A simple method to estimate parametric claim size distributions from grouped data

Joachim Brix and *Dietmar Pfeifer* (Hamburg)

1. Introduction

Reinsurance brokers and companies are frequently faced with the problem that claim size data obtained for actuarial analysis are usually processed in grouped form, and mostly even only available for the larger claim size layers. The statistical estimation of appropriate claim size distributions for the total portfolio – say with the aim of forecasting probable maximum losses as upper quantiles of that distribution – is then a difficult task which cannot be performed with the usual elementary statistical tools, although some useful recommendations can be found in the literature such as moment and modified maximum likelihood methods (cf. e.g. [3], section 4.3.A), modified minimum-distance methods (cf. e.g. [3], section 3.3 and section 4.3.A), linear regression methods in the particular case of Pareto distributions (cf. [4], section 3.3.3(c)), or particular methods in the case of lognormal distributions (cf. [2], section 1.4.3). For a similar discussion with respect to extreme value distributions, see [1].

In this paper, we want to show that such an analysis can, however, be more simply performed for most parametric classes of claim size distributions using certain non-linear regression techniques for densities that are nowadays implemented in several statistical software packages, such as STATISTICA. The powerfulness of this method will be demonstrated using both artificial as well as real data from fire, windstorm and health care losses.

2. The mathematical background

Throughout the paper we shall assume that the claim size distribution to be estimated is of parametric form, with a density $f(x;\theta)$ being continuous on its support and a (possibly multidimensional) parameter $\theta \in \Theta$ where Θ denotes the underlying parameter set. Further we assume that the data are grouped in a certain number m of pairwise disjoint layer intervals $L_i = (a_i, b_i]$, $i = 1, \ldots, m$; note, however, that we do not necessarily assume that these intervals are adjacent. Let $m_i = (a_i + b_i)/2$ denote the midpoint of each inverval, and k_i denote the number of claims falling in the layer band L_i. By the mean value theorem of classical analysis, we have

$$\int_{L_i} f(x;\theta)\,dx = (b_i - a_i)\, f(\xi;\theta) \approx (b_i - a_i)\, f(m_i;\theta)$$

where ξ is a suitable intermediate point in L_i. On the other hand, under the assumption of stochastic independence of the data generating random variables, distributed as X, say, an application of the law of large numbers shows

$$\int_{L_i} f(x;\theta)\,dx = P(X \in L_i) \approx \frac{k_i}{n}$$

where P denotes the underlying probability measure and n is the total number of claims. We thus obtain the simple approximation formula

$$f(m_i;\theta) \approx \frac{k_i}{n(b_i - a_i)}, \quad i = 1, \ldots, m.$$

If one defines – depending on the data given – a suitable loss function, e.g.

$$L(\theta) = \left(w\left(\frac{k_i}{nd_i}\right) - w(f(m_i;\theta)) \right)^2, \quad \theta \in \Theta,$$

with $d_i = b_i - a_i$, and a suitable weight function w, a parameter estimation for $\theta \in \Theta$ can be performed using a non-linear regression technique using the loss function L. A great advantage of this method over those outlined in the introduction is that we do not make any use of the underlying cumulative distribution function, which is generally not expressible in closed form, e.g. for lognormal or gamma distributions. For the problem under consideration, $w = \lg$ (the logarithm with base 10) has turned out to be quite efficient, since this enforces a more accurate approximation in particular in the tails of the distribution, which is especially desirable from the viewpoint of reinsurance.

3. Some practical example

The following table contains the grouped data k_i from 2000 simulations of $LN(\mu, \sigma)$-lognormally distributed random variables with $\mu = 1$ and $\sigma = 2$ in the column named KI, with a total of $m = 12$ layer bands. The column named KI_N_DI contains the transformed data $k_i/n/d_i$.

Table 1

NUMERIC VALUES	1 AI	2 BI	3 MI	4 DI	5 KI	6 KI_N_DI	7 KI_DI
	0.0	1.0	.5	1.0	604	.30200000	604.00000
	1.0	5.0	3.0	4.0	637	.07962500	159.25000
	5.0	10.0	7.5	5.0	260	.02600000	52.000000
	10.0	20.0	15.0	10.0	191	.00955000	19.100000
	20.0	50.0	35.0	30.0	178	.00296667	5.9333333
	50.0	100.0	75.0	50.0	67	.00067000	1.3400000
	100.0	150.0	125.0	50.0	26	.00026000	.52000000
	150.0	200.0	175.0	50.0	14	.00014000	.28000000
	200.0	500.0	350.0	300.0	16	.00002667	.05333333
	500.0	750.0	625.0	250.0	4	.00000800	.01600000
	750.0	1000.0	875.0	250.0	1	.00000200	.00400000
	1000.0	4500.0	2750.0	3500.0	2	.00000029	.00057143
SUM case 1-12					2000		

Using the module Nonlinear Estimation in STATISTICA with the above user-specified loss function

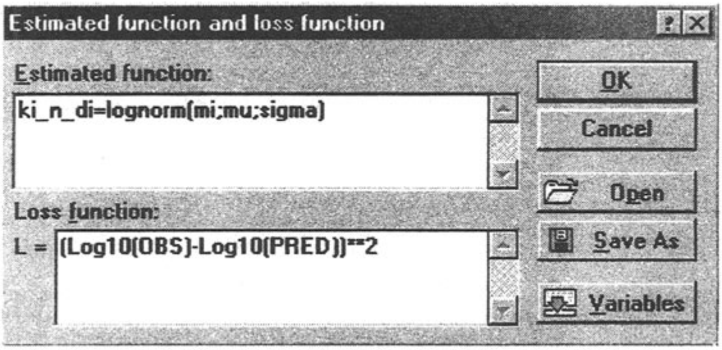


Figure 1

and the Hooke-Jeeves-pattern move procedure (which has turned out to be one of the most stable numerical procedures for our problem, among the choices offered by STATISTICA)

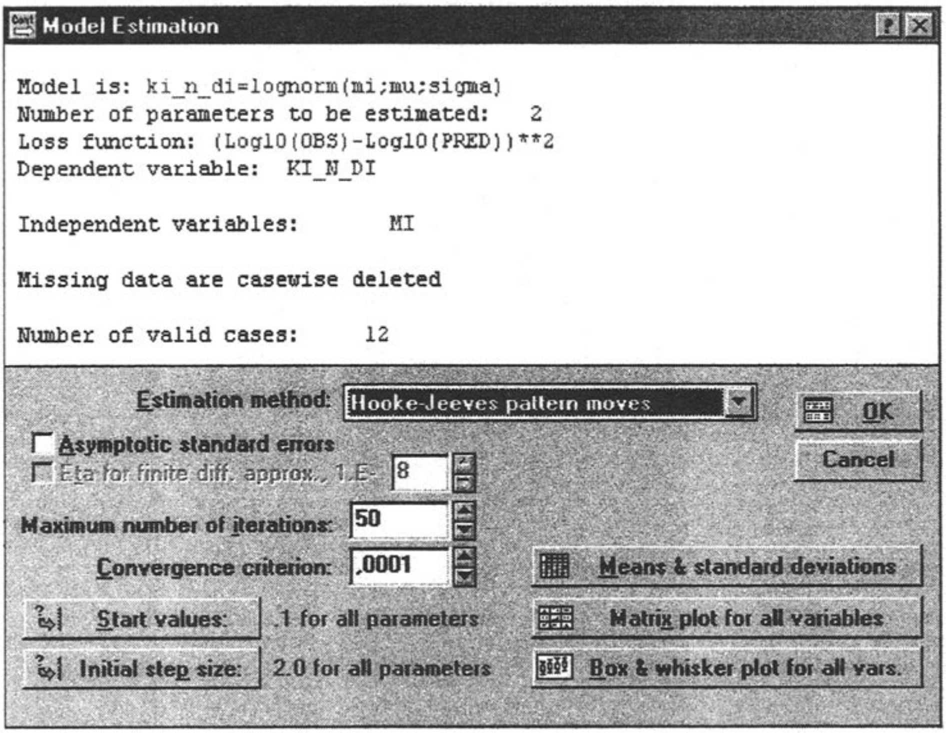


Figure 2

we obtain the following estimates for μ and σ:

$$\hat{\mu} = 0{,}9823075, \qquad \hat{\sigma} = 2{,}009623$$

and the estimated density plot (in log-log-scale, cf. [2], p. 94, Fig. 1.4.3.6):

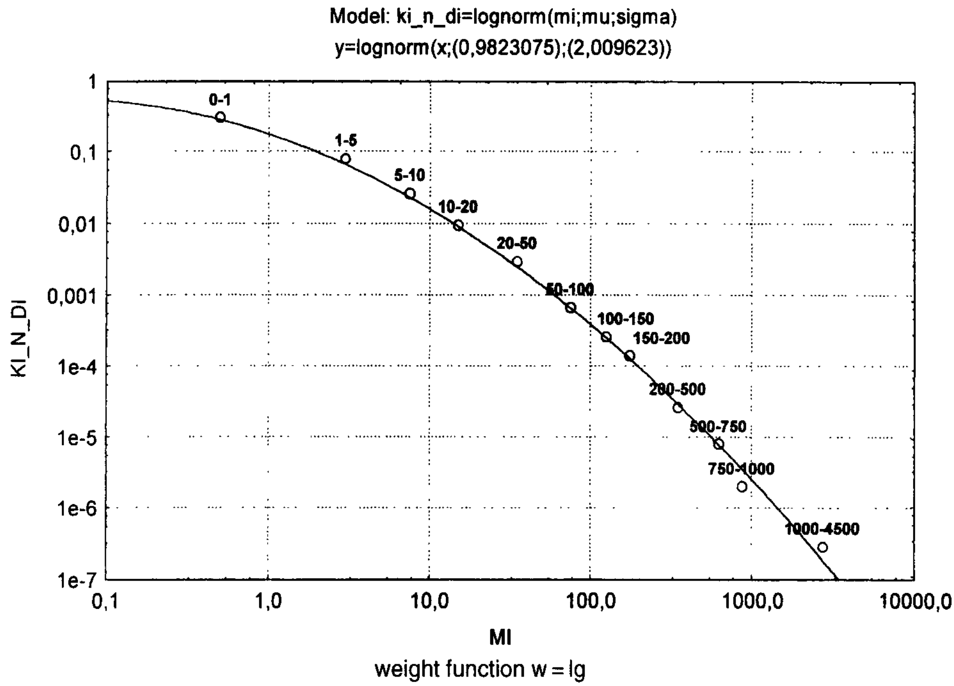


Figure 3

Seemingly, the fit to the actual distribution is extremely good. The following graph shows the fitting result when only the upper 7 layers are used (i.e. only 130 out of 2000 original data!):

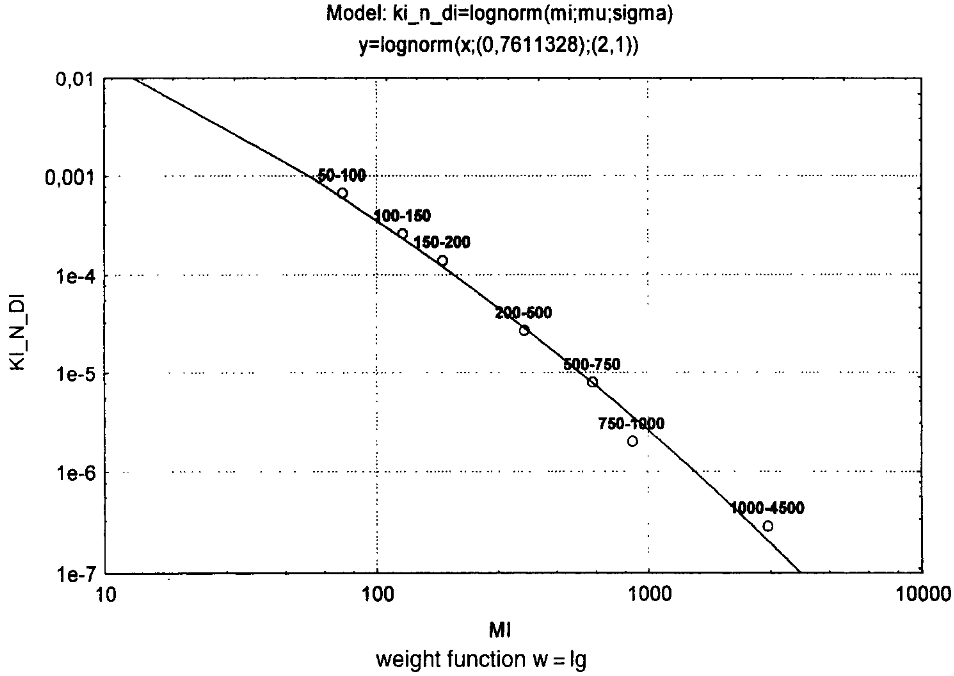


Figure 4

with the estimates

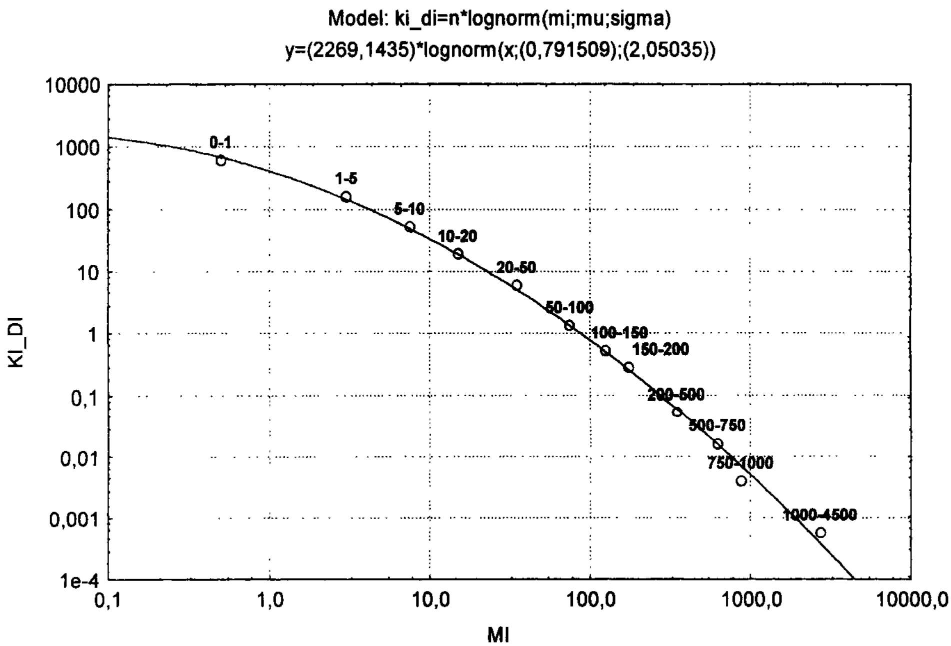
$$\hat{\mu} = 0{,}7611328, \qquad \hat{\sigma} = 2{,}1$$

which are still reasonably close to the original parameters in spite of the fact that only 6,5% of the available information was used.

The problem of distribution fitting becomes still a little bit more complicated if the total number of claims is unknown, which is sometimes the case if only data for the upper layer bands are available. In this situation, the number n of observations can formally be added as another component to the parameter θ, i.e. the loss function will now be

$$L^*(\theta, n) = \left(w\left(\frac{k_i}{d_i}\right) - w(n \cdot f(m_i;\theta)) \right)^2, \quad \theta \in \Theta, \quad n \in \mathbb{N},$$

where w is again a suitable weight function. This approach avoids the otherwise necessary consideration of conditional distributions, which would require an inclusion of the cumulative distribution function in the loss function. For the full data set, a corresponding analysis with w = lg gives the following picture (note that column KI_DI in the table above contains the ratios k_i/d_i):



weight function w = lg

Figure 5

with still acceptable estimates

$$\hat{n} = 2269, \qquad \hat{\mu} = 0{,}791509, \qquad \hat{\sigma} = 2{,}05035\,.$$

Note, however, that the use of w = lg sometimes does not produce stable results, if too little of the layer bands are given and n is large. This is due to the fact that the products

$n \cdot f(m_i; \theta)$ increase with n, so that the weight function w = lg will level out even significant differences between the fit function and the data. E.g. in the example above, no reasonable parameter estimates – especially for n – are obtained if more than the first layer band is removed from the analysis. A general possibility of improvement here consists in a different choice of the weight function w. Good results are usually obtained if the lg-function is replaced by $\sqrt[4]{\ }$. The following graph shows the corresponding results, where again only the 7 upper layer bands were considered for the analysis.

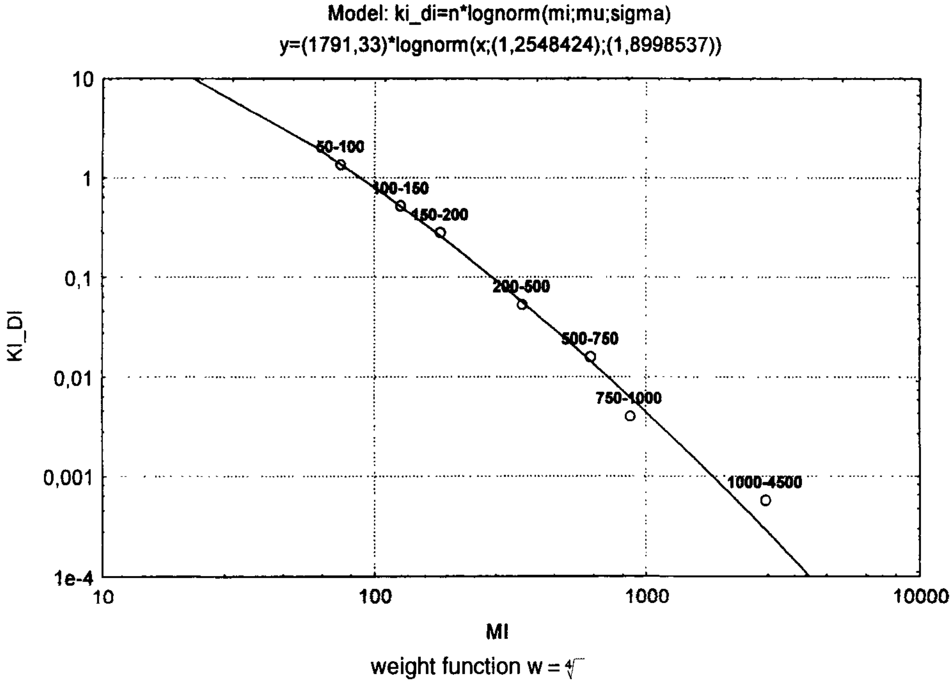


Figure 6

The corresponding parameter estimates are here

$$\hat{n} = 1791, \quad \hat{\mu} = 1{,}2548424, \quad \hat{\sigma} = 1{,}8998537$$

which still is a reasonably good result.

4. Worked examples from actuarial practice

In this section, we firstly want to present the outcome of an analysis of the above type for an existing portfolio with fire and windstorm losses, resp., from the year 1997. The data have been kindly provided by AON Re Jauch & Hübener, Hamburg. According to actuarial experience, the fire claim data are fitted by a lognormal, the windstorm data by a Fréchet distribution with cumulative distribution function.

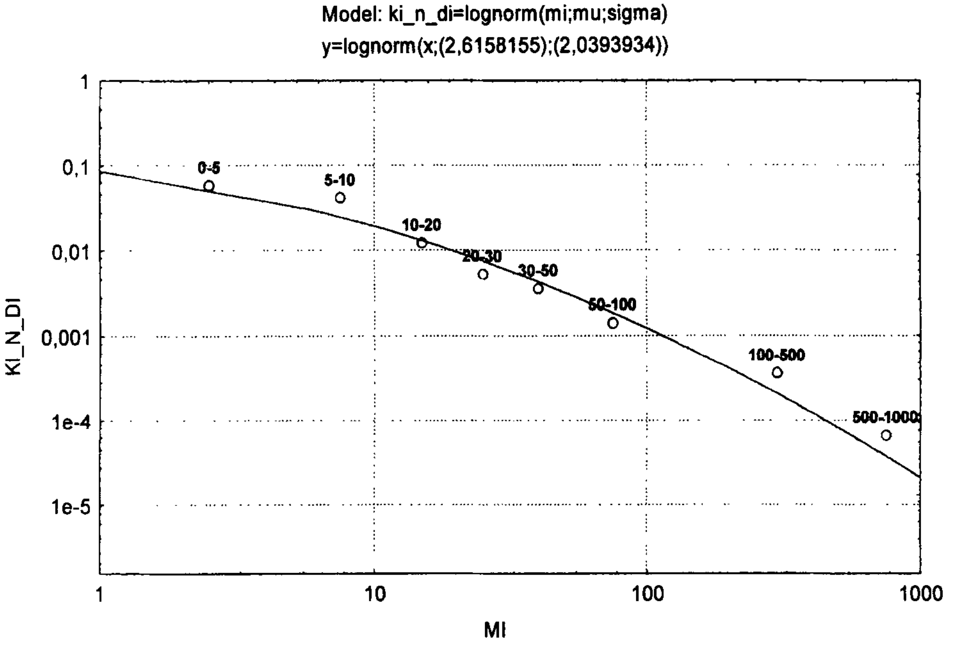
$$F(x) = e^{-(Ax)^{-\alpha}}, \qquad x > 0$$

with shape parameter α and scale parameter A (see e.g. [5] for the problem of distribution fitting for windstorm losses).

Table 2

NUMERIC VALUES	1 AI	2 BI	3 MI	4 DI	5 KI	6 KI_N_DI	7 KI_DI
	0,0	5,0	2,5	5,0	620	,05868434	124,00000
	5,0	10,0	7,5	5,0	440	,04164695	88,000000
	10,0	20,0	15,0	10,0	257	,01216280	25,700000
	20,0	30,0	25,0	10,0	110	,00520587	11,000000
	30,0	50,0	40,0	20,0	150	,00354946	7,5000000
	50,0	100,0	75,0	50,0	148	,00140085	2,9600000
	100,0	500,0	300,0	400,0	307	,00036323	,76750000
	500,0	1000,0	750,0	500,0	70	,00006626	,14000000
	1000,0	2000,0	1500,0	1000,0	11	,00000521	,01100000
SUM case 1-9					2113		

fire claims in 1000 DEM, number of claims = 2113



estimated lognormal density in log-log-scale, weight function w = lg

Figure 7

with estimates

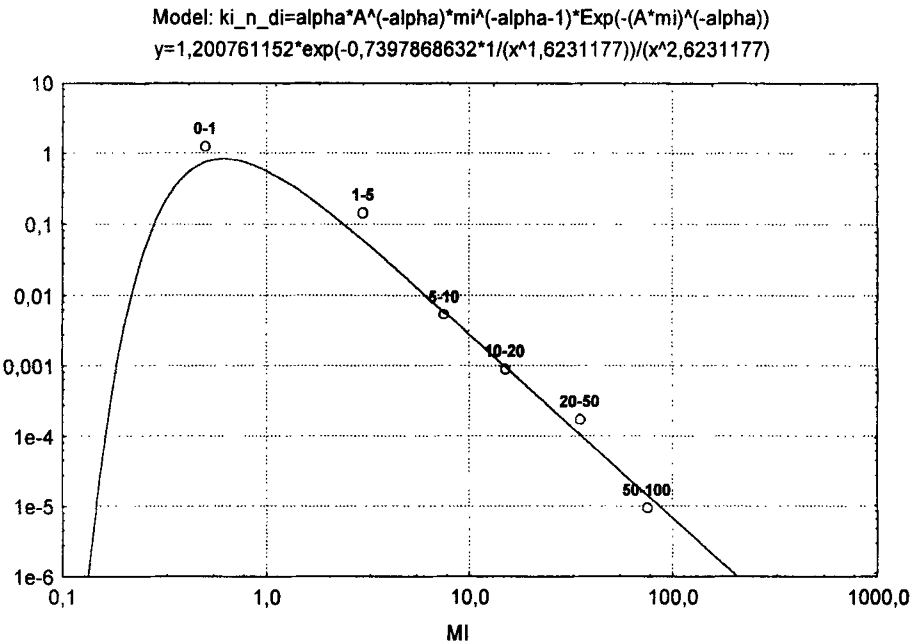
$$\hat{\mu} = 2{,}6158155, \quad \hat{\sigma} = 2{,}0393934$$

For the windstorm losses, we have the following data.

Table 3

NUMERIC VALUES	1 AI	2 BI	3 MI	4 DI	5 KI	6 KI_N_DI	7 KI_DI
	0,0	1,0	,5	1,0	2678	1,2673923	2678,0000
	1,0	5,0	3,0	4,0	1210	,14316138	302,50000
	5,0	10,0	7,5	5,0	57	,00539517	11,400000
	10,0	20,0	15,0	10,0	19	,00089920	1,9000000
	20,0	50,0	35,0	30,0	11	,00017353	,36666667
	50,0	100,0	75,0	50,0	1	,00000947	,02000000
SUM case 1-6					3976		

windstorm losses in 1000 DEM, number of claims = 3976



estimated Fréchet density in log-log-scale, weight function w = lg

$\hat{\alpha} = 1{,}6231177, \quad \hat{A} = 0{,}8934019$

Figure 8

In both cases, the results seem to be quite satisfactory from the practical point of view.
The following investigation refers to the analysis performed in [1], concerning health care data, which are given in the following table.

Table 4

NUMERIC VALUES	1 AI	2 BI	3 MI	4 DI	5 KI	6 KI_N_DI	7 KI_DI
	0,0	5,0	2,5	5,0	1835	,17368670	367,00000
	5,0	10,0	7,5	5,0	1663	,15740653	332,60000
	10,0	20,0	15,0	10,0	1101	,05210601	110,10000
	20,0	40,0	30,0	20,0	717	,01696640	35,850000
	40,0	60,0	50,0	20,0	252	,00596309	12,600000
	60,0	80,0	70,0	20,0	103	,00243729	5,1500000
	80,0	100,0	90,0	20,0	56	,00132513	2,8000000
	100,0	150,0	125,0	50,0	42	,00039754	,84000000
	150,0	200,0	175,0	50,0	14	,00013251	,28000000
	200,0	300,0	250,0	100,0	13	,00006152	,13000000
SUM case 1-10					5796		

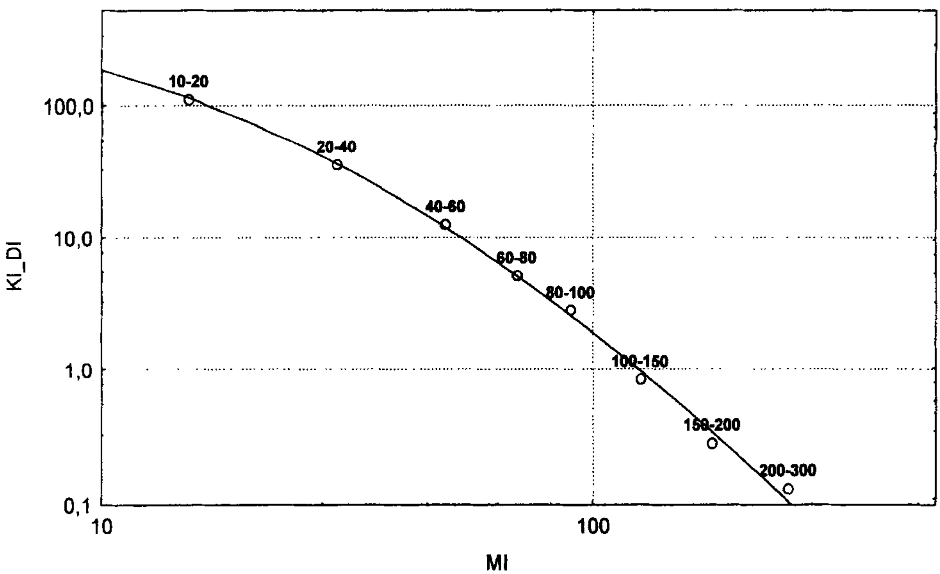
health care claims in 1000 DEM, number of claims = 5796

In [1], the analysis was performed with the 8 upper layer bands, fitting a generalized Pareto distribution with cumulative distribution function of the form

$$F_{\xi,\beta}(x) = 1 - (1 + \xi x/\beta)^{-1/\xi}, \quad x \geq 0$$

with shape parameter $\xi > 0$ and scale parameter $\beta > 0$. Note that we have only changed the endpoint of the last layer band from ∞ to 300. The following graph shows the result of an analysis with our method.

Model: ki_di=5796/beta/(1+xi*mi/beta)**(1+1/xi)
y=5796/(9,969185)/(1+(0,372664)*x/(9,969185))**(1+1/(0,372664))



estimated generalized Pareto density in log-log-scale, weight function $w = \sqrt[4]{\ }$

Figure 9

The estimates here are

$$\hat{\xi} = 0{,}372664, \qquad \hat{\beta} = 9{,}969185$$

of which only the scale parameter estimate $\hat{\beta}$ differs essentially from the corresponding estimate $\hat{\beta} \approx 17{,}5$ in [1]. However, our estimates are in better coincidence with the data given in table 4, as can be seen from the following table:

Table 5

NUMERIC VALUES	1 AI	2 BI	3 KI	4 KI_NEW	5 KI_OLD
	0.0	5.0	1835	2136.3546	1386.5239
	5.0	10.0	1663	1187.8452	976.08056
	10.0	20.0	1101	1175.9868	1232.1678
	20.0	40.0	717	797.48790	1137.9963
	40.0	60.0	252	251.66485	469.85392
	60.0	80.0	103	105.29814	228.67129
	80.0	100.0	56	52.194457	124.52167
	100.0	150.0	42	52.633740	136.63812
	150.0	200.0	14	17.810559	49.700474
	200.0	300.0	13	11.699080	33.789316
SUM case 1-10			5796	5788.9753	5775.9433

Here the column KI_NEW contains the expected number of claims in the corresponding layer band according to our estimates, while KI_OLD contains the expected number of claims in the corresponding layer band according to the estimates in [1], obtained with the χ^2-method (cf. also Abb. 2 in [1]). It is clearly seen that the deviation between actual claim numbers and expected claim numbers is much less with our method than with the methods in [1] which are based on the cumulative distribution function instead of the density, as in our case. In particular, with the results in [1] the tail of the (fitted) distribution is obviously overestimated, resulting in slightly too high premiums.

5. Final remarks

The density based method for fitting claim size distributions to grouped data presented in this paper is not only fast but seemingly also produces good or even better results in comparison with other methods based on the cumulative distribution function. In particular, it is possible to fit claim size distributions to incomplete data sets, either with given total number of claims, or without (being probably less accurate then), which is of special importance to all kind of reinsurance applications.

REFERENCES

[1] *Furrer, Hansjörg:* Methoden der Extremwerttheorie zur Bestimmung eines Einzelschaden-Exzedenten im Krankenversicherungsbereich. Blätter der DGVM XXIV (1999); 87–102.

[2] *Mack, Thomas:* Schadenversicherungsmathematik. Schriftenreihe Versicherungsmathematik, Heft 28; Verlag Versicherungswirtschaft e.V., Karlsruhe; 1997.

[3] *Hoog, Robert V., Klugman, Stuart A.:* Loss Distributions; Wiley, N.Y.; 1984.
[4] *Daykin, C. D., Pentikäinen, T., Pesonen, M.:* Practical Risk Theory for Actuaries; Chapman & Hall, London; 1994.
[5] *Pfeifer, Dietmar:* A statistical model to analyse natural catastrophe claims by means of record values. Proceedings of the XXVIIIth International ASTIN Colloquium, Cairns, Australien 1997; 45–57.

Zusammenfassung

Eine einfache Methode zur Schätzung parametrischer Schadensverteilung aus gruppierten Daten

In der Rückversicherungspraxis wird man häufig mit dem Problem gruppierter Daten konfrontiert, die zudem meist auch unvollständig – d.h. nur in höheren Schadenbändern – vorliegen. Standard-Verfahren der mathematischen Statistik zur Schätzung der zugrundeliegenden Verteilung lassen sich dann in der Regel nicht ohne weiteres anwenden. In diesem Aufsatz soll daher gezeigt werden, wie unter Verwendung nicht-linearer Regressionsmethoden für Dichteschätzungen, die heutzutage in vielen Statistik-Software-Produkten vorhanden sind, eine solche Analyse doch relativ einfach durchgeführt werden kann. Die Stärken dieses Verfahrens werden sowohl anhand simulierter Daten als auch anhand konkreter Schadenfälle aus der Feuer-, Sturm- und Krankenversicherung veranschaulicht.

Summary

A simple method to estimate parametric claim size distributions from grouped data

In the praxis of reinsurance the problem often occurs that claim size data are usually processed in grouped form, and mostly even only available for the larger claim size layers. The statistical estimation of appropriate claim size distributions for the total portfolio is then a difficult task which cannot be performed with the usual elementary statistical tools. In this paper, we want to show that such an analysis can, however, be simply performed for most parametric classes of claim size distributions using certain non-linear regression techniques for densities that are nowadays implemented in several statistical software packages. The powerfulness of this method is demonstrated using both artificial as well as real data from fire, windstorm and health care losses.