RATE MAKING AND LARGE CLAIMS

Patrizia Gigante
Liviana Picech
University of Trieste (Italy)

Luciano Sigalotti
University of Udine (Italy)

XXXIII INTERNATIONAL ASTIN COLLOQUIUM
Cancún, 21st March 2002

LARGE CLAIMS

- Have a strong influence on the estimated tariff:
  - dominating effect on the estimation process;
  - distortion in the pricing analysis.

- How to handle large claims?
  - to reduce their impact on the evaluations;
  - to investigate their dependence on tariff characteristics.

<u>The actuarial approach:</u>

1. select a truncation point or *trimming point* (i.e. the amount over which a claim should be considered as a large claim);
2. top-slice all claims at the truncation point;
3. fit a tariff model to top-sliced data;
4. add on loads to premiums to allow for excess costs over the truncation point.

Two main problems:

- how to select the trimming point;
- how to calculate the loading for large claims.

Some contributions in the actuarial literature:

- Some statistical methods have been developed within the context of *Credibility Theory* (Gisler (1980), Bühlmann et al. (1982));
- Benabbou, Partrat (1994): assume a *mixture model for the claim amount distribution* and, given the trimming point, determine maximum likelihood estimates of the two conditional distributions and the weights of the mixture; the "large-claim" component is assumed independent of the tariff characteristics.

In this paper, we apply the statistical techniques developed within the *Extreme Value Theory* (Embrechts, et al. (1997), McNeil, (1997), …):

- as a possible approach to the trimming point selection;
- to estimate the expected claim amount exceeding the trimming point.

Assumption: mixture model for the claim amount distribution.

A MIXTURE MODEL FOR THE RISK PREMIUM

Let $E(X_i)$ be the risk premium for a policy in tariff class *i*:

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

where: N_i is the claim number

Z_i is the claim amount

R is the trimming point

- The expected claim amount is a mixture of:
  - $E(Z_i | Z_i \le R)$ ("ordinary" claims)
  - $E(Z_i | Z_i > R)$ ("large" claims).

Questions:

- could EVT be usefully applied to choose the trimming point *R*?
- how could we estimate the claim distributions for "ordinary" and "large" claims?
- how could we estimate the weights of the mixture?

A NUMERICAL EXAMPLE

172.161 policies observed over one year (data from a motor insurance portfolio of an Italian insurance company).

Information:

- Sex of the insured: 1 for female, 2 for male;
- Age of the insured (grouped into 8 levels);
- Chief town: 1 if 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;
- Claim amounts.

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

✓ $E(N_i)$ ESTIMATION:

- GLM with Poisson distribution and logarithmic link function
- Selected tariff variables:

  Age, Fuel, Chief town, KW Power and Mass.

- Interactions:

  Chief town and KW Power,
  Age and Chief town.

AN APPLICATION OF EXTREME VALUE THEORY TO CLAIM ANALYSIS

- ✓ Choice of the trimming point R
- ✓ $E(Z_i | Z_i > R)$ estimation

We will obtain a model for the distribution of the claim amount exceeding a *threshold u*

$\Rightarrow$ $E(Z_i | Z_i > R)$ can be calculated $\forall\, R \geq u$

Software: S-PLUS and Library EVIS by A. McNeil (www.math.ethz.ch/~mcneil/software.html)

n=12,662 claims

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
$\widehat{x}_{0.99}$	27,000,000
$\widehat{x}_{0.995}$	46,390,000
$\widehat{x}_{0.999}$	183,475,000

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

Let $$H_\xi(z) = \begin{cases} \exp\left\{-(1+\xi\ z)^{-1/\xi}\right\} & \xi \neq 0, \quad 1+\xi\ z > 0 \\ \exp\left\{-e^{-z}\right\} & \xi = 0 \end{cases}$$

a *Generalised Extreme Value* (GEV) distribution.

<u>Pickands, Balkema, de Haan Theorem</u>

For any $\xi \in \Re$, the distribution of a random variable *Z* belongs to the maximum domain of attraction of a GEV distribution H_ξ iff it exists a positive function $\sigma(u)$ such that

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

where: z_0 is the right endpoint of the distribution of *Z*;

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

$$G_{\xi,\sigma}(z) = \begin{cases} 1-(1+\xi\, z/\sigma)^{-1/\xi} & \xi \neq 0 \\ 1-\exp(-z/\sigma) & \xi = 0 \end{cases} \qquad \text{with:} \qquad \begin{array}{ll} z \geq 0 & \text{if } \xi \geq 0 \\ 0 \leq z \leq -\dfrac{1}{\xi} & \text{if } \xi < 0 \end{array}$$

is a *Generalised Pareto Distribution (GPD)*: ξ is the shape parameter,
σ is the scale parameter.

From a practical point of view:

- many probability distributions belong to the maximum domain of attraction of a GEV distribution: e.g. distribution models of the claim amounts (Embrechts et al. (1997));

  $\Rightarrow \quad Z-u \,|\, Z>u$ can be approximated by a GPD, over a sufficiently high threshold *u*.

- The parameters ξ and σ can then be estimated, for instance, by the maximum likelihood method, using all the observations exceeding *u*.

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

  if we take

  $P\{Z\le u\}=F_n(u)$, the empirical distribution function

  $F_u(z)$ equal to the approximating distribution $G_{\xi,\sigma}(z) \quad \Rightarrow \quad F_u(z-u)=G_{\xi,\sigma}(z-u)$

  $\Rightarrow \quad \hat{F}(z)=(1-F_n(u))G_{\xi,u,\sigma}(z)+F_n(u)=G_{\xi,\tilde{\mu},\tilde{\sigma}}(z), \quad z>u,$ (McNeil (1997))

  where: $G_{\xi,\mu,\sigma}(z)=G_{\xi,\sigma}(z-\mu)$ (*three-parameter GPD*) and

$$\tilde{\sigma}=\sigma(1-F_n(u))^{\xi} \qquad \tilde{\mu}=\begin{cases} u-\dfrac{\sigma(1-F_n(u))^{\xi}}{\xi}\left[(1-F_n(u))^{-\xi}-1\right] & \xi\neq 0 \\ u+\sigma\log(1-F_n(u)) & \xi=0 \end{cases}.$$

❖ Are our data heavy-tailed?

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

$G_{0,1}^{-1}$ inverse of the exponential distribution function

$z_{(1)} \geq \ldots \geq z_{(n)}$ ordered claim amounts.

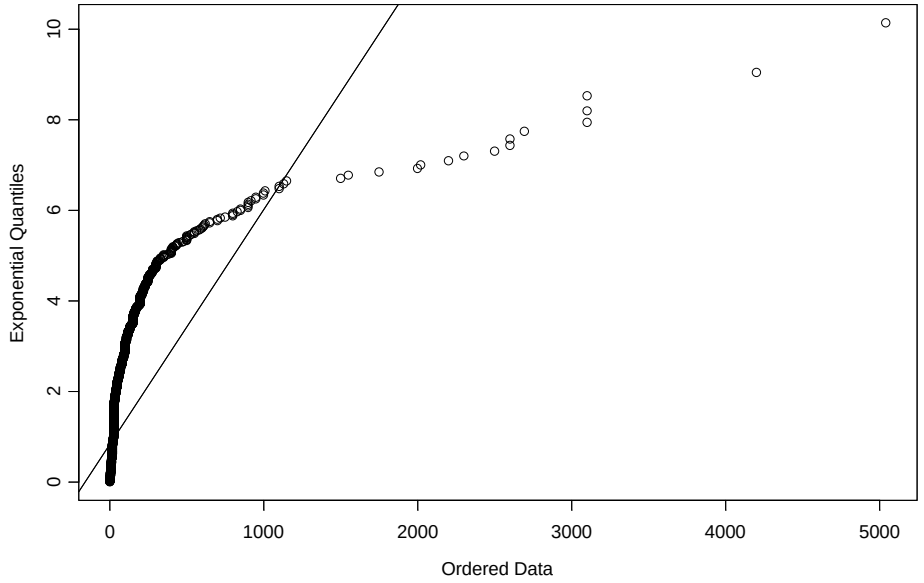

Figure 2: QQ-plot against the exponential distribution.
(Scale on the x-axis: 1=100,000 ITL)

❖ How can we choose the threshold u?

Let

$$e(u) = E\left(Z - u \middle| Z > u\right)$$ the *mean excess function of Z*

sample mean excess plot: $$\left\{ \left(z_{(k)}, e_{\mathrm{n}}\left(z_{(k)}\right)\right) \middle|\ k = 1, \ldots, \mathrm{n} \right\}$$

where

$$e_n(u) = \frac{1}{n\left[1 - F_n(u)\right]} \sum_{z_k > u} \left(z_k - u\right)$$ *empirical mean excess function*

⇒ Heavy tail behaviour, if the points show an upward trend, (see Embrecths et al. (1997), Hogg, Klugman (1984)).

⇒ GPD with $\xi < 1$ could be a model to describe the data in the area above *u*, if the pattern of the plot is approximately a straight line with positive slope above *u*, (*u* can be chosen as threshold).

In fact, if $Z \sim G_{\xi,\mu,\sigma}(z)$ $\Rightarrow$ $e(u) = \dfrac{\sigma + \xi(u - \mu)}{1 - \xi}$ linear in *u*,

with $\sigma + \xi(u - \mu) > 0$ and $u < z_0$.

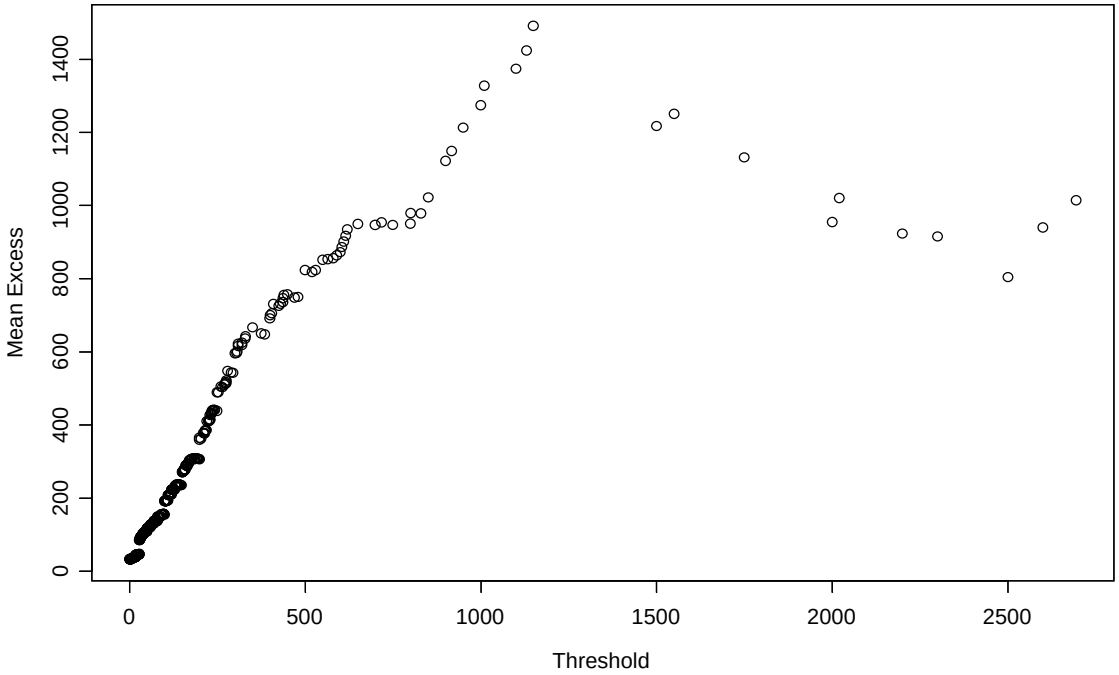

Figure 3: Empirical mean excess function.
(Scale on the x-axis: 1=100,000 ITL)

Further investigations:

- Estimates of the shape parameter ξ of the GPD for $Z-u|Z>u$, for different values of u.

➢ Hill-plot: $\left\{\left(k, \hat{\alpha}_{k,n}^{(H)}\right): k = 2,\ldots, n\right\}$

where $\hat{\alpha}^{(H)} = \hat{\alpha}_{k,n}^{(H)} = \left(\frac{1}{k} \sum_{j=1}^{k} \log z_{(j)} - \log z_{(k)} \right)^{-1}$ and $\alpha = \xi^{-1}$, with $\xi < 1$, $u < z_0$.

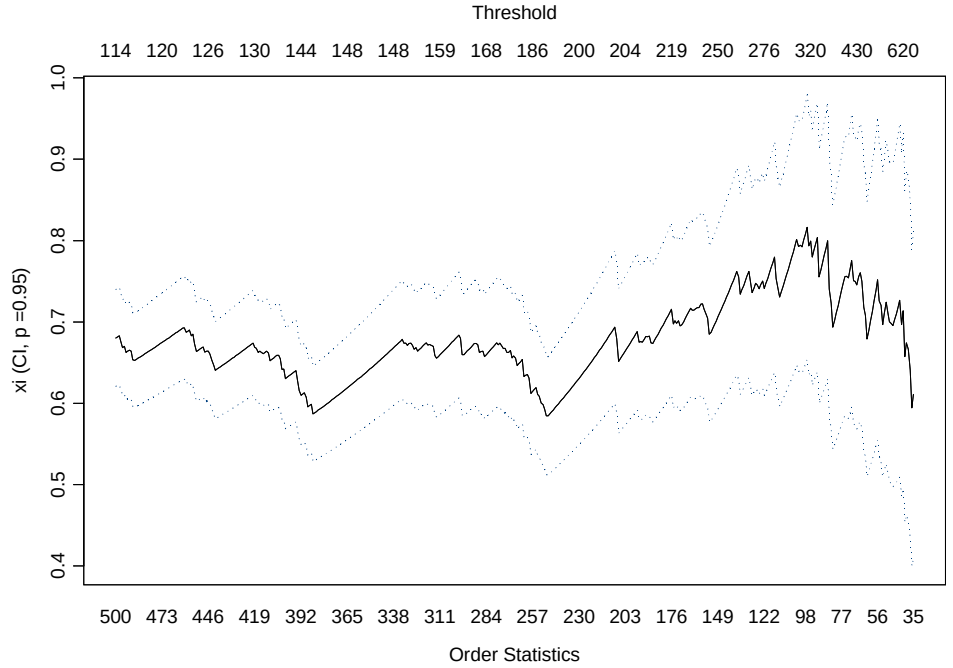

Figure 4: Hill-plot.
(Scale of the threshold on the upper x-axis: 1=100,000 ITL)

⇒ a reasonable choice for *u*: $250 < u < 300$ (exceedances: 137 and 102)

➢ Maximum likelihood estimates

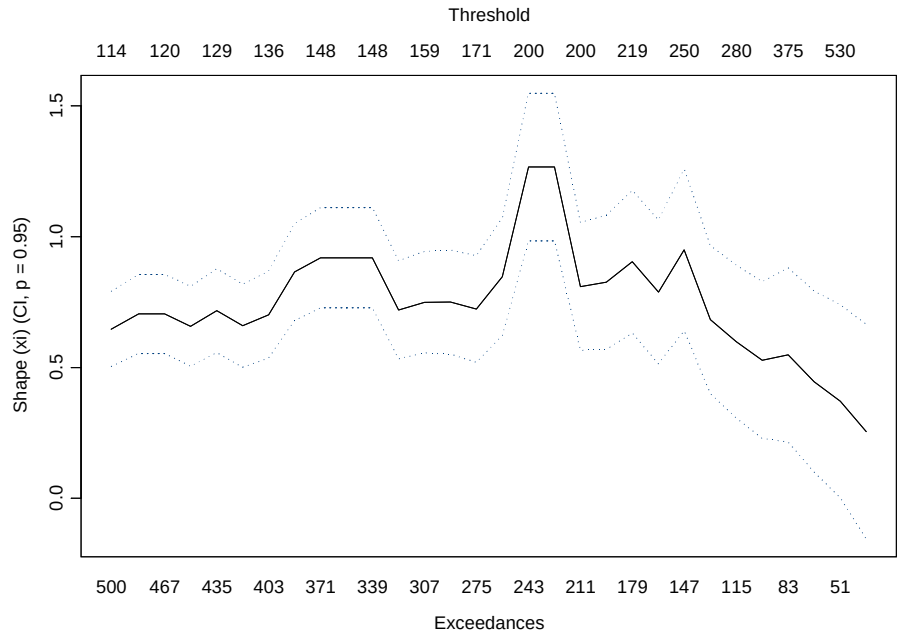

Figure 5: maximum likelihood estimates of the shape parameter.
(Scale of the threshold on the upper x-axis: 1=100,000 ITL)

$\Rightarrow$ stable estimates: $u \in (200, 250)$

❖ Quantiles of $\hat{F}(z) = G_{\xi,\tilde{\mu},\tilde{\sigma}}(z)$:

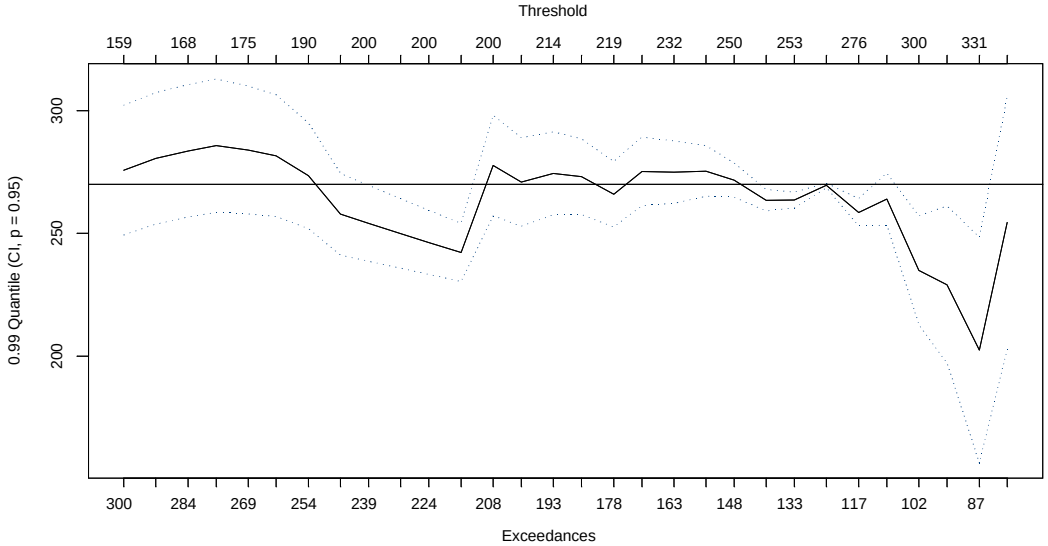

Figure 6: 0.99 quantile estimates.
(Scale of the threshold on the upper x-axis: 1=100,000 ITL)

$\Rightarrow$ stable estimates of the 0.99 quantile: $u \in (210, 250)$

➢ Thresholds 230 and 250 seem convenient.

❖ Quantiles and $E_u(Z-R)_+ = E\big(\max\big(0, Z-R\big)\big),\ R \geq u,$

where $Z \sim \hat{F}(z) = G_{\xi,\tilde{\mu},\tilde{\sigma}}(z)$

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	0.780	274.79	450.61	1506.97	8.4571	7.2345	5.2460
250	137	0.667	267.14	463.93	1487.49	6.3990	5.1513	3.1019
300	102	0.523	234.86	472.60	1505.29	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.

➢ Thresholds $u = 230 \Rightarrow$ conservative evaluations;

➢ Thresholds $u = 250 \Rightarrow$ estimates closer to the empirical values.

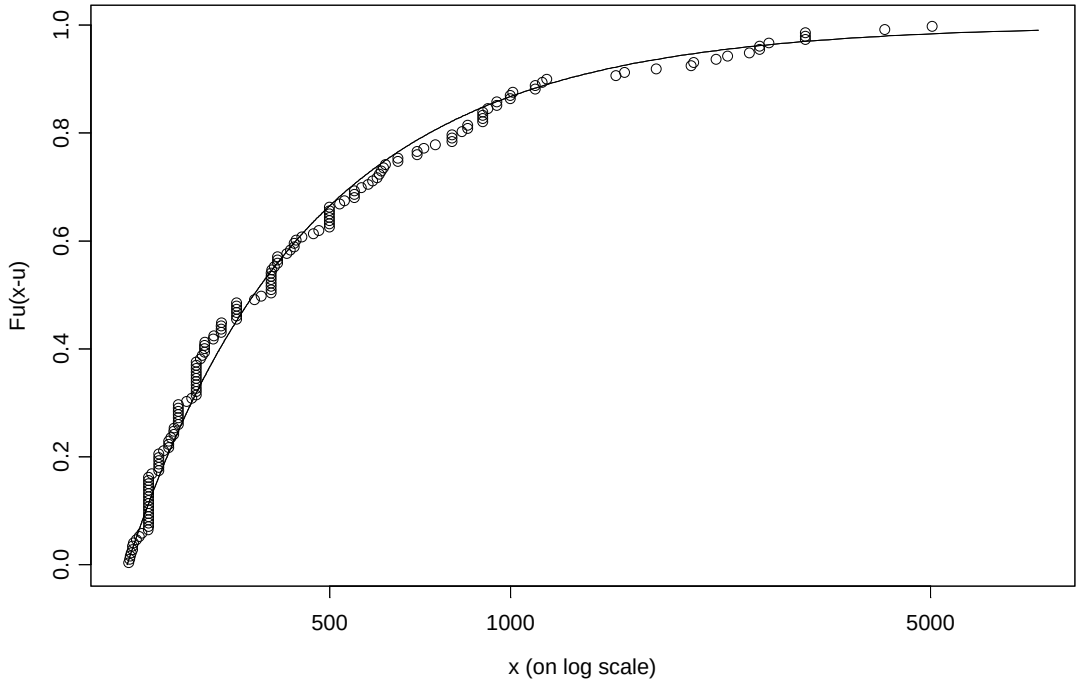

Figure 7: GPD fitted from the threshold 230.

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

✓ $E(Z_i | Z_i > R)$ ESTIMATION

Let ξ_u $(\xi_u < 1)$ and σ_u the estimated parameters of the GPD from the threshold *u*

$$\Rightarrow \quad E_u(Z_i | Z_i > R) = \frac{\sigma_u + \xi_u (R - u)}{1 - \xi_u} + R$$

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*.
(*u*, *R* and $E_u(Z_i | Z_i > R)$ are expressed in ITL divided by 100,000).

➢ $E(Z_i | Z_i > R)$ are assumed <u>independent of the tariff variables</u>.

ESTIMATES OF THE WEIGHT OF THE MIXTURE

✓ $P(Z_i > R)$ estimation

Two approaches:

➢ independence of the tariff characteristics:

a) a balance condition on the portfolio
b) observed frequency of the exceedances

➢ dependence on the tariff characteristics: <u>investigated by GLM</u>

$Z^{(j)}$ the claim amount of the *j*-th claim in the portfolio

$\left|Z^{(j)} > R\right|$ the response variable in a GLM

❖ GLM with Binomial distribution and logit link function:

- $R = 230$ Selected tariff variables: Mass and Chief town (p = 4%); Mass (p = 2%).
- $R = 250$ Selected tariff variable: KW Power.
- $R = 500$ No variables are selected at a significant level.

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

✓ $P(Z_i > R)$ ESTIMATION: $R = 230$

GLM: Mass and Chief town.

➢ If only one factorial tariff variable is selected the estimated probabilities $P(Z_i > R)$ are the observed frequencies.

✓ $P(Z_i > R)$ estimation: $R = 250$ and $R = 500$

a credibility like approach

A "CREDIBILITY MODEL" TO ESTIMATE THE WEIGHTS OF THE MIXTURE

Let

n_i the observed number of claims in tariff class i, i =1,…,s,

$Z_i^{(j)}$ the claim amount of the j-th claim in tariff class

$$X_{ij} = \left| Z_i^{(j)} > R \right|$$

Hypothesis on the process $X_{i1}, X_{i2}, \ldots$ (tariff class i):

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

For the different classes we assume that:

- the processes $\left(\Theta_i, X_{i1}, X_{i2}, \ldots, X_{in_i}\right)$ are independent;
- $\Theta_1, \ldots, \Theta_s$ are identically distributed;
- for any i, the variables $X_{ih} \left| \Theta_i = \theta \right.$ are identically distributed.

➢ Prior evaluation of $P(Z_i > R)$:

The probabilities that a claim amount would exceed the trimming point are the same in all tariff classes:

$$\mu = E\left[E\left(X_{ih} \middle| \Theta_i\right)\right]$$

➢ These probabilities are updated by taking account of the observed frequencies in the different tariff classes:

$$p_i = (1-\alpha_i)\mu + \alpha_i \bar{x}_i \qquad \text{linear credibility formula (Bühlmann (1967))}$$

where

$$\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}\ , \quad \text{with } x_{ij} \text{ the observed value of } X_{ij}\ , \qquad \alpha_i = \frac{n_i}{n_i + k}\ , \quad \text{with } k = \frac{v}{a}$$

and

$$v = E\left[\text{var}\left(X_{ih} \middle| \Theta_i\right)\right] \qquad a = \text{var}\left[E\left(X_{ih} \middle| \Theta_i\right)\right].$$

Estimators (e.g. Klugman, Panjer, Willmot (1998))

for μ :

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

where:

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

for v :

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

for a :

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

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

✓ $P(Z_i > R)$ ESTIMATION:

- $R = 230$ Selected tariff variable: Mass
- $R = 250$ Selected tariff variable: KW Power.

Mass	u=230	μ=0.012923	KW Power	u=250	μ=0.010795
i	$\bar{x}_i$	p_i	i	$\bar{x}_i$	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.

- ✓ $P(Z_i > R)$ ESTIMATION: $R = 230$
  - ➢ “ glm ” : GLM (Mass and Chief town)
  - ➢ “ f ” : observed frequencies (Mass)
  - ➢ “ c ” : credibility (Mass)
- ✓ $P(Z_i > R)$ ESTIMATION: $R = 250$
  - ➢ “ f ” : observed frequencies (KW Power)
  - ➢ “ c ” : credibility (KW Power)
- ✓ $P(Z_i > R)$ ESTIMATION: $R = 500$
  - ➢ “ c ” : credibility (KW Power)
  - ➢ “ constant ” : independent of the tariff characteristics

SOME NUMERICAL APPLICATIONS TO MOTOR INSURANCE PRICING

(P1) $E(X_i) = E(N_i)\,E(Z_i)$

✓ $E(Z_i)$ ESTIMATION:

- GLM with Gamma distribution and logarithmic link function
- Selected tariff variables:

  Age, Fuel, Chief town and Mass.

(P2) $E(X_i) = E(N_i) \cdot \left[P(Z_i \le R)E(Z_i | Z_i \le R) + P(Z_i > R)E(Z_i | Z_i > R)\right]$

✓ Choice of the trimming point R

- threshold $u = 230$ and trimming point $R = 230$
- threshold $u = 250$ and trimming point $R = 250$
- threshold $u = 250$ and trimming point $R = 500$

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

✓ $E(Z_i | Z_i \le R)$ ESTIMATION:

- GLM with Gamma distribution and logarithmic link function
- Selected tariff variables:

  Sex, Age and Mass.

  - upper unlimited support of the Gamma distribution:

    ⇒ assign a proper link function

    ⇒ negligible probability on the right tail.

u	R	Premium model	"ordinary" claims	"large" claims	Total earned premiums
		P1			42,373
230	230	P2 glm	31,646	15,439	47,085
230	230	P2 f	31,647	15,424	47,072
230	230	P2 c	31,645	15,541	47,186
250	250	P2 f	32,307	12,141	44,448
250	250	P2 c	32,308	12,143	44,451
250	500	P2 c	35,150	8,865	44,015
250	500	P2 constant	35,149	9,003	44,152
		Total observed claim amount			42,430

Table 5: earned premiums / 1,000,000.

- **threshold** $u = 230$ **:** expected total claim amount "considerably overestimated"
- **threshold** $u = 250$ **:** a more "reasonable overestimation"
- **limited liability** should be taken into account
- the global effect of **different choices of the weights** in the mixture model is moderate

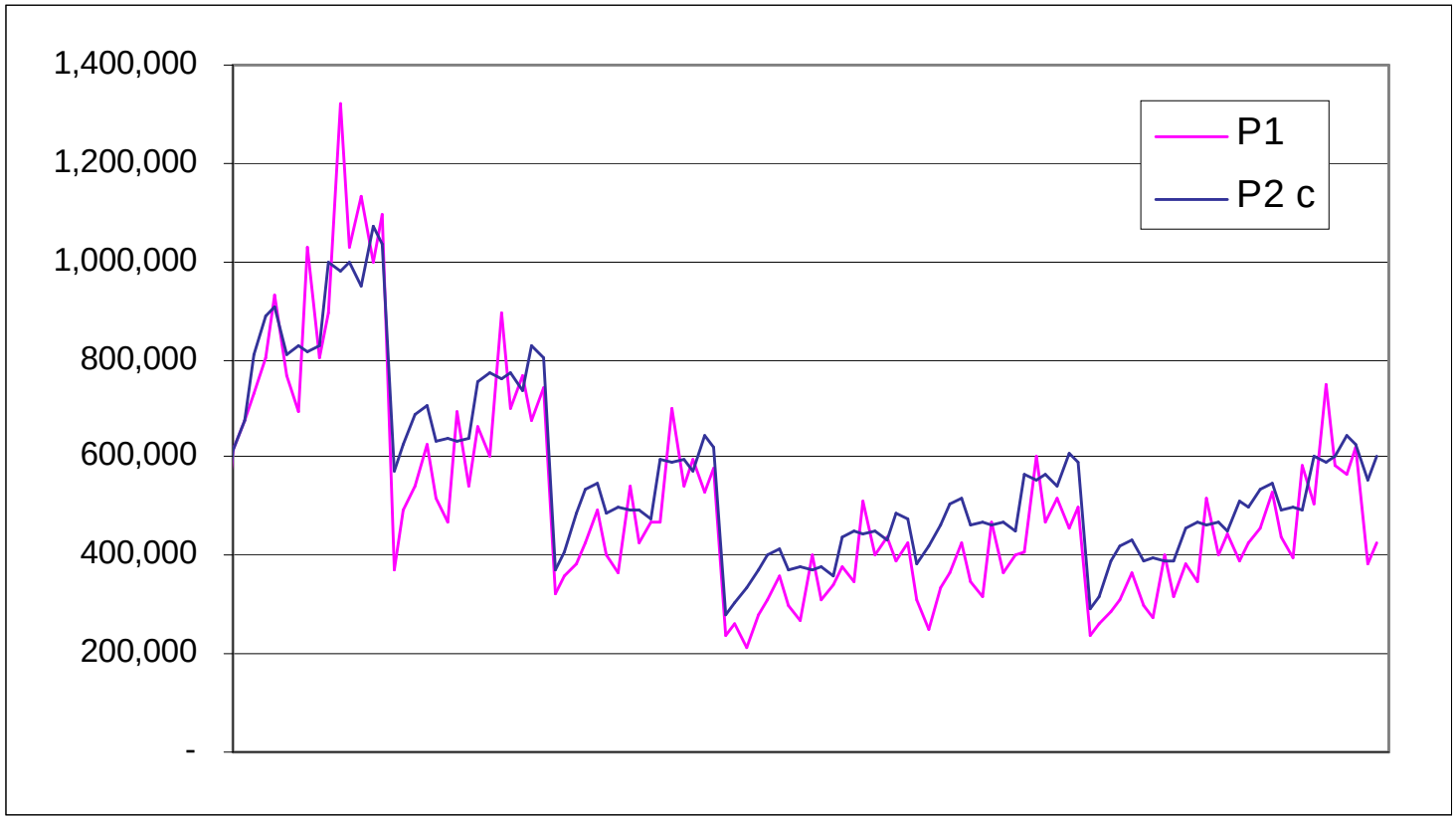

Figure 8: premiums P1 and P2 c, for some tariff classes defined by Sex=1, Chief town=2, Fuel=2, and ordered by Age, KW Power and Mass.

- Premiums P1 show much more notable fluctuations then P2.

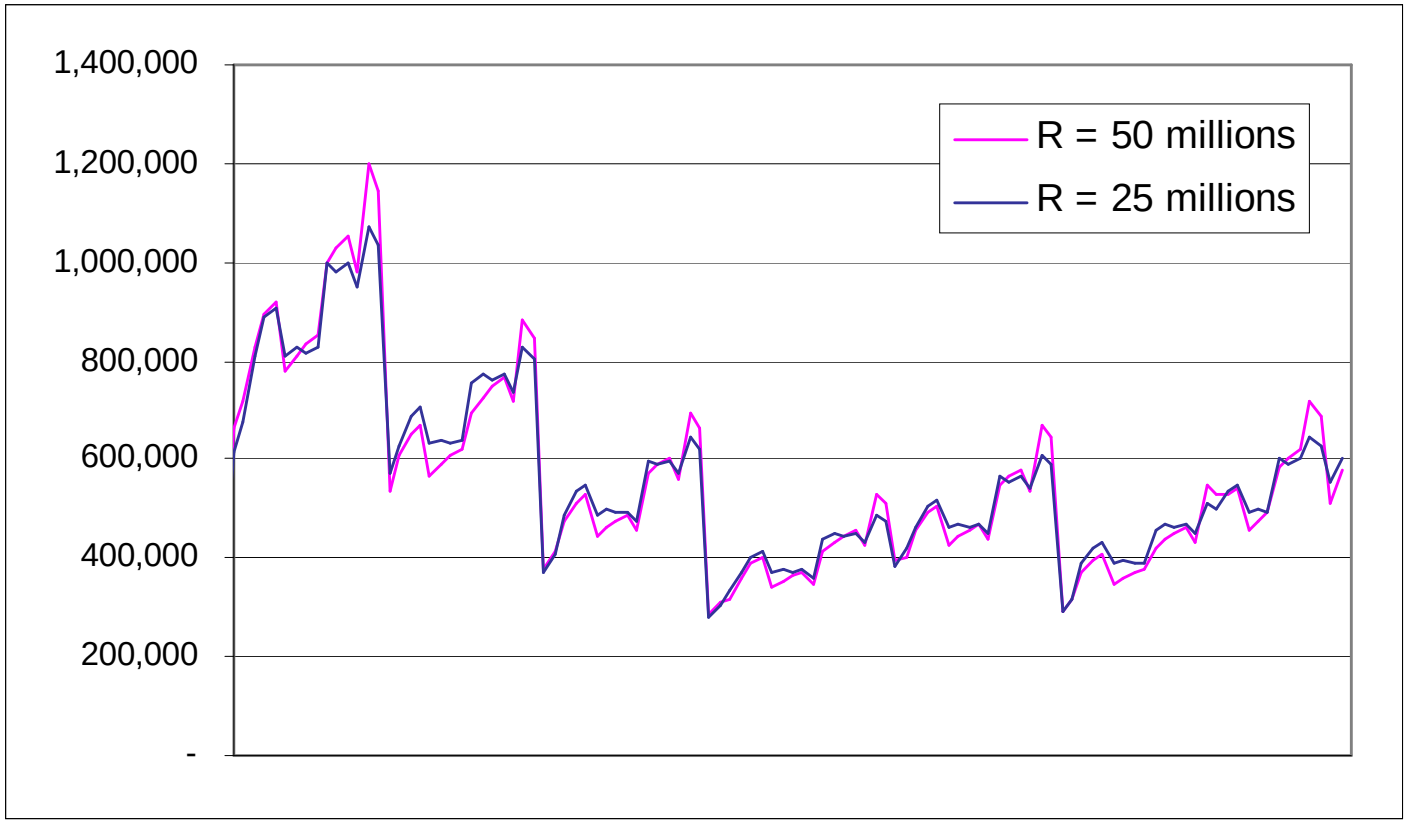

Figure 9: premiums P2 c, trimming points R = 25 millions and R = 50 millions. Tariff classes defined by Sex=1, Chief town=2, Fuel=2, and ordered by Age, KW Power and Mass.

- Premiums P2 with trimming point R=50 millions show again some fluctuations: some high claim amounts are included in the "ordinary" claim component.

CONCLUSIONS

- From these evaluations it seems 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.