A Stochastic Model to Estimate the Amount of IBNR Claims Using Micro-data

Luciene G. de Souza
Ph. D. student, PUC-Rio, DEE
lucisouzarj@gmail.com

Álvaro Veiga
Associated Professor, PUC-Rio, DEE
alvf@ele.puc-rio.br

Keywords: IBNR, Run-off triangle, Micro-data, Linear model, Truncated distributions, EM Algorithm.

Abstract

This article proposes a new micro-data stochastic model to estimate the quantity of IBNR claims. The model is stated in terms of the distribution of the time-to-report of claims with same occurrence day, conditional to the total number of claims occurred in that day. The main difficulty lies in the truncated data from the number of claims and the corresponding times to report. The model considers a mixture of exponential distributions to describe the time to report. We develop an EM algorithm to obtain the maximum likelihood estimates of parameters for the case of a simple exponential model and use an ordinary non-linear search algorithm for the mixture model. We present a case study with DPVAT —Compulsory Limited Motor Third-Party Liability —Brazilian insurance data where we compare the forecast ability of the proposed model to other micro-data models and run-off triangle based methods.

1 Introduction

There is a large literature on techniques for calculating IBNR reserves. Taylor et al. (2003) classifies the methods as deterministic or stochastic, as dynamic or static and as phenomenological or micro-structural, with optimal or heuristic parameters estimation. From the combination of these characteristics, models are organized in a evolution diagram akin to a Darwinian tree of life. The observed evolution goes from deterministic to stochastic, heuristic to optimal estimation, static to dynamic and phenomenological to micro-structural. Two characteristics are emphasized: dynamism and micro-structural structure. The authors stress that the triangle in which traditional methodologies are based present just a summary, while raw data can provide much more information.

Following the aforementioned classification, the model proposed in this article is stochastic, with optimal parameters estimation, dynamic and based on micro-data. The model aims to estimate the quantity of IBNYR claims, or true IBNR

claims or, further, Pure IBNR claims. The main variables of the model are: number of claims occurred in each day, time between occurrence and report to the insurer and number of reported claims until the last date in the sample.

Existing literature on micro-data based models is very limited. Our work borrows elements from the approaches proposed by Parodi (2013), Weissner (1978) and Antonio and Plat (2012), but presents crucial differences. Firstly, our approach can be viewed more as framework to conjointly model the problem variables, allowing the assignment of a variety of different distributions both to the number of claims and the time to report. Also, we develop maximum likelihood estimation procedures that conjointly take all information about the variables into account. Finally, we adopt a time series approach and compare performance according to their forecasting abilities.

In this article, we develop our approach by assuming two different distributions to model the time report of claim. In the simplest case, we use a simple exponential distribution as in (Weissner, 1978). For this particular case, we develop a very efficient EM algorithm. In a second version, we adopt a mixture of two exponential distributions to model the time to report. For this case, we develop a non-linear search procedure.

The approach was put into test in a forecast experiment with DPVAT —Compulsory Limited Motor Third-Party Liability —data, with 11 years of reported claims in a daily basis. This is a highly non-stationary phenomenon due frequent changes in the regulation. Also, the level of public awareness of accident victims coverage by DPVAT insurance has changed greatly due to massif publicity. The result is that the number of reported claims has been increasing constantly.

In the experiment, we adopt an adaptive procedure in order to take into account the time variability of the parameters, using gliding estimation windows. The performances of our approach and of some triangle-based methods —organized in (Schmidt and Zocher, 2008) —were compared over forecasting horizons from one to three years. In this particular application, our model showed superior performance for short forecasting horizons, but otherwise no evidence of superiority when compared traditional methods for longer horizons.

The remainder of the paper is organized as follows: section 2 discusses and presents out modeling approach as well as two versions of the model and the corresponding estimation procedures. Section 3 presents the performance measures. Section 4 presents a real world application of the model and compares it to the traditional triangle-based approaches and section 5 concludes the paper.

2 Methodology

The proposed model was initially inspired in the model presented by Weissner (1978), who treats the problem of observed data truncation. Weissner proposes to fit truncated distributions to the delay data using maximum likelihood. However, the likelihood function that he considers does not include the fact that the amount of claims already reported until the moment of truncation is also a random variable. Another method that inspired the proposed model was presented by Antonio and Plat (2012) and it does not address the problem of truncation. Thus,

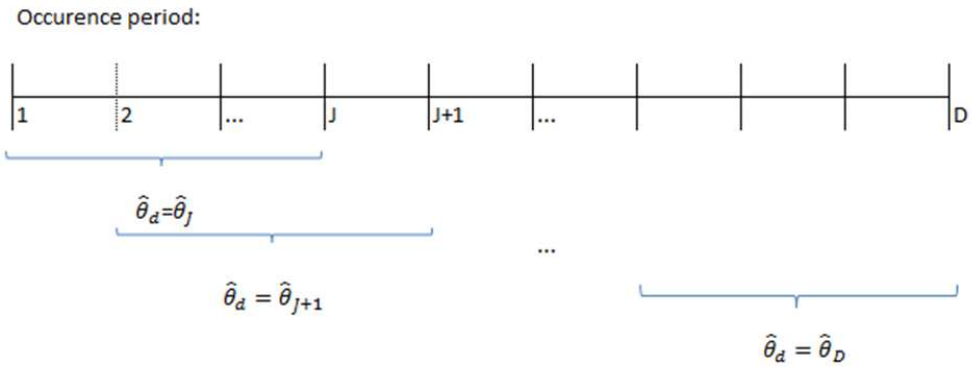
the estimated delay distribution must submit a artificially high likelihood for short delays. This problem was treated by Parodi (2013) that presents a way to fix it through the relationship between the complete distribution and the truncated distribution of delays. However, this correction can require very complicated calculations depending on the distribution adopted for adjustment.

In this work, these issues were treated in an integrated way. The number of reported claims was modeled by a binomial distribution and the reporting delay was modeled as a truncated distribution. Moreover, the total number of occurred claims is modeled by the Poisson distribution, that directly gives us the distribution of the amount of IBNR claims. The choice of Poisson distribution for modeling the total amount of occurred claims and Binomial distribution for modeling the number of reported claims until the moment of truncation, in addition to find justification in the fundamental concepts of each distribution, save a relationship that enables the accounts of the likelihood function adopted. The choice of the delays distribution is free. In this work we consider exponential distribution and a mixture of exponential distributions to delays modeling.

2.1 Formalization

Let d be a variable that represents the occurrence period of claim and D the maximum observable occurrence period in the sample, with $d = 1, \ldots, D$. Now consider a window of occurrence periods with length $J \leq D$ and occurrence periods contained in this window identified by $t = 1, \ldots, J$. The last occurrence period of this window will always be one of the observable occurrence period, by this way this window can contain since the occurrence periods interval $d = 1, \ldots, J$ until $d = D - J + 1, \ldots, D$. Then it is a sliding window that will run all observable occurrence days of the data to obtain the new parameter estimates of the proposed model as we traverse these days of occurrence. The figure 1 illustrates the sliding window of data used to estimate the vector of all needed parameters to model the phenomenon for each occurrence day, $\widehat{\theta}_d$:

Figure 1: Sliding data window



Notation:

T_t: the maximum observable delay of reporting of the claims occurred at the period t;
N_t: the number of claims occurred at each period t;
K_t: random variable that represents the number of claims occurred at the period t reported until T_t;
$\Gamma_t = (\Gamma_{t,1}, \ldots, \Gamma_{t,N_t})$: random vector of all observable delays of the claims occurred at the period t;
$\Gamma_{I,t} = (\Gamma_{t,1}, \ldots, \Gamma_{t,K_t})$: random vector of all non-observable delays of reported claims occurred at the period t;
$\Gamma_{II,t} = (\Gamma_{t,K_t+1}, \ldots, \Gamma_{t,N_t})$: random vector of all delays of the claims yet to be reported occurred at t;

$T = (T_1, \ldots, T_J)$: random vector of all maximum observable delays;
$N = (N_1, \ldots, N_J)$: random vector of the total quantity of claims occurred by period;
$K = (K_1, \ldots, K_J)$: random vector of all reported claims by occurrence period;
$\Gamma = (\Gamma_1, \ldots, \Gamma_J)$: random vector of all delays;

n_t: non observable number of the total of claims occurred at t(ultimate);
k_t: observable number of reported claims until T with origin at t;
$\tau_t = (\tau_{t,1}, \ldots, \tau_{t,N_t})$: random vector of all delays of all the claims occurred in the period t;
$\tau_{I,t} = (\tau_{t,1}, \ldots, \tau_{t,K_t})$: random vector of all delays of all the claims reported occurred in the period t;
$\tau_{II,t} = (\tau_{t,K_t+1}, \ldots, \tau_{t,N_t})$: random vector of all delays of claims yet to be reported occurred at t;

λ: vector of parameters of the delay distribution;
γ_t: vector of parameters of the distribution of the number of claims occurred at t,N_t.

The joint distribution of random variables delay, number of claims reported and total number of claims occurred in the period t:

$$\begin{aligned} f_{\Gamma_t,K_t,N_t}(\tau_t, k_t, n_t; \lambda, \gamma, T_t) = & f_{\Gamma_t/K_t=k_t,N_t=n_t}(\tau_t; \lambda, T_t) \\ & \times f_{K_t/N_t=n_t}(k_t; \lambda, T_t) \\ & \times f_{N_t}(n_t; \gamma) \end{aligned} \tag{1}$$

The parameters estimation process adopted in this work depends on the delays distribution choice. For the exponential delay distribution the parameters estimates are found by maximizing likelihood function using the EM algorithm. The model of delays as a mixture of two exponential distributions has the parameters estimated through maximization of the likelihood function using a nonlinear search algorithm implemented in Matlab.

2.2 Exponential delay model with estimation by EM algorithm

If all possible delay of reporting for claims occurred in J periods were observable we would have complete data and the likelihood function of the delay distribution parameters and the total number of claims distribution for claims occurred in these J periods would be:

$$\begin{aligned} L\left(\lambda,\gamma/\Gamma,K,N;T\right) &= \prod_{t=1}^{J} f_{\Gamma_t,K_t,N_t}\left(\tau_t,k_t,n_t;\lambda,\gamma,T_t\right) \\ &= \prod_{t=1}^{J} f_{\Gamma_t/K_t=k_t,N_t=n_t}\left(\tau_t;\lambda,T_t\right) f_{K_t/N_t=n_t}\left(k_t;\lambda,T_t\right) f_{N_t}\left(n_t;\gamma\right) \end{aligned} \tag{2}$$

where:
$f_{\Gamma_t/K_t=k_t,N_t=n_t}\left(\tau_t;\lambda,T_t\right)$: delay distribution given the quantity of claims reported until T_t, k_t, and the total of claims occurred at t, n_t;
$f_{K_t/N_t=n_t}\left(k_t;\lambda,T_t\right)$: the distribution of the probability of observe K_t claims notified until T_t;
$f_{N_t}\left(n_t;\gamma\right)$: distribution of total of claims occurred at t.

This model contains many non-observable components. The total quantity of occurred claims, N_t, is a non-observable component of the model. Other non-observable components are the delays of claims that was not yet reported. The number of claims to be reported is the IBNR quantity that we wish to estimate. To maximize the likelihood function we can use the EM(Expectation−Maximization) algorithm.

The EM Algorithm was proposed by DempsterDempster et al. (1977) and it is a iterative method to find estimates of maximum likelihood to parameters of a statistical model, when the model depends on non-observable variables. The interaction of EM alternates between a E-step of calculation of expectation, which creates an expectation function of the log-likelihood evaluated using the current values of the parameters estimates, and a M-step of maximization, which calculates the parameters by the maximization of the log-likelihood expected function obtained in the E-step. These parameters estimates are then used to determine the distribution of the latent variables in the next E-step. The equations that represent the E-step and M-step are:

E–step:

$$Q\left(\theta,\theta^{(i)}\right) = E\left[l\left(\theta/X,Y\right) \mid X=x,\theta^{(i)}\right] \tag{3}$$

M-step:

$$\theta^{(i+1)} = \arg\max_{\theta} Q\left(\theta,\theta^{(i)}\right) \tag{4}$$

where,

$\theta^{(i)}$: vector of current parameters estimates of the model of interest;
θ : parameters vector to be estimated.

After the definition of theoretical densities that correspond to the functions $f_{T_t/K_t=k_t, N_t=n_t}(\tau_t; \lambda, T_t)$, $f_{K_t/N_t=n_t}(k_t; \lambda, T_t)$ and $f_{N_t}(n_t; \gamma)$, we can obtain the equations to estimate the vector θ composed by delay parameters and total number of claims occurred by period.

The theoretical densities defined are: the Poisson density with parameter γ for variable N_t, exponential density with parameter λ for the delays τ_t and the binomial density that depends on both parameters γ e λ for the variable K_t. So, the vector θ is composed by λ and γ. Doing all required calculations we obtain:

The estimator of N_t(the quantity of claims occurred in the period t):

$$\widehat{n}_{t,\{Nt\geq k_t;\lambda^{(i)},\gamma^{(i)}\}} = \gamma^{(i)} F_{\tau_t}(T_t; \lambda^{(i)}) + k_t \tag{5}$$

where $F_{\tau_t}(T_t; \lambda^{(i)})$ is the probability of the delay be greater than the maximum observable delay T_t. So, the estimator of total quantity of claim for each occurrence period t is well defined as a percentage of the expected value of the distribution of total quantity of occurred claims which is expected that will be reported with greater delay than the maximum observable delay, T_t, added to the amount of claims occurred in t and reported with a delay less than or equal to T_t.

Calculations made to obtain the estimators $\widehat{n}_{t,\{Nt\geq k_t;\lambda^{(i)},\gamma^{(i)}\}}$, shown above, can be found in Souza (2013).

The update equation to estimate γ is given by:

$$\gamma^{i+1} = \frac{1}{J}\sum_{t=1}^{J} \widehat{n}_{t,\{Nt\geq k_t;\lambda^{(i)},\gamma^{(i)}\}} \tag{6}$$

The estimator of γ is just the average of estimates of total of claims occurred in each window time, $t = 1, \ldots, J$.

The update equation for estimation of λ:

$$\frac{1}{\lambda^{i+1}} = \frac{J\bar{\tau}_\cdot + \sum_{t=1}^{J}\{\left(T_t + \frac{1}{\lambda^{(i)}}\right)(\widehat{n}_{t,\{Nt\geq k_t;\lambda^{(i)},\gamma^{(i)}\}} - k_t)\}}{J\bar{n}} \tag{7}$$

where, $\bar{\tau}_\cdot = \frac{1}{J}\sum_{t=1}^{J}\sum_{j=1}^{k_t}\tau_{t,j}$ e $\bar{n} = \frac{1}{J}\sum_{t=1}^{J}\widehat{n}_{\{N_t\geq k_t;\lambda^(i),\gamma^(i)\}}$.

The equation above is presented in terms of inverse of λ to be more interpretive. Therefore, the update equation of the estimate λ is the inverse of weighted average between the observable delays and estimated delays for claims not reported.
The update equations will be used in the iterative process of EM algorithm. To start the process the definition of initial values, $\gamma^{(0)}$ and $\lambda^{(0)}$, is needed. The iterations are discontinued when $\lambda^{(i+1)} - \lambda^{(i)} < \varepsilon$ and $\gamma^{(i+1)} - \gamma^{(i)} < \xi$, with ε and ξ as small as you want.

The calculations to obtain the EM algorithm update equations can be found in Souza (2013).

2.3 Model of delays as mixture of exponential distributions

Further on exponential distribution, other distributions can be adjusted to delay distribution in the proposed model. It was tested the fit of the mixture of exponential, in addition to exponential distribution for delays. This mixture is defined by:

$$f_{\tau_t}(\tau_t; \lambda_1, \lambda_2, \alpha) = \alpha\lambda_1 e^{-\lambda_1\tau_t} + (1-\alpha)\lambda_2 e^{-\lambda_2\tau_t} \tag{8}$$

also used in the conditional form to observable data:

$$f_{\tau_t/\tau_t \leq T_t}(\tau_t; \lambda_1, \lambda_2, \alpha) = \frac{\alpha\lambda_1 e^{-\lambda_1\tau_t} + (1-\alpha)\lambda_2 e^{-\lambda_2\tau_t}}{1 - \alpha e^{-\lambda_1 T_t} - (1-\alpha)e^{-\lambda_2 T_t}} \tag{9}$$

The likelihood function for incomplete data of the proposed model is:

$$L(\lambda, \gamma) = \prod_{t=1}^{J} \left\{ \prod_{i=1}^{k_t} \frac{f_{\tau_t}(\tau_{t,i}; \lambda)}{1 - F_{\tau_t}(T_t; \lambda)} \right\} \sum_{n=k_t}^{\infty} \frac{n!}{k_t!(n-k_t)!} (1 - F_{\tau_t}(T_t; \lambda))^{k_t} (F_{\tau_t}(T_t; \lambda))^{n-k_t} \frac{\lambda^n e^{-\gamma}}{n!} \tag{10}$$

where $F_{\tau_t}(T_t; \lambda)$ is the probability of delay be greater than the maximum observable delay T_t.

The log-likelihood function to be maximized is:

$$l(\lambda, \gamma) = \sum_{t=1}^{J} \sum_{i=1}^{k_t} ln f_\tau(\tau_{t,i}, \lambda) + \sum_{t=1}^{J} [k_t ln\gamma + F_{\tau_t}(k_t; \lambda)\gamma - lnk_t! - \gamma] \tag{11}$$

The maximization of the log-likelihood function concerns to parameter γ produces the following estimator:

$$\widehat{\gamma} = \frac{k_\cdot}{J - \sum_{t=1}^{J}(F_{\tau_t}(T_t; \lambda))} \tag{12}$$

where, $k_\cdot = \sum_{t=1}^{J} k_t$. The derivatives of the log-likelihood function in relation to each delay distribution parameter are:

$$\frac{\partial l(\lambda, \gamma)}{\partial \lambda_1} = \sum_{t=1}^{J} \sum_{i=1}^{k_t} \frac{\alpha e^{-\lambda_1\tau_{t,i}}(1 - \lambda_1\tau_{t,i})}{\alpha\lambda_1 e^{-\lambda_1\tau_{t,i}} + (1-\alpha)\lambda_2 e^{-\lambda_2\tau_{t,i}}} - \sum_{t=1}^{J} T_t \alpha e^{-\lambda_1 T_t}\gamma \tag{13}$$

$$\frac{\partial l(\lambda, \gamma)}{\partial \lambda_2} = \sum_{t=1}^{J} \sum_{i=1}^{k_t} \frac{(1-\alpha) e^{-\lambda_2\tau_{t,i}}(1 - \lambda_2\tau_{t,i})}{\alpha\lambda_1 e^{-\lambda_1\tau_{t,i}} + (1-\alpha)\lambda_2 e^{-\lambda_2\tau_{t,i}}} - \sum_{t=1}^{J} T_t (1-\alpha) e^{-\lambda_2 T_t}\gamma \tag{14}$$

$$\frac{\partial l(\lambda, \gamma)}{\partial \alpha} = \sum_{t=1}^{J} \sum_{i=1}^{k_t} \frac{\lambda_1 e^{-\lambda_1\tau_{t,i}} - \lambda_2 e^{-\lambda_2\tau_{t,i}}}{\alpha\lambda_1 e^{-\lambda_1\tau_{t,i}} + (1-\alpha)\lambda_2 e^{-\lambda_2\tau_{t,i}}} + \sum_{t=1}^{J} (e^{-\lambda_1 T_t} - e^{-\lambda_2 T_t})\gamma \tag{15}$$

If the derivatives above we equalized to zero, is notable that is not possible to isolate the parameters of the delay distribution in order to obtain an analytical expression for each estimator. So, to find the estimates of these parameters, the log-likelihood function above was maximized in relation to each parameter by nonlinear algorithm fmincon of Matlab software, using the derivatives above to compose the gradient. The parameter γ was estimated iteratively, by following steps:
1- An initial value to γ is chosen;
2- With the value of γ fixed, the parameters of the delay distribution are find according to the explanation above;
3- A new γ is found using the estimated parameters of delay distribution;
4- If the difference between the new γ and the previous γ is greater than a certain ε, return to step 2.

2.4 Update of parameter estimates X data truncation

For the last occurrence days, there is few observed data to fit delay curves. Furthermore, the observable delays are very short. These facts make it difficult to estimate a distribution that represents the delays that will be observed in these days by the maximum likelihood method. According to Al-Athari (2008) the maximum likelihood estimator of exponential distribution parameter only exists if the sample average is less than a half of the term until the truncation of data. Because of it, the estimate of λ was replaced by his last estimated value when the expected average, $1/\hat{\lambda}$, becomes larger than a half of truncation term. When the delay distribution is a mixture of exponential, a similar rule is adopted. When the expected mean of one of the exponential distributions of the mixture exceeds a half of the term truncation of the occurrence day, the value estimated to the λ and to the α that combine the estimated distributions of previous day is repeated until the last occurrence day of the data.

2.5 Quantity of IBNR claims estimator

According to the model specification, the total amount of occurred claims in every period d, N_d, follows a Poisson distribution with parameter γ_d. However, the distribution of N_d given the known information in the last observable instant is unknown. Since the amount N_d given the known information is estimated by his expectancy and given that $(N_d - K_d) \mid \{N_d \geq k_d, K_d = k_d\} \sim Poisson(\gamma_d F_{\tau_d}(T_d; \lambda_d))$, we have:

$$\begin{aligned}\widehat{N}_d =& E\left[N_d \mid N_d \geq k_d, K_d = k_d\right] \\ =& E\left[(N_d - K_d) \mid N_d \geq k_d, K_d = k_d\right] + E\left[K_d \mid K_d = k_d\right] \\ =& \widehat{\gamma}_d F_{\tau_d}(T_d; \lambda_d) + k_d\end{aligned} \tag{16}$$

where $\widehat{\gamma}_d$ is the estimate of γ when the data window is composed by occurrence days $t = d - J + 1, \ldots, d$, with $J \leq d \leq D$. The demonstration of the distribution

of $(N_d - K_d) \mid \{N_d \geq k_d, K_d = k_d\}$ is also found in Souza (2013).
The amount of IBNR claims in the occurrence period d will be estimated by:

$$\begin{aligned}\widehat{Q}_d =& \widehat{N}_d - k_d \\ =& \widehat{\gamma}_d F_{\tau_d}(T_d; \lambda_d) + k_d - k_d \\ =& \widehat{\gamma}_d F_{\tau_d}(T_d; \lambda_d)\end{aligned} \tag{17}$$

Therefore, the total IBNR amount $\widehat{Q}$ is estimated by

$$\sum_{d=1}^{D} \widehat{Q}_d \tag{18}$$

2.6 Distribution of the amount of IBNR claims

Since the total amount of IBNR claims is a sum of independent variables with Poisson distribution, Q_d, the distribution of total amount of IBNR claims, Q, is also a Poisson variable with mean equal to the sum of average of variables Q_d. So, its confidence interval is directly calculated by Poisson percentiles with average and variance equal to the total amount of IBNR claims estimated. The amount of IBNR claims per reporting period is also Poisson distributed with mean equal to the sum of Poisson variables means that compose it. The distribution for this amount, per reporting period, is important in the evaluation of the confidence interval obtained for each notice period that was excluded from the sample to assess the quality of model prediction. Demonstrations about amount of distribution, per notice period, can be found in Souza (2013).

3 Measures to evaluate the quality of forecasts

To evaluate the quality of the forecasts, the measures: MAE(Mean Absolute Error), MAPE (Mean Absolute Percentage Error) e RMSE (Root Mean Squared Error) will be used.Let n be the last observable reporting period in the sample used for the application of the methods and A_{n+h} the quantity of reported claims on the period $n + h$ from the maximum occurrence and maximum delay of report n. From the *run-off* triangle we can obtain estimates of these quantities of claims until $n - 1$ reporting periods after n. Keeping these conditions, however maintaining a reasonable mass of data for adjusting the methods tested here, the last H reporting periods will be removed from the data that will be used to apply methods in order to forecast and evaluate the quality of forecasts of these periods. Therefore, for an horizon $h = 1, \ldots, H$ with $H = 1, \ldots, n - 1$ we have:

$$MAE = \frac{1}{H} \sum_{h=1}^{H} \mid A_{n+h} - \widehat{A}_{n+h} \mid \tag{19}$$

$$MAPE = \frac{1}{H} \sum_{h=1}^{H} \mid \frac{A_{n+h} - \widehat{A}_{n+h}}{A_{n+h}} \mid \times 100 \tag{20}$$

$$RMSE = \sqrt{\frac{1}{H}\sum_{h=1}^{H}(A_{n+h} - \widehat{A}_{n+h})^2} \tag{21}$$

where $A_{n+h} = \sum_{t=h+1}^{n} Q_{t,n-t+1+h}$ and $\widehat{A}_{n+h} = \sum_{t=h+1}^{n} \widehat{Q}_{t,n-t+1+h}$.

Among them, the measure MAE is considered as the most relevant because it exhibits the same scale of the original data.

4 Application

4.1 Data

For a case study we use the diary data of quantity of DPVAT insurance claims occurred in a window of 11 years of occurrence and notices, years 2001 to 2011. The table 1 contains the average reporting delays(by day) of claims observed in each occurrence year in the sample. Besides this information it is also presented the maximum observable delay and an average of the total of claims occurred in each day of the sample of claims reported until dec/2011, by year of occurrence.

Table 1: Descriptive Statistics

Occurrence Year	Average Delay	Maximum Delay	Average Quantity
2001	176,5	4.016	90,1
2002	176,3	3.651	95,8
2003	194,1	3.286	91,0
2004	224,0	2.921	93,4
2005	190,9	2.555	97,1
2006	173,1	2.190	96,6
2007	166,6	1.825	103,8
2008	133,2	1.460	106,0
2009	116,6	1.094	104,1
2010	96,4	729	113,0
2011	59,3	364	89,4

The distance between the maximum observable delays and the average observable delays is significantly large in all years. However, until the year of 2007 the average observable delay does not seem to be much affected by the reduction of observability of the delays. Although, from the occurrence year 2008 to the last occurrence year, the average delay is far from any average delays of previous years, reaching an average delay of almost 60 days. This average around 60 days is not reasonable. The delay distributions of recent occurrence periods are so affected by the non-observability of long delays and to complete this information is a task that should be considered by any way in the methods and/or models used.

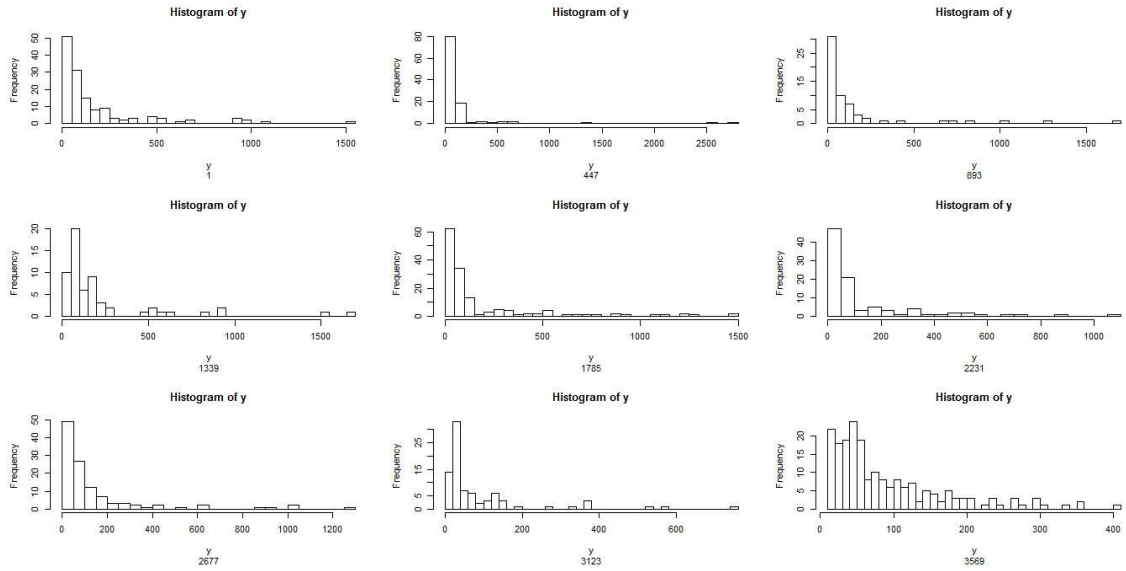
4.2 Delay distribution and frequency of claims

In the figure 2 the histograms show the distribution of observed delays in 9 days of the sample. These 9 days are well distributed among all of the occurrence days of 11 observed years (4.017 days). These empirical distributions are truncated in the right side because only claims reported until the last date can be observed. Thus, the older the day of occurrence, the lower the truncation of the distribution of delays. From the selection of these days we can analyze since a more complete empirical delay distribution (the first observed occurrence day) to an incomplete. Among the presented histograms, the most recent occurrence day, with the most incomplete distribution, is the day 3569(10/09/2010), who can present delays until 448 days of delay. The table 2 shows an identification of the selected days(y):

Table 2: Selected occurrence days(y)

Day(y)	Date of Occurrence
1	01/01/2001
447	03/23/2002
893	06/12/2003
1339	08/31/2004
1785	11/20/2005
2231	02/09/2007
2677	04/30/2008
3123	07/20/2009
3569	10/09/2010

Figure 2: Histogram of delays observed in 9 selected days



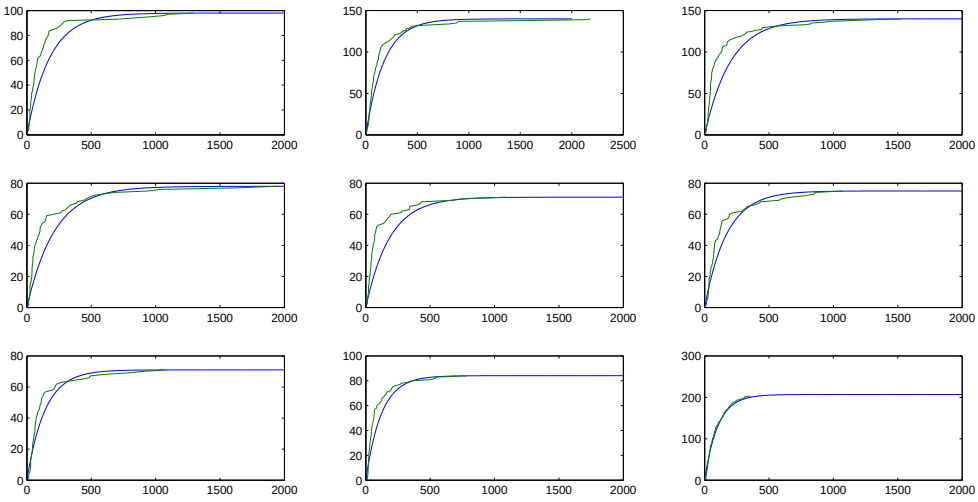
We can observe a similarity between these distributions and the exponential distribution, however the missing data by the right. Since for recent occurrence days only short delays are observable, the last graph does not seem so similar to an exponential distribution. Another distribution that could be used is log-normal

function which could reach the less frequency of delays on the left of the modal one shown in the histograms of the days $y = 1339$ and $y = 3123$.

First, it was adjusted a truncated exponential distribution for the delay data using the methodology proposed in Weissner (1978). The same methodology was used for the first fit of a mixture of exponential distributions. The parameter λ for each occurrence day t was estimated using the data until 364 occurrence days before t. Thus, the sliding window of data utilized for this estimation has length equal 365 days.

The figure 3 shows the graphs of the accumulated exponential distribution adjusted to the delay data of the 9 days selected versus accumulated curves of observed delays.

Figure 3: Exponential curves adjusted - 9 selected days



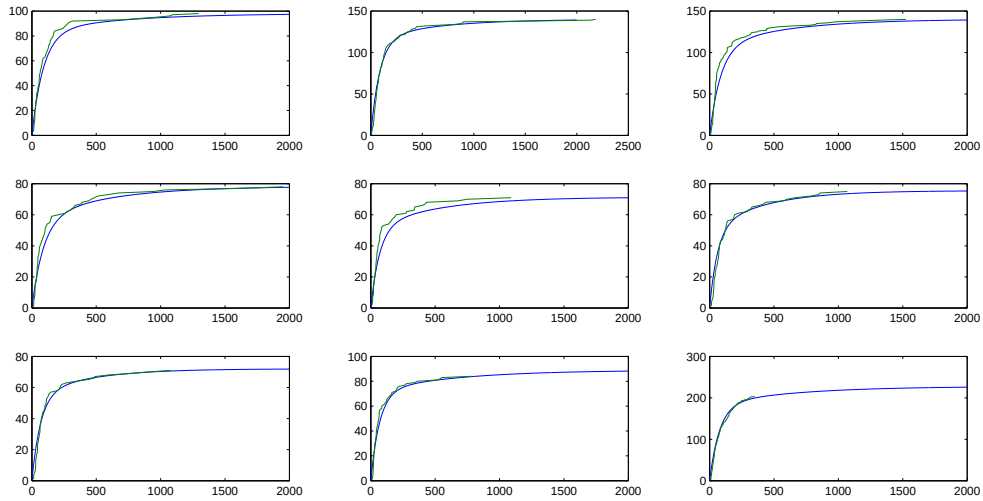
We can see that fitted exponential curves are close of the empirical curves and there is a strong development of the reports that the fitted curves cannot reach. At the same time, we have a development of long delays that could not be reached if the mean of the adjusted exponential was inferior. This is why we adjust to the delays the curve formed by a mixture of two exponential distributions. We can interpret this distribution as if there are two groups of claimants of this insurance: one that, in general, quickly reports the claim, soon after occurrence, and another group that takes a long time to report the claim, being this group, a group with rare reports with short delays. Follows in the figure 4 the graphics with curves of the fitted mixture of exponential and empirical curves.

We can notice that in these latest curves it was possible to reach the initial strong development shown by empirical curves and by the adjusted parameters values we can note the possibility of include in the model the two groups of claimants mentioned above, inclusive the percentage of participation of each one.

The graphs 5, 6, 8 and 10 show the expected values of delays in each day of occurrence present on the data base.

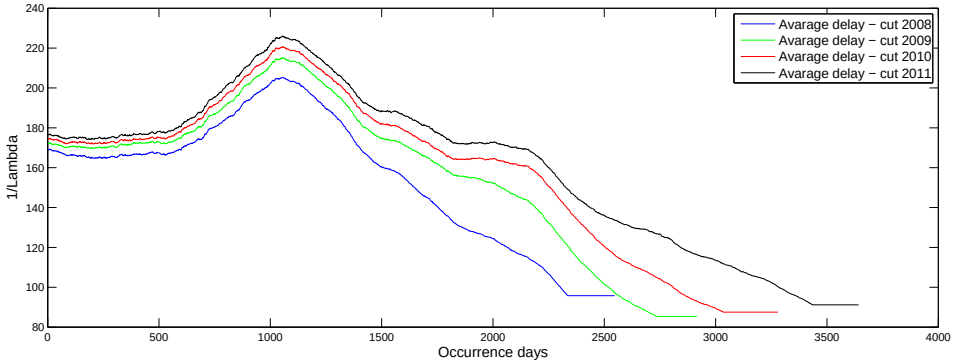
To evaluate the effect of data truncation on the estimates of delay distribution parameter, the model was fitted eliminating until the last three years of reporting

Figure 4: Mixture of exponential Curves adjusted - 9 selected days



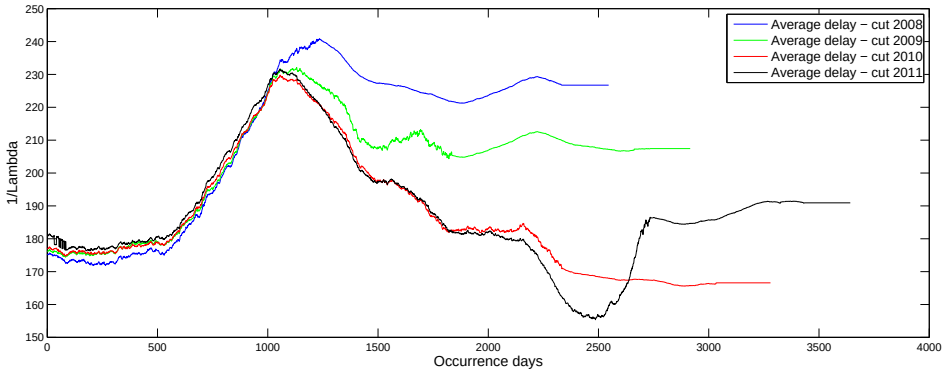
of available data. The figures 5 and 6 show the expected delays given the adjusted delay distribution per occurrence day as a simple exponential and a mixture of exponential.

Figure 5: Expected delays - Exponential (Method proposed by Weissner)



We can see that in the end of each curve there is a constant value repeated. This is due to the treatment explained in the section 2.4. Each group of data (occurred and reported claims until one, two or three periods before the last period in the complete sample or all of observable claims) used for different adjustments has the estimates of distribution parameters kept constant from different days. In the case of fit mixture distribution of exponential, each λ involved in distribution delays is fixed from a different point. The figure 6 shows the expected delays obtained from a combination of two exponential that composes delay distributions. However, the estimated λ of each component in the mixture of exponential are very distinct, one of them generates an expected delay around 600 days while another one generates an expected delay around 80 days. In the estimation process, the λ regarding the distribution with higher expected delay is the first to be fixed, because he generates expected values that beats half of deadline of data truncation faster than in the

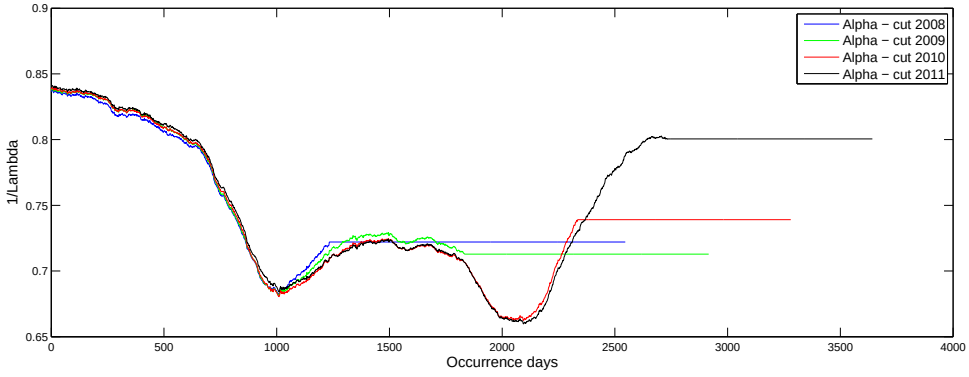
Figure 6: Expected delays - Mixture of exponential (Method proposed by Weissner)



other distribution. In this moment the α is also fixed. As the estimation of the λ of another distribution was not fixed yet, the graph of expected delay calculated from all parameters of the delay distribution is not constant from this fixation point until the other λ becomes constant too. It is possible to observe that, although we consider in the likelihood the data truncation adjusting to them a conditional distribution, the estimated parameters still varies strongly in accordance with truncation.

In the Figure 7 we can see the evolution of the estimates of the parameter α. This parameter combines two exponential distributions and it is about 85% in the first occurrence days studied. During the period between 2002(occurrence day around 500) and 2008 (occurrence day around 2500), the estimated value of α is smaller, indicating a growth of the frequency of long delays in this period, after it, he starts to rise reaching 80%.

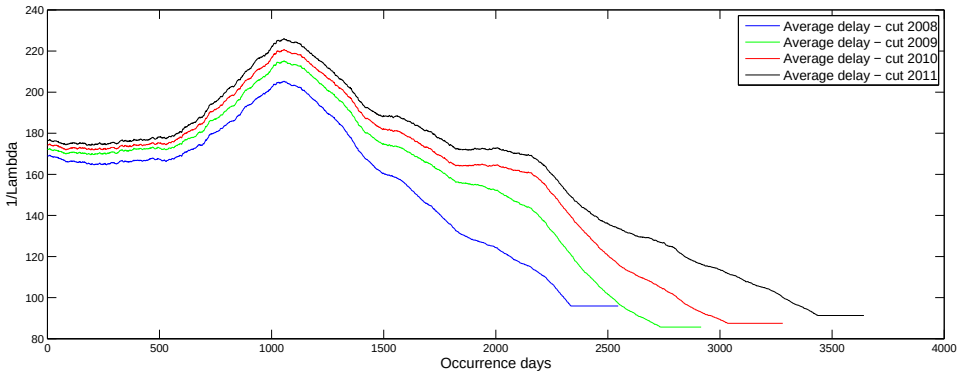
Figure 7: Parameter alpha - Delays as a mixture of exponential (Method proposed by Weissner)



From the adjusted delays curves, we can get a first estimate of the expected total quantity of claims occurred in each observable day as Weissner (1978) proposed. From this quantity we can obtain the estimate of IBNR quantity. The likelihood proposed by Weissner is not complete, because he ignores that the number of observed claims until the last date in the data is also a random

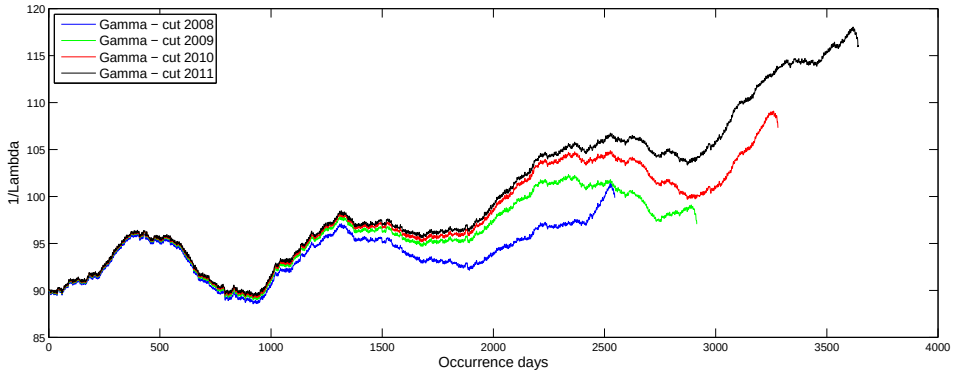
variable. The proposed model in this document considers that this amount have binomial distribution with parameters that depends on the parameters of the delay distribution, on the maximum observable delay and on the parameter of the distribution of the total number of claims occurred in each day. The figures 8 and 9 shows the graphs of the new expected values of delays and the graph of parameter γ relating to the proposed model, estimating them, the λ and γ by the application of EM algorithm.

Figure 8: Expected delays - Exponential distribution (Method proposed in this article)



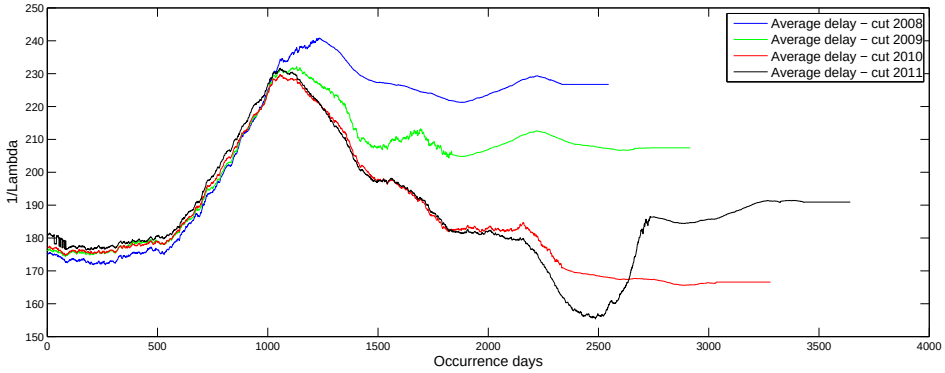
We observe that the graph of expected delays obtained from the estimation of λ (figure 10) is very similar to the one obtained independently of the adjust for estimation of parameter γ. Even in this model, the bias caused by truncation still occurs: as more truncate are the delays shorter is the estimated average delay.

Figure 9: Parameter gamma - Exponential Delays (Method proposed in this article)



The exponential curve fitted jointly with the distribution of the number of occurred claims cannot reach the fast development of delay's curve too, as said in the beginning of this section. Therefore, a joint estimation considering the delay's distribution as a mixture of exponential using the showed likelihood function optimized from a search nonlinear algorithm implemented in Matlab was also realized. The figure 10 presents the expected delays of the adjusted curves for each occurrence day. The figure 10 refers to curves adjusted to claims reported until the end of the years 2008 to 2011, as legend identify.

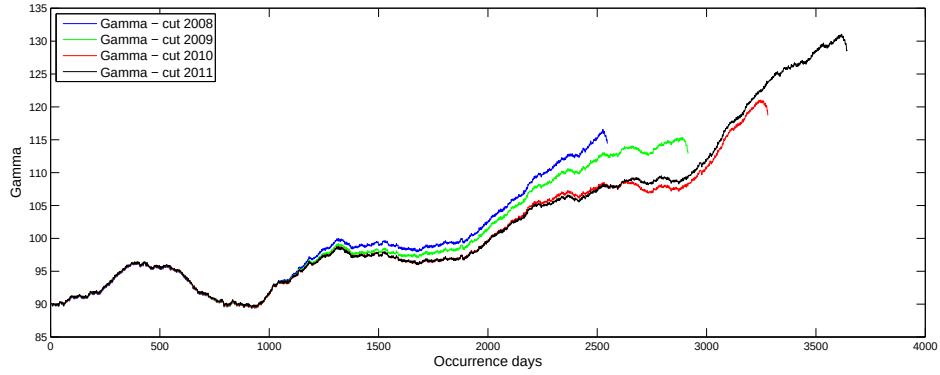
Figure 10: Expected values of delays - Mixture of exponential(Method proposed in this article) - Samples with and without the data of last observed reporting years



The expected values of delays showed in the figure 10 do not show the bias of truncation showed in the figure 8. This occurs due the application of rule described in 2.4. When data of claims with reporting date until dec/2008 is used, the estimate of parameter λ of mixture of exponential which represents longer delays and the estimate of the parameter α begins to be repeated around the occurrence day 1400.

Using data of reported claims reported until dec/2009 this repetition of the last estimate starts from the occurrence day 1800, after the falling of the expectancy of the adjusted distribution to the first exponential of the mixture of exponential. The same occurs when we use data with claims reported until dec/2010, the repetition of parameters estimates occurs after the falling of the expectancy of the adjusted distributions. Only with data of claims reported until dec/2011 it is possible to generate estimates of parameter that give us a increasing expected delay until the estimates start to be repeated.

Figure 11: Parameter gamma - Delays as mixture of exponential (Method proposed in this article)



From the graphs 9 and 11 we observe that the joint parameters estimation of the distributions incorporated in the model influences the parameters estimates obtained. The estimates of the parameter of total amount of claims distribution, γ, vary according to the specification and estimation of parameter of the delay distribution. These estimates of γ seem to be less sensitive to truncation when

the estimation is made jointly with delay distribution parameters as a mixture of exponential. We also see that both graphs exhibit similar movements for the estimate of the parameter γ during the periods of occurrence studied. Around the period of occurrence 1000(sep/2003) both lines starts an upward trend of occurred amount of claims in this insurance, already well publicized in the national media.

The adjusted parameters utilizing data of eleven observed years except the last month of report (data from jan/2001 to nov/2011) until those obtained with the use of data without the last three months of report (data from jan/2001 to sep/2011) have a development very similar to the development of the parameters obtained by the adjusting by the complete data, with the eleven observed years.

With the estimate of parameters, the estimates of total claims amount per occurrence were calculated. The figures 12 and 13 show the relation between the amount of observed claims k_d, the total amount of claims estimated by the model which only considers the distribution of delays as a mixture of exponential (with no association with any distribution of another variables of the process) and the model proposed in this work who also models the total amount of occurred claims per period, N_d, in accordance with Poisson distribution.

Figure 12: Ultimate x k_d x γ - Mixture of exponential (Method proposed by Weissner)

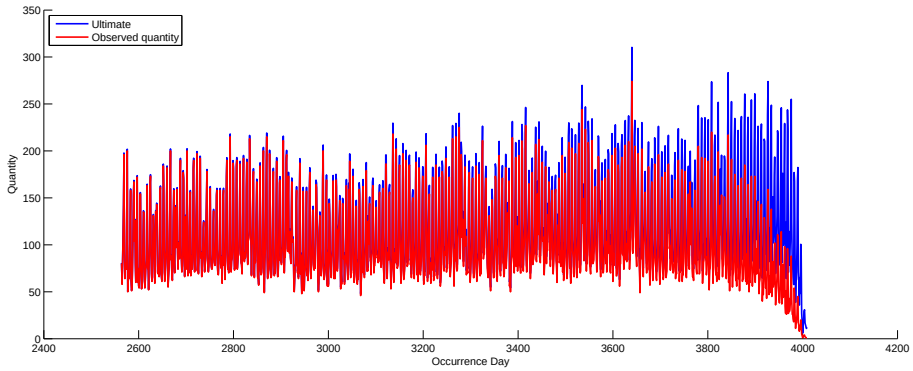
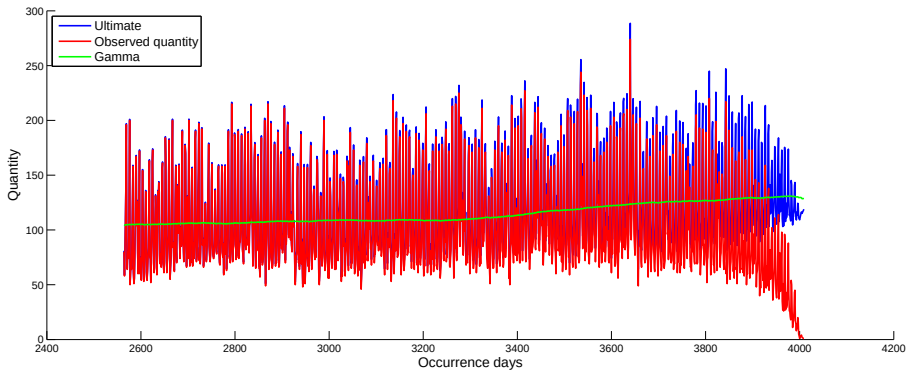


Figure 13: Ultimate x k_d x γ - Mixture of exponential(Method proposed in this article)



The figures 12 and 13 show the amount of observed claims and the total amount

estimated for each occurrence day starting from the day 2500(nov/2007) by the last day in the base of data, 4017(dec/2011). Based on them, we can observe a problem in the total amount of claims estimated by the method proposed by Weissner (blue portion at the right extremity on the graph). Since the estimate of this amount by this method, is made by the application of a factor calculated from the delays distribution on the observed amount of claims k_d, the total amount of claims estimated is influenced by the abrupt fall of the number of observed reports in the last occurrence periods, which is not reasonable, because there is no justification for a so big drop in the total number of occurred claims estimated in a so short range of time.

The method proposed in this article is robust in concerning of this drop because it has a percentage of total amount of claims estimated by the distribution of N_d as an estimative of the total quantity of claims occurred in d. We can observe in the figure 13 that in the last periods the level of total estimated amount does not suffer the drop observed in figure 12. The γ obtained is also reasonable regarding to total historical amount, capturing the growing of expected amount N_d even in the last periods of occurrence. For evaluate the quality of model prediction, the amount of IBNR reported in the periods excluded of the sample utilized to adjust the model was estimated too. The estimated amount used for comparison with the predictions made by the proposed models uses triangles without tail adjustment. Hence, the tail effect was eliminated of the amount predicted by Micro-data method covered here.

4.3 Forecast errors and estimates of the quantity of IBNR claims

To forecast the last 1, 2 and 3 reporting years out of sample of model fitting :

Table 3: Error Measures of Extended B-F - annual forecasts

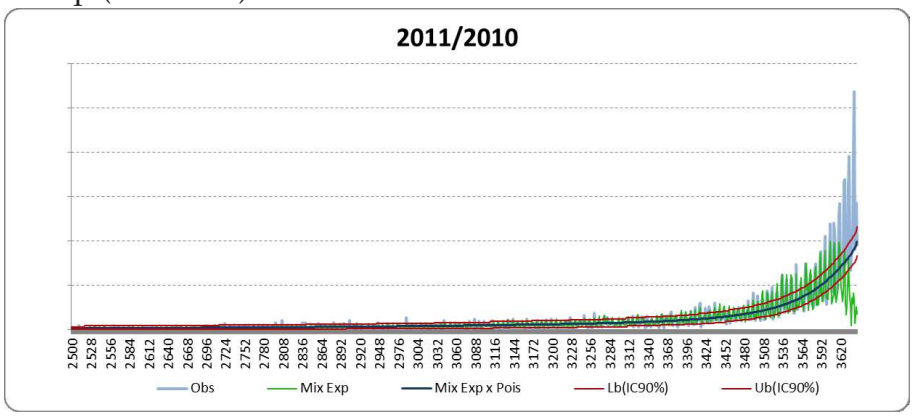
Extended B-F Methods	Until Dec/2008			Until Dec/2009			Until Dec/2010			IBNR
	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	2011
Ultimate LD(Pan Devel.)	24%	866	946	25%	1,872	2,150	8%	1,089	1,089	24,984
Ultimate Pan(Pan Devel.)	23%	852	932	26%	1,904	2,183	8%	1,197	1,197	25,131
Ultimate Pan*(Pan Devel.)	23%	852	932	26%	1,904	2,183	8%	1,197	1,197	25,131
Ultimate LD(CL Devel.)	25%	930	1,021	26%	1,936	2,232	9%	1,237	1,237	25,164
Ultimate Pan(CL Devel.)	24%	898	984	26%	1,946	2,240	9%	1,303	1,303	25,243
Ultimate Pan*(CL Devel.)	25%	920	1,011	27%	1,975	2,274	9%	1,359	1,359	25,342
Ultimate Mack(Mack Devel.)	26%	978	1,079	27%	2,013	2,324	9%	1,367	1,367	25,333
Ultimate Pan(AD Devel.)	31%	1,093	1,179	31%	2,205	2,488	11%	1,609	1,609	26,264
Ultimate LD(AD Devel.)	32%	1,188	1,292	32%	2,292	2,595	12%	1,702	1,702	26,389
Ultimate Pan*(AD Devel.)	32%	1,190	1,296	33%	2,355	2,661	13%	1,857	1,857	26,630
Mixture of Exp.(Proposed)	43%	1,102	1,142	24%	1,160	1,160	14%	2,082	2,082	19,856
Mixture of Exp.(Weissner)	46%	1,229	1,259	22%	856	924	20%	2,878	2,878	19,721
Ultimate CC(Pan Devel.)	35%	1,522	1,771	42%	3,120	3,615	22%	3,248	3,248	28,015
Ultimate AD(Pan Devel.)	36%	1,551	1,809	42%	3,161	3,662	23%	3,310	3,310	28,104
Ultimate CC(CL Devel.)	37%	1,577	1,837	42%	3,173	3,685	23%	3,375	3,375	28,151
Ultimate AD(CL Devel.)	37%	1,601	1,868	43%	3,209	3,727	24%	3,433	3,433	28,231
Ultimate CC(AD Devel.)	44%	1,812	2,072	48%	3,492	3,996	26%	3,767	3,767	29,350
Ultimate AD(AD Devel.)	44%	1,812	2,072	48%	3,492	3,996	26%	3,767	3,767	29,350
Exponential(Proposed)	75%	2,853	3,158	63%	3,355	3,377	40%	5,832	5,832	9,781
Exponential(Weissner)	75%	2,844	3,145	66%	3,627	3,685	44%	6,389	6,389	9,057

In the table 3 the columns with header 'Until Dec/2008' present the measures of errors of predictions of amount of IBNR claims to be reported until dec/2009,

dec/2010 and dec/2011 using observable data until dec/2008, in other words, until three steps ahead, without effect of the tail for claims occurred in 2001 to be comparable to the triangle methods used here. The columns with header 'Until Dec/2010' show the measures of errors of predictions of the amount of claims to be reported until dec/2010 and until dec/2011 and the columns with header 'Until Dec/2009' presents the measures of errors for predictions of amount of claims to be reported until dec/2011, so one step forward the "cut" of data base.

We can note that the model with exponential distribution is not appropriate to the data studied here. His performance is the worst among the presented methods. The best predictions among the methods applied to micro-data are performed by the model proposed in this work with distribution of delays following a mixture of exponential. The model proposed by Weissner with delays distribution according to a mixture of exponential has measurements of errors very close to the model proposed in this article. However, analyzing the figure 14, where the observed amount of claims in 2011 given the observations by dec/2010 and the predictions of both models are represented, we see that the predictions obtained by Weissner are not consistent with the amount observed. In the model of Weissner the amount of IBNR is a percentage of the observed amount in each occurrence period, as the observed amount decreases in the last occurrence periods, the predictions of the IBNR amount also decreases, which is not consistent with the reality. We see that the number of IBNR claims reported during the period of interest grow extremely in the recent occurrence periods, behavior that is captured by the proposed model in this article.

Figure 14: Observed Quantities x Estimates from Mixture of Exp.(Proposed) x Mixture of Exp.(Weissner)



The micro-data method with delays distribution as a mixture of exponential is not the one that presents has smallest errors, but they are competitive with traditional methods. There are methods that use traditional data format that presents prediction errors larger than the micro-data methods with mixture of exponential presented here, as Cape Cod method with the use of Chain Ladder development factors.

Table 4: Error Measures of Extended B-F - monthly forecasts

Extended B-F Methods	Until Sep/2011			Until Oct/2011			Until Nov/2011			IBNR
	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	2011
Mixture of Exp. (Proposed)	9%	140	191	8%	238	286	4%	129	129	19,856
Exponential(Proposed)	6%	133	141	11%	343	407	9%	307	307	9,781
Mixture of Exp.(Weissner)	9%	143	185	8%	277	381	10%	324	324	19,721
Ultimate LD(Pan Devel.)	27%	596	612	17%	462	462	12%	380	380	24,013
Exponential(Weissner)	6%	135	142	13%	408	475	13%	427	427	9,057
Ultimate Pan(Pan Devel.)	23%	510	526	19%	506	507	14%	467	467	24,033
Ultimate Pan*(Pan Devel.)	23%	510	526	19%	506	507	14%	467	467	24,033
Ultimate Pan(AD Devel.)	26%	562	574	21%	559	560	16%	512	512	23,928
Ultimate Pan(CL Devel.)	26%	559	573	21%	560	560	16%	515	515	23,019
Ultimate LD(CL Devel.)	35%	770	791	24%	651	651	17%	563	563	23,381
Ultimate Mack(Mack Devel.)	36%	783	803	25%	665	665	18%	580	580	23,506
Ultimate LD(AD Devel.)	36%	790	810	25%	673	673	18%	588	588	24,453
Ultimate AD(Pan Devel.)	32%	693	709	25%	667	668	26%	853	853	28,022
Ultimate CC(Pan Devel.)	32%	693	709	25%	668	669	26%	853	853	28,025
Ultimate CC(CL Devel.)	34%	740	754	27%	714	715	27%	887	887	26,913
Ultimate CC(AD Devel.)	35%	752	764	27%	724	725	28%	894	894	27,981
Ultimate AD(AD Devel.)	35%	752	764	27%	724	725	28%	894	894	27,981
Ultimate AD(CL Devel.)	35%	748	761	27%	723	724	28%	897	897	26,982
Ultimate Pan*(CL Devel.)	39%	866	892	34%	913	918	28%	907	907	25,421
Ultimate Pan*(AD Devel.)	40%	885	910	35%	936	941	29%	939	939	26,717

To forecast the last 1, 2 and 3 reporting months out of sample of model fitting :

In the tables 3 and 4 the columns with header 'Until Sep/2011' present the measures of errors of predictions of amount of IBNR claims to be reported until oct/2011, until nov/2011 and until dec/2011 using the observed data until sep/2011, in other words, until three steps ahead, without effect of the tail to be comparable to the triangle methods used here. The remaining columns are similar to those, with 2-step or a step forward, as was explained to the tables of measurements for annual predictions. The micro-data methods present a performance far superior to most traditional methods, thus showing to be more suitable than traditional forecasts for shorter periods of development in this case study. Methods of micro-data are robust with respect to variability in data found at this level, while traditional methods are sensitive to short periods of use. The method proposed in this article showed the best results, with the lower MAPE among all other forecast methods until 2 steps ahead. The MAPE that are associate to forecast until 3 steps ahead of this model is low and very close to others micro-data methods' MAPE.

4.4 Confidence intervals obtained in this proposed approach

The tables 5 and 6 contains the observed quantities and the confidence interval of 90% for variables A_τ, claims amount reported in year τ, whose estimates generated forecast errors that were presented in the tables 3 and 4.

In table 5 we see that none of the confidence interval contains the observed amount in this period of report. This could point us that the Poisson distribution, assigned to these variables, maybe, can be replaced for another one with higher variability compared to its average. However, the Poisson parameter may have

Table 5: CI(90%) x Observed quantity by out of sample reporting year

	Last year in the sample of model fitting								
Forecast horizon	2008			2009			2010		
	L.Bound	Obs	U.Bound	L.Bound	Obs	U.Bound	L.Bound	Obs	U.Bound
2009	12,384	13,460	12,824						
2010	4,120	2,722	4,376	12,844	11,926	13,292			
2011	2,236	1,406	2,426	4,160	3,109	4,417	12,191	14,491	12,628

been underestimated. Which could be resolved with the best adjust for the delay distributions through treatment of truncation bias or even with the replacement of the distribution formed by a mixture of exponential exchanged for another one.

Table 6: CI(90%) x Observed quantity per month of reporting out of the sample

	Last month in the sample of model fitting								
Forecast horizon	Sep/2011			Oct/2011			Nov/2011		
	L.Bound	Obs	U.Bound	L.Bound	Obs	U.Bound	L.Bound	Obs	U.Bound
Oct/2011	3,081	3,195	3,266						
Nov/2011	2,177	2,179	2,333	2,892	3,378	3,072			
Dec/2011	1,683	1,430	1,821	2,183	2,181	2,340	3,012	3,232	3,195

In table 6 some confidence intervals contains the observed amount of the reported periods presented. The results in tables 5 and 6 would indicate the need for adjustments in the used model. On the other hand, the assessment of the confidence intervals obtained for this variable is not robust, because we have few available observations for this variable. To evaluate more values of this variable using the available data we would have to eliminate more data of the sample used for the model adjustment, but this could weaken the adjust of model.

According the last column in the table 3 the total estimated amounts of IBNR claims by traditional methods are between 24.000 and 29.000 claims. The methods with minor errors generate smaller prediction of total quantities.

In table 4 we see that using monthly periods to calculate IBNR through the traditional methods instead of using annual periods generates new estimates to total amount of IBNR, slightly different to those obtained per annual data, in general, lower. The method proposed in this work generates an estimation of the total amount of IBNR of 19.856 claims, is a lower forecast, although very close of the predictions of the traditional methods that showed smaller errors.

The confidence interval of 90% of total IBNR, in according to the proposed method in this work is [19.625; 20.088].

5 Conclusions:

This paper has presented a micro-data modeling approach to the problem of forecasting the quantity of IBNR claims. The model describes the conjoint distribution of the principal variables: the number of occurred claims, the number of reported claims and time to report. Different distributions can be assigned both to the number of occurred claims and the time to report. In the present paper, we have explored two distributions for the time to report: a simple exponential

distribution and a mixture of two exponential distributions. We developed a maximum likelihood estimator for both cases. For the first one, we could write an EM algorithm. For the second case, we construct a non-linear search procedure. The approach was tested in a forecasting exercise using individual data from DPVAT claims.

Although the use of micro-data based models did not showed a clear superior forecasting performance, they show a promising analytic potential since they allows the construction of confidence intervals for the forecasting very naturally, they can be used with different assumptions for the distributions of the principal variables and they can be extended to accept external explanatory variables. The case study presented consistent results with respect of information published in the media about this insurance.

References

Al-Athari, M. M. (2008). Estimation of the mean of truncated exponential distribution. *Journal of Mathematics and Statistics*, 4(4):284–288.

Antonio, K. and Plat, R. (2012). Micro-level stochastic loss reserving for general insurance.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B Methodological*, 39(1):1–38.

Parodi, B. P. (2013). Triangle-free reserving : a non-traditional framework for estimating reserves and reserve uncertainty.

Schmidt, K. D. and Zocher, M. (2008). The Bornhuetter-Ferguson Principle. *Variance Journal*, 2(1):85–110.

Souza, L. (2013). Comparação de métodos de micro-dados e de triângulo run-off para previsão da quantidade ibnr. url: www.maxwell.lambda.ele.puc-rio.br.

Taylor, G., McGuire, G., and Greenfield, A. (2003). Loss reserving: past, present and future. *ASTIN Colloquium*, (109).

Weissner, E. W. (1978). Estimation of the distribution of report lags by the method of maximum likelihood. *PCAS LXV*.