Lithuanian Journal of Statistics *2012, vol. 51, No 1, pp. 36-50*
Lietuvos statistikos darbai *2012, 51 t., Nr. 1, 36–50 p.*
www.statisticsjournal.lt

ON PREMIUM ESTIMATION USING THE C&RT/POISSON MODEL AND ITS EXTENSIONS

Meelis Käärik, Ants Kaasik
Institute of Mathematical Statistics, University of Tartu, Tartu, Estonia
Address: J. Liivi 2, Tartu, 50409, Estonia
E-mail: meelis.kaarik@ut.ee, ants.kaasik@ut.ee

Received: June 2012 Revised: September 2012 Published: November 2012

Abstract. Premium estimation is a key concept in insurance mathematics. Estimation of the mean and variance of a total claim amount of a portfolio can be considered as necessary prerequisites for this. In turn, dividing the portfolio into homogeneous subportfolios can be considered as a first step towards finding those estimates. We consider the problem of estimating the claim intensity and propose a regression trees based approach for clustering the portfolio into homogeneous subportfolios in a situation where the durations of the policies differ and overdispersion is present. Several other generalizations are discussed. A case study involving Estonian casco insurance is included.

Keywords: actuarial mathematics, collective risk model, premium calculation, classification and regression trees, overdispersion.

1. Introduction

Premium calculation is one of the fundamental issues in insurance mathematics with several proposed methods and models. In this article we focus on one of the classical ideas, namely claim distribution modelling. Modelling claim distributions allows us to use certain risk models and to divide the expected claims proportionally between risks (policies). It also connects two important problems in insurance mathematics: estimation of the total claim amount and finding the insurance premium. Also, applying a distribution based model allows us to describe the behaviour of the claims process with a reasonably small number of parameters. It is possible to model the total claim amount either directly or through the distributions of the number of claims and individual claim amounts (see, e.g., Klugman et al., 2004). We choose the latter model because it has several well-known advantages like the possibility to take into account the influence of inflation, deductibles, reinsurance and more. The weakest point of this setup is obtaining a homogeneous risk portfolio, which is a rare case in practice, but one can certainly use some clustering techniques to group similar risks. The number of different clustering algorithms is extensive (see, e.g., Hastie et al., 2009). In this paper we choose classification and regression trees (C&RT) clustering because of its easy applicability and a straightforward connection to the compound Poisson model, which is of special interest to us.

The problem of (pure) premium estimation can be divided into two main tasks: estimation of individual claim sizes and estimation of the number of claims. The first issue, i.e. analyzing models suitable for describing individual claims, has been of interest in some recent studies (see, e.g., Käärik & Umbleja, 2010, 2011; Käärik & Kadarik, 2012; Käärik & Žegulova, 2012). The current paper can be seen as a follow-up to these studies with the main emphasis on the

claim number distribution. More precisely, we utilize the classical (compound) Poisson model and analyze the effect of the following two common factors on this model:

- difference of durations of insurance periods,
- overdispersion in portfolio.

The (compound) Poisson model is a preferred choice because of its convenient additive properties (the sum of independent Poisson-distributed random variables is also Poisson-distributed, similar properties carry over to the compound Poisson distribution). The latter property is of importance when moving from the individual risk level to a class or portfolio level. With the compound Poisson model the estimation of each individual risk in homogeneous classes gives immediately an estimate for the distribution of the total claim amount as well. So it is clearly an appealing choice, but obviously the nature of the Poisson distribution involves a restriction that the expectation and variance of the claim number are approximately equal. If there is considerable overdispersion, one has to take this into account in order to get an acceptable model. Another aspect of interest is the duration of insurance policies: in most classical models the duration is assumed to be constant for simplicity, but in practice this is rarely the case, thus we are looking for a model that takes the different lengths of insurance periods into account as well.

The paper is organized as follows. In Section 2 we recall the properties of the classical collective risk model and the basics of the C&RT clustering method. In section 3 we turn our attention to the problem of potentially different durations of insurance periods. The question of allowing overdispersion in the proposed model is addressed in section 4, where the quasi-likelihood framework leads us to an overdispersed Poisson model. In section 5 we offer possible generalizations to handle overdispersion. A practical application is described in section 6 and final comments and conclusions are given in section 7.

2. Preliminaries

2.1. Classical collective risk model

Let us start with the classical collective risk model (see, e.g., Kaas et al., 2009; Klugman et al., 2004). The total claim amount S is given in the form

$$S = \sum_{j=1}^{N_*} Z_j,$$

where N_* is the frequency (number of claims) in a given period (say, in a year) and Z_j are severities (individual claim sizes). We assume also that the individual claims $Z_1, Z_2, \ldots$ are independent and also independent from the number of claims N_* (i.e. the claim count does not depend on the claim size).

The main idea of premium calculation using the collective risk model is the following:

- cluster the portfolio into homogeneous subportfolios,
- estimate the frequency and severity in each subportfolio,
- using the collective model, estimate the total claim amount in each subportfolio and divide the expected claim amount (proportionally) between policies.

Naturally there are several different clustering techniques that can be used and the choice of method depends on the particular problem and on personal preferences. In this paper we choose the method of classification and regression trees (C&RT), see next subsection for more details.

Assume now that our (total) portfolio is divided into homogeneous subportfolios, i.e., in each subportfolio we have risks with i.i.d. frequencies and i.i.d. severities. Consider an arbitrary homogeneous subportfolio. In order to not complicate the notation, let us denote the total claim amount for that subportfolio by S and apply the same model for all subportfolios. Let there be n policies in that subportfolio and let N_i and Y_i denote the frequency and the total claim amount corresponding to risk (policy) i ($i = 1, \ldots, n$), respectively, and let Z_{ij} denote the claim size corresponding to j-th claim from i-th policy ($j = 1, \ldots, N_i$). Depending on the level of generalization we may omit some indices, i.e. write N instead of N_i, Y instead of Y_i and Z_j or Z instead of Z_{ij} if the particular indices are not relevant.

Then the total claim amount in each subportfolio is given by $S = \sum_{j=1}^{N_*} Z_j$ and the following equalities hold

$$ES = EN_* \cdot EZ, \tag{1}$$

$$VarS = EN_* \cdot VarZ + (EZ)^2 \cdot VarN_*. \tag{2}$$

If we move to the individual policy level, the claim amount for i-th policy is given by

$$Y_i = \sum_{j=1}^{N_i} Z_{ij}$$

and from (1) and (2) the expectation and variance of i-th policy are:

$$EY_i = EN_i EZ, \tag{3}$$

$$VarY_i = EN_i VarZ + (EZ)^2 VarN_i. \tag{4}$$

Since $N_* = \sum_{i=1}^{n} N_i$, we also have $EN_* = nEN$ and $VarN_* = nVarN$, i.e. the expectation and variance of the frequency N_* for a (sub)portfolio are proportionally divided between individual policies.

In conclusion, given that the portfolio can be divided into homogeneous subclasses and for each class we can find the estimate for the total claim amount, the individual pure premium is found as the pure premium for the class divided by the number of policies in that class.

Remark. We recall that in previous discussion we assumed that all the policies have same duration. In practice this is rarely the case, therefore we need to take the durations into account in order to find a more realistic model. This is one of the key aspects we are focusing on in this study.

There are three classical choices for the frequency distribution: binomial, Poisson and negative binomial. While each of them has its benefits, the Poisson model is clearly the most common, because of its applicability and extendibility. It is known that the Poisson model fits well if $EN \approx VarN$, while $EN > VarN$ supports the binomial and $EN < VarN$ the negative binomial. As in practice overdispersion is the more common problem to handle, we are more interested in the latter.

2.2. C&RT clustering

C&RT (classification and regression trees) methodology allows us to produce an "if-then" type set of conditions based on auxiliary variables (one can think of them as risk factors so possible candidates could be variables related to the object insured, like its value, but could also be related to the owner of the insured object, like the data of previous insurance contracts). As with all clustering algorithms the amount and the quality of the auxiliary data determines the clustering result.

Because of the differing length of policies it is easier to think that the subportfolios are more homogeneous if two policies belonging to the same class have as similar claim intensities as possible. The easiest way to produce subclasses is to divide the policies into two classes (leaves) using some auxiliary variable. Then we can proceed with the new subclasses and divide them (using perhaps some other auxiliary variable). Technically it would be possible to continue until each subclass contains only a single policy but usually it is beneficial to stop earlier. In the following we describe the algorithm based on Therneau and Atkinson (1997).

Suppose we have defined the deviance of a model allowing us to compare models pairwise (see Section 3 for precise definition). Let $D(T)$ stand for the deviance of model T. Let us also assume that the number of auxiliary variables available is m.

Then we proceed stepwise. At each step we choose a risk factor X_j, $j \in \{1, \ldots, m\}$ so that when distributing the observations into two subclasses (according to that risk factor) the deviance is decreased the most. Thus there are two choices at each step: choose a subclass and a risk factor with the discrimination rule. At the i-th step there are i candidates for a class and m candidates for a risk factor. The number of possible splitting rules for a particular combination of risk factor and class depends on the set of values that the risk factor has in that class. If the risk factor is continuous and has k distinct values then the amount of possible splits is $k-1$. If the risk factor is categorical and has k distinct values then the amount of possible splits is $2^{k-1}-1$. This also means that categorical risk factors with many levels are likely to be selected because of the large number of possible splits.

More precisely, suppose $i-1$ steps have been completed, $i \in \{1, 2, 3, \ldots\}$ and the current model (tree) is T_{i-1}. This tree has i leaves or, equivalently, the model has i classes. Let $\mathcal{I}_j$ be the set of indices (of the policies) belonging to the j-th class, $j \in \{1, \ldots, i\}$. Suppose that class j has more than one policy and for every risk factor X_k, $k \in \{1, \ldots, m\}$ the possible values are indexed by $\mathcal{J}_{jk}$. Possible splits are denoted by θ_s^{jk} (class j, risk factor X_k and rule s), where

$$s \in \begin{cases} 1, \ldots, 2^{|\mathcal{J}_{jk}|-1} - 1, & \text{if } X_k \text{ is categorical,} \\ 1, \ldots, |\mathcal{J}_{jk}| - 1, & \text{if } X_k \text{ is continuous.} \end{cases}$$

By applying the splitting rule θ_s^{jk} to the current tree T_{i-1}, we get a new tree that has one more leaf. Denote the deviance of the new tree by $D(T_i^{\theta_s^{jk}})$. The goal is to find the splitting rule θ_s^{jk}, which minimizes the deviance $D(T_i^{\theta_s^{jk}})$. We call this (not necessarily unique) split optimal. After finding the optimal split we have completed i steps and we can continue with a tree T_i that has $i+1$ leaves.

Usually we demand that each split would decrease the deviance by at least some fixed amount. If we can't find such a splitting rule then we stop. We can express this idea as the "price of a leaf" and redefine the deviance as

$$D_\alpha(T) = D(T) + \alpha|T|, \tag{5}$$

where $\alpha \geq 0$ is the penalization parameter and $|T|$ is the number of leaves of tree T. In reality this means that "adding an additional leaf to the tree costs α units". Now the goal is to minimize the function (5) for a fixed α over the set of all subtrees[1] of the maximal tree T_∞ . Thus we want to find the subtree with minimal deviance $D_\alpha(T)$. We call this tree optimal. The optimal tree depends on α: if $\alpha = 0$, the addition of leaves is "for free" and the optimal tree will be T_∞. On the other hand when we increase α, the number of leaves on an optimal tree decreases. The optimal tree may not be unique but it is known that if for some α there exist trees T^a and T^b such that both are subtrees of T_∞ and $D_\alpha(T^a) = D_\alpha(T^b)$, then one of them is a subtree of the other and we choose the one with fewer leaves as the optimal one (Breiman et al., 1984). It also holds that when we increase the penalization parameter then the optimal tree will either be the same or a subtree of it. This means that we can always find the optimal tree by pruning the tree T_∞. Cross-validation is used for determining a reasonable value for α. If needed, a minimum limit can be set for subclasses – this makes sure that we will have enough data at the subportfolio level for estimation.

From the practical point of view an important aspect is that C&RT methodology is also able to handle missing data by defining surrogate splits – that is, in addition to the best primary split, every tree node may also be split on one or more other variables with nearly the same results.

3. C&RT/Poisson model with different insurance periods

Let us now extend the classical model to the case where the insurance periods have different lengths. Let n_i denote the number of claims of i-th policy, t_i the corresponding insurance period, and let n_{ij} denote the number of claims of policy i in time unit j (e.g., day or year, depending on model[2]). In that case we can express n_i as follows: $n_i = \sum_{j=1}^{t_i} n_{ij}$.

We define the deviance of the model T as

$$D(T) = \sum_{i=1}^{n} \left(\lambda_{[i]} t_i - n_i \log(\lambda_{[i]} t_i)\right), \tag{6}$$

where $\lambda_{[i]}$ is the average number of claims (in a time unit) in the class containing the i-th policy (as this value minimizes the deviance). Thus we have

$$\lambda_{[i_1]} = \lambda_{[i_2]} = \ldots = \lambda_{[i_n]} = \frac{\sum_{j=1}^{n} n_{i_j}}{\sum_{j=1}^{n} t_{i_j}} \tag{7}$$

if policies i_1, i_2, …, i_n (and no others) belong to the class in question.

Let us assume that we are dealing with a homogeneous subportfolio, i.e. the claim numbers (per time unit) n_{ij} are independent realizations of the random variable N_{ij} (or simply $N_{[i]}$ if the exact time j is not relevant), for all $i = 1, \ldots, n$ and $j = 1, \ldots, t_i$. Then each quantity n_i is an independent realization of the random variable $N_i = \sum_{j=1}^{t_i} N_{ij}$. Applying this model to Formulae (3) and (4) we get the following estimates for the expectation and variance of the claim amount for risk i ($i = 1, \ldots, n$):

$$EY_i = t_i EN_{[i]} EZ, \tag{8}$$

$$VarY_i = t_i EN_{[i]} VarZ + t_i (EZ)^2 VarN_{[i]}. \tag{9}$$

[1] Subtrees of tree T are all those trees that we can form by pruning T, i.e. omitting leaves (and nodes that have turned into leaves).

[2] In mathematical models it is convenient to use as small a time unit as possible, which in practice usually means that daily numbers are considered. For illustrative purposes, though, it is better to use some larger time interval (e.g., a year) to achieve intuitively more understandable quantities.

For the aforementioned three distributions, the following relations between N_{ij} and N_i hold:

- if $N_{ij} \sim Po(\lambda)$, then $N_i \sim Po(\lambda t_i)$;
- if $N_{ij} \sim NBin(\alpha, p)$, then $N_i \sim NBin(\alpha t_i, p)$;
- if $N_{ij} \sim Bin(n, p)$, then $N_i \sim Bin(nt_i, p)$.

Let us now assume that the frequency in time unit N_{ij} in a given (sub)portfolio is Poisson-distributed with parameter λ. Then the frequency N_i (during the insurance period for i-th policy) is Poisson-distributed with parameter λt_i, where t_i is the length of the corresponding insurance period (basically one can think of Poisson process model with intensity λ). Thus the formula for the score function is given by

$$s(\lambda) = \frac{1}{\lambda} \sum_{i=1}^{n} n_i - \sum_{i=1}^{n} t_i. \tag{10}$$

and the maximum likelihood estimate for parameter λ is found from

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} n_i}{\sum_{i=1}^{n} t_i}. \tag{11}$$

Notice that the maximum likelihood estimate is the same that is used for clustering by the C&RT method (see Formula (7)).

4. C&RT/overdispersed Poisson model

There are different options to handle overdispersion. One can choose a certain mixed Poisson model (e.g. negative binomial), apply some regression-type model, or use the so-called overdispersed Poisson model. In the latter case, an actual distribution is not fixed, only the variance-mean relationship is determined through an overdispersion parameter φ by $VarN = \varphi EN$. This allows us to use the classical Poisson model, with the exception that the parameter estimation is done using the so-called quasi-likelihood framework. The properties of the quasi-likelihood function are similar to those of the regular log-likelihood function, the only difference is that the value of the quasi-likelihood function is not a log-likelihood of any actual probability distribution (see e.g. Wedderburn (1974) or McCullagh (1983) for more details). Besides this technical difference (from the perspective of our setup) we can apply the same ideas that we used for Poisson model in previous section.

In general, the score function $s(\lambda)$ for such a construction can be given as follows:

$$s(\lambda) = \sum_{i=1}^{n} \frac{\partial \lambda_i}{\partial \lambda} \frac{n_i - \lambda_i}{\varphi \lambda_i}, \tag{12}$$

where the relation between the individual parameters λ_i and the theoretical global parameter λ is determined by a particular model.

Let us now assume that the parameters λ_i are defined by a general rate parameter λ and by the length of the period t_i as follows: $\lambda_i = \lambda t_i$. Then the formula for the score function is an obvious generalization of Formula (10),

$$s(\lambda) = \sum_{i=1}^{n} t_i \frac{n_i - \lambda t_i}{\varphi \lambda t_i} = \frac{1}{\varphi\lambda} \sum_{i=1}^{n} (n_i - \lambda t_i) \tag{13}$$

and the maximum likelihood estimate for λ is still given by (11). Thus the C&RT methodology is feasible when overdispersion is present, as well. The practical implication is that the cross-validation estimate of the penalization parameter α (see Formula (5)) will typically be larger as we cannot estimate the homogenuous subportfolios as precisely and this causes larger classes in the optimal tree (i.e. tree has fewer leaves).

The estimate for the overdispersion parameter φ is found from

$$\hat{\varphi} = \frac{1}{n-1} \sum_{i=1}^{n} \frac{(n_i - \hat{\lambda}_i)}{\hat{\lambda}_i}, \tag{14}$$

where the estimate of λ_i is assumed to be found using (12) or directly from (11) depending on the setup (see Wedderburn (1974) for more details).

Remark. Small amounts of overdispersion (if $\hat{\varphi}$ is close to one) are usually of little concern. The question of when is $\hat{\varphi}$ so large that an overdispersed model must be used depends on many factors, but clearly one must consider an overdispersed model if $\hat{\varphi} > 2$. See also Hilbe (2007), Cameron, Trivedi (2008) and Tutz (2012) for more details.

In the case of the overdispersed Poisson model, Formulae (8) and (9) for the expectation and variance of the severity of the risk i are the following:

$$EY_i = \lambda t_i EZ, \tag{15}$$

$$VarY_i = \lambda t_i VarZ + \lambda\varphi t_i (EZ)^2 = \lambda t_i (VarZ + \varphi (EZ)^2). \tag{16}$$

Let us now assume that the individual frequencies in our subportfolios are Poisson distributed and apply the overdispersed Poisson model. A natural question is what model does the total claim number for the sum of claims for those two subportfolios follow. The following Lemma shows that if the overdispersion in the frequency data does not influence the distribution of severities (which is quite a natural assumption), the compound overdispersed Poisson model retains the closedness properties of the Poisson model.

Lemma 1. *Let us assume that the random variables $N_1, \ldots, N_k$ follow the overdispersed Poisson model, i.e. $EN_i = \lambda_i$ and $VarN_i = \varphi_i EN_i$, for $i = 1, \ldots, k$. Then the sum $N_* = \sum_{i=1}^{k} N_i$ follows the same model with Poisson parameter $\lambda_* = \sum_{i=1}^{k} \lambda_i$ and overdispersion parameter $\varphi_* = \frac{1}{\lambda_*} \sum_{i=1}^{k} \lambda_i \varphi_i$. In the case when $\varphi_i = \varphi$ for $i = 1, \ldots, k$, we have $\varphi_* = \varphi$.*

The proof is straightforward, details are omitted.

5. Handling overdispersion in a more general framework

Although the C&RT/overdispersed Poisson model proposed in Section 4 is simple and easily applicable, it is not always sufficient to handle the overdispersion, especially in cases where the overdispersion is considerably large (see, e.g., Hilbe, 2007). This motivates us to seek for more possible generalizations of the claim number process to solve the issue of overdispersion. As previously, we still would like the model to retain certain closedness properties the Poisson model has.

5.1. Mixed Poisson model

Assume that the frequency N_i is a Poisson distributed random variable with a parameter Λ_i, which is also a random variable. Assume also that the corresponding expectation is known, $E\Lambda_i = \lambda_i$. Obviously, taking $\Lambda_i \equiv \lambda_i$ reduces to regular Poisson model. The random variable Λ_i is called a mixing variable and we say the frequency N_i follows a certain mixed Poisson distribution. Clearly, a proper mixing can take care of the overdispersion. But our question of main interest is does such a model retain the nice properties that the Poisson model has, i.e. if and under what conditions is the sum of mixed Poisson distributions again a mixed Poisson distribution. A similar question can be asked about the compound process.

To answer these questions, we recall the multiplicative mixed Poisson model as described by Daykin et al. (1994). Here the mixing variable Λ_i is given through two components: the expected value λ_i and a properly scaled mixing value Q_i such that $EQ_i = 1$. Then the conditional distribution of N_i is $(N_i|Q_i = q_i) \sim Po(\lambda_i q_i)$. We can first establish that indeed

$$EN_i = E(E(N_i|Q_i)) = E(\lambda_i Q_i) = \lambda_i \tag{17}$$

and

$$\begin{aligned} VarN_i &= E(Var(N_i|Q_i)) + Var(E(N_i|Q_i)) = E(\lambda_i Q_i) + Var(\lambda_i Q_i) \\ &= \lambda_i + \lambda_i^2 Var(Q_i), \end{aligned} \tag{18}$$

showing that this is a suitable model to handle overdispersion. As previously, the expectation and variance of the severity for the i-th policy are found simply by substituting the outcome of Formulae (17) and (18) into Formulae (3) and (4).

Now, it has been proved that under such a construction the sum of mixed Poisson variables $N_i, i = 1, \ldots, k$ (where k is arbitrary) is a mixed Poisson-distributed if the summands N_i are either mutually independent or depend on each other only through their mixing variables Q_i. The generalization to the compound mixed Poisson case is not that straightforward. It turns out that the sum of independent compound mixed Poisson random variables is not in general a compound mixed Poisson variable. Still, the sum of compound mixed Poisson random variables all having same mixing variable Q, but otherwise independent summands, is again a mixed Poisson random variable with mixing variable Q. For more details see Daykin et al. (1994).

Another problem is interpretability. We can imagine that the data generation algorithm consists of generating the (unobserved) realization of Λ_i and then generating the realization of a Poisson random variable (which is observed) with parameter equal to the (unobserved) realization generated previously. With this algorithm it is of course reasonable to think that the parameters of the mixing distribution are determined by the auxiliary variables. While the realizations of the mixing distribution are unobserved we can follow the idea proposed by Karlis (2005) that employs the EM algorithm. This allows us to calculate the conditional means of the mixing variable given the data which then could be used as data from the mixing distribution.

5.2. Negative binomial model

As already mentioned, the negative binomial model is a valid model if the variance exceeds the expectation and it is one of the classical choices to describe the claim number in a (sub)portfolio. Consider now the aspect of our main interest, i.e. if and when then the sum of negative binomials or compound negative binomials is also a negative binomial or compound negative binomial,

respectively. Unfortunately the negative binomial distribution does not have this property in general. The sum of two independent negative binomial random variables $N_i \sim NBin(\alpha_i, p_i)$, $i = 1, \ldots, k$ (for some k) is again negative binomial if all parameters p_i are equal, $p_i \equiv p$. The same holds if we apply the model with different insurance periods. Still, if we would like to apply this model for the whole portfolio, we should assume that in all subportfolios the frequency N follows a negative binomial distribution with fixed parameter p, which is obviously a very strong restriction.

Since the negative binomial distribution can be represented as a Poisson mixture with a Gamma mixing distribution, all the properties from the previous section carry over to negative binomial distribution as well. More precisely, it is a known fact that mixing the Poisson distribution with mixing variable $\Lambda_i \sim \Gamma(\alpha_i, \frac{\alpha_i}{\lambda_i})$ (or, equivalently, with the "normalized" mixing variable $Q_i \sim \Gamma(\alpha_i, \alpha_i)$) results in a negative binomial random variable, $N_i \sim NBin(\alpha_i, \frac{\lambda_i}{\alpha_i+\lambda_i})$, see, e.g., Johnson et al. (1994). Straightforward calculations (either using Formulae (17) and (18) or the properties of the negative binomial distribution) yield that in this case

$$EN_i = \lambda_i$$

and

$$VarN_i = \lambda_i + \frac{\lambda_i^2}{\alpha_i}.$$

Still, the interpretability of the parameters is an issue for the negative binomial case as well as for all mixed Poisson cases. One way to overcome this problem is to apply the ideas proposed by Karlis (2005) for the negative binomial case. Using this approach we can formulate the following two-step algorithm. In first step we find the conditional means of the variates from the Gamma distribution and in second step we use these as data and divide the portfolio into subportfolios making use of the auxiliary variables.

Recall also that the negative binomial distribution can be formulated as a compound Poisson distribution with the summands having a logarithmic distribution. Such a model is appealing since there are simple explicit maximum likelihood estimates for both Poisson and logarithmic distributions, unfortunately in our setup there is no clear interpretation of the parameters of such a model, especially for C&RT clustering.

5.3. Poisson regression in subportfolios

In case the subportfolios obtained by clustering do not have homogeneous structure or the homogeneous subportfolios turn out to be too small to form a basis of any statistical analysis, one may also apply certain regression models to take care of the heterogeneity. Note that our previous constructions can be seen as regression with an empty (intercept only) model.

Consider the following Poisson regression model in a chosen subportfolio

$$\lambda_i = \exp(\mathbf{x}_i^T \boldsymbol{\beta}), \tag{19}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)$ is the vector of coefficients and $\mathbf{x}_i$ is the i-th row of the auxiliary data matrix. The same formula holds for the overdispersed Poisson model.

Now, the score functions of the two models we are interested in have the following forms (see, e.g., Tutz, 2012, McCullagh et al. 1989):
for the Poisson model we get

$$s(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{x}_i \frac{\partial \lambda_i}{\partial \boldsymbol{\beta}} \frac{n_i - \lambda_i}{\lambda_i} \tag{20}$$

and for the overdispersed Poisson model the formula is

$$s(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{x}_i \frac{\partial \lambda_i}{\partial \boldsymbol{\beta}} \frac{n_i - \lambda_i}{\varphi \lambda_i}. \tag{21}$$

The estimate for the overdispersion parameter φ is found from

$$\hat{\varphi} = \frac{1}{n-k-1} \sum_{i=1}^{n} \frac{n_i - \hat{\lambda}_i}{\hat{\lambda}_i},$$

where the estimate $\hat{\lambda}_i = \exp(\mathbf{x}_i^T \hat{\boldsymbol{\beta}})$ is assumed to be found using (21). See also Tutz (2012) for more details.

In conclusion, the estimate of the severity for the i-th risk (N_i) is given by

$$\widehat{EN}_i = \hat{\lambda}_i$$

in both the Poisson model and the overdispersed Poisson model case. Similarly, the estimate for the variance is

$$\widehat{VarN}_i = \hat{\lambda}_i \text{ or } \widehat{VarN}_i = \hat{\varphi}\hat{\lambda}_i$$

in the Poisson and the overdispersed Poisson cases, respectively.

Remark. Note that assuming the intercept only model, i.e. $k = 0$, we have $\mathbf{x}_i = 1$ and Formula (19) simplifies to $\lambda_i = \exp(\beta_0)$. Taking also $\lambda = \lambda t_i$, it can be seen that Formulae (10) and (13) follow immediately from Formulae (20) and (21).

6. Case study: Estonian casco insurance

The proposed methodology was used for premium calculation in an Estonian insurance company. Different risks like glass breakage risk, traffic accident risk, theft risk and more were considered. The data covered 7 years and several important characteristics about the vehicle like the value, type, make, model and year of manufacture were available. Several characteristics about the owner of the vehicle (including sex, age and more) were also typically available.

We utilized the following simple algorithm for finding the premium through the estimation of claim distributions:

1. Divide the original portfolio into homogeneous subportfolios using C&RT clustering.
2. In each subportfolio estimate the claim frequency:
   (a) choose the suitable model or models to be fitted (Poisson, overdispersed Poisson, mixed Poisson, etc.);
   (b) find the maximum likelihood estimates for the parameters of the chosen model (see Formulae (11), (14));
   (c) if the sample expectation and variance can be considered equal, use the Poisson model (see Section 3), if there is overdispersion apply the overdispersed Poisson model (see Section 4) or more complex models depending on the nature of the particular problem (see Section 5);
3. Estimate the claim severity in each subportfolio:

(a) choose the suitable class of distributions to be fitted;

(b) for each distribution find the maximum likelihood estimates for the parameter(s) and choose the distribution that fits best;

4. Find the pure premium for each policy using Formulae (8) and (9).

5. Apply the risk loading corresponding to the chosen premium calculation principle.

By C&RT clustering the most relevant variables that divided the risks into classes were the value and the type of the vehicle together with the year of manufacture (or age of the vehicle) when dealing with severities. Lognormal, Pareto, Gamma and Weibull distributions were used as theoretical candidates. For traffic accident damage the most suitable distribution was lognormal, for glass damage we got the best results with the lognormal distribution in some classes and with the Gamma distribution in other classes.

When dealing with the claim frequencies, the age and the year of manufacture of a vehicle were the most important variables. Also, the sample mean and variance were of a similar size is most classes (subportfolios) with the variance usually slightly larger, but not large enough to reject the Poisson model.

To give a better illustration of the method used but to keep it brief, we present an arbitrary subbranch of the glass breakage risk model obtained by the C&RT clustering (see Figure 1). Keep in mind that (due to confidentiality of the data) the given subbranch does not include all the clustering conditions, only the last two splits are shown. In other words, the policies assigned to this subbranch are already separated from other policies using some other (and potentially more relevant) criteria based on the auxiliary variables. The chosen subbranch has 3 leaves corresponding to 3 subportfolios, each leaf has the estimated claim intensity (per year) attached to it together with the actual number of claims observed in the respective subportfolio and the total number of policies (with possibly different durations) assigned to that subportfolio. One can see that there is considerable difference between the obtained estimate for intensity and the ratio which does not take the lengths insurance periods into account.

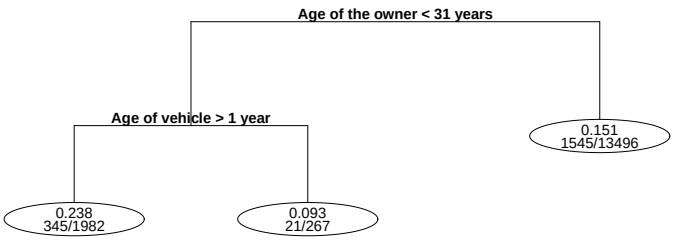

Figure 1: Example of a C&RT (sub)branch

From Figure 1 we can also see that in the given branch the factors determining the clustering criteria are the age of the owner and the age of the vehicle. The classical principle for decision trees is that items meeting the condition are moved to the left and others to the right. Thus in the presented figure the left-most leaf (with a claim intensity of 0.238 claims per year) consists of policies of which the owner is less than 31 years of age and the vehicle is more than one year old.

Let us now study the subportfolio determined by this left-most leaf of the C&RT branch in more detail. The fit of the proposed Poisson distribution to the claim number in this subportfolio can be seen in Figure 2. The bars represent the frequencies of claim numbers in given sample and the dots are the corresponding values from the proposed Poisson distribution.

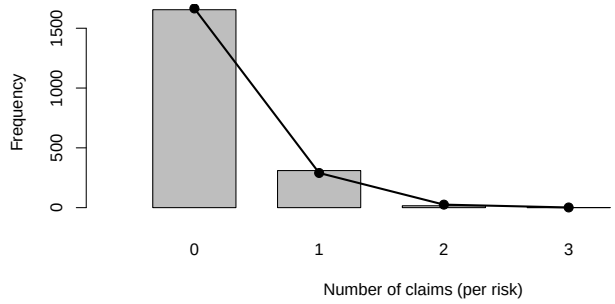

Figure 2: Fitting a Poisson distribution to the number of claims per risk

Fitting the empirical distribution of severity (individual claim size) by the theoretical candidate distributions is shown in Figure 3. From the proposed candidate distributions, the lognormal distribution has the best fit with the sample histogram, Gamma distribution is performing slightly worse. The results of GOF-tests support this conclusion.

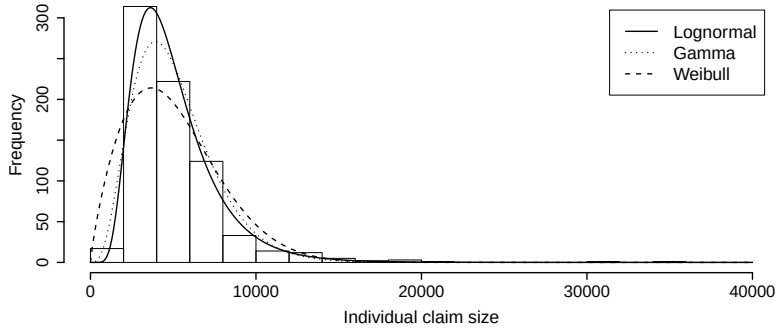

Figure 3: Fitting distributions to individual claim size

When the amount of data is large but the premium calculation should include arriving data as soon as possible (i.e. be dynamic) then it is possible to perform Step 1 of the given algorithm (defining the rules for clustering into subportfolios) less often while the fast operations (recalculating the mean and variance estimates) could be carried out whenever new policies are added to the subportfolios (if necessary).

Most models and calculations in this case study were made using R statistical software packages *actuar* and *rpart* (Dutang et al., 2008, Therneau et al., 2010, R Development Core Team, 2012).

7. Conclusions

The estimation of certain claim distributions and making decisions based on these estimates is an old but pressing topic in actuarial mathematics and raises various different problems and questions. In this study we followed a simple algorithm for finding the premium through the estimation of claim distributions (see Section 6), with the main emphasis on the clustering of the original portfolio into homogeneous subportfolios using the C&RT method, and the estimation of the claim frequency.

While being complicated non-linear models in practice, C&RT models have very good interpretability and are in general very fast to fit which might be crucial with large data-sets. The speed is gained due to the greedy algorithm employed. On the flip-side this means that the solution might not be the global optimum. Minimal assumptions and good prediction accuracy are also strong points of the methodology while a practical problem from an insurance point of view is the fact that the nature of the model means that two neighbouring observations with very close values of the auxiliary variables may have radically different predictions (should they end up in different subportfolios).

The collective risk model is a classical model and provided that the clustering to homogeneous subportfolios is achievable, several problems like the estimation of the total claim amount distribution and the calculation of the pure premium for each risk are easily solved. Here the main attention is obviously on the case when the claim number process is a Poisson process, which has especially nice properties in regard of the sums of distributions and compound distributions belonging to the same class as the summands. These properties hold also in cases with different insurance periods, the generalization being straightforward, the changes in estimation mostly technical. The main problem related to the Poisson model is that it has only one parameter and therefore lacks the dynamics to handle possible overdispersion. There are various alternatives available, but most generalizations either lose some of the key properties of the Poisson model or raise other issues (e.g. the interpretation of parameters in the mixed Poisson case). In this sense the C&RT/overdispersed Poisson is the best choice as it takes into account possible overdispersion while retaining the good properties of Poisson model.

The proposed methodology was applied to a real-life problem from casco insurance and the obtained estimates were reasonably good, especially considering the simplicity of the model used.

Acknowledgements

The work is supported by Estonian Science Foundation Grants No 7313 and No 8802 and by Targeted Financing Project SF0180015s12. The authors also thank the referees for their helpful comments and suggestions.

References

[1] Breiman., L., Friedman, J., Olshen, R., and Stone, C. 1984: *Classification and regression trees*, Belmont, Wadsworth.

[2] Cameron, A.C., Trivedi, P.K., 2008: *Regression Analysis of Count Data*, Cambridge, Cambridge University Press.

[3] Daykin, M., Pentikäinen, T., and Pesonen, M. 1994: *Practical risk theory for actuaries*, London, Chapman & Hall.

[4] Dutang, C., Goulet, V., and Pigeon, M. 2008: actuar: An R Package for Actuarial Science. *Journal of Statistical Software*, **25**(7), 1–37.

[5] Hastie, T., Tibshirani, R., and Friedman, J. H. 2009: *Elements of Statistical Learning: Data Mining, Inference and Prediction (Second Edition)*, New York, Springer-Verlag.

[6] Hilbe, J.M. 2007: *Negative binomial regression*, Cambridge, Cambridge University Press.

[7] Johnson, N.L., Kotz, S., and Balakrishnan, N. 1994: *Continuous Univariate Distributions, Vol. 1 (Second Edition)*, New York, Wiley.

[8] Käärik, M., and Kadarik, H. 2012: Statistical inference with the limited expected value function. In: *Multivariate Statistics with Applications: Proceedings of IX Tartu Conference on Multivariate Statistics & XX International Workshop on Matrices and Statistics.[to appear]*

[9] Käärik, M., and Umbleja, M. 2010: Estimation of claim size distributions in Estonian traffic insurance. In: *Selected Topics in Applied Computing. Proceedings of Applied Computing Conference (ACC '10), Timisoara, Romania.*, 28–32.

[10] Käärik, M., and Umbleja, M. 2011: On Claim Size Fitting and Rough Estimation of Risk Premiums based on Estonian Traffic Insurance Example. *International Journal of Mathematical Models and Methods in Applied Sciences*, **5**(1), 17–24.

[11] Käärik, M., and Žegulova, A. 2012: On estimation of loss distributions and risk measures. *Acta et Commentationes Universitatis Tartuensis de Mathematica*, **16**.*[to appear]*

[12] Kaas, R., Goovaerts, J., Dhaene, J., and Denuit, M. 2009: *Modern Actuarial Risk Theory Using R*, Heidelberg, Springer.

[13] Karlis, D. 2005: EM algorithm for mixed Poisson and other discrete distributions. *Astin bulletin*, **35**(1), 3–24.

[14] Klugman, S., Panjer, H., and Willmot, G. 2004: *Loss Models: From Data to Decisions (Second Edition)*, New York, Wiley.

[15] McCullagh, P., 1983: Quasi-likelihood functions. *The Annals of Statistics*, **11**(1), 59–67.

[16] McCullagh, P., Nelder, J.A. 1989: *Generalized Linear Models (Second Edition)*, London, Chapman & Hall.

[17] R Development Core Team, 2012: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: `http://www.R-project.org`

[18] Therneau, T.M., and Atkinson, E.J. 1997: An introduction to recursive partitioning using the RPART routines. Mayo Foundation. Available at: `http://www.mayo.edu/hsr/techrpt/61.pdf`

[19] Therneau, T.M., Atkinson, E.J., and Ripley, B. 2010: rpart: Recursive Partitioning. Available at: `http://CRAN.R-project.org/package=rpart`

[20] Tutz, G. 2012. *Regression for Categorical Data*, Cambridge, Cambridge University Press.

[21] Wedderburn, R.W.M., 1974: Quasi-likelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika* **61**, 439–447.

ĮMOKŲ VERTINIMAS TAIKANT C&RT/PUASONO MODELĮ IR JO APIBENDRINIMAS

Meelis Käärik, Ants Kaasik

Santrauka. Įmokų vertinimas yra pagrindinis draudimo matematikos objektas. Portfelio žalų sumos vidurkio ir dispersijos įverčiai gali būti pasitelkiami siekiant apskaičiuoti draudimo įmoką. Savo ruožtu draudimo portfelio išskaidymas į homogeninius subportfelius galėtų būti pirmas žingsnis norint įvertinti žalų sumos vidurkį ir dispersiją. Straipsnyje nagrinėjamas žalų intensyvumo parametro vertinimo uždavinys, kuriam spręsti pasiūlomas regresijos medžiais paremtas portfelio padalijimui į homogeninius subportfelius metodas, kai draudimo liudijimų trukmė yra skirtinga ir susiduriama su didelės dispersijos problema. Taip pat pristatoma keletas modelio apibendrinimų. Straipsnio pabaigoje pateikiamas praktinis modelio panaudojimo pavyzdys Estijos transporto priemonių (KASKO) draudimo produktui.

Reikšminiai žodžiai: aktuarinė matematika, kolektyvinės rizikos modelis, įmokų apskaičiavimas, klasifikavimo ir regresijos medžiai (C&RT), didelės dispersijos problema.