per insured using the two-dimensional conditional distribution. Finally, Section 5 concludes the paper.

2 Evaluation of two-dimensional conditional distribution

Yeo and Johnson [17] introduced the following power transformation of the standard normal random variable z:

$$\tilde{\varepsilon} = \begin{cases} \left[(z+1)^{\lambda^+} - 1\right]/\lambda^+ & ,(z \geq 0, \lambda^+ \neq 0) \\ log(z+1) & ,(z \geq 0, \lambda^+ = 0) \\ -\left[(-z+1)^{\lambda^-} - 1\right]/\lambda^- & ,(z < 0, \lambda^- \neq 0) \\ -log(-z+1) & ,(z < 0, \lambda^- = 0) \end{cases} \tag{1}$$

In Equation (1), the variable $\tilde{\varepsilon}$ is said to have a power-normal distribution with parameters λ^+ and λ^-.

Suppose $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_k)^T$ is a vector of random variables and the i-th variable ε_i is given by

$$\varepsilon_i = \sigma_i[\widetilde{\varepsilon_i} - E(\widetilde{\varepsilon_i})]/[Var(\widetilde{\varepsilon_i})]^{1/2} \tag{2}$$

where $\sigma_i > 0$ is a constant and $\widetilde{\varepsilon_i}$ has a power-normal distribution with parameters λ_i^+ and λ_i^-. Furthermore let $\boldsymbol{\mu}$ be a $k \times 1$ vector of constants and $\boldsymbol{H}$ a $k \times k$ orthogonal matrix. Then

$$\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{H\varepsilon} \tag{3}$$

is said to have a $k-$dimensional power-normal distribution with parameters $\boldsymbol{\mu}, \boldsymbol{H}, \lambda_i^+, \lambda_i^-, \sigma_i, 1 \le i \le k$.

When the values of the initial $k-2$ components of $\boldsymbol{y}$ are given, it is possible to find numerically the conditional distribution of the last two components. The required procedure is as follows:

(1) From the observed values of y_{k-1} and y_k, find the minimum value m_i and maximum value M_i of $y_i, i = k-1, k$. For $i = k-1, k$, a feasible range for the values of y_i may then be taken to be $R_i = [(1-c)m_i, (1+c)M_i\ ]$ where c is a fractional value. The range R_i may be partitioned into N intervals of length $h_i = [(1+c)M_i - (1-c)m_i]/N$.

(2) Find a $(k-1)-$ dimensional power-normal distribution for $(y_1, y_2, \dots, y_{k-1})^T$.

(3) When $y_1, y_2, \dots, y_{k-2}$ are given, find a conditional distribution for y_{k-1}. Let the conditional probability density function (pdf) evaluated at $y_{k-1} = y_{k-1}^{(i_1)} = (1-c)m_{k-1} + (i_1 - 1)h_{k-1}$ be denoted by $g_{i_1}^{(1)}, i_1 = 1, 2, \dots, N$.

(4) Find a $k-$ dimensional power-normal distribution for $\boldsymbol{y} = (y_1, y_2, \dots, y_k)^T$.

(5) When the $k-1$ initial values of $\boldsymbol{y}$ are $y_1, y_2, \dots, y_{k-2}, y_{k-1}^{(i_1)}$, find a conditional distribution for y_k. Let the conditional pdf evaluated at $y_k = y_k^{(i_2)} = (1-c)m_k + (i_2 - 1)h_k$ be denoted by $g_{i_2}^{(2)}$.

(6) When $y_1, y_2, \dots, y_{k-2}$ are given, the conditional joint pdf of (y_{k-1}, y_k) evaluated at $\left(y_{k-1}^{(i_1)}, y_k^{(i_2)}\right)$ is then given by $g_{i_1,i_2} = g_{i_1}^{(1)} g_{i_2}^{(2)}$.

3 Two-dimensional prediction region

When $y_1, y_2, \dots, y_{k-2}$ are given, the conditional joint pdf g_{i_1,i_2} of (y_{k-1}, y_k) evaluated at $\left(y_{k-1}^{(i_1)}, y_k^{(i_2)}\right)$ can be equated approximately to a two-dimensional power-normal distribution by using the following procedure adapted from Pooi [18]:

(1) Find

$$\hat{E}\left(y_{k-1}^{j_1}\, y_k^{j_2}\right) = \textstyle\sum_{i_1=1}^N \sum_{i_2=1}^N \left[y_{k-1}^{(i_1)}\right]^{j_1} \left[y_k^{(i_2)}\right]^{j_2} g_{i_1,i_2} h_{k-1} h_k, \tag{4}$$

for $j_1 \ge 0, j_2 \ge 0, j_1 + j_2 \le 2$ and obtain the estimated variance-covariance matrix $\widehat{\boldsymbol{M}}$ of y_{k-1} and y_k.

(2) Find the matrix $\boldsymbol{V}$ formed by the eigenvectors of $\widehat{\boldsymbol{M}}$.

(3) Let

$$\binom{S_{k-1}}{S_k} = \boldsymbol{V}^T \binom{y_{k-1} - \hat{E}(y_{k-1})}{y_k - \hat{E}(y_k)} \tag{5}$$

and $S_i^{(i_1,i_2)}$ be the value of S_i evaluated at $\left(y_{k-1}^{(i_1)}, y_k^{(i_2)}\right)$, $i = k-1, k$.

(4) Find

$$\hat{E}\left(S_i^j\right) = \textstyle\sum_{i_1=1}^N \sum_{i_2=1}^N \left[S_i^{(i_1,i_2)}\right]^j g_{i_1,i_2} h_{k-1} h_k, i = k-1, k; 1 \le j \le 4. \tag{6}$$

(5) Express S_i in term of a random variable $\tilde{\varepsilon}_i$:

$$S_i = \sigma_i \frac{\tilde{\varepsilon}_i - E(\tilde{\varepsilon}_i)}{[Var(\tilde{\varepsilon}_i)]^{1/2}}, i = k-1, k \tag{7}$$

where $\sigma_i = \left[\hat{E}(S_i^2)\right]^{1/2}$,

and $\tilde{\varepsilon}_i$ has a power-normal distribution with parameters λ_i^+ and λ_i^-. The parameters λ_i^+ and λ_i^- are chosen such that the j-th moment of S_i equals $\hat{E}\left(S_i^j\right), 1 \leq j \leq 4$.

A nominally $100(1-\alpha)\%$ prediction region for (y_{k-1}, y_k) may then be expressed as

$$\left[(y_{k-1}, y_k) : z_1^2 + z_2^2 \leq \chi_{2,\alpha}^2, y_{k-1} > 0 \text{ and } y_k > 0\right] \tag{8}$$

where $\chi_{2,\alpha}^2$ is the $(1-\alpha)$-quantile of a chi square distribution with two degrees of freedom and z_i is given by z in Equation (1) with $\tilde{\varepsilon}, \lambda^+$ and λ^- changed respectively to $\tilde{\varepsilon}_i, \lambda_i^+$ and λ_i^-.

The conditional joint pdf g_{i_1,i_2} of (y_{k-1}, y_k) also provides an alternative method for finding a prediction interval for the variable $y^* = y_{k-1} \times y_k$ given by the product of y_{k-1} and y_k. The alternative method is as follows:

(a) Find

$$\hat{E}\left(y^{*j}\right) = \sum_{i_1=1}^N \sum_{i_2=1}^N \left[y_{k-1}^{(i_1)} y_k^{(i_2)}\right]^j g_{i_1,i_2} h_{k-1} h_k,\ 1 \leq j \leq 4. \tag{9}$$

(b) From the first four moments of y^* in (a), express y^* as

$$y^* = \sigma^* \frac{\tilde{\varepsilon}^* - E(\tilde{\varepsilon}^*)}{[Var(\tilde{\varepsilon}^*)]^{1/2}} \tag{10}$$

where $\sigma^* = [Var(y^*)]^{1/2}$ and $\tilde{\varepsilon}^*$ has a power-normal distribution of which the parameters λ^{*+} and λ^{*-} are chosen such that the first four moments of y^* are equal to those in (a).

A nominally $100(1-\alpha)\%$ prediction interval for y^* is then given by

$$\left(y^* : -z_{\alpha/2} \leq z^* \leq z_{\alpha/2}, y^* > 0\right) \tag{11}$$

where $z_{\alpha/2}$is the $(1-\alpha/2)$-quantile of the standard normal distribution and z^* is given by z in Equation (1) with $\tilde{\varepsilon}, \lambda^+$ and λ^- changed respectively to $\tilde{\varepsilon}^*, \lambda^{*+}$ and λ^{*-}.

When the given $y_1, y_2, \ldots, y_{k-2}$ are respectively equal to the initial $k-2$ components of one of the observed data point $(y_1, y_2, \ldots, y_{k-2}, y_{k-1}, y_k)$, then the prediction region (or interval) in Equation (8) (or Equation (11)) may be referred to as an in-sample prediction region (or interval).

But when the given $y_1, y_2, \ldots, y_{k-2}$ do not form the initial $k-2$ components in any of the observed data point $(y_1, y_2, \ldots, y_{k-2}, y_{k-1}, y_k)$, then the prediction region (or interval) in Equation (8) (or Equation (11)) may be referred to as an out-of-sample prediction region (or interval).

4 Swedish third-party motor insurance

The Swedish third-party motor insurance data for 1977 ([1] and [19]) contain the number of insured in policy years (N_y), the number of claims (N_c), the total value of payments (C_t) when the kilometres travelled per year (K), geographical zone (Z), no claims bonus (B) and make of the car (M) are given. After assigning the binary codes $(Z_1, Z_2, \dots, Z_6)$ and $(M_1, M_2, \dots, M_8)$ respectively to the qualitative variables Z and M, we can obtain a total of 2,183 observed values of the vector $\widetilde{\boldsymbol{y}} = \left(K, Z_1, Z_2, \dots, Z_6, B, M_1, M_2, \dots, M_8, N_y, N_c, C_t\right)$.

Under the category specified by (K, Z, B, M), the estimated probability P_c that an insured will make a claim during the policy year may be approximated by $P_c = N_c/N_y$ while the average claim size may be estimated by $C_a = C_t/N_c$.

As the probability P_c is defined over an interval of one year, we may also refer to this probability as the average claim rate.

By following the method in Section 3, the observed values of the vector $\boldsymbol{y} = (K, Z_1, Z_2, \dots, Z_6, B, M_1, M_2, \dots, M_8, P_c, C_a)$ may be used to construct a prediction region for the average claim rate and average claim severity.

Based on the 2183 observed values of $\boldsymbol{y}$, an 18-dimensional power-normal distribution for $\boldsymbol{y}$ is formed. When the value of $y^{(j)} = (K, Z_1, Z_2, \dots, Z_6, B, M_1, M_2, \dots, M_8)$ is given by the j-th row of the observed values of $\boldsymbol{y}$, a nominally 95% in-sample prediction region is found by using the method in Section 3 for (P_c, C_a). We may refer to the prediction region which corresponds on the j-th row of the observed value of $\boldsymbol{y}$ as the j-th prediction region. The probability that the prediction region will cover the observed value of (P_c, C_a) is called the coverage probability of the prediction region. Among the first 100 prediction regions, it is found that 90 of them cover the observed (P_c, C_a). Thus, an estimated coverage probability of the prediction region is 0.9 which is not too far from the targeted value of 0.95.

The following figures show two examples of prediction region for (P_c, C_a).

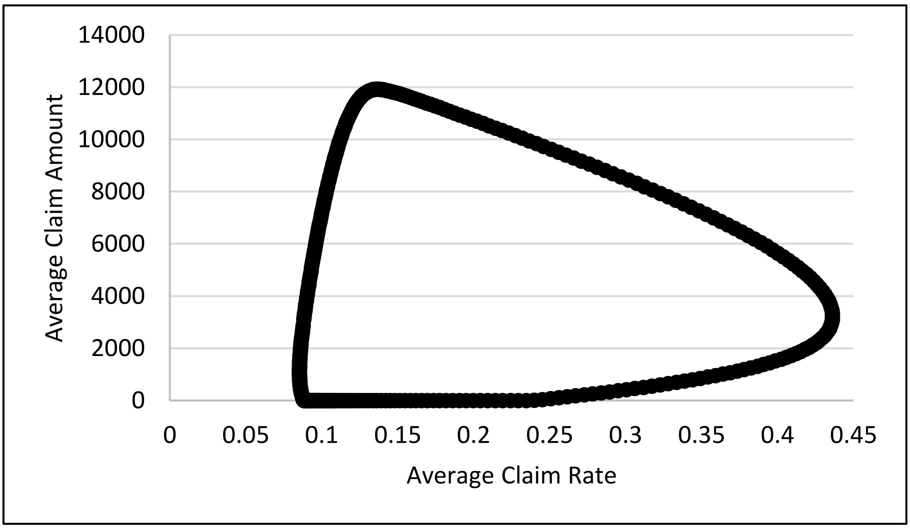

Fig. 1. The first prediction region for (P_c, C_a) $[K = 1, Z = 1, B = 1, M = 1]$.

The prediction region in Fig. 1 shows that the average claim rate is likely to be within 0.09 and 0.43. Furthermore, when the average claim intensity is large, the claim is likely to be around 3,000 with a relatively small range of variation. But when the average claim rate becomes smaller, the range of variation of the claim becomes larger.

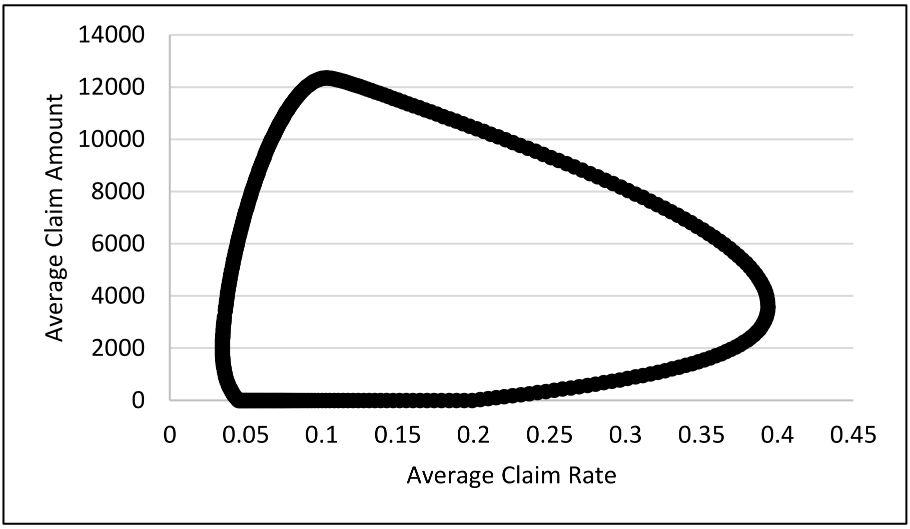

Fig. 2. The 28th prediction region for (P_c, C_a) $[K = 1, Z = 1, B = 4, M = 1]$.

The prediction region in Fig. 2 shows that when the no claims bonus is changed from 1 to 4 while the values of K, Z and M remain unchanged, the average claim rate tends to be smaller while the average claim amount appears to be about the same as before.

Thus, when the values of K, Z, B and M are given the prediction region gives an idea of the possible ranges of average claim rate and average claim amount.

By using the method given in Section 3, a nominally 95% prediction interval for the claim amount per insured can be found when the values of K, Z, B and M are given. The lower and upper limits of the first 100 in-sample prediction intervals are shown in Fig. 3. The estimated coverage probability and average length found by using these 100 in-sample prediction intervals are 0.98 and 1,318.17 respectively. Fig. 4 shows the corresponding 100 in-sample prediction intervals for claim amount per insured given in Pan et al. [10]. The estimated coverage probability and average length based on the prediction intervals in Fig. 4 are 0.99 and 1,476.08 respectively.

Thus, compared to the 100 prediction intervals presented in this paper, those given in Pan et al. [10] have comparable estimated coverage probability, but longer average length.

To investigate the performance of out-of-sample prediction regions and intervals, we initially form a table of 2,183 rows with the values of its j-th row denoted by $\boldsymbol{y}^{(j)}$. Next we choose a particular value j^* of j and consider that the initial 16 components $y_1^*, y_2^*, \dots, y_{16}^*$ of $\boldsymbol{y}^{(j^*)}$ are the given values and we wish to predict (y_{17}^*, y_{18}^*).

We choose $N_c = 200$ rows from the remaining 2,180 rows in the table such that the k-th chosen row $\tilde{\boldsymbol{y}}^{(k)}$ is such that the distance $\left[\sum_{m=1}^{16}\left(\tilde{y}_m^{(k)} - y_m^*\right)^2\right]^{1/2}$ is the k-th smallest value among the distances computed using the remaining 2,180 rows.

The chosen $N_c = 200$ rows may be used as the data for getting an 18-dimensional power-normal distribution. The methods in Sections 2 and 3 may next be used to construct the out-of-sample prediction region for (y_{17}^*, y_{18}^*) (or interval for $y_{17}^* \times y_{18}^*$).

By choosing $j^* = 1, 2, \dots, 100$, we can obtain 100 out-of-sample prediction regions and another 100 out-of-sample prediction intervals. The estimated coverage probabilities of the prediction regions and intervals are found to be 0.85 and 0.96 respectively, while the average length of the prediction intervals is 1,083.61. Fig. 5 shows that the 100 out-of-sample prediction intervals tend to have shorter lengths than those given in Figs. 3 and 4. The shorter prediction intervals may be attributed to the choice of data points of which the initial 16 components are close to those of the given $y_1^*, y_2^*, \ldots, y_{16}^*$.

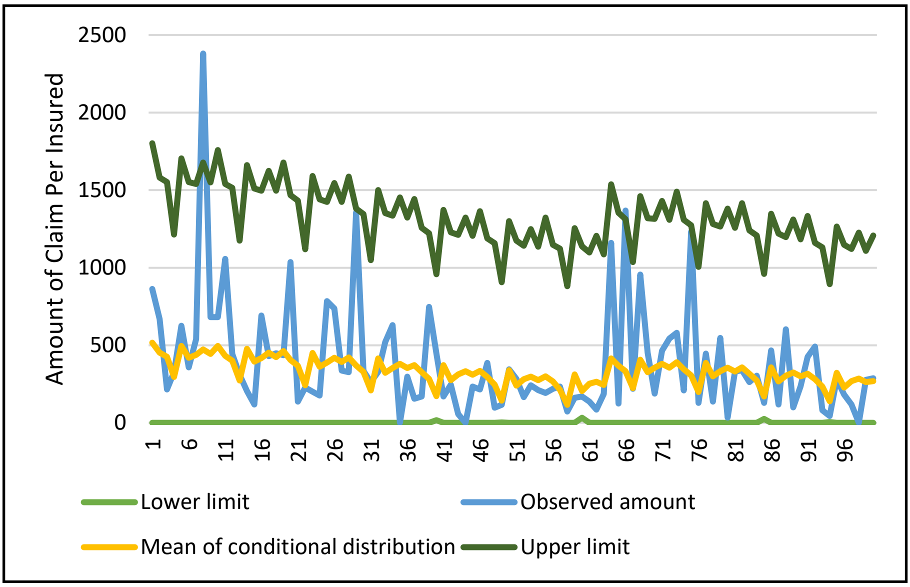

Fig. 3. Lower and upper limits of the first 100 prediction intervals for the amount of claim per insured ($\alpha = 0.05$, in-sample prediction).

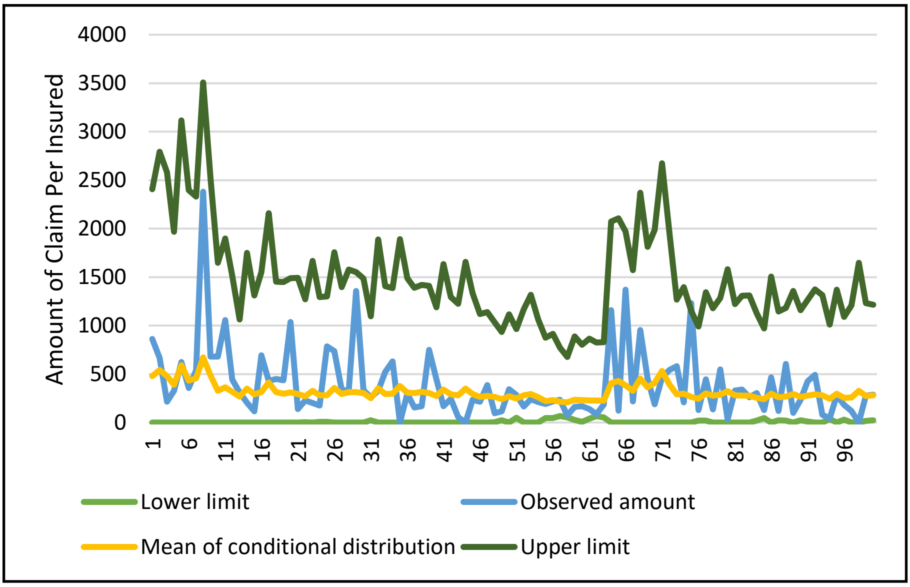

Fig. 4. Lower and upper limits of the first 100 prediction intervals for the amount of claim per insured given in Pan et al. [10] ($\alpha = 0.05$).

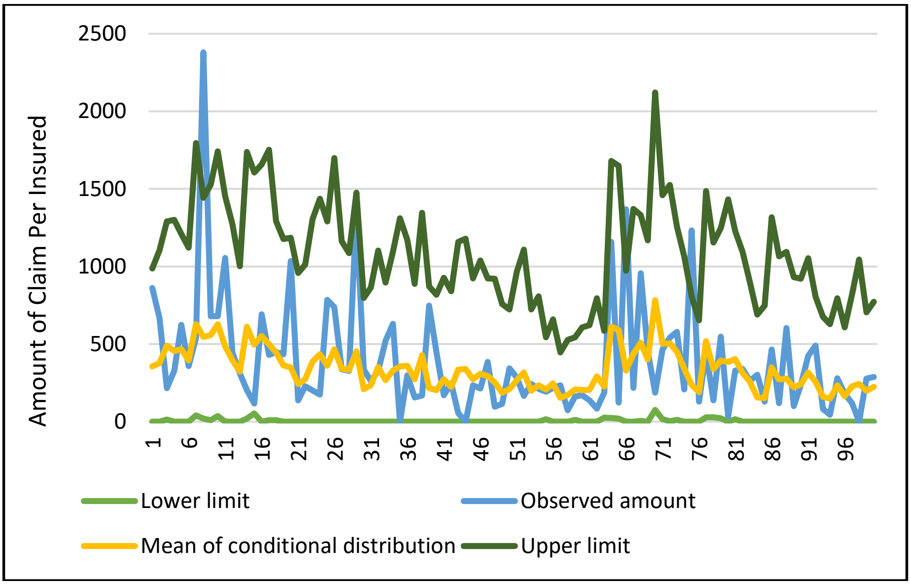

Fig. 5. Lower and upper limits of the first 100 prediction intervals for the amount of claim per insured ($\alpha = 0.05$, out-of-sample prediction).

5 Concluding remarks

Given the profile of a customer, we are interested in the probability that the customer will make a claim (P_c), and the size of the claim (C_a). As the prediction region for (P_c, C_a) provides a set of likely values of (P_c, C_a), it may be used to perform the customer risk assessment. The method in his paper may also be used when the profile of a customer includes the data collected via smart sensors on the driving behaviour of the customer.

References

1. D.F. Andrews, A.M. Herzberg, Data: a collection of problems from many fields for the student and research worker (Springer Verlag, 1985)
2. A. Christmann, Empirical risk minimization for car insurance data (2006)
3. W.Y. Loh, Handbook of data visualization 447 – 469 (2008)
4. M-G, Marcos, A. Christmann, *Insurance: an R-program to model insurance data*, Technical Report 49, Department of Statistics, University of Dortmund, Germany (2004)
5. M.C.P. De Souza, B. Jørgensen, Scand Actuar. J. 69 – 93 (1994)
6. B. Jørgensen and G.K. Smyth, ASTIN Bull. **32**(1), 143 – 157 (2002)
7. D.A. Andersen, W.H. Bonat, EJASA **10**(02), 384 – 407 (2017)
8. Y. Yang, W. Qian, H. Zou, J. Bus. Econ. Stat. **36**(3), 456 – 470 (2018)
9. W. Qian, Y. Yang, H. Zou, J. Comput Graph Stat. **25**(2), 606 – 625 (2016)
10. W.Y. Pan, H.C. Soo, A.H. Pooi, *Determination of motor insurance rates*, in Proceedings of the International Conference on Mathematics, Statistics, and Financial Mathematics, ICMSFM, 18-19 November 2014, Malaysia (2014)
11. E.W. Frees, P. Wang, Insur. Math. Econ. **38**(2), 360 – 373 (2006)

12. S. Gschl*ö*ßl, C. Czado, Scand. Actuar. J. **2007**(3), 202 – 225 (2007)
13. E.W. Frees, J. Gao, M.A. Rosenberg, N. Am. Actuar. J. **15**(3), 377 – 392 (2011)
14. C. Czado, R. Kastenmeier, E.C. Brechmann, A. Min, Scand. Actuar. J. **2012**(4), 278 – 305 (2012)
15. N. Kr*ä*mer, E.C. Brechmann, D. Silvestrini, C. Czado, Insur Math Econ. **53**(3), 829 – 839 (2013)
16. J. Garrido, C. Genest, J. Schulz, Insur. Math. Econ. **70**, 205 – 215 (2016)
17. I.K. Yeo, RA Johnson, Biometrika **87**(4), 954 – 959 (2000)
18. A.H. Pooi, Appl. Math. Sci. **6**(115), 5735 – 5748 (2012)
19. M. Hallin, J-F. Ingerbleek, Scand. Actuar. J. 49 – 64 (1983)