“Rate making and large claims”

P. Gigante, L. Picech, L. Sigalotti

Summary

In this paper, we deal with the problem of taking account of large claims in rate making procedure. Both Generalised Linear Models and Extreme Value Theory are applied in order to build a model to evaluate the fair premiums in a tariff system. A numerical application concerning motor insurance is developed.

« Tarification et grands sinistres «

P. Gigante, L. Picech, L. Sigalotti

Résumé

L'objet de cet article est de considérer le problème des "grands" sinistres dans les méthodes de tarification des risques. Aufin de tenir compte de ce genre de sinistres on propose une méthode d'évaluation de la prime pure, basée sur les Modèles Linéaires Généralisés et sur la Théorie des Valeurs Extremes. Les résultats sont appliqués a un portefeuille d'assurance automobile.

"RATE MAKING AND LARGE CLAIMS"

Patrizia Gigante, Liviana Picech
University of Trieste

Luciano Sigalotti
University of Udine

1. INTRODUCTION

It is well-known (Van Eeghen et al. (1983), Czernuszewicz et al. (1998)) that the estimated tariff for a general insurance coverage is highly affected by the presence of large claims in the data. In fact, the occurrence of a particularly high claim in a tariff class may have a dominating effect on the estimation process and distort the pricing analysis. These large claims need therefore being handled, both to reduce their impact on the evaluations and to investigate their possible dependence on the tariff variables.
In the actuarial practice (Czernuszewicz et al. (1998)), it is generally accepted that large claims are funded for by a contribution from all risks in the portfolio. To calculate the amount of this contribution, one can operate as follows:

- select a truncation point or trimming point; it can be defined as the amount over which a claim should be considered as a large claim;
- top-slice all claims at the truncation point;
- fit a tariff model to top-sliced data;
- add on loads to premiums to allow for excess costs over the truncation point.

Two main problems arise: how to select the truncation point and how to calculate the loadings for large claims. In the actuarial practice, there are several approaches for the selection of the truncation point, for instance:

- choose an amount equal to a given rate of the earned premiums or of the total claim amount;
- fix a conveniently high quantile of the empirical claim distribution (e.g. 95%);
- choose as truncation point the reinsurance retention limit of a possible reinsurance treaty.

In the actuarial literature, statistical methods for the truncation point selection have been developed for instance within the context of Credibility Theory (Gisler (1980), Bühlmann et al. (1982)). As for the loading evaluation, in some models it is assumed that large claim occurrence does not depend on the tariff classes, then a constant amount is added to the premiums of all classes. On the other side, one could observe that the frequencies of large claim occurrence differ in the tariff classes (Van Eeghen et al. (1983)). In this case, different loading amounts could be determined in accordance with the evaluation of different probabilities of large claim occurrence in the tariff classes.
In this paper, statistical techniques developed within the Extreme Value Theory (Embrechts, et al. (1997), McNeil, (1997)) are applied as a possible approach to the trimming point selection and in order to estimate the expected claim amount exceeding this point. For each tariff class the premium is then given by the product between the expected claim number and the expected claim amount, where the last one is a mixture of the expected claim amounts below and above the trimming point. The weights of the mixture are the probabilities that one incurred claim be below and, respectively, above the trimming point and are estimated in such a way to allow for large claim dependence on the tariff classes to be taken into account. The expected claim numbers and the expected claim amounts below the trimming point are estimated by the largely used Generalised Linear Models

(McCullagh, Nelder (1989)). As for the weights of the mixture, we discuss some estimation methods: empirical estimates, fitted values throughout regression models and a smoothing model inspired by Credibility Theory.
Numerical examples on a motor insurance data file illustrate the proposed approach. All the numerical evaluations have been performed in S-PLUS.

2. PRELIMINARIES

In a-priori pricing, the observations on a portfolio of risks are used in order to detect and select the a-priori observable characteristics that mainly influence the risk propensity. The aim is to subdivide the portfolio into homogeneous risk classes, the so called tariff classes, and to determine the insurance premium for each class. Under the hypothesis of compound distribution, the pure premium for a risk in tariff class *i*, $E(X_i)$, is given by:

(1) $$E(X_i) = E(N_i)\,E(Z_i)$$

where $E(N_i)$ and $E(Z_i)$ are the expected claim number and the expected claim amount. The tariff can be obtained by estimating, for any *i*, $E(N_i)$ and $E(Z_i)$.

In the actuarial practice, the generalised linear models (GLMs) are largely used (see for instance Brockmann, Wright (1992)) to estimate regression models for the claim number and the claim amount. However, as mentioned in the introduction, before applying this model a pre-processing of the data is necessary to manage the very high claim amounts. In fact, they can have a relevant effect on the estimates of $E(Z_i)$ in the classes in which they have occurred. To show this aspect, we have built a tariff according to (1) keeping all the observed claim amounts as they were reported in the data file.
The data are drawn from a motor insurance portfolio of an Italian company and consist of 172.161 policies observed over one year. For each policy, the following information are available:

- Sex of the insured: 1 for female, 2 for male;
- Age of the insured (grouped into 8 levels);
- Chief town: 1 means that the insured lives in a chief town, 2 otherwise;
- KW Power of the vehicle (grouped into 5 levels);
- Fuel: 1 for petrol supplied vehicles; 2 for diesel cars;
- Mass of the vehicle (grouped into 10 levels);
- Time exposure;
- Number of claims incurred;
- Total claim amount.

In addition, for each policy having reported claims, the claim amounts have been detected. (1)
The Poisson and the Gamma distributions, both with logarithmic link function, have been used to model the claim numbers and the claim amounts, respectively. For the claim numbers, the following variables have been selected: Age, Fuel, Chief town, KW Power and Mass. The estimated regression model includes also the interactions: Chief town and KW Power, Age and Chief town. As for the claim amounts, the selected variables are: Age, Fuel, Chief town and Mass. By applying this tariff to our portfolio of risks, the total earned premiums (42,373 millions ITL) would be slightly lower than the observed total claim amount (42,430 millions ITL). Note that, by using the Gamma distribution, in presence of few but noticeably high claims, the claim amount distribution could be underestimated. Moreover, the pure premiums are considerably affected by large claims, as can be seen in Figure 1.

We note the well-known pattern of the premiums at varying the Age classes: high premiums for young drivers, then decreasing and then again slightly increasing for old drivers. In each age class the premiums show notable fluctuations. Looking, for instance, at the first age class, the fluctuations are due to the presence in our data file of very high claim amounts in some tariff classes having the Mass levels 6 and 8. For this reason we have high premiums when Mass=6, lower when Mass=7, higher again when Mass=8 and lower when Mass=9.

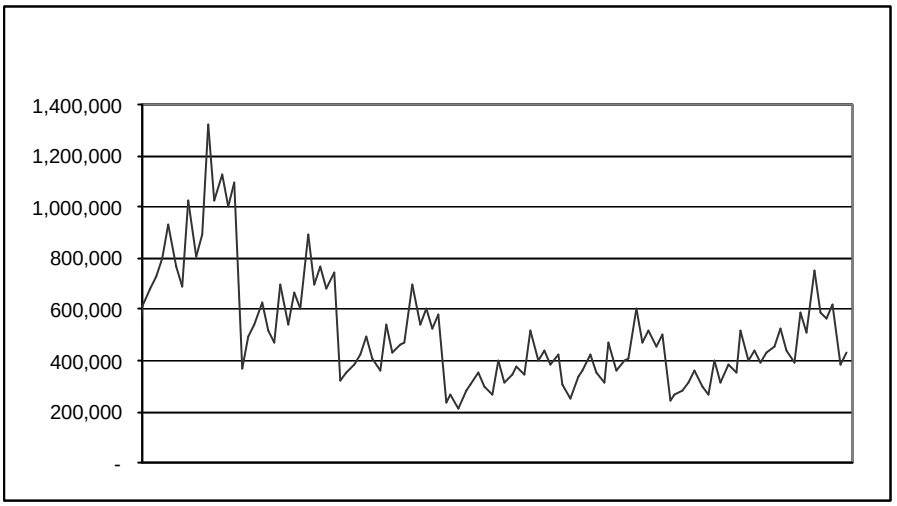

Figure 1: premiums for some tariff classes defined by Sex=1, Chief town=2, Fuel=2, and ordered by Age, KW Power and Mass.

In order to take account of the large claims, it seems natural to assume for the claim distribution a mixture model: one distribution for "ordinary" claims (with amounts not particularly high) and one heavy-tailed distribution for large claims. Obviously, *the trimming point R* distinguishing ordinary and large claims, the respective probability distributions and the weights of the mixture have to be determined, by taking also account of the risk classification. The pure premium can then be described by the following model:

(2) $$E(X_i) = E(N_i)\left[P(Z_i \le R)E(Z_i|Z_i \le R) + P(Z_i > R)E(Z_i|Z_i > R)\right].$$

The expected claim amount is a mixture of two components: the one concerning ordinary claims $E(Z_i|Z_i \le R)$ and the other one for large claims $E(Z_i|Z_i > R)$.

In pricing, mixture models for the claim amount distribution have been applied by many authors. For instance, Bühlmann et al. (1982) afford in this way, within the Credibility Theory, both the problem of choosing the trimming point and the one of evaluating the premium; Benabbou, Partrat (1994) determine, for a given level of the trimming point, maximum likelihood estimates of the two conditional distributions and the weights of the mixture, by assuming the independence of the "large-claim" component from the a-priori tariff characteristics.

In this paper we consider a mixture model for the claim amount distribution and analyse some methodologies for the estimation of the different components in (2).

3. AN APPLICATION OF EXTREME VALUE THEORY TO CLAIM ANALYSIS

In order to choose the trimming point and to estimate the probability distribution of the claim amount exceeding this point we take advantage of some methodologies developed within Extreme Value

Theory (EVT). These methodologies, allow us to obtain an analytical model for the distribution of the claim amount exceeding a threshold u. In this way, for any trimming point $R \geq u$ we can calculate the expected values $E(Z_i | Z_i > R)$ which appear in (2).

The Pickands, Balkema, de Haan Theorem provides a useful result for the estimation of large claim distributions. In fact, it shows that that for any $x \in \Re$, the distribution of a random variable Z belongs to the maximum domain of attraction of a generalised extreme value distribution iff it exists a positive function $s(u)$ such that

$$\lim_{u \to z_0} \sup_{0<z<z_0-u} \left| F_u(z) - G_{x,s(u)}(z) \right| = 0$$

where F_u is the distribution function of the *conditional excesses* $Z - u \,|\, Z > u$,

$$G_{x,s}(z) = \begin{cases} 1 - (1 + x\, z/s)^{-1/x} & x \neq 0 \\ 1 - \exp(-z/s) & x = 0 \end{cases}$$

with $z \geq 0$ if $x \geq 0$ and $0 \leq z \leq -\frac{1}{x}$ if $x < 0$ and z_0 is the right endpoint of the distribution of Z.

The distribution $G_{x,s}(z)$ is named generalised Pareto distribution (GPD). It is characterised by two parameters, the shape ξ and the scale σ.

We just recall that the class of probability distributions belonging to the maximum domain of attraction of a generalised extreme value distribution is remarkably wide and it includes almost all the distributions that are commonly used to model the claim amount distribution (see Embrechts et al. (1997)). From a practical point of view, the quoted theorem suggests that over a sufficiently high threshold u the conditional excesses of a claim amount distribution can be approximated by a generalised Pareto distribution. The parameters x and s can then be estimated, for instance, by the maximum likelihood method, using all the observations exceeding u.

Note that if, for $z > u$, we take $F_u(z)$ equal to the approximating distribution $G_{x,s}(z)$, then $F_u(z-u) = G_{x,s}(z-u) = G_{x,u,s}(z)$ which is a three parameter generalised Pareto distribution.

Given that

$$F(z) = P\{Z \leq z\} = (1 - P\{Z \leq u\}) F_u(z-u) + P\{Z \leq u\}, \quad z > u,$$

if we take $P\{Z \leq u\} = F_n(u)$, where $F_n(u)$ is the empirical distribution function evaluated at u, then the tail of the distribution of Z is estimated by

$$\hat{F}(z) = (1 - F_n(u)) G_{x,u,s}(z) + F_n(u) = G_{x,\tilde{u},\tilde{s}}(z), \quad z > u, \tag{3}$$

where $G_{x,\tilde{u},\tilde{s}}(z)$ is the distribution function of a three parameter GPD with shape parameter ξ, and convenient scale and location parameters $\tilde{s}$ and $\tilde{u}$ (see McNeil (1997)).

We recall two more results concerning GPD. If a random variable Z has the generalised Pareto distribution $G_{x,m,s}(z)$, its expectation is finite iff $x < 1$ and it is given by $E(Z) = m + s/(1-x)$. Moreover, the mean excess function of Z, $e(u) = E(Z - u | Z > u)$, is linear in u and it is given by

$$e(u) = \frac{s + x(u-m)}{1-x}$$

with $s + x(u-m) > 0$ and $u < z_0$.

These results are useful to analyse the fitted model with respect to the data and, preliminarily, to investigate the choice of the threshold u. Indeed this is a crucial point in the application of the EVT methodology. At a general level, we can say that the threshold needs to be sufficiently high in order to fulfil the applicability conditions of the theorem, however it cannot be too high so that an

acceptable number of observations be available to estimate the parameters. Graphical analysis are proposed in literature as tools of investigation when choosing the threshold: e.g. the empirical mean excess function, the pattern of the parameter estimates as a function of the threshold, the plot of the estimated quantiles.
In the following, we are applying these analyses to our data file containing the claim amounts caused by the 172,161 risks, in one year. Some very low amounts have been excluded, so that the data file consists of n=12,662 figures (compared to 12,691 claims). The numerical evaluation have been performed in S-PLUS by means of the Library EVIS (www.math.ethz.ch/~mcneil).
We report in Table1 some statistics that summarise the data characteristics. A remarkable positive asymmetry is shown.

n	12,662
minimum	53,000
1° quartile	800,000
median	1,802,000
average	3,351,000
3° quartile	2,836,000
maximum	504,000,000
$\hat{x}_{0.99}$	27,000,000
$\hat{x}_{0.995}$	46,390,000
$\hat{x}_{0.999}$	183,475,000

Table1: summary statistics on the data file.
$\hat{x}_p$ is the empirical p quantile.

To check whether a heavy tail distribution is suitable to describe our data we can look at the QQ-plot of the quantiles of the empirical distribution against the ones of the exponential distribution:

$$\left\{\left(z_{(k)}, G_{0,1}^{-1}\left(\frac{n-k+1}{n+1}\right)\right); k=1,...,n\right\}$$

where $G_{0,1}^{-1}$ is the inverse of the distribution function of the exponential distribution with parameter 1 and $z_{(1)} \geq K \geq z_{(n)}$ are the ordered claim amounts. In Figure 2, the concave departure from the linear shape shows that the tail of our data is heavier than the one of the exponential distribution.
Another useful graphical tool is the sample mean excess plot:

$$\left\{ \left(z_{(k)}, e_n\left(z_{(k)}\right)\right) \middle| \; k = 1,...,n \right\}$$

where

$$e_n(u) = \frac{1}{card\,\Delta_n(u)} \sum_{k \in \Delta_n(u)} (z_k - u),$$

with $\Delta_n(u) = \{k \mid k = 1, K, n, z_k > u\}$.
The empirical mean excess function $e_n(u)$ is a sample version of the mean excess function $e(u)$. If the points show an upward trend, this is a sign of heavy tail behaviour (see Embrecths et al. (1997), Hogg, Klugman (1984)). In particular, if the pattern of the plot is approximately a straight line with positive slope above a point u, this is an indication that a GPD with $x < 1$ could be a model to describe the data in the area above u and that u can be chosen as threshold.

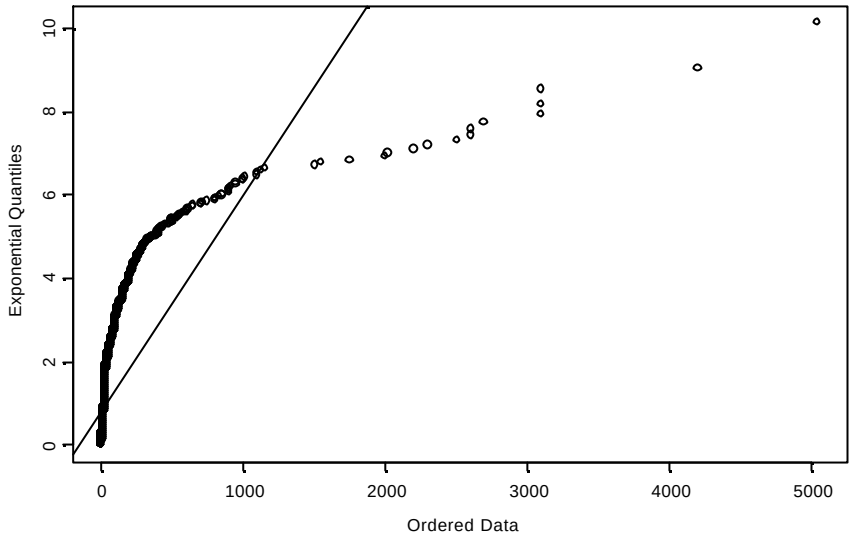

Figure 2: QQ-plot against the exponential distribution.
(Scale on the x-axis: 1=100,000 ITL)

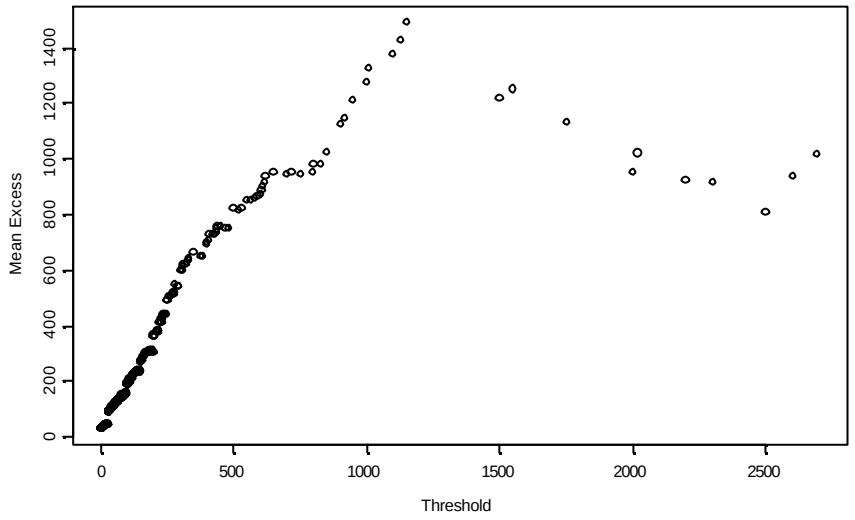

Figure 3: Empirical mean excess function.
(Scale on the x-axis: 1=100,000 ITL)

Looking at Figure 3, excluding the points at the very high levels of u, which are calculated on few data and show an erratic pattern, the plot supports the choice of a generalised Pareto with positive shape parameter in the tail area. However, we cannot clearly single out one threshold level; in fact, different values of u between 200 and 400 could be suitable.
Further investigations on the choice of the threshold are necessary and, following some suggestions in literature (see Embrechts et al. (1997), McNeil (1997)), we estimate the shape parameter ξ of the generalised Pareto distribution for $Z - u | Z > u$ for different values of u.

As a preliminary analysis, we estimate ξ by the Hill estimator. If k, the number of the exceedances above the threshold, is properly chosen also with respect to n, the Hill estimate of the parameter $a = x^{-1}$, with $x < 1$, is:

$$\hat{a}^{(H)} = \hat{a}_{k,n}^{(H)} = \left(\frac{1}{k} \sum_{j=1}^{k} \log z_{(j)} - \log z_{(k)} \right)^{-1}.$$

The graphical analysis based on the Hill estimator is generally summarised in the Hill-plot:

$$\left\{ \left(k, \hat{a}_{k,n}^{(H)} \right) : k = 2, \ldots, n \right\}.$$

Looking at Figure 4, we note that for $u > 300$ the asymptotic confidence intervals of the estimates are quite width and the estimates are not stable. When u is between 250 and 300 the estimates are based on a number of exceedances reasonably high (137 and 102, respectively) and they seem rather stable: a value of u in this interval seems to be a compromise between the bias and the variance of the estimator.

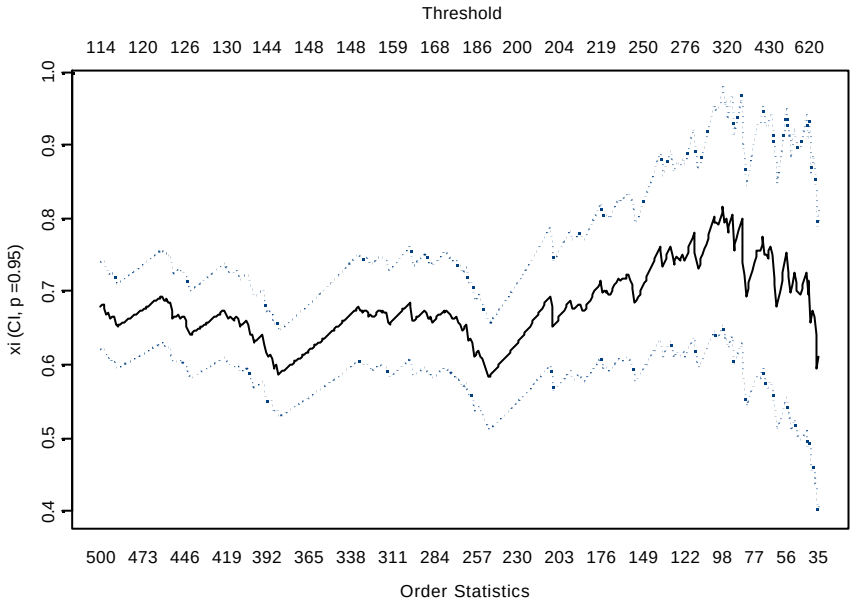

Figure 4: Hill-plot.
(Scale of the threshold on the upper x-axis: 1=100,000 ITL)

We continue the analysis with the maximum likelihood estimates of the shape parameter ξ which are shown in Figure 5. Looking for intervals of stability, such that the choice of a slightly different threshold will not produce a completely different fit of the distribution over the threshold, we note that the pattern fluctuate notably near the thresholds 150 and 200. This might be due to the high number of values all equal to 150 and 200 in our data file, which could correspond to reserved claims. In the interval between 210 and 250 (k=195 and k=137, respectively) the estimates seem to be rather stable. For this reason, the central value 230 (k =164) will be considered in the following.
Another analysis concerns the comparison of the 0.99 empirical quantile, reported in Table 1, with those of the probability distributions estimated by (3) for different values of u. In Figure 6, the quantile of the estimated distributions with threshold between 210 and 250 are rather stable and they are also quite close to the empirical one, while the quantile of the estimated distribution with threshold equal to 300 is remarkably lower than the empirical one. From these considerations, the thresholds 230 and 250 seem convenient.

Now, on the ground of previous analyses, we determine the maximum likelihood estimates of the parameters ξ and σ of three generalised Pareto distributions with thresholds 230, 250 and for the sake of comparison also with the threshold 300.
As an example of the resulting fit, we report in Figure 7 the estimated distribution function $F_u(z-u)$, where u=230. Similar fits are obtained with the other two estimated distributions.

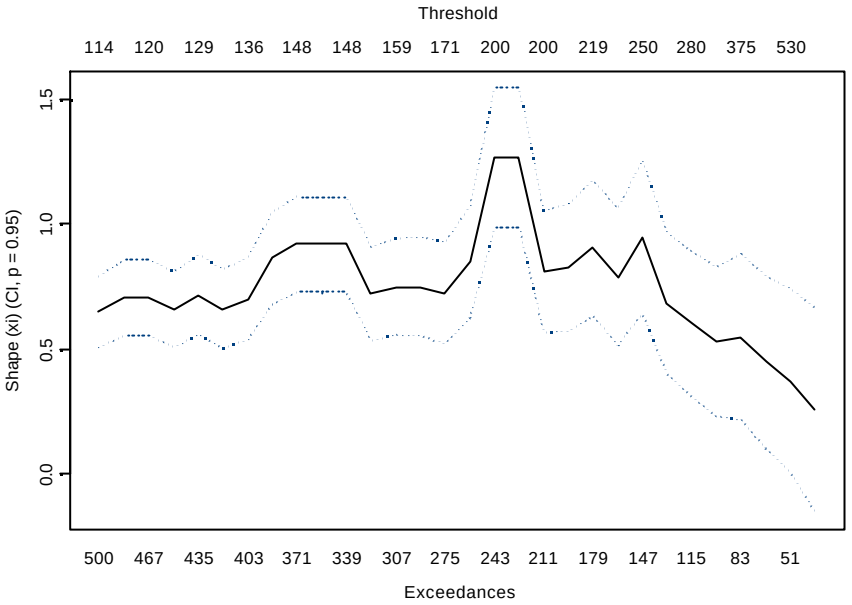

Figure 5: maximum likelihood estimates of the shape parameter.
(Scale of the threshold on the upper x-axis: 1=100,000 ITL)

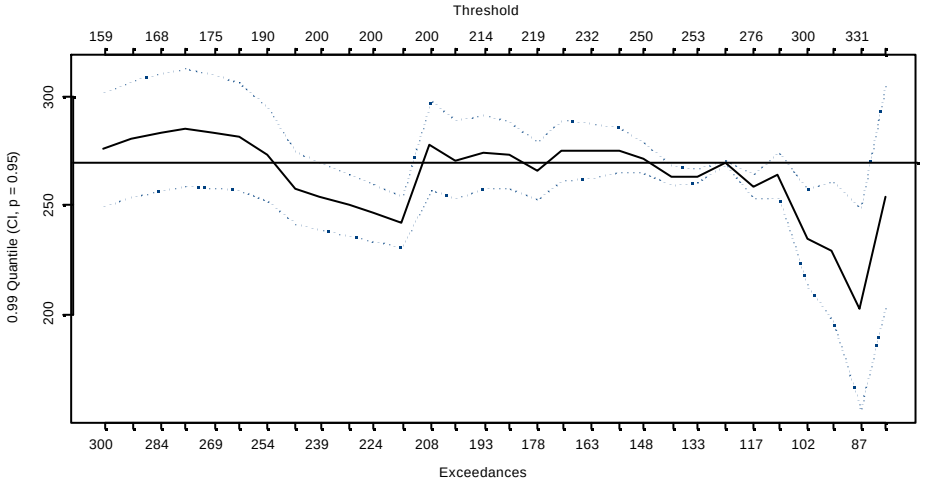

Figure 6: 0.99 quantile estimates.
(Scale of the threshold on the upper x-axis: 1=100,000 ITL)

Table 2 contains the estimates of the shape parameter and some quantiles of the three distributions. Note that the estimated 0.99 and 0.995 quantiles are not so different from the empirical ones. On the contrary, the 0.9999 empirical quantile is considerably higher then the estimated ones. However, the analysis of this quantile is not particularly significant because it is based on very few data.

From the pricing point of view, it is interesting to compare the expected values $E_u(Z-R)_+$, $R \geq u$, at different values of R, with the corresponding sample values (see McNeil (1997), Beirlant et al. (2001)). Here $E_u(Z-R)_+$ denotes the expectation of $\max\{0, Z-R\}$ taken with respect to the GPD estimated from the threshold u. We observe in Table 2 that, for the three considered thresholds u, the expected values are higher then the sample ones when R=300 and R=500. The differences are notably dependent on the threshold u and hence on the related value of the shape parameter. For instance, when u=230 we get the higher estimate of ξ and the most conservative evaluation. In the case R=1500, the empirical value is higher than the one estimated from the threshold u=300 which results from the lowest value of x.

Since the threshold 230 leads to quite conservative evaluations and threshold 250 gives estimates closer to the empirical values, we decide to go on with the analysis choosing these two levels. The threshold 300 seems to be unsuitable for the purpose of a prudential pricing.

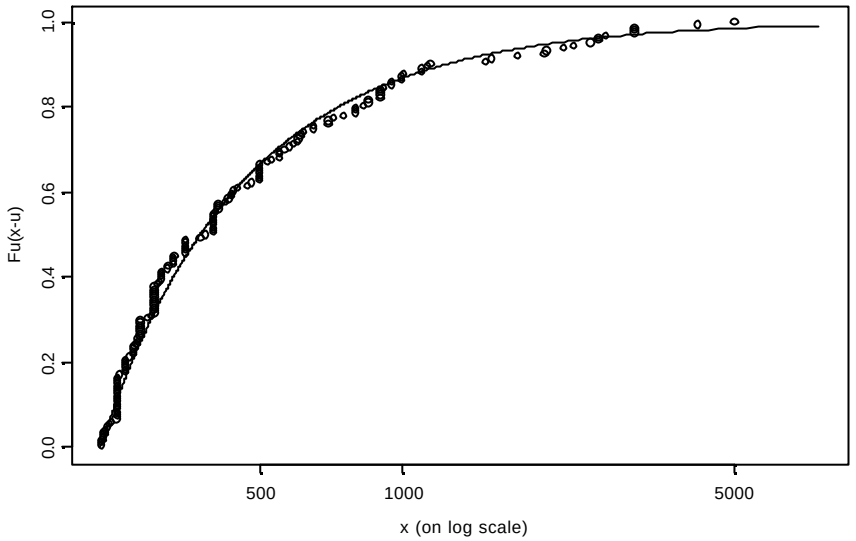

Figure 7: GPD fitted from the threshold 230.

u	k	ξ_u	0,99	0,995	0,9999	$E_u(Z\text{-}300)_+$	$E_u(Z\text{-}500)_+$	$E_u(Z\text{-}1500)_+$
230	164	0,78004	274,7946	450,6125	1506,9663	8,4571	7,2345	5,2460
250	137	0,66695	267,1415	463,9284	1487,4891	6,3990	5,1513	3,1019
300	102	0,52327	234,8622	472,6039	1505,2903	5,3839	4,1565	1,9966
Empirical values			270	463,9	1834,75	4,7957	3,5732	3,0409

Table2: estimated ξ, percentiles, expected values of the excesses, for different threshold u.

4. ESTIMATES OF THE WEIGHTS OF THE MIXTURE

In this paragraph we deal with the problem of estimating, for each tariff class i, the probabilities $P(Z_i > R)$ that a claim amount exceeds a chosen trimming point R.

If we assume that these probabilities do not depend on the tariff characteristics, a common estimate for all the tariff classes could arise e.g. from a balance condition on the portfolio or, alternatively, from the incidence on all claims of those exceeding the trimming point (the observed frequency of the exceedances).

On the contrary, we can carefully investigate on the dependence of the probabilities $P(Z_i > R)$ on the tariff characteristics in order to take properly account of the incidence of large claims in the tariff model. This can be done, for instance, by the GLM methodology.

Let $Z^{(j)}$ be the claim amount of the *j*-th claim in the portfolio and consider as response variable the indicator $\left|Z^{(j)} > R\right|$, $j = 1, 2, \ldots$. We can assume for these variables the Binomial distribution and, after choosing a suitable link function, a selection procedure leads to identify the significant tariff variables. When two or more tariff variables are selected, we can take the probabilities estimated by the GLM technique. When only one factorial variable is selected, the GLM estimates are just the observed frequencies in the tariff classes described by this unique tariff variable; hence no smoothing effect is produced.

Estimating the probabilities $P(Z_i > R)$ by the observed frequencies, could not be an advisable solution, in particular in those classes with very few data. An alternative solution is given by a model suggested by the Credibility Theory, which allows to take account of the specific experience of each tariff class but, on the other side, produces a smoothing effect due to the whole experience all over the portfolio.

Let $X_{ij} = \left|Z_i^{(j)} > R\right|$, with $Z_i^{(j)}$ the claim amount of the *j*-th claim in tariff class *i*, *i* =1,…,*s*, *s* the number of tariff classes.

We take the following hypothesis for the tariff class *i*. For the process $X_{i1}, X_{i2}, \mathrm{K}$ assume that:

- the probability distribution of the process depends on a random parameter Θ_i;
- conditioned to Θ_i, the random variables $X_{i1}, X_{i2}, \mathrm{K}$ are i.i.d.

For the different classes we assume that

- the processes $\left(\Theta_i, X_{i1}, X_{i2}, \mathrm{K}\ , X_{in_i}\right)$, with n_i the observed number of claims in tariff class *i*, *i* =1,…,*s*, are independent;
- $\Theta_1, \ldots, \Theta_s$ are identically distributed;
- for any *i*, the variables $X_{ih} \mid \Theta_i = ?$ are identically distributed. We denote

$$m = E\left[E\left(X_{ih} \mid \Theta_i\right)\right],\ v = E\left[\mathrm{var}\left(X_{ih} \mid \Theta_i\right)\right] \text{ and } a = \mathrm{var}\left[E\left(X_{ih} \mid \Theta_i\right)\right].$$

The model assumptions imply that, in our prior judgement, the probabilities that a claim amount would exceed the trimming point are the same for all tariff classes. The observed frequencies of claim amounts exceeding the trimming point, in the different tariff classes, allow us to update the prior estimates by taking account of the experience in each class.

Using a linear credibility formula (Bühlmann (1967)) the probability estimate p_i of $P(Z_i > R)$ is:

$$p_i = (1 - a_i)m + a_i \bar{x}_i$$

where

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij} \text{ , with } x_{ij} \text{ the observed value of } X_{ij}$$

$$a_i = \frac{n_i}{n_i + k} \text{ , with } k = \frac{v}{a}.$$

Setting

$$\overline{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij} \text{ and } m = \sum_{i=1}^{s} n_i ,$$

we consider the following estimators for the parameters *m*, *v*, and *a*, respectively (see e.g. Klugman, Panjer, Willmot (1998)),

$$\overline{X} = \frac{1}{m}\sum_{i=1}^{s} n_i \overline{X}_i ,$$

$$V = \frac{\sum_{i=1}^{s}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_i)^2}{\sum_{i=1}^{s}(n_i - 1)} ,$$

$$A = \left(m - \frac{1}{m}\sum_{i=1}^{s} n_i^2 \right)^{-1} \left[\sum_{i=1}^{s} n_i (\overline{X}_i - \overline{X})^2 - V(s-1) \right].$$

The estimates p_i are evaluated by adjusting the estimate of *m*, obtained from data of all the collective, to reflect the experience on each tariff class.
By means of GLM, we have applied a selection procedure to our data file both for the thresholds 23 and 25 millions. In both cases we have selected only one tariff variable, Mass and KW Power, respectively. The probabilities estimated using the above described model are reported in Table 3.

Mass	*u*=230	μ=0.012923	KW Power	*u*=250	μ=0.010795
i	$\overline{x}_i$	p_i	*i*	$\overline{x}_i$	p_i
1	0.007329	0.009802	1	0.004505	0.007184
2	0.005908	0.010044	2	0.011241	0.011195
3	0.014885	0.014354	3	0.007530	0.008485
4	0.014037	0.013598	4	0.014191	0.013239
5	0.014085	0.013673	5	0.027778	0.013841
6	0.009751	0.011378			
7	0.008209	0.010192			
8	0.023018	0.018441			
9	0.011070	0.011946			
10	0.018667	0.014521			

Table 3: observed frequencies and probability estimates.

5. SOME NUMERICAL APPLICATIONS TO MOTOR INSURANCE PRICING

In this paragraph, we report some numerical applications of the proposed approach to the pricing of motor vehicle insurance. The data are those described in paragraph 2 and already used in the previous evaluations.
We are interested into applying the rating model (2) in which the amounts of large claims and their incidence in the tariff classes are taken into account explicitly. We are going to discuss the effects on these premiums, denoted by P2, of different levels of the threshold and of the trimming point; in details, the three selected combinations are shown in Table 4. We have assumed that the expected values $E(Z_i | Z_i > R)$ in (2), do not depend on the tariff variables. Even though this hypothesis can appear quite restrictive, it can be accepted in force of some analysis developed by means of GLMs.

In fact, in our data, the probability distribution of large claims shows a weak dependence on at most one tariff variable.

Being x_u and s_u the estimated parameters of the GPD from the threshold u, we have:

$$E_u(Z_i | Z_i > R) = \frac{s_u + x_u(R-u)}{1-x_u} + R$$

since in our case $x_u < 1$.

The results are reported in Table 4.

u	R	ξ_u	σ_u	$E_u(Z_i \mid Z_i > R)$
230	230	0.7800395	156.2871	940.52348
250	250	0.6669510	211.8857	886.19978
250	500	0.6669510	211.8857	1636.84007

Table 4: parameter estimates and expected claims over the trimming points R. (The amounts u, R and $E_u(Z_i | Z_i > R)$ are expressed in ITL divided by 100,000).

The component "ordinary" claim, $E(Z_i | Z_i \le R)$ in (2), has been estimated by a GLM with Gamma distribution function and logarithmic link function. By applying a selection procedure, the tariff variables Sex, Age and Mass have been selected.

Note that, whereas the Gamma distribution has an upper unlimited support, the response variable $Z_i | Z_i \le u$ is limited with values in the interval $[0,u]$. For this reason, one could assign a proper link function, defined on $[0,u]$ and having values in $\Re$. However, since the probability assigned by the Gamma on the right tail can be considered negligible, supported by the numerical results, we feel confident that the Gamma distribution and the logarithmic link function are reasonable assumptions.

As for the weights of the mixture in (2), we have applied the GLM methodology to investigate on the possible dependence of the incidence of "large" claims on the tariff characteristics. For this purpose, as mentioned in paragraph 4, we have assumed a Binomial distribution for the response variable and the logit link function. The model selection is clearly affected by the threshold level. If it is fixed at 23 millions, both Mass and Chief town are selected at a 4% significance level. If the significance level is dropped to 2%, only Mass is selected. Remember that in this case the fitted values are equal to the observed frequencies so that the credibility model, described in paragraph 4, seems to be much more suitable. For this reason, when we consider the threshold 23 millions, we have three different choices for the weights of the mixture: a two-variable GLM regression model (to which we will refer to as "glm"); the observed frequencies ("f") and the credibility weights ("c").

If we fix the threshold at 25 millions, only the KW Power is selected and in this case we can take as weights in the mixture, either the frequencies or the credibility weights.

When the trimming point is 50 millions, since only few claim amounts exceed this value, a statistical selection process on the tariff variables would not give significant results. For this reason, one can assume that the occurrence of one large claim does not depend on the tariff characteristics. In this case, we can assign as weight in the mixture, for instance, the observed frequency of claims over 50 millions (we will refer to this case as "constant"). Another possibility is to make, in any case, the weights depend on some tariff characteristics; since the KW Power is selected when the trimming point is 25 millions, one can, e.g. decide to assign the credibility weights varying with the KW Power class of the vehicle.

We report in Table 5 the total earned premiums evaluated for our portfolio of risks by applying the above mentioned premium models. P1 denotes the premiums calculated according to (1) in

paragraph 2. The total premiums are subdivided into the two components: the one financing "ordinary" claims and the other financing "large" claims.

u	*R*	Premium model	"ordinary" claims	"large" claims	Total earned premiums
		P1			42,372,634,500
230	230	P2 glm	31,646,473,878	15,438,929,545	47,085,403,424
230	230	P2 f	31,647,140,256	15,424,495,338	47,071,635,595
230	230	P2 c	31,644,823,102	15,541,277,506	47,186,100,608
250	250	P2 f	32,306,771,311	12,140,937,067	44,447,708,378
250	250	P2 c	32,307,856,391	12,143,178,477	44,451,034,869
250	500	P2 c	35,150,303,036	8,865,352,677	44,015,655,713
250	500	P2 constant	35,149,588,016	9,002,620,480	44,152,208,496
		Total observed claim amount			42,429,570,073

Table 5: earned premiums.

We can see that the global effect of different choices of the weights in the mixture model (2) is moderate, much more important is the probability distribution of the conditioned excesses. In fact, when we set the threshold at 230, as already remarked in paragraph 3, we get the expected total claim amount considerably overestimated, with respect to the total observed claim amount. If the threshold is 250 the overestimation is more reasonable. Anyway, we have reported the premiums evaluated with the threshold u=230, to enlighten how different estimates of the weights do not produce, on the whole, substantially different results.

Looking at the premiums obtained with the threshold u=250, we can appreciate the effect of different choices of the trimming point R. In fact, if R=250 the "large" claim component amounts to about the 27% of whole premiums, whereas if R=500 the incidence of this component decreases to about 20%. Note that the evaluations in the last two columns refer to policies with unlimited liability. However, in practice, we often face policies having limited liability. In this case the large claim component could be considerably reduced.

To illustrate the different effects of the proposed models on the premium evaluations in the different classes, we report some graphs.

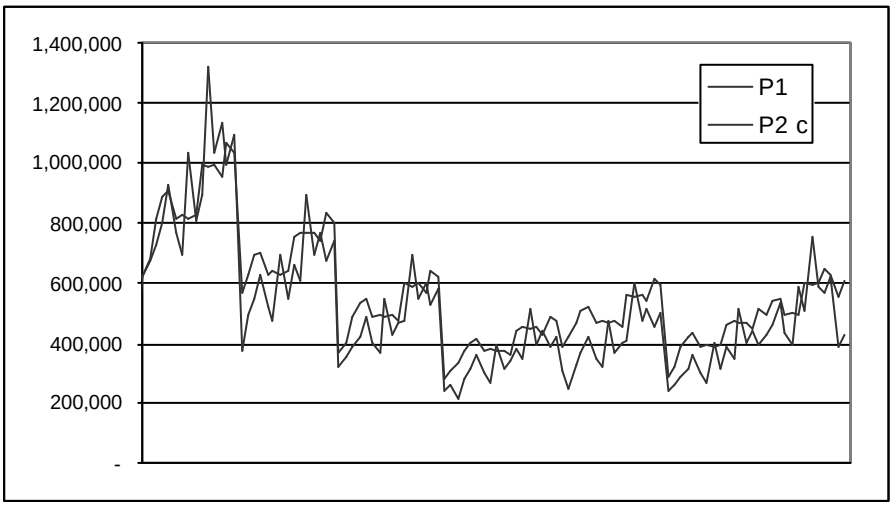

Figure 8: premiums P1 and P2 c, for some tariff classes defined by Sex=1, Chief town=2, Fuel=2, and ordered by Age, KW Power and Mass.

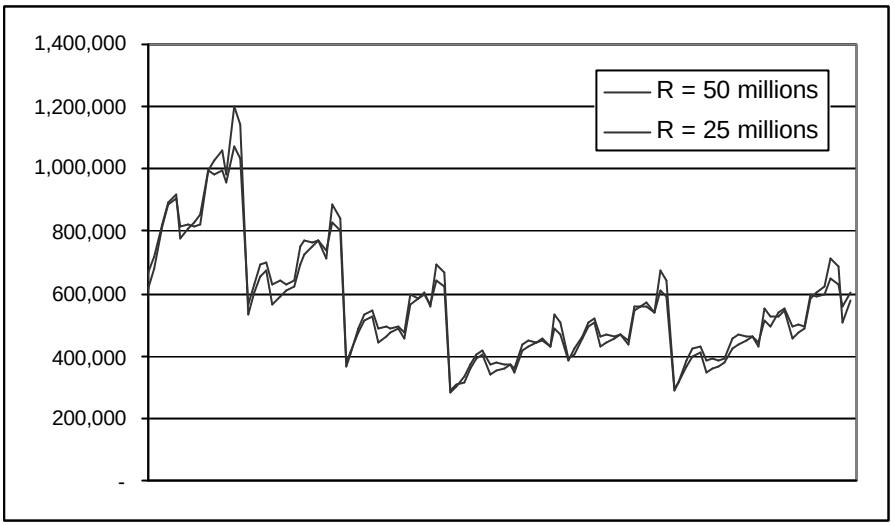

Figure 9: premiums P2 c, trimming points R = 25 millions and R = 50 millions. Tariff classes defined by Sex=1, Chief town=2, Fuel=2, and ordered by Age, KW Power and Mass.

In Figure 8, we note thate the premiums P1 show much more notable fluctuations then P2. As we would have expected, premiums P2, with credibility weights, show a smoother pattern. Taking the observed frequencies, instead of the credibility weights, the results are quite similar: in some tariff classes the smoothing effect could be slightly reduced.
As shown in Figure 9, if the trimming point is raised from R=25 millions to R=50 millions, the premiums show again some fluctuations and this is due to the fact that in the component "ordinary" claims we have again some high claim amounts. The fluctuations are, generally, less pronounced than those of premium P1. If the weights of the mixture are taken constant (P2 constant) and R=50 millions, the pattern of the premiums is very similar to that one given by P2 c.
The results of the evaluations and the reported graphs show that setting the trimming point equal to the threshold (in our example 25 millions) fulfils the aim of building a tariff in which the smoothing reduces the impact of large claims conveniently. This suggests that the EVT methodology could be effectively applied not only to estimate the tail of the loss distribution but also to choose the trimming point for rate making purposes.
As closing remarks, we would like to point out that the integrated use of sound statistical tools such as EVT, GLM and Credibility, allows to take conveniently account of large claims, by considering both their amounts and the influence of the tariff variables on their occurrence. In this way we can achieve a-priori tariff models that provide careful evaluations but at the same time show features of flexibility which make them appropriate for a practical use.

NOTES:

* This research work was partially supported by Regione Autonoma Friuli-Venezia Giulia (research project: Modelli matematici innovativi per lo studio dei rischi finanziari e assicurativi).

(1) The data file has been prepared by Mariella Rossi, who also made some explorative analysis while working on her Thesis in Actuarial Statistics "Tariffazione R.C.A.: sinistri eccezionali e classi tariffarie", A.A. 1999-2000.

REFERENCES

BEIRLANT J., MATTHYS G., DIERCKX G. (2001), Heavy-tailed distributions and rating, *ASTIN Bulletin* **31**, 37-58.

BENABBOU Z., PARTRAT C. (1994), "Grands" sinistres et lois mélanges, *Transactions of the XXV ASTIN Colloquium*, Cannes.

BROCKMAN M.J., WIGHT T.S. (1992), Statistical motor rating: making effective use of your data, *Journal of the Institute of Actuaries* **119**, 457-543.

BÜHLMANN H. (1967), Experience rating and credibility, *ASTIN Bulletin* **4**, 199-207.

BÜHLMANN H., GISLER A., JEWELL W.S. (1982), Excess claims and data trimming in the context of credibility rating procedures, *Mitteilungen Vereinigung Schweizerischer Versicherungs-mathematiker*, 117-147.

CZERNUSZEWICZ A., COUPER A., ORR J., SEAGO W., UPSON J., HOWARD N., MITCHLL N., MC PEARSON J. (1998), Reserving and Pricing for Large Claims, *General Insurance Convention & ASTIN Colloquium, Glasgow, Scotland.*

EMBRECHTS P., KLÜPPELBERG C., MIKOSCH T. (1997), *Modelling extremal events for insurance and finance*, Springer Verlag, Berlin.

GISLER A. (1980), Optimum Trimming of Data in the Credibility Model, *Mitteilungen der Vereininigung Schweizerischer Versicherungsmathematiker*, Vol. 80, N. 3, pp 313-325.

HOGG R.V., KLUGMAN S.A. (1984), *Loss distributions*, Wiley Series in Probability and Mathematical Statistics, New York.

KLUGMAN S.A., PANJER H.H., WILLMOT G.E. (1998), *Loss models. From data to decisions*, Wiley Series in Probability and Mathematical Statistics, New York.

MCCULLAGH P., NELDER J.A., (1989), *Generalized linear models*, Chapman and Hall, New York.

MCNEIL A. (1997), Estimating the tails of the loss severity distributions using extreme value theory, *ASTIN Bulletin* **27**, 117-137.

VAN EEGHEN J., GREUP E.K., NIJSSEN J.A. (1983), *Rate Making*, Survey of Actuarial Studies, No. 2, Nationale-Nederlanden N.V., The Netherlands.