PREDICTION OF RBNS AND IBNR CLAIMS USING CLAIM AMOUNTS AND CLAIM COUNTS

BY

RICHARD VERRALL, JENS PERCH NIELSEN
AND ANDERS HEDEGAARD JESSEN

ABSTRACT

A model is proposed using the run-off triangle of paid claims and also the numbers of reported claims (in a similar triangular array). These data are usually available, and allow the model proposed to be implemented in a large variety of situations. On the basis of these data, the stochastic model is built from detailed assumptions for individual claims, but then approximated using a compound Poisson framework. The model explicitly takes into account the delay from when a claim is incurred and to when it is reported (the IBNR delay) and the delay from when a claim is reported and to when it is fully paid (the RBNS delay). These two separate sources of delay are estimated separately, unlike most other reserving methods. The results are compared with those of the chain ladder technique.

1. INTRODUCTION

There are a number of stochastic models that can be used to estimate reserves in non-life insurance mathematics; see Schmidt (2007) for an extensive literature list. Wüthrich and Merz (2008), England and Verrall (2002) and Taylor (1986) provide useful overviews of stochastic claims reserving. Most of these models have been designed to deal with data which have been aggregated in some way, as this often makes the presentation of the data convenient. However, this aggregation of the data may lead to a loss of information that in some cases can give relatively poor estimation and prediction of the outstanding liabilities. This has been the subject of some recent papers on reserving: for example, Taylor and McGuire (2004) uses a generalized linear model framework to model the characteristics of individual claims. Norberg (1993, 1999) sets out a framework for the claims occurrence, reporting and payment process, at an individual claims level. Also relevant in this context is Norberg (1986). Models based on individual claims data tend to be very detailed, often rather complex and use extensive data to estimate parameters. For the practising actuary however, they have certain limitations: in particular, they are difficult to implement because the use of data at an individual level is particularly

Astin Bulletin 40(2), 871-887. doi: 10.2143/AST.40.2.2061139 © 2010 by Astin Bulletin. All rights reserved.

computationally challenging. Furthermore, very large and elaborate data sets are often hard to get in insurance companies, and it is often the case that a model will only get used in a practical situation if it can be applied to a wide variety of data sets across a wide variety of business lines.

It can be seen from this that there is a difficult choice to be made of whether to use individual data, which is theoretically appealing but computationally difficult, or whether to use aggregate data, which is much easier to deal with but from which some (possibly important) information has been lost.

In this paper, we propose a stochastic model for loss reserving which is based on incremental reported claim numbers N_{ij} and paid amounts X_{ij} and which serves to predict RBNS and IBNR claims separately. We take a similar approach as Ntzoufras and Dellaportas (2002) and Wright (1990), in that we build a model for aggregate paid claims from basic principles at the level of individual data. We believe that the use of the aggregated counts data, which is readily available in most actuarial offices, can improve the reserving accuracy. Other interesting possibilities of adding extra data compared to the simple chain ladder method are Liu and Verrall (2009a, 2009b), Ntzoufras and Dellaportas (2002), Schnieper (1991), Verrall (1990) and Verrall and England (2005).

Including the data on incurred counts enables us to model the payment patterns for RBNS claims. In this way, we separate the reporting delay (in the incurred counts) from the payment delay (in the paid claims). In contrast, widely used methods such as the chain ladder technique simply include all sources of delay in a single development pattern.

The chain ladder technique was originally introduced without a stochastic model specified using heuristic reasoning to estimate the sum of Incurred But Not Reported (IBNR) claims and RBNS claims. In Hachemeister and Stanard (1975), Mack (1991), Neuhaus (2004) and Renshaw and Verrall (1994) stochastic models have since been formulated that lead to the same estimates as the chain ladder method. In all these papers, the models take the data as given and do not attempt to build a model based on the commonly accepted compound Poisson framework, used elsewhere in risk theory. It could be argued that the over-dispersed Poisson model could be interpreted in this way (see, for example, England and Verrall, 1999), but this was not the original approach taken. In this paper we derive a model which is an approximation to an exact model based on more detailed data, and which is a compound Poisson model. Separate models are defined for IBNR and RBNS claims, allowing for the prediction of IBNR and RBNS claims separately. In this way, we take a similar approach to Bühlmann et al (1980) (for example), who also split the reserve into two elements.

The paper is set out as follows. In Section 2 we define the notation and describe the data which we will assume is available. In Section 3, the theoretical development is given, working from assumptions at the level of individual data, which shows how the claims development is split into the IBNR and RBNS delays. In Section 4, the model which we will actually apply is given, as an approximation to the more detailed models for individual data. Section 5 considers prediction and Section 6 examines results based on the model. The conclusion is contained in Section 7.

2. Data and notation

In deciding which data to use, the two main considerations are that this data should be readily available for most practical actuaries and that this extra data should have the potential to provide improved estimates of the RBNS and IBNR reserves. In general, we assume that a run-off triangle consists of the random variables $\Delta_m = \{X_{ij} : (i,j) \in \mathcal{A}_m\}$ where $\mathcal{A}_m = \{(i,j) \in \mathbb{N} \times \mathbb{N}_0 : 1 \leq i + j \leq m\}$.

$$\begin{array}{llll} X_{10} & X_{11} & \cdots & X_{1,m-2} \; X_{1,m-1} \\ X_{20} & X_{21} & \cdots & X_{2,m-2} \\ \vdots & \ddots & & \\ X_{m0} & & & \end{array}$$

The random variables in Δ_m could represent either the paid or the incurred data. For the paid data, X_{ij}, $(i,j) \in \mathcal{A}_m$, are the total claims incurred in period i which are paid with j periods delay from when they were incurred. For the incurred data, X_{ij} are total claims incurred in period i with delay j . In this case, the claims which are reported but not paid a claim estimate is included, rather than the actual payment. In a practical context, a decision has to be made as to whether to use paid or incurred data (or to try to use both). The advantage of using only paid data is that Δ_m then contains no human judgement: it is "real data". However, it is possible that the case estimates, which are included in the incurred data, contain useful information about future payments. On the other hand, the inclusion of claim estimates is debatable since these are not "real data", and there may be political or business related considerations which make the individual claim estimates unreliable. There is further some variability that is disregarded, since the claim estimates and the actual paid amount often differ. Finally, claims estimates appear as paid at the wrong point in time which can disrupt the cashflow modeling.

For these reasons the approach taken in this paper is to use the triangle of paid claims, which is easily accessible in most companies. Thus, X_{ij} is the total claims incurred in period i and paid with j periods delay. We combine this paid triangle with a second triangle, in the same format as the paid triangle above, containing the number of incurred claims. Note that these data are obtained from the incurred claims, and therefore use some of the information not used when just the aggregate paid claims are used. The random variables are denoted by $\aleph_m = \{N_{ij} : (i, j) \in \mathcal{A}_m\}$ where N_{ij} represents the number of claims incurred in period i and reported with j periods delay (in period $i + j$) for $(i, j) \in \mathcal{A}_m$. It would also have been possible to consider the number of payments. This would remove one difficulty with the incurred claims (as discussed below), in that some reported claims end up without a payment being made — the zero-claims. However, the use of the number of payments can lead to a number of other difficulties. For example, the number of payments is rarely easily accessible in insurance companies. The number of reported claims on the other hand is usually relatively easy to obtain.

We therefore assume that two triangles of data are available: the triangle of aggregate paid claims and the triangle of the number of incurred claims. As mentioned in the introduction, restricting the analysis to just these triangles to some extent complicates the statistical analysis, and it would be better, from this point of view at least, to assume that data was available at whatever level of detail was required. A disadvantage of this would be that the estimation of the models would become much more computationally intense, and the models could not be used when the data requirements were not satisfied.

Thus, a compromise about the data has been made, but it will be seen in Section 5 that by just including the count data for the incurred claims, it is possible to improve significantly on the chain ladder technique without completely giving up the well known chain ladder idea.

In the next section we define the model for Δ_m and $\aleph_m$ using some unobservable random variables. The structure which are intended to mimic the models from Norberg (1993, 1999), but using a discrete time framework.

3. Modelling IBNR and RBNS claims

In this section a micro model is introduced, using a number of (in practise often) unobservable random variables. Based on this micro model, a compound poisson interpretation of IBNR and RBNS claims (at the individual level) is derived. The aim of specifying a model for the individual claims is to derive a suitable model for the aggregated data which is assumed to be available.

Consider the kth claim of the N_{ij} claims incurred in period i and reported with j periods delay. Usually a claim is not paid immediately upon notification to the insurance company. The final claim amount is generally paid with some waiting time from notification, often due to general consideration of the case, legal issues, collection of further information concerning the case, etc. In other words, there is a delay from a claim being reported until it is fully paid. The claims that have been reported but are not yet paid are the so called RBNS claims (or Incurred But Not Enough Reported (IBNER) claims). The related delay in payment is referred to as the RBNS delay.

Denote by N_{ijk}^{paid} the part of the N_{ij} claims which are (fully) paid with k periods delay (after being reported), $k = 0, \ldots, d$. Here $k = 0$ corresponds to a claim being paid in the same period as it was reported whereas $k = d$ is the maximal possible RBNS delay in the model. d could be chosen using information from the underlying data or the judgement from a claims handler.

The aggregate paid claims will depend on the number of claims paid in each development period, rather than the number of reported claims. The number of claims incurred in period i and (fully) paid with j periods delay after being reported is denoted by N_{ij}^{paid}, where

$$N_{ij}^{paid} = N_{ij0}^{paid} + N_{i,j-1,1}^{paid} + \cdots + N_{i,j-\min\{j,d\},\min\{j,d\}}^{paid} = \sum_{k=0}^{\min\{j,d\}} N_{i,j-k,k}^{paid}$$

for $(i,j) \in \mathcal{A}_m$.

Thus, the lifetime of a claim is divided into two: the IBNR delay and the RBNS (and IBNER) delay. These two sources of delay are modelled separately. The IBNR delay is considered when a model is specified for the reported numbers of claims since the outstanding numbers for this triangle are the claims still to be reported. For the reported claims, the RBNS delay can be considered by specifying a model for the number of claims paid, given the numbers of reported claims. In other words, we specify a model for $N_{ijk}^{paid} \mid N_{ij}$.

To begin with the numbers of reported claims, it is assumed that the number of claims incurred in period i and reported with j periods delay, N_{ij}, are independently distributed and have an over-dispersed Poisson distribution. It is well-known that for this model simplest way to obtain the predictions of future incurred claims is to use the chain ladder technique (see, for example, Hachemeister and Stanard, 1975 and Renshaw and Verrall, 1994, for proofs of this well known result). Thus, a straightforward way to obtain forecasts of the numbers of claims that will be reported for future delay periods is simply to apply the chain ladder technique to the triangle of the numbers of reported claims. In this way, the IBNR delay can be estimated.

For the RBNS delay, it is necessary to consider the delay in a reported claim being paid, and also consider the issue of a claim being paid in partial payments. For the estimation of the outstanding claims, the distribution of the claim severities is also required. In this paper, it is assumed that claims are settled with a single payment, which simplifies the theory, estimation and data questions considerably. Although it may often be the case that there is often more than one payment per claim, including this in the model leads to a much more complicated approach. Also, we believe that the simplified approach used in this paper should provide some useful and realistic insights into the different sources of the claims delay patterns. Finally, data are often not available on the development of the payment patterns and definitions on payments may differ from one insurance company to the other, or even within the same insurance company. With this assumption,we can now specify the distribution of $N_{ijk}^{paid} \mid N_{ij}$. Given N_{ij} we assume that the distribution of the numbers of paid claims follows a multinomial distribution. i.e.

$$\left(N_{ij0}^{paid}, \ldots, N_{ijd}^{paid}\right) \sim \text{Multi}(N_{ij}; p_0, \ldots, p_d)$$

for $(i,j) \in \mathcal{A}_m$ and $p_0 + \cdots + p_d = 1$ where $p_i \in (0,1)$, $0 \leq i \leq d$.

The aggregated incremental claims can be obtained by summing the individual payments:

$$X_{ij} = \sum_{k=1}^{N_{ij}^{paid}} Y_{ij}^{(k)} \tag{1}$$

for $(i, j) \in \mathcal{A}_m$. Here $Y_{ij}^{(k)}$, $(i, j) \in \mathcal{A}_m$, $k \geq 1$, denotes an individual claim payment. The final part of the model is to specify the distribution of the individual claims, $Y_{ij}^{(k)}$. It is assumed that these are independent of the numbers of claims, and do not depend on the IBNR delay or the RBNS delay. It is also assumed that they are independently and identically distributed. We recognise that the assumption that the claim payments are identically distributed is unlikely to be valid in practice: in particular, the sizes of the payments are likely to depend on the delays. However, the model with this basic assumption provides a reasonable starting point and ways in which this can be relaxed can be explored in further developments of the approach.

With these distributional assumptions, the likelihood function can be written as

$$\begin{aligned}
\mathcal{L}_{\aleph_m, \triangle_m} &= \mathcal{L}_{\aleph_m} \times \mathcal{L}_{\triangle_m | \aleph_m} \\
&= \left(\prod_{i=1}^{m} \prod_{j=0}^{m-i} P\big(N_{ij} = n_{ij}\big) \right) \\
&\quad \times \left(\prod_{i=1}^{m} f_{X_{i0}, \ldots, X_{i,m-i} | N_{i0}, \ldots, N_{i,m-i}} \big(x_{i0}, \ldots, x_{i,m-i} \,|\, n_{i0}, \ldots, n_{i,m-i} \big) \right).
\end{aligned}$$

Since $\mathcal{L}_{\aleph_m}$ and $\mathcal{L}_{\Delta_m | \aleph_m}$ are not functions of the same parameters, it is sufficient to maximize $\mathcal{L}_{\aleph_m}$ and $\mathcal{L}_{\Delta_m | \aleph_m}$ separately to maximize $\mathcal{L}_{\Delta_m, \aleph_m}$. As stated above, the likelihood function of $\aleph_m$ can be maximised using the chain ladder method.

Having defined the general framework, the next section formulates an approximation to this the model and discusses its possibilities and limitations.

4. Approximating the Likelihood with an over-dispersed Poisson Distribution

As was stated in the introduction, the process we have followed is to derive a model as far as possible, based on very basic unobservable random variables, and then approximate the model as closely as possible to motivate a model for the data available. The previous section has looked at the process that generates the claims, and we now look at the resulting model for the aggregate data and derive an approximation which is easier to use in practice. The approximation to the log-likelihood function for Δ_m given $\aleph_m$ is based on the theory of generalised linear models (see, for example, McCullagh and Nelder, 1989). The approach taken is to construct a quasi-log likelihood which (for generalised linear models) requires just the first two moments, $E[X_{ij}|\aleph_m]$ and $V[X_{ij}|\aleph_m]$.

$$\begin{aligned}
E\left[X_{ij}\,|\,\aleph_m\right] &= E\left[E\left[X_{ij}\,|\,N_{ij}^{paid}\right]|\,\aleph_m\right]\\
&= E\left[E\left[\sum_{k=1}^{N_{ij}^{paid}} Y_{ij}^{(k)}\,|\,N_{ij}^{paid}\right]|\,\aleph_m\right]\\
&= E\left[N_{ij}^{paid}E\left[Y_{ij}^{(k)}\right]|\,\aleph_m\right]\\
&= E\left[N_{ij}^{paid}\,|\,\aleph_m\right]E\left[Y_{ij}^{(k)}\right]
\end{aligned}$$

$$\begin{aligned}
V\left[X_{ij}\,|\,\aleph_m\right] &= E\left[V\left[X_{ij}\,|\,N_{ij}^{paid}\right]|\,\aleph_m\right] + V\left[E\left[X_{ij}\,|\,N_{ij}^{paid}\right]|\,\aleph_m\right]\\
&= E\left[V\left[\sum_{k=1}^{N_{ij}^{paid}} Y_{ij}^{(k)}\,|\,N_{ij}^{paid}\right]|\,\aleph_m\right] + V\left[N_{ij}^{paid}E\left[Y_{ij}^{(k)}\right]|\,\aleph_m\right]\\
&= E\left[N_{ij}^{paid}V\left[Y_{ij}^{(k)}\right]|\,\aleph_m\right] + V\left[N_{ij}^{paid}E\left[Y_{ij}^{(k)}\right]|\,\aleph_m\right]
\end{aligned}$$

Since the claim severities are assumed to be independently, identically distributed, we can write $E[Y_{ij}^{(k)}] = \mu$ and $V[Y_{ij}^{(k)}] = \sigma^2$. Hence

$$V\left[X_{ij}\,|\,\aleph_m\right] = E\left[N_{ij}^{paid}\,|\,\aleph_m\right]\sigma^2 + V\left[N_{ij}^{paid}\,|\,\aleph_m\right]\mu^2 \tag{2}$$

It can be seen that the distribution of $N_{ij}^{paid}|\aleph_m$ plays a crucial role in the first two moments of $X_{ij}|\aleph_m$. This distribution can be derived by considering sums of the numbers of claims paid with the appropriate length of delay, since (as shown in section 3) $N_{ij}^{paid} = \sum_{k=0}^{\min\{j,d\}} N_{i,j-k,k}^{paid}$. Considering first the mean,

$$\begin{aligned}
E\left[N_{ij}^{paid}\,|\,\aleph_m\right] &= E\left[\sum_{k=0}^{\min\{j,d\}} N_{i,j-k,k}^{paid}\,|\,\aleph_m\right]\\
&= \sum_{k=0}^{\min\{j,d\}} E\left[N_{i,j-k,k}^{paid}\,|\,\aleph_m\right] && (3)\\
&= \sum_{k=0}^{\min\{j,d\}} N_{i,j-k}\,p_k && (4)
\end{aligned}$$

Hence

$$E\left[X_{ij} \mid \aleph_m\right] = \sum_{k=0}^{\min\{j,d\}} N_{i,j-k}\, p_k\, \mu. \tag{5}$$

Assuming that the numbers of claims paid from different origin years are uncorrelated,

$$\begin{aligned} V\left[N_{ij}^{paid} \mid \aleph_m\right] &= V\left[\sum_{k=0}^{\min\{j,d\}} N_{i,j-k,k}^{paid} \mid \aleph_m\right] \\ &= \sum_{k=0}^{\min\{j,d\}} V\left[N_{i,j-k,k}^{paid} \mid \aleph_m\right] &(6) \\ &= \sum_{k=0}^{\min\{j,d\}} N_{i,j-k}\, p_k\left(1-p_k\right) &(7) \end{aligned}$$

Hence,

$$\begin{aligned} V\left[X_{ij} \mid \aleph_m\right] &= \sum_{k=0}^{\min\{j,d\}} N_{i,j-k}\, p_k \sigma^2 + \sum_{k=0}^{\min\{j,d\}} N_{i,j-k} p_k (1-p_k)\mu^2 \\ &= \sum_{k=0}^{\min\{j,d\}} N_{i,j-k} p_k\left(\sigma^2 + \mu^2(1+p_k)\right) \\ &\approx \sum_{k=0}^{\min\{j,d\}} N_{i,j-k} p_k\left(\sigma^2 + \mu^2\right) \end{aligned}$$

With this approximation,

$$V\left[X_{ij} \mid \aleph_m\right] \approx \left(\sigma^2+\mu^2\right) \sum_{k=0}^{\min\{j,d\}} N_{i,j-k}\, p_k = \varphi E\left[X_{ij} \mid \aleph_m\right] \tag{8}$$

where $\varphi = \dfrac{\sigma^2+\mu^2}{\mu}$.

Since the variance is proportional to the mean, an over-dispersed Poisson model can be used, with the model for the mean being

$$E\left[X_{ij} \mid \aleph_m\right] = \sum_{k=0}^{\min\{j,d\}} N_{i,j-k}\, p_k\, \mu = \sum_{k=0}^{\min\{j,d\}} N_{i,j-k}\, \psi_k \tag{9}$$

where $\psi_k = \mu p_k$. Now

$$\sum_{k=0}^{d} \psi_k = \sum_{k=0}^{d} \mu p_k = \mu \sum_{k=0}^{d} p_k = \mu \tag{10}$$

and hence $p_k = \frac{\psi_k}{\sum_{k=0}^{d} \psi_k}$. Also, the first two moments of the claim severity distribution can be derived using $\mu = \sum_{k=0}^{d} \psi_k$ and $\sigma^2 + \mu^2 = \varphi\mu = \varphi\sum_{k=0}^{d} \psi_k$. Thus,

$$\mu = \sum_{k=0}^{d} \psi_k \tag{11}$$

and

$$\sigma^2 = \varphi \sum_{k=0}^{d} \psi_k - \left(\sum_{k=0}^{d} \psi_k \right)^2 \tag{12}$$

To summarise, the chain ladder technique will be applied to the triangle of the numbers of reported claims, an over-dispersed Poisson model is fitted to the paid claims traingle, with mean

$$E\left[X_{ij} \,|\, \aleph_m\right] = \sum_{k=0}^{\min\{j,d\}} N_{i,j-k}\, \psi_k$$

to obtain the maximum likelihood estimates of the parameters, $\hat{\psi}_0, \hat{\psi}_1, \ldots, \hat{\psi}_d$. The maximum likelihood estimates of the parameters required to obtain predictions of RBNS and IBNR claims can then be obtained as follows:

$$\hat{p}_k = \frac{\hat{\psi}_k}{\sum_{k=0}^{d} \hat{\psi}_k} \tag{13}$$

$$\hat{\mu} = \sum_{k=0}^{d} \hat{\psi}_k. \tag{14}$$

This section has derived the model which will be applied to estimate the IBNR and RBNS delays and to predict the outstanding claims. The following section considers how this model can be used to estimate the RBNS and IBNR reserves.

5. Prediction

This section considers how the model outlined in section 4 can be used to predict outstanding claims. A key feature of the model described in this paper is that it is possible to separate the reserves for RBNS and IBNR claims.

In the actuarial literature, a lot of attention has been given to the derivation of stochastic models for existing deterministic models. In fact, we would argue that too much attention has been paid to the minutiae of the philosophical underpinning of methods such as the chain ladder technique. We believe that it is more important to investigate ways to improve the insights and inferences that can be obtained from stochastic methods that look at the problem in different ways from the standard deterministic methods.

The model for the triangular arrays $(\aleph_m, \Delta_m) = \{(N_{ij}, X_{ij}) : (i,j) \in \mathcal{A}_m\}$ can be extended in a natural way to the random variables N_{ij}, $(i,j) \in \mathcal{B}_m$ and X_{ij}, $(i,j) \in \mathcal{C}_m$, where

$$\mathcal{B}_m = \{(i,j) \in \mathbb{N}_0^2 : 1 \le i \le m,\ 0 \le j \le m-1\}$$

and

$$\mathcal{C}_m = \{(i,j) \in \mathbb{N}_0^2 : 1 \le i \le m,\ 0 \le j \le m+d-1\}.$$

The random variables thus appear in this format

$$\begin{array}{ccc} N_{10} & \cdots & N_{1,m-1} \\ N_{20} & \cdots & N_{2,m-1} \\ \vdots & & \vdots \\ N_{m0} & \cdots & N_{m,m-1} \end{array} \qquad \begin{array}{ccc} X_{10} & \cdots & X_{1,m+d-1} \\ X_{20} & \cdots & X_{2,m+d-1} \\ \vdots & & \vdots \\ X_{m0} & \cdots & X_{m,m+d-1} \end{array}$$

The estimates of outstanding claims are obtained by summing the predicted values of incremental claims. We therefore require a prediction for the expected incremental paid claims, X_{il}, where $l > n - i$, given $(\aleph_m, \Delta_m)$. The future values of the paid claims will be made up of two separate elements, and the prediction methods are different for each of these. The RBNS claims arise from the claims which have already been reported: in other words, they come from the values of N_{ij} in $\aleph_m$ which are already known. The IBNR claims arise from the claims which are yet to be reported: they come from the future values of N_{ij} and, in this paper, these are predicted using the chain ladder technique. Thus, the expected RBNS claims are

$$\mu \sum_{k=i-m+j}^{\min\{j,d\}} p_k N_{i,j-k} \tag{15}$$

and the expected IBNR claims are

$$\mu \sum_{k=0}^{\min\{i-m+j-1,d\}} p_k \widehat{N}_{i,j-k} \tag{16}$$

where we have emphasised the difference from the RBNS claims by using the notation $\widehat{N}_{i,j-k}$ which is a forecast of number of reported claims.

The estimates of the RBNS and IBNR claims can be obtained by substituting in the estimates of μ and p_k, (13) and (14). This is illustrated in the following section. Also of interest are prediction errors and predictive distributions. There are a number of ways to approach this: analytically, bootstrapping or Bayesian methods. Given the relative complexity of the formulae for the RBNS and IBNR claims, we do not believe that analytical expressions for the prediction errors would be straightforward to derive. Also, prediction errors in themselves are of limited practical usefulness: predictive distributions are really required for capital setting and solvency requirements. For these reasons, we would recommend that bootstrapping or Bayesian methods are preferable: see, for example, England and Verrall (2006). In this paper, we concentrate on the properties of the parameter estimates, the reserve estimates, and the implications of the proposed method in terms of understanding the characteristics of the delays in greater detail. Prediction errors and predictive distributions are not considered any further here: this will be considered in future research when the new methodology has been developed further.

6. Data study

The application of the model is illustrated in this section, using data from Royal & Sun Alliance. The data relate to a portfolio of motor policies, and in this example the auto third part liability (TPL) data is considered. The reason for choosing this data set is that we expect there to be reasonably long settlement delays (RBNS delays). This could be of particular interest as the methodology developed in this paper explicitly models the RBNS delay. The data displayed in Table 1 is inflation corrected, so that

$$X_{ij} := \frac{Y_{ij}}{\delta_{i+j}}$$

where Y_{ij}, $(i,j) \in \mathcal{A}_{10}$, are the observed payments and δ_i is an inflation index, $1 \le i \le 10$. In a full analysis of a dataset such as this, the inflation index could be modeled independently, for example by a time series which should then be used in the prediction, $X_{ij}\,\delta_j$ for $j \ge 10 - i + 1$. For the purpose of this paper, we assume that the claims inflation has already been estimated, and we concentrate on modeling the inflation corrected payments, Δ_{10}, which are shown in Table 1.

The incurred counts are shown in Table 2.

For these data, as sometimes occurs in practice, some of the reported claims are settled with no payment. This can occur if, for example, there is consideration about who carries responsibility for a claim, fraud or similar. These are referred to as zero claims, and are, by nature, different from the

TABLE 1

THE PAID RUN-OFF TRIANGLE, X_{ij}, $(i,j) \in \mathcal{A}_{10}$, FOR THE AUTO TPL DATA.

$i\backslash j$	0	1	2	3	4	5	6	7	8	9
1	451288	339519	333371	144988	93243	45511	25217	20406	31482	1729
2	448627	512882	168467	130674	56044	33397	56071	26522	14346	
3	693574	497737	202272	120753	125046	37154	27608	17864		
4	652043	546406	244474	200896	106802	106753	63688			
5	566082	503970	217838	145181	165519	91313				
6	606606	562543	227374	153551	132743					
7	536976	472525	154205	150564						
8	554833	590880	300964							
9	537238	701111								
10	684944									

TABLE 2

THE NUMBER OF REPORTED CLAIMS, N_{ij}, $(i,j) \in \mathcal{A}_{10}$, FOR THE AUTO TPL.

$i\backslash j$	0	1	2	3	4	5	6	7	8	9
1	6238	831	49	7	1	1	2	1	2	3
2	7773	1381	23	4	1	3	1	1	3	
3	10306	1093	17	5	2	0	2	2		
4	9639	995	17	6	1	5	4			
5	9511	1386	39	4	6	5				
6	10023	1342	31	16	9					
7	9834	1424	59	24						
8	10899	1503	84							
9	11954	1704								
10	10989									

claims which result in a payment. They can be dealt with in a number of different ways: for example, it would be possible to include in this set of claim numbers a separate class for those which are settled at zero. An alternative approach which leads to the same model is to use a claim severity distribution which is of mixed type. In this paper, we use the latter approach and use a claims severity distribution which a discrete probability that a claim is zero combined with a continuous claim size distribution. The probability that a claim is settled at zero is denoted by Q, $Q \in [0,1)$, and Q is assumed to be a known constant for $(i,j) \in \mathcal{A}_m$. Thus, the distribution of claim payments is such that $P(Y_{ij}^{(k)} = 0) = Q$ and the density of $Y_{ij}^{(k)} | Y_{ij}^{(k)} > 0$ is denoted by f. The most natural assumption for f is that it is a Gamma distribution. However, for the purposes of this paper, we only consider the first two moments of this distribution. The expert advice from the company has been used to determine the fraction of reported zero-claims, Q, and the maximal possible RBNS delay, $d \leq 10$. In this case $(Q, d) = (0.2, 7)$.

Note that the mean and variance of the (non-zero) claim distribution, f, can be obtained from the estimates of μ and σ^2, using:

$$\mu = E\left[Y_{ij}^{(k)}\right] = (1 - Q)E\left[Y_{ij}^{(k)} \mid Y_{ij}^{(k)} > 0\right]$$

and

$$\sigma^2 = V\left[Y_{ij}^{(k)}\right] = (1 - Q)V\left[Y_{ij}^{(k)} \mid Y_{ij}^{(k)} > 0\right] + Q(1 - Q)\left(E\left[Y_{ij}^{(k)} \mid Y_{ij}^{(k)} > 0\right]\right)^2.$$

Thus, the estimate of $E\left[Y_{ij}^{(k)} \mid Y_{ij}^{(k)} > 0\right]$ is

$$\frac{\sum_{k=0}^{d} \hat{\psi}_k}{1 - Q}$$

and the estimate of $V\left[Y_{ij}^{(k)} \mid Y_{ij}^{(k)} > 0\right]$ is

$$\frac{\sum_{k=0}^{d} \hat{\psi}_k \left[(1 - Q)\hat{\phi} - \sum_{k=0}^{d} \hat{\psi}_k\right]}{(1 - Q)^2}$$

The chain ladder technique has been applied to the data in Table 2, in order to estimate the numbers of IBNR claims. The over-dispersed Poisson distribution derived in Section 4 has then been applied and the prediction of IBNR and RBNS claims has been conducted as proposed in Section 5. The chain ladder technique applied to the triangle of the reported numbers of claims provides estimates of the development factors, and it is straightforward to convert these into the distribution of the IBNR delay. The distribution of the IBNR delay gives the proportion of ultimate number of claims which is expected to be reported in each development period. The development pattern and distribution of the IBNR delay are shown in Table 4.

For the paid claim amounts the RBNS delay is given by the estimates of the parameters $p_0, p_1, \ldots, p_7$, which are shown in Table 5.

Notice that IBNR delay (the last line in Table 4) and the IBNS delay (in Table 5) both sum to 1.

The estimates of the mean and variance of an individual (non-zero) claim severity are 203.01 and 3,496,125.

As was mentioned at the beginning of this section, the TPL claims are expected to have relatively long settlement delays (RBNS delays) as bodily injury claims often take a long time to settle, and this can seen from the left

TABLE 4

Development factors and distribution of the IBNR delay for the numbers of reported claims

j	Development Factor	IBNR Delay
0	0.8752	
1	1.1353	0.1184
2	1.0038	0.0038
3	1.0009	0.0009
4	1.0003	0.0003
5	1.0003	0.0003
6	1.0002	0.0002
7	1.0001	0.0001
8	1.0003	0.0003
9	1.0004	0.0004

TABLE 5

Maximum likelihood estimates of p_l, $0 \leq 1 \leq 7$

j	0	1	2	3	4	5	6	7
p	0.3637	0.2881	0.1134	0.0852	0.0661	0.0358	0.0255	0.0222

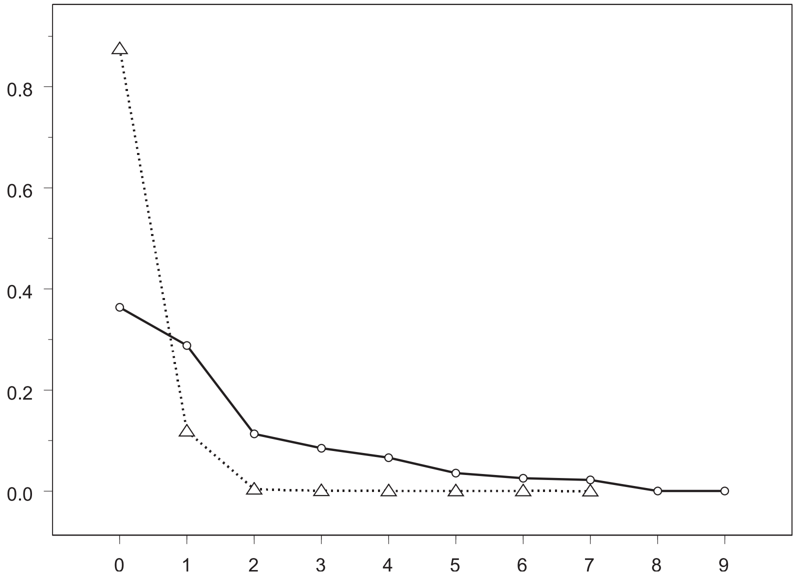

Figure 1: The dotted line represents the IBNR delay and the solid lines represents the RBNS delay.

hand plot in Figure 1, since a long time after the majority of claims has been reported there are still some significant payments. Also notice that p_0 may be relatively small. As claims happen, on average, in the middle of the year there is on average only half a year to receive the final payment in order to finish in the category of claims related to p_0. All other delay periods are full years.

Figure 1 compares the IBNR delay (from Table 4) with the RBNS delay (from Table 5).

The average IBNR delay estimated is 0.14 years whereas the average RBNS delay is 1.52 years. Hence the RBNS reserve is expected to be about ten times as large as the IBNR reserve because the individual claims are assumed iid.

Using these parameter estimates, the IBNR and RBNS can be obtained using the expressions in section 5. Table 5 shows the IBNR reserves (from claims which are not yet reported), the RBNS reserves (reported claims not yet paid), and the total reserve.

TABLE 6

THE ROW WISE RESERVE ESTIMATES SPLIT INTO IBNR AND RBNS CLAIMS, TOGETHER WITH THE CHAIN LADDER ESTIMATES

i	IBNR	RBNS	TOTAL	CHAIN LADDER
2	628	605	1,233	1,685
3	1,350	4,514	5,863	29,379
4	1,510	43,623	45,133	60,638
5	1,967	94,526	96,493	101,158
6	2,579	171,633	174,212	173,802
7	3,168	299,136	302,304	249,349
8	5,349	509,334	514,684	475,992
9	14,280	852,144	866,423	763,919
10	254,499	1,135,678	1,390,177	1,459,860
Total	285,329	3,111,192	3,396,521	3,315,779

As discussed above, the IBNR delay is (on average) shorter than the RBNS delay, and hence the RBNS reserve is expected to be larger then the IBNR reserve. The actual estimates divide the reserves such that the RBNS reserve takes up 91.6% of the total reserve and the IBNR only 8.4%: roughly 10 : 1 as suggested above. The chain ladder reserves include both the IBNR and a part of the RBNS claims, but it is not possible to split them.

7. CONCLUSION

This paper has developed a new stochastic model for claims reserving, which has a number of advantages over the standard approaches based on a single triangle of data (such as the chain ladder technique). A significant extra element in the results is that the sources of the delay in the claims process are split into the IBNR and RBNS components. We believe that this approach has the potential to make real improvements in the practical approaches to reserving, and the data study in Section 6 illustrates this.

The approach taken in this paper steers a middle course between the crude methods based on a single triangle and the very detailed methods based on data at the individual claim level. We believe that, in a practical context, this is a completely realistic approach: it does not throw away a lot of useful information, as the chain ladder technique does, but nor does it make very heavy extra computational demands.

The basic model described in this paper may be useful for some sets of data, but we would suggest that more development is needed, particularly in order to relax the assumption that the claims are identically distributed. Also, it would be useful to develop full predictive distributions using, for example, bootstrapping.

References

Bühlmann, H., Schnieper, R. and Straub, E. (1980) Claims reserves in casualty insurance based on a probability model. *Mitteilungen der Vereinigung Schweizerischer Versicherungsmathematiker*.

England, P.D. and Verrall, R.J. (1999) Analytic and bootstrap estimates of prediction errors in claims reserving. *Insurance: Mathematics and Economics*, **25**, 281-293.

England, P.D. and Verrall, R.J. (2002) Stochstic claims reserving in general insurance (with discussion). *British Actuarial Journal*, **8**, 443-544.

England, P.D. and Verrall, R.J. (2006) Predictive Distributions of Outstanding Liabilities in General Insurance. *Annals of Actuarial Science*, **1**, 221-270.

Hachemeister, C.A. and Standard, J.N. (1975) *IBNR claims count estimation with static lag functions*. 12th Astin Colloquium, IAA, Portimao, Portugal.

Liu, H. and Verrall, R. (2009a) Predictive Distributions for Reserves which Separate True IBNR and IBNER Claims. *Astin Bulletin*, **39**, 35-60.

Liu, H. and Verrall, R. (2009b) A Bootstrap Estimate of the Predictive Distribution of Outstanding Claims for the Schnieper Model. To appear in *Astin Bulletin*.

Mack, T. (1991) A simple parametric model for rating automobile insurance or estimating IBNR claims reserves. *Astin Bulletin*, **21(1)**, 93-108.

McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models, second edition*. Chapman & Hall, London.

Neuhaus, W. (2004) On the estimation of outstanding claims. *Australian Actuarial Journal*, **10**, 485-518.

Norberg, R. (1986) A contribution to modelling of IBNR claims. *Scandinavian Actuarial Journal*, 155-203.

Norberg, R. (1993) Prediction of outstanding liabilities in non-life insurance. *Astin Bulletin*, **23(1)**, 95-115.

Norberg, R. (1999) Prediction of outstanding claims: Model variations and extensions. *Astin Bulletin*, **29(1)**, 5-25.

Ntzoufras, I. and Dellaportas, P. (2002) Bayesian modeling of outstanding liabilities incorporating claim count uncertainty. *North American Actuarial Journal*, **6(1)**, 113-137.

Renshaw, A.E. and Verrall, R. (1994) *The stochastic model underlying the chain ladder technique*. Actuarial research paper no. 63.

Schmidt, K.D. (2007) A bibliography on loss reserving. Available on: http://www.math.tu-dresden.de/sto/schmidt/dsvm/reserve.pdf

SCHNIEPER, R. (1991) Separating true IBNR and IBNER claims. *Astin Bulletin*, **21**, 111-127.

TAYLOR, G.C. (1986) *Claims reserving in non-life insurance*. North Holland.

TAYLOR, G.C. and MCGUIRE, G. (2004) Loss reserving with GLM – a case study. CAS discussion paper program, 327-391.

VERRALL, R.J. (1990) Bayes and empirical Bayes estimates for the chains ladder. *Astin Bulletin*, **20**, 217-243.

VERRALL, R.J. (1991) Chain ladder and maximum likelihood. Journal of the Institute of Actuaries, **118**, 489-499.

VERRALL, R.J. and ENGLAND, P.D. (2005) Incorporating expert opinion into a stochastic model for the chain-ladder technique. *Insurance, Mathematic and Economics*, **37**, 355-370.

WRIGHT, T.S. (1990) A stochastic method for claims reserving in general insurance. *Journal of the Institute of Actuaries*, **117**, 677-731.

WÜTHRICH, M.V. and MERZ, M. (2008) *Stochastic Claims Reserving Methods in Insurance*. Wiley.

RICHARD VERRALL
Faculty of Actuarial Science and Insurance
Cass Business School
City University
London
E-mail: r.j.verrall@city.ac.uk

JENS PERCH NIELSEN
Faculty of Actuarial Science and Insurance
Cass Business School
City University
London

ANDERS HEDEGAARD JESSEN
Department of Mathematical Sciences
University of Copenhagen