BAYESIAN RISK ANALYSIS

Richard L. Smith [1] and Dougal J. Goodman [2]

July 4 2000

Summary

The solvency of an insurance company depends critically on the size and frequency of large claims. Assessment of insurance risk therefore depends on accurate modelling of the distribution of large claims. In this paper, insurance claim data obtained from a large company are analysed to determine the distribution of tail values. It is shown that the distribution is indeed extremely long-tailed. A Bayesian computation is made of the predictive distribution of the total loss likely to be incurred by the company over a future one-year period. The data set contains two outliers, the only two claims in the data which represent losses of an entire facility. There are thus some grounds for treating these as members of a separate population. However, if these two observations are omitted, the fitted distribution is still very long-tailed and the predictions of total loss are still high. Further, the fact that the data consist of a mixture of different types of claim is taken into account through a hierarchical model framework, and it is shown that the assessed risk is smaller if this feature is taken into account. Finally, we consider the effect of possible trends in the data. There is a peak in the number of claims corresponding to the worldwide crisis in the insurance industry of the early 1990s. This is statistically significant when modelled as a quadratic time-trend, but this does not appear to be a reasonable basis for future extrapolation. A better approach is via a random effects model to incorporate year-to-year variation. This is successfully incorporated within the hierarchical model, but there is little evidence of any persistent trend, and the assessed risk is little different if year effects are taken into account than in the case where year effects are omitted. The whole paper is presented as a demonstration of the merits of combining established models for extreme values with modern statistical techniques including Bayesian inference, hierarchical models and Monte Carlo sampling.

---

[1] Department of Statistics, University of North Carolina, Chapel Hill, N.C. 27599-3260, USA, email rls@email.unc.edu. This work was carried out in part at the Isaac Newton Institute during a programme on Nonlinear and Nonstationary Signal Processing, July–December 1998. Supported by NSF grant DMS-9705166, EPSRC visiting fellowship GR-K99015 and by a grant from the TSUNAMI Initiative to the Isaac Newton Institute.

[2] Director, Foundation for Science and Technology, 78 Buckingham Gate, London SW1E 6PE, UK, e-mail dougal.goodman@btclick.com. This work was carried out while at BP Insurance and the British Antarctic Survey, Cambridge, England.

1 Introduction

The viability of the insurance industry depends on probabilistic calculations of risk. Although the majority of claims fall within the scope of standard actuarial calculations, there are a small number of very large claims for which current actuarial practice does not have a satisfactory answer. For such claims, it is essential to have as accurate a characterisation as possible of the probability distribution of very large values. This is particularly the case when there is no *a priori* limit on the size of claims which may be entertained, as is typically the case in the reinsurance industry. The issue has come into particular focus in recent years with a series of environmental disasters such as Hurricane Andrew and the Mississippi floods, and consequent financial failures in the insurance industry.

In this paper, the distribution of very large claims, and its consequences for overall losses, are examined with reference to a large data set obtained from a well-known international company. To protect the confidentiality of the source, the data are coded so that both the units of money and the exact time frame covered by the data are not revealed, but apart from that, the data are actual claims, over a 15-year period, above a threshold level set at 0.5 units of currency. The units of currency have been adjusted for inflation.

There are a total of 425 claims classified into one of seven types. Fig. 1 presents the data graphically in a number of ways. Fig. 1(a) depicts the time (horizontal axis) and amount (vertical axis) of all the individual claims. Fig. 1(b) is a cumulative (CUSUM) plot of *numbers* of claims against time, without taking account of the sizes of claim. This looks like a straight line, indicating no evidence of any trend in the frequency of claims. Fig. 1(c) is a similar plot based on cumulative *sizes* of claim up to any particular time. This is distorted by the presence of two very large claims, both of which occurred in year 8, corresponding to which there is a jump shift in the plot, but other than that this plot is also linear in appearance. Fig. 1(d) is explained later.

One issue that arises is the distinction between "claims" and "losses". A catastrophic event, such as a hurricance, may result in a number of claims but it would not be reasonable to treat these, for statistical purposes, as independent random variables. In the industry, it is usual to aggregate claims from a single cause into "losses", which are effectively independent. In the present data set, there is no direct indication of when different claims correspond to the same loss. There is indirect evidence in that the date associated with any particular claim is the date on which the original loss occurred, not the date on which the claim was filed, so there are reasonable grounds for thinking that claims of the same type originating on the same day are in fact due to a common loss and should therefore be aggregated. In fact, when the data are aggregated in this way, the total number of claims is only reduced from 425 to 393, it seems unlikely that the analysis which follows would be substantially different if the claims/loss distinction was ignored completely. Losses of the same class or type such as fire, flood, liability etc. which can occur on the same day may be correlated if they are known to be at the same site or in the same region if the loss is weather related. Frequently a reinsurer will not have a full description of the data from the primary insurer and will be working with aggregated losses. In the following analysis,

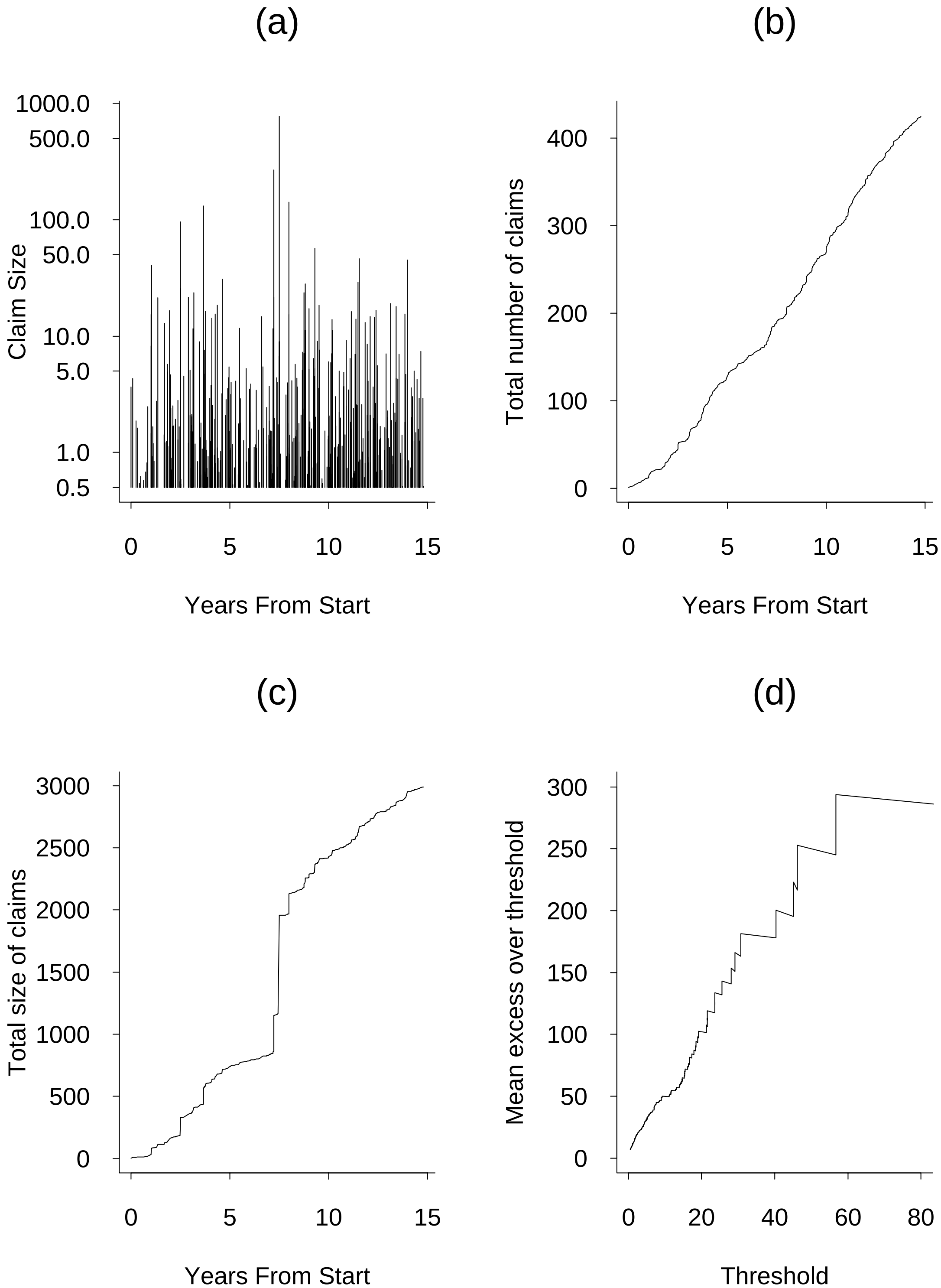

Fig. 1: Plots of raw data. Plot(a) shows the time and magnitude of all 425 claims above the base level 0.5. (b) and (c) are CUSUM plots based respectively on the number of claims and the total amount of claims above 0.5. (d) shows a mean excess over threshold plot.

claims of the same type on the same day were aggregated, so the analysis which follows is based on the reduced set of 393 losses. We shall continue to refer to these as "claims" because most of them are single claims, and we reserve the word "loss", at least in the present paper, to refer to the total loss suffered by the company over a specified period of time.

Within this framework, a number of questions arise:–

(a) What is the probability distribution of the claims, with particular reference to the very large claims?

(b) Is there any trend towards rising or falling claims, or any seasonal variation which might affect our assessment of the probability distribution?

(c) What is the influence of the two very large claims (henceforth called outliers)?

(d) Is there any significant difference in the distributions associated with different types of claim?

(e) What are the implications of all these issues for the overall risk on the company?

With reference to the two outliers, there is of course a reason for concern in any case where some observations are much larger than the rest of the sample. (For the record, the largest five claims are 776.19, 268.00, 141.95, 131.05, 95.76.) There is particular concern about the use of extreme value methods, which we are going to apply in the rest of the paper, when there is a possibility that these represent some separate process (for example, the distinction between normal wind speeds and hurricanes). In the present case, these concerns are well-founded, because the two largest claims are the only ones in the data set which represent "total losses", i.e. when a complete unit was lost. Ideally one would like to do a separate analysis of claims resulting from total losses, but with only two such claims available, this is not practicable. The two outliers will therefore be combined with the rest of the data for most of the analysis, but their separate origins do need to be borne in mind in interpreting the results.

The remainder of the paper aims to answer questions (a)–(e). Section 2 discusses the probability distribution of large claims with particular reference to Pareto and generalised Pareto tail approximations. One conclusion from this is that the distribution of claims is indeed very long-tailed, even without the two outliers. An alternative version of the method is also given, by which one can better assess the stability of parameters across different thresholds, and which also provides a basis for trend-testing.

Section 3 then considers the total loss suffered by the company (i.e. the total over all claims, not "total loss" in the more technical sense of two paragraphs ago) on the assumption that this is the quantity of direct interest to the company. In particular, we aim to answer the question "How large a reserve would be needed to cover all losses in a particular year, with probability $1/N$?" for various values of N. This analysis is based on a Bayesian calculation of the predictive distribution, using Monte Carlo simulation.

Section 4 extends the analysis to consider both the effect of different types of claim, and of possible trends resulting in different distributions for different years. There is evidence in the data that both effects are important, but it is difficult to be definitive about this because of the small number of large claims of any given type or in any single year. However, an alternative approach which allows both type-of-claim effects and year effects to be taken into account is via a hierarchical Bayesian model, and this is developed in Section 5. The comparison between different models, and their consequences for the assesed risk of a very large loss, are developed in some detail.

Finally, Section 6 presents a summary and conclusions.

2 The Distribution of Very Large Claims

Although the main analysis in this paper is Bayesian, our initial analyses are from a classical (frequentist) perspective. This serves partly to show what is possible using known extreme value techniques, and also to highlight some of the limitations of those techniques which we shall seek to overcome using a Bayesian viewpoint.

Modern methods of extreme value analysis are based on exceedances over high thresholds. Denoting the threshold by u, the conditional distribution of excesses over u is modelled by the generalised Pareto distribution (GPD):–

$$\Pr\{Y \leq u+y \mid Y > u\} \approx 1 - \left(1 + \frac{\xi y}{\sigma}\right)_+^{-1/\xi}, \quad y \geq 0 \tag{1}$$

($x_+ = \max(x, 0)$) where $\sigma > 0$ is a scale parameter and ξ a shape parameter. The GPD is formally justified as a limiting distribution, valid as u approaches the upper end of the distribution of Y, in which case it is applicable to very wide classes of underlying distributions of Y (Pickands 1975, Davison and Smith 1990). The three cases $\xi < 0$, $\xi = 0$ and $\xi > 0$ correspond to different types of tail behaviour. The case $\xi < 0$ arises in distributions where there is finite upper bound on the claims which are possible. Although it might be thought that this case would apply (in the sense that no company has infinite liability), in practice we would expect to detect such a limit only if there was a tendency for claims to cluster near the upper limit, which is not the case here. Therefore, we do not consider this possibility further. The second case, $\xi = 0$, typically arises in cases with an exponentially decreasing tail (in this case we take a formal limit $\xi \to 0$ in (1), resulting in the exponential distribution $1 - e^{-y/\sigma}$). This arises, not only when the distribution of Y is indeed exponential, but from many other common distributions such as gamma, Weibull, normal, lognormal etc. Some of these distributions (e.g. gamma, lognormal) are quite commonly used for actuarial data, so we might expect to find the estimated value of ξ close to 0 in practice. However the third case, $\xi > 0$, is of more concern because this corresponds to a genuinely "long-tailed" distribution. This case arises whenever $\Pr\{Y > y\} \sim cy^{-\alpha}$ as $y \to \infty$ for some positive constants c and α. This therefore corresponds to what is usually called the Pareto tail (hence the terminology "generalised Pareto" to extend the distribution to the cases when $\xi \leq 0$); the relation between ξ and α is $\xi = 1/\alpha$. When $0 < \alpha < 2$ the distribution is also tail equivalent to an α-stable distribution.

A critical issue in practice is the selection of an appropriate threshold u. If this is set too high, there will not be enough data over the threshold to calculate good estimates of σ and ξ. However we do not want u to be too low: there is no point in basing the estimates on claims which are too small to be considered "large claims", and to do so could induce a bias associated with lack of fit of the GPD.

One diagnostic device which has been developed (Davison and Smith 1990) is the "mean excess over threshold" plot. For each possible threshold, we compute the mean of all excesses over the threshold. For example, associated with threshold 100 there are four excesses (676.19, 168.00, 41.95 and 31.05), and the mean of these is 229.3. Fig. 1(d) shows a plot of mean excess over the threshold, plotted against the threshold itself, as this runs from the minimum value 0.5 up to 80. If the data really follow a GPD, then this plot should stay close to a straight line of slope $\xi/(1-\xi)$, provided $\xi < 1$. We can see that over most of the plot, the mean excess over the threshold does indeed appear to rise linearly, suggesting that the GPD may be a good fit over much of the range of the data. The apparent exception to this is at the right-hand end of the plot, but in fact this is not such a significant matter because in this region there are very few data points — the mean excess is computed from a very small number of exceedances and hence has a lot of sampling variability. On the basis of this plot, the evidence in favour of the GPD seems good.

The GPD is fitted to exceedances over a variety of thresholds in Table 1. The method of fitting is numerical maximum likelihood. The main feature of this table is that, with the exception of the last row based on just six exceedances, all the values of ξ are very large (> 0.8), and some are in the range $\xi \geq 1$ for which even the mean of the distribution is infinite. (If $\frac{1}{2} \leq \xi < 1$ the mean of the claim size distribution is finite, but the variance is infinite.) Thus we are clearly in the "Pareto tail" case, with a very long-tailed distribution.

Threshold	Number of Exceedances	Mean Excess	σ	ξ
0.5	393	7.11	1.02	1.01
2.5	132	17.89	3.47	0.91
5	73	28.9	6.26	0.89
10	42	44.05	10.51	0.84
15	31	53.60	5.68	1.44
20	17	91.21	19.92	1.10
25	13	113.7	33.76	0.93
50	6	206.6	150.8	0.29

Table 1. Summary of generlised Pareto fits.

This method of analysing the extreme values has the advantage of being straightforward to implement, but there are a number of disadvantages when considering broader features of the distribution, such as whether there are trends, a topic to which we return in Section 4. An alternative approach, introduced by Smith (1989), is via a point process representation of the exceedances. The idea is to view all claims exceeding a given level x as a Poisson process of intensity

$$\Lambda_x = \left(1 + \xi \frac{x-\mu}{\psi}\right)_+^{-1/\xi}. \tag{2}$$

It is assumed that this formula is valid over all $x \geq u$ for some given threshold u — in fact a more detailed description of the process allows for both "claim times" and "claim amounts" in excess of u to be treated as a two-dimensional nonhomogeneous Poisson process. This is equivalent to the GPD in the following sense: if (2) is valid for all $x \geq u$, and if Y denotes an arbitrary exceedance of the process (i.e. an arbitrary claim bigger than u) then (1) holds with ξ the same as in (2), and $\sigma = \psi + \xi(u - \mu)$. However, so long as the model remains valid, the parameters μ, ψ and ξ are independent of the threshold. Thus the approximate equality of these parameter estimates over a variety of thresholds is one indication that the model is fitting correctly.

Another, and more important, advantage of the point process formulation is that the model extends easily to the case where the parameters μ, ψ and ξ are not constants, but dependent on time. In this case we write μ_t, ψ_t and ξ_t where t is time, and model the point process of exceedances over x as a *nonhomogeneous* Poisson process of intensity

$$\Lambda_t(x) = \left(1 + \xi_t \frac{x-\mu_t}{\psi_t}\right)_+^{-1/\xi_t}. \tag{3}$$

Applications of this include looking for trends in the data, and models including covariates. For example, Smith (1989) allowed μ_t to vary with time to examine whether high-level ozone exceedances in Houston had varied significantly with time. Smith and Shively (1995) extended this analysis by writing μ_t and $\log \psi_t$ as linear combinations of meteorological covariates with unknown coefficients. This is important in the context of atmospheric ozone, because it is well known that ozone production is highly dependent on the underlying weather conditions. In the present paper, we shall use the model (3) to look for trends. So far, the possible effect of different covariates has not been taken into account, though the idea of trying to use meteorological or other kinds of environmental covariates, as a means of assessing to what extent environmental change could influence insurance claims (e.g. the possibility of increased risk of flooding arising from global warming) would appear to be an intriguing possibility.

The model (2) or (3) may be estimated by numerical maximum likelihood, as described in Smith (1989). For numerical convenience, the parameter ψ is replaced by $\log \psi$ in virtually all subsequent analysis. Table 2 shows the results applied to a number of thresholds. One feature to note here is that the parameter estimates are very stable across different thresholds, certainly as far as $u = 10$, in the sense that differences between successive values of the parameter estimates are small in comparison with their standard errors. The fits for values above $u = 10$ show a sharp increase in the standard errors of $\log \widehat{\psi}$ and $\widehat{\xi}$, and therefore seem less reliable.

Threshold	Number of Exceedances	μ	$\log \psi$	ξ
0.5	393	26.5 (4.4)	3.30 (0.24)	1.00 (0.09)
2.5	132	26.3 (5.2)	3.22 (0.31)	0.91 (0.16)
5	73	26.8 (5.5)	3.25 (0.31)	0.89 (0.21)
10	42	27.2 (5.7)	3.22 (0.32)	0.84 (0.25)
15	31	22.3 (3.9)	2.79 (0.46)	1.44 (0.45)
20	17	22.7 (5.7)	3.13 (0.56)	1.10 (0.53)
25	13	20.5 (8.6)	3.39 (0.66)	0.93 (0.56)

Table 2. Estimation of the model (2) with constant μ, ψ and ξ. Numbers in parentheses are standard errors.

Smith and Shively (1995) introduced a number of diagnostic devices to examine the fit of this model. One idea is based on what we shall call Z-statistics: if $\Lambda_t(x)$ is given by (3), let

$$Z_k = \int_{T_{k-1}}^{T_k} \Lambda_s(u) ds \tag{4}$$

where T_k denotes the time of the k'th exceedance of u ($T_0 = 0$, assumed to be the starting time of the observations). If the model is correct, then $Z_1, Z_2, ...,$ will be independent exponentially distributed random variables with mean 1. In practice we usually work in discrete time (as here, where all claims of the same type on a single day have been aggregated) so we replace the integral in (4) by a sum, but the exponential distribution should still be approximately correct provided the inter-claim times are not too small.

The Z-statistics are an indication of how closely the exceedances of a fixed level u are represented by a nonhomogeneous Poisson process, but they do not test the GPD assumption for the distribution of excess over the threshold. This can be done via W-statistics: with T_k the k'th exceedance time and Y_k the corresponding value, let

$$W_k = \frac{1}{\xi_{T_k}} \log \left[1 + \xi_{T_k} \frac{Y_k - u}{\psi_{T_k} + \xi_{T_k}\{u - \mu_{T_k}\}} \right].$$

Then $W_1, W_2, ...,$ are also independent exponential random variables with mean 1, if the model is correct.

There are various plots we can do to decide whether these assumptions are in fact supported by the data. These include:–

(i) Scatter plots of Z_k or W_k against T_k. Systematic variation of the Z or W values with time would indicate a trend not accounted for in the model.

(ii) Q-Q plots, i.e. arrange the Z_k or W_k values in increasing order and plot them against the expected order statistics under the exponential distribution. If the plot stays close to a straight line through the origin of unit slope, we may conclude that the exponential distribution is a good fit.

(iii) Serial correlation plots — an indicator of whether successive values of Z_k or W_k are in fact independent.

Fig. 2 shows all three of these plots computed for both the Z and W statistics, based on all exceedances over threshold $u = 5$ and the parameter estimates from the homogeneous model in Table 2. The choice of $u = 5$ was somewhat arbitrary, but following our earlier remarks this seems a reasonable compromise between trying to fit the model to the whole data set, in which case the truly large claims would not have such a strong influence, and setting the threshold too high when the estimates become unstable. The time plots (a) and (d) are shown together with a crude smoothed curve through the scatterplot, computed using the SPlus `lowess` procedure. There is some concern about plot (a), since there seems to be a slight dip in the curve around year 4 and a more prominent dip around year 11. This could indicate an increase in the number of claims in those years (note that small Z values mean small time intervals between claims, i.e. an increase in the rate of claims). In plots (b) and (e), the points will lie close to the diagonal straight line if the model is correct. The interesting thing to note here is that with the W-statistics, where there are two large values corresponding to the two very large claims already discussed, the observed values at the right-hand end of the plot are still only just above the expected values, suggesting that these may not be outliers at all, but just the kind of values that would have been expected under such a long-tailed distribution. The serial correlation plots (c) and (f), shown together with nominal 95% confidence intervals at $2/\sqrt{N}$, where N is the number of exceedances, show no cause for concern.

It is possible to re-fit both the GPD and point process models with the two outliers omitted. For example, fitting the GPD with $u = 5$, we obtain $\widehat{\sigma} = 6.83$ (standard error 1.50), $\widehat{\xi} = 0.59$ (0.20) compared with the earlier estimates of 6.26 and 0.89. The main thing to note here is that the estimate of ξ has dropped quite a bit, though we still have a very long-tailed distribution on our hands (e.g. the variance is still infinite). Thus although the estimates are changed somewhat if we remove the outliers, this does not explain away our basic concerns about a long-tailed distribution. In the more complicated models developed

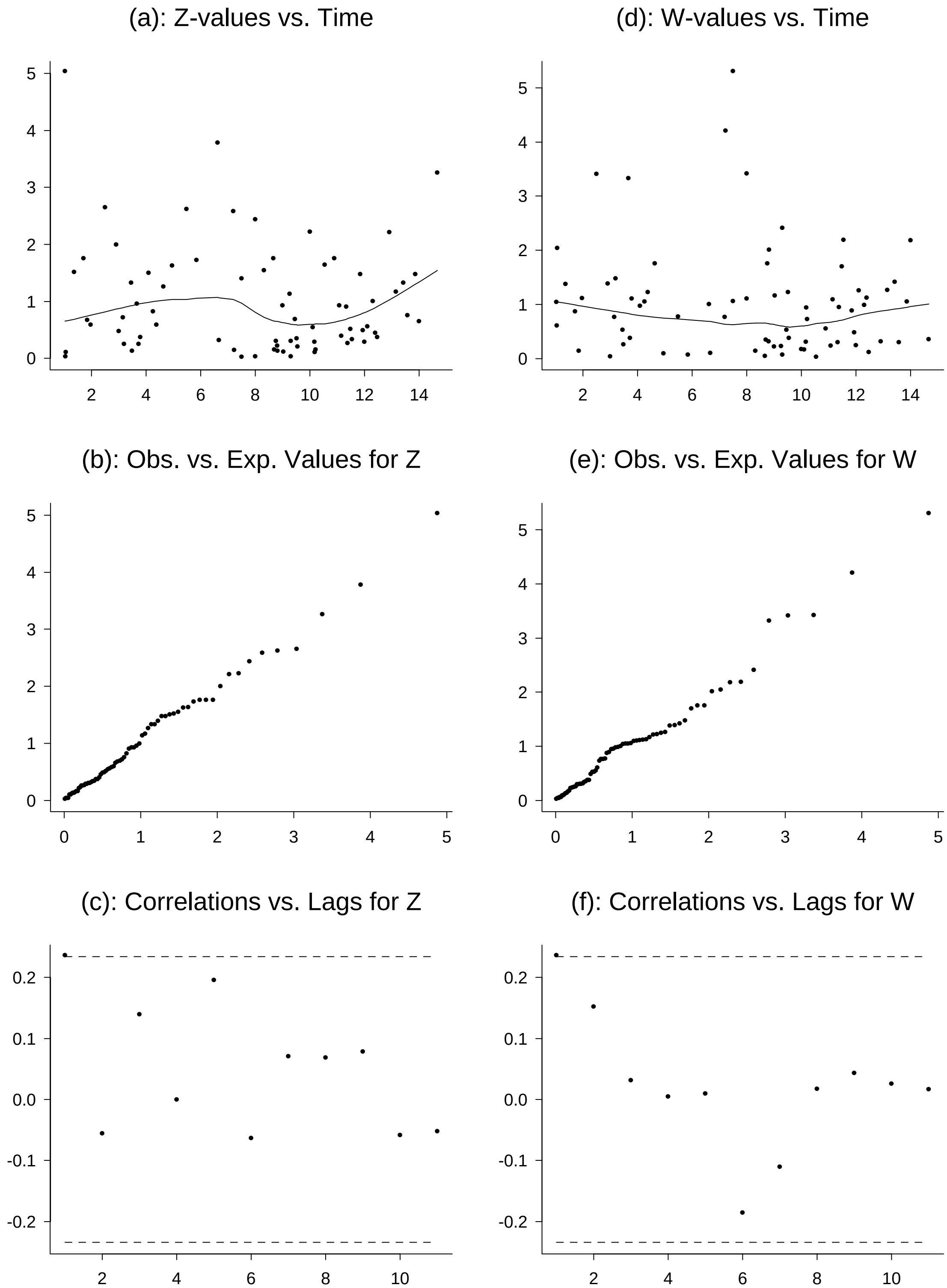

Fig. 2: Diagnostic plots. (a)–(c) show timeplot (with Lowess smoother), Q-Q plot and serial correlation plot (with 95% confidence limits) for the Z statistics. (d)–(f) show the same plots for the W statistics.

later in the paper, the same comparison of results with and without the outliers will be repeated a number of times, to indicate just how influential they are.

The overall conclusion at this stage, however, is that with the exception of the possibility that a trend might exist, the model (2) appears to fit very well.

3 The Predictive Distribution of Future Losses

Let us now assume that the model (2) is correct. What implications does this have for future losses suffered by the company?

The simplest way to deal with this issue is by simulation. Using the GPD, we can simulate the values of all claims over the threshold u over a fixed period of time, say 1 year. By repeating the simulation many times, we obtain the distribution of future loss.

One objection to this approach is that it ignores all claims below the threshold. It seems unlikely that this could influence the distribution of very large losses, but to accommodate the small claims, it is possible to do an independent simulation of claims in the range $(0.5, u)$. (Recall that the original data set is restricted to claims above 0.5.) For this purpose a bootstrap simulation has been used: each simulated "claim" is obtained by sampling with replacement from the distribution of existing claims in the range $(0.5, u)$. This method of simulation seems to work well for claims within a finite range, but it would not work for the whole of the distribution, because the bootstrap method can only produce simulated claims within the range of the existing data set, and therefore could never allow properly for future claims outside the range of the current data set. This is a well-known difficulty with applying bootstrap methods to extreme value problems.

There is a second and more fundamental difficulty with the naïve simulation approach. We have estimated extreme value parameters μ, ψ and ξ, but it is unreasonable to treat these as known quantities, when they have in fact been estimated from a limited amount of data with considerable uncertainty as represented by the standard errors. In particular, the uncertainty of the parameter ξ could have major implications for the distribution of extremely large losses.

We dwell on this point, for a moment, because this is one point where the approach advocated in this paper differs sharply from the conventional actuarial approach. Conventially, the estimation of an appropriate distribution for claim sizes is treated as a separate problem from the calculation of future loss events. One of the main points of the present paper is that these features cannot be separated but must be treated as components of a single overall estimation problem.

The proposed solution is Bayesian. We calculate the posterior joint distribution of $(\mu, \log\psi, \xi)$, assuming a prior distribution which is uniform over a very large set (effectively, the whole of $\mathcal{R}^3$). The parameters are estimated by a Gibbs-Hastings-Metropolis simulation algorithm in which the posterior densities of the three parameters are estimated by updating one at a time, perturbing each parameter by a small random amount and using a Hastings-Metropolis step to decide whether to accept or reject the perturbed value (Smith and Roberts 1993). The predictive distribution of "future loss" is then averaged over the posterior distribution of $(\mu, \log\psi, \xi)$. We iterate the Gibbs-Hastings-Metropolis algorithm over a total of M_0 cycles, and once every M_1 cycles we simulate one value of the total loss. This gives a sample of M_0/M_1 simulated values from the predictive distribution, with which we form an empirical c.d.f. In all the examples in this paper we have taken $M_0 = 10^6$, $M_1 = 100$ so we end up with a sample of 10,000 values from the predictive distribution. Although no formal tests of convergence of the simulation algorithm have been applied, inspection of data plots and comparisons of results from different (independent) simulations suggests that convergence issues are not in any sense a problem with this procedure.

As an example, four independent runs were performed of the entire simulation, starting from different random number seeds. Fig. 3 shows the posterior densities (estimated using the SPlus `density` routine) for the parameters μ, $\log\psi$ and ξ, plus the predictive distributions of "total loss", from four independent runs of the simulation. There is little variability between simulations. However, the posterior distribution of ξ shows that a substantial proportion of the posterior distribution (about 38%) lies in the "extremely long-tailed" region $\xi \geq 1$, while the estimates of loss associated with various probabilities (Fig. 3(d)) serve to quantify the effect of the extreme long-tailedness of the distribution on the losses of the company.

Up to this point, however, the analysis has all been based on an assumption of a homogeneous process in which all the claims are effectively i.i.d. random variables. We now look at this assumption a little more closely.

4 Claim-Type and Trend Effects

So far, the analysis has not taken any account of the fact that the data are classified into seven different types of claim. Also, there has been some suggestion of a trend in the data, which we need to investigate further. Once again, we perform some initial non-Bayesian analyses, which serve to motivate a Bayesian hierarchical modelling approach, which is pursued in detail in section 5.

Let us first consider the effect of different claim types. The issue here is that it is possible that the seven different types have completely different distributions associated with them, so that analysing them all as one common distribution will produce distorted estimates of the true tail parameters.

A simple way to deal with this is to fit a separate GPD to each of the claim types. The difficulty here is that, for some of the types, the number of large claims is very small. For example, with $u = 5$ the number of exceedances of the seven different types is

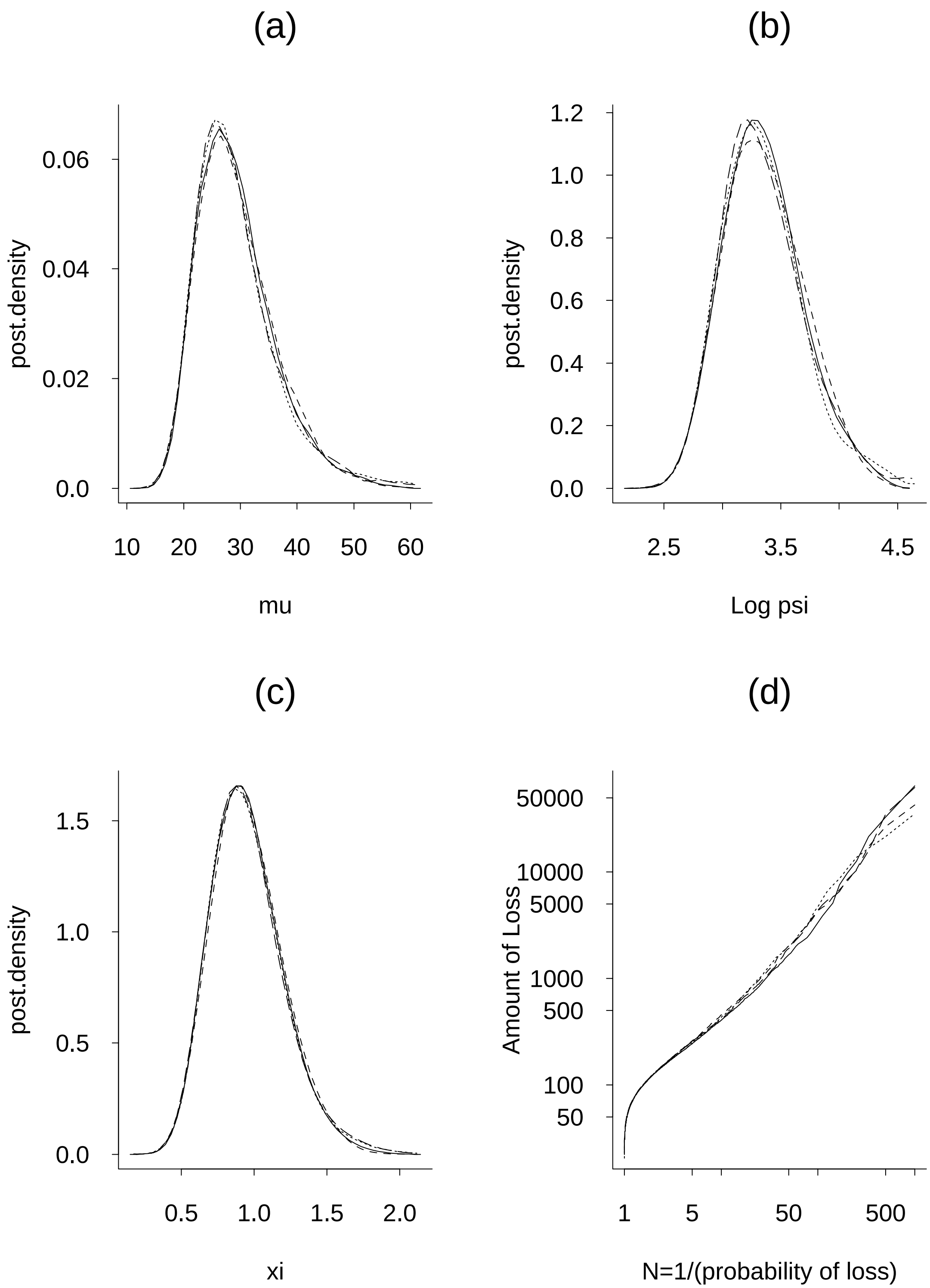

Fig. 3: Posterior densities of μ, $\log\psi$ and ξ, under the homogeneous Bayesian model based on exceedances over $u = 5$, together with the predictive loss curves in (d). Four independent runs of the same simulation are shown, to allow comparison between different simulations.

$38+4+12+3+9+7+0=73$ while for $u=2.5$ it is $64+9+17+8+23+11+0=132$. In neither case is it possible to say anything at all about type 7, for which there are only two (relatively small) claims in the entire data set. Henceforth, we ignore type 7, reducing the number of types actually considered to 6. Even in that case, there are not many claims of types 2, 4 and 6, but the situation is rather better if we set $u=2.5$ rather than $u=5$. Therefore, we now use $u=2.5$ as the main threshold for reference.

Fitting separate GPDs to the six types, based on $u=2.5$, we have estimated σ values $(4.5, 1.5, 4.8, 2.5, 2.0, 5.7)$ and ξ values $(0.90, 1.66, 0.97, 0.97, 0.59, -0.09)$. This immediately suggests that type 5 and 6 have smaller ξ values than the others, and some further exploration of different models suggests that a reasonable one is to group the types into two classes, one consisting of types 1,2,3,4 and the other of types 5,6. The first class yieds parameter estimates $\widehat{\sigma}=3.98$ (standard error 0.81) and $\xi=0.98$ (0.20). The second has $\widehat{\sigma}=3.07$ (1.01) and $\widehat{\xi}=0.32$ (0.29).

One can formally test for different models using deviance statistics. For example, if we let model I denote the model with common GPDs for all types, model II with separate GPDs for all types, and model III the last-named model in which there are just two classes, then we find that the deviance statistic (twice the difference of log likelihood ratios) for testing I against II is 11.92 with 10 degrees of freedom, for a p-value of 0.29. This suggests that the separate-parameters model is not statistically significant, but this may be due to model II being overparametrised relative to the amount of data available. The test of I against III yields deviance statistic 6.82 with 2 degrees of freedom, p-value 0.03. This is much better and suggests that we do have a significant result in this case. However we need to be careful here as well — there is a strong hint of post-hoc selection in model III which the quoted p-value does not allow for.

This whole analysis is unsatisfactory because of the small number of claims in certain types, and this makes the interpretation of the deviance statistics difficult. However, the analysis serves to draw attention to the possible importance of a claim-type effect, and in section 5, we shall pursue this in more detail.

Now let us consider trends in the data. We have already seen in Fig. 2(a) that there is evidence of a dip in the Z-statistics, meaning a rise in total claim rate, around year 10. The total claim rate increase is probably due to a change of strategy in the company towards insurance purchase. A thorough review of why insurance was purchased was made around year 10 which led to a close examination by asset managers of what they had been claiming. Changes of policy wording, marketing campaigs, or changes of the balance of the portfolio of insureds can lead to non-stationary behaviour in the loss history. However, for the moment, we revert to the case in which all seven claim types are treated as part of a homogeneous process.

A simpler way to look at this is just to plot the number of exceedances in each year. Fig. 4 shows this for a variety of thresholds, together with `lowess` plots to indicate the trend. Given the high year-to-year variability, the results are not conclusive, but they do point to a peak around years 10-11, with a secondary peak around year 4.

If we fix ψ and ξ and let μ_t be a linear function of time t, then the result as measured by deviance statistics is not statistically significant (p-vaue 0.24 compared with model (2)). However if we allow a quadratic trend in μ_t, the result is significant (p-value 0.03). The fit is not substantially improved by adding further cubic and quartic terms. Similar results are obtained by fixing μ and letting $\log \psi$ vary with time: a linear trend is not significant, but a quadratic trend is. The fit is not substantially improved by letting μ and $\log \psi$ vary together, nor by allowing ξ_t to vary as well. In passing, it may be noted that some models in which the parameters varied sinusoidally *within* each year were tried, in the hope of identifying seasonal effects, but this approach did not yield any statistically significant results.

One possible interpretation of these results is that the appearance of a trend might be a consequence of the fact there are two outliers near the middle of the data. However, when these outliers are removed and the models re-fitted, the results are very similar. For example, with constant ψ_t and ξ_t and a quadratic trend in μ_t, the p-value against the case of constant μ was 0.04. With constant μ and $\log \psi_t$ a quadratic function of t, the evidence is even slightly stronger (p-value 0.03). Thus the evidence for a quadratic trend is not due to the outliers. Further evidence of this is that all the fitted quadratic curves are consistent with a peak in the rate of claims around year 10, somewhat later than the two outliers but consistent with the general rise in insurance claims of the early 1990s. This therefore suggests that we ought to take the trend seriously.

There remains the question of how to model the trend. Fitting a quadratic function of time is a useful device for determining that a trend exists, but it is hardly a satisfactory basis for future extrapolation (which would imply, for instance, that the rate of claims will continue to fall after 1996). A more reasonable approach is to allow an additional random effect to represent the variation from year to year. We now consider a way to do this, as part of the hierarchical modelling approach already mentioned.

5 Hierarchical Models

In Section 4, it was argued that both type-of-claim effects and year effects are potentially important to the analysis, but a simple-minded attempt to implement this (for example, by performing separate analyses for each of the six types) runs into difficulties because of the small number of claims of certain types. In this section a Bayesian approach to this problem is proposed, with the types allowed to be different, but linked through a hierarchical model. The approach is first developed without year effects, and then extended to include year effects as well.

The hierarchical structure is as follows:

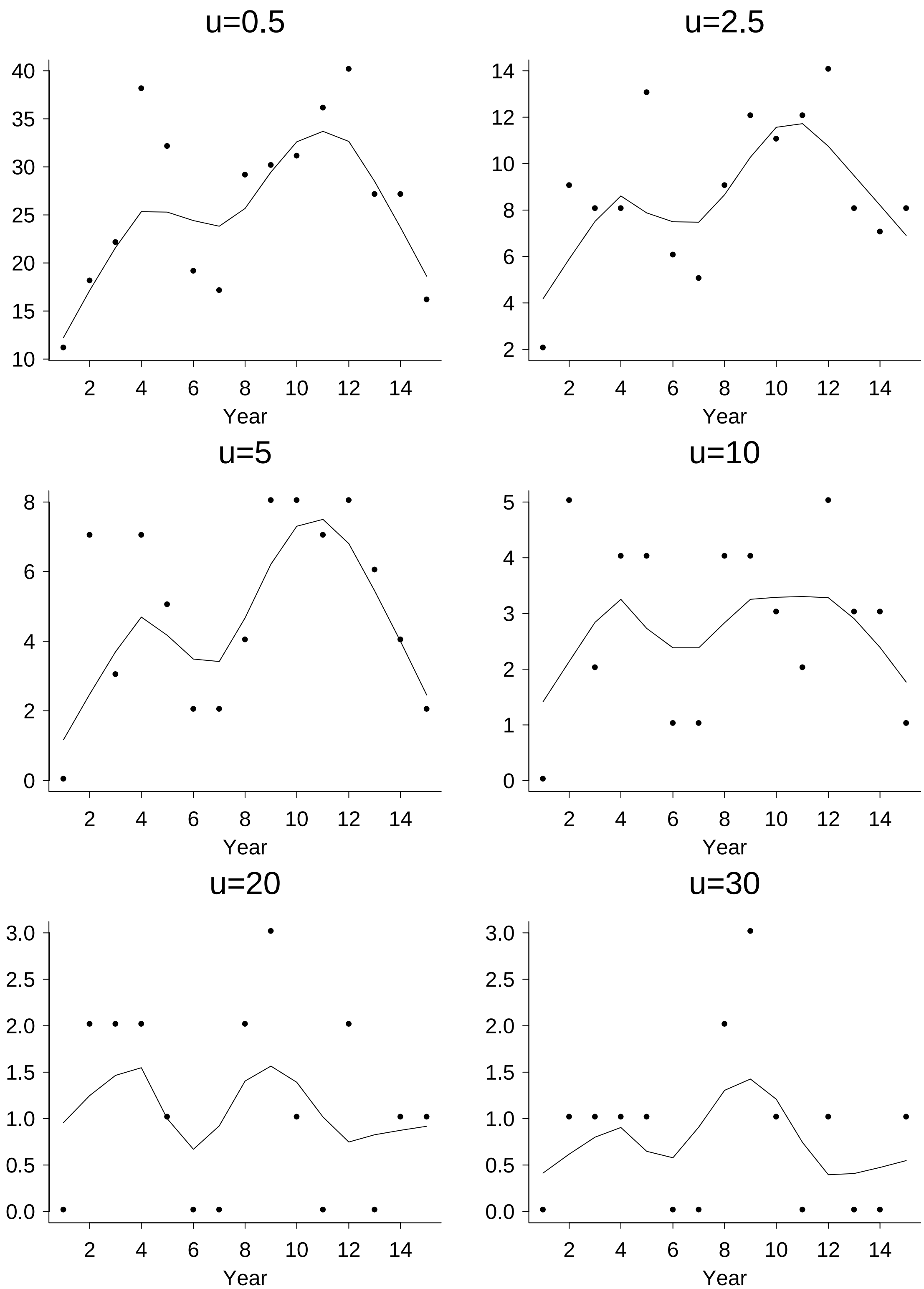

Fig. 4: Numbers of exceedances above various thresholds, plotted by year, with lowess smoothed curves.

Level I. Parameters $m_\mu,\ m_\psi,\ m_\xi, s_\mu^2,\ s_\psi^2,\ s_\xi^2$ are generated from a prior distribution (specified later).

Level II. Conditional on the parameters in Level I, parameters $\mu_1, ..., \mu_J$ (where J is the number of types) are independently drawn from $N(m_\mu, s_\mu^2)$, the normal distribution with mean m_μ, variance s_μ^2. Similarly, $\log\psi_1, ..., \log\psi_J$ are drawn independently from $N(m_\psi, s_\psi^2)$, $\xi_1, ..., \xi_J$ are drawn independently from $N(m_\xi, s_\xi^2)$.

Level III. Conditional on Level II, for each $j \in \{1, ..., J\}$, the point process of exceedances of type j is generated from the Poisson process defined by (2) with parameters $\mu_j,\ \psi_j,\ \xi_j$.

The prior distribution for (m_μ, s_μ^2) is of the well-known "Gamma-Normal" type: let τ_μ be drawn from a Gamma distribution with shape parameter α and scale parameter β and, given τ_μ, define $s_\mu^2 = 1/\tau_\mu$ and let $m_\mu \sim N(\nu, \frac{1}{\kappa\tau_\mu})$. We represent this distribution as $(m_\mu, s_\mu^2) \sim GN(\alpha, \beta, \nu, \kappa)$. Similarly, (m_ψ, s_ψ^2) and (m_ξ, s_ξ^2) are independently drawn from the same distribution. We fix $\alpha = \beta = \kappa = 0.001$ and $\nu = 0$ to represent a proper but very diffuse prior distribution.

It can easily be verified that the GN family is conjugate: given $\mu_1, ..., \mu_J$ and defining $\bar\mu = \frac{1}{J}(\mu_1 + ... + \mu_J)$, the posterior distribution of (m_μ, s_μ^2) is $GN(\alpha', \beta', \nu', \kappa')$ where

$$\begin{aligned}
\alpha' &= \alpha + \frac{J}{2}, \\
\beta' &= \beta + \frac{1}{2}\frac{J\kappa}{J+\kappa}(\bar\mu - \nu)^2 + \frac{1}{2}\sum_j (\mu_j - \bar\mu)^2, \\
\nu' &= \frac{\kappa\nu + J\bar\mu}{\kappa + J}, \\
\kappa' &= \kappa + J.
\end{aligned}$$

Similar formulae apply, for course, for the ψ and ξ parameters.

The estimation procedure is then a Markov chain Monte Carlo (MCMC) simulation algorithm in which the values of the unknown quantities in Levels I and II are updated as follows:

1: (Gibbs step) Given current values of $\{\mu_j, \psi_j, \xi_j,\ j = 1, ..., J\}$, new values of the Level I parameters $m_\mu,\ m_\psi,\ m_\xi, s_\mu^2,\ s_\psi^2,\ s_\xi^2$ are generated from the appropriate posterior distributions.

2: (Hastings-Metropolis step) Given both the observed data and the current values of $m_\mu,\ m_\psi,\ m_\xi, s_\mu^2,\ s_\psi^2,\ s_\xi^2$, each of the parameters $\{\mu_j, \psi_j, \xi_j,\ j = 1, ..., J\}$ is updated by generating a small random perturbation and accepting or rejecting according to the Hastings-Metropolis criterion.

The procedure is repeated for up to 10^6 iterations, as in Section 3. Moreover, at the end of each 100 cycles, a new value for "future loss" is generated conditionally on the current values of μ_j, ψ_j, ξ_j for each j, and the results aggregated over all classes. Thus we obtain a MCMC sample from the predictive distribution, just as in Section 3. Time series plots of the Level I and Level II parameters suggest that the MCMC algorithm settles down to its stationary distribution quickly.

This model may be further extended to include a year effect, as follows. Suppose the extreme value parameters for type j in year k are not μ_j, ψ_j, ξ_j but $\mu_j + \delta_k, \psi_j, \xi_j$. We fix $\delta_1 = 0$ to ensure identifiability, and let $\{\delta_k,\ k > 1\}$ follow an AR(1) process:

$$\delta_k = \rho\delta_{k-1} + \eta_k, \quad \eta_k \sim N(0, s_\eta^2).$$

This can be fitted using the same kind of Gamma-Normal prior as used for the other parameters: if the prior distribution of (ρ, s_η^2) is $GN(\alpha, \beta, 0, \kappa)$, then the posterior given $\delta_2, ..., \delta_K$ is $GN(\alpha', \beta', \nu', \kappa')$ where

$$\begin{aligned}
\alpha' &= \alpha + \frac{K-1}{2}, \\
\beta' &= \beta + \frac{1}{2}\sum_2^K (\delta_k - \rho\delta_{k-1})^2, \\
\nu' &= \frac{\sum_2^K \delta_{k-1}\delta_k}{\kappa + \sum_2^K \delta_{k-1}^2}, \\
\kappa' &= \kappa + \sum_2^K \delta_{k-1}^2.
\end{aligned}$$

To update $(\delta_2, .., \delta_K)$ within level II of the MCMC algorithm, each value in turn is updated, using its conditional distribution given $(\delta_{k-1}, \delta_{k+1})$ (or just δ_{k-1} in the case $k = K$) combined with the likelihood for the observed data in year k given δ_k, in a Hastings-Metropolis updating step.

This analysis was performed using all large claims above threshold $u = 2.5$. Fig. 5 shows crude boxplots of the posterior distributions of the Level II parameters, in which the central black dot for each parameter represents the posterior median for that parameter (taken directly from MCMC output) and the horizontal bars represent the first and third quartiles of the posterior distribution. Plots (a) to (c) show that there is wide variability in the μ_j and ψ_j parameters for the different types, but not for the ξ_j parameters. This suggests that our earlier conclusion, in which the values of ξ_j seemed markedly different for different types, was not correct, but there is a general masking effect causing the common value of ξ to be estimated larger when type effects are ignored than when they are included. In the analysis of section 3, the posterior mean of ξ was 0.955 and the posterior probability of $\{\xi \geq 1\}$ was 0.38. Here, the posterior means of each of the ξ_j parameters are respectively .741, .749, .749, .726, .718, .701, and the posterior probabilities of $\{\xi_j \geq 1\}$ are .045,.077,.075,.062,.059,.052. The distributions are still very long-tailed, but at least we now have some assurance that we are dealing with distributions of finite mean! Recall that claim type j has infinite mean if $\xi_j \geq 1$.

The plots based on $\{\delta_k\}$ in Fig. 5(d) show some common features with the plots in Fig. 4. In particular, there appear to be large values in years 5, 11 and 12, but there is also a large overlap in the posterior distributions for different years, which casts doubt on the statistical significance of the year effects. Another feature is that there is little evidence either of any long-term trend in the δ_k's, or even of persistence from one year to the next — the median values of δ_k drop right back to near 0 in the last three years of the data, while the posterior density of ρ (Fig. 6) is nearly flat over the range $(-0.5, 1)$, suggesting that we could just as well have taken $\rho = 0$. A side comment here is that ρ was not constrained *a priori* to lie in the interval $(-1, 1)$, since in this kind of analysis there is no need to make a stationarity assumption. Nevertheless, nearly all of the posterior mass lies within this range, which adds further reassurance that we are dealing with a process without long-term trends.

Thus our conclusions as far as year effects are concerned is that there are reasons to think that the 'blip' around years 11-12 was genuine, but it does not persist into subsequent years, and there is no evidence of any long-term trend, to the extent that the scope of this data set allows us to judge that.

Fig. 7 shows the "loss curves" (predictive distributions of future loss suffered over a one-year period) under various model assumptions. The seven curves plotted are based on:

A. Homogeneous model of Section 3, using threshold $u = 5$. (This is the average of the four curves in Fig. 3(d).)

B. Same as A, but based on $u = 2.5$ (to be consistent with later hierarchical analyses). There is little difference between curves A and B, which is reassuring because it shows that the largely arbitrary choice of threshold does not greatly influence the results.

C. Same as A, but omitting the two largest observations. This curve is well below A and B, showing that the outliers are certainly "influential observations" even if they are not genuine outliers, but it still represents a very long-tailed distribution. The choice between A and C depends on whether it is desired to include "total losses" (in the sense in which this expression was used in Section 1) in the assessment or not. There appear to be no statistical grounds for excluding the outliers, but this is based on just two "total losses" in the data set, and more data on these may suggest that it is unreasonable to aggregate both total and partial losses into a single distribution.

D. Hierarchical model, $u = 2.5$, no year effects, outliers left in the data.

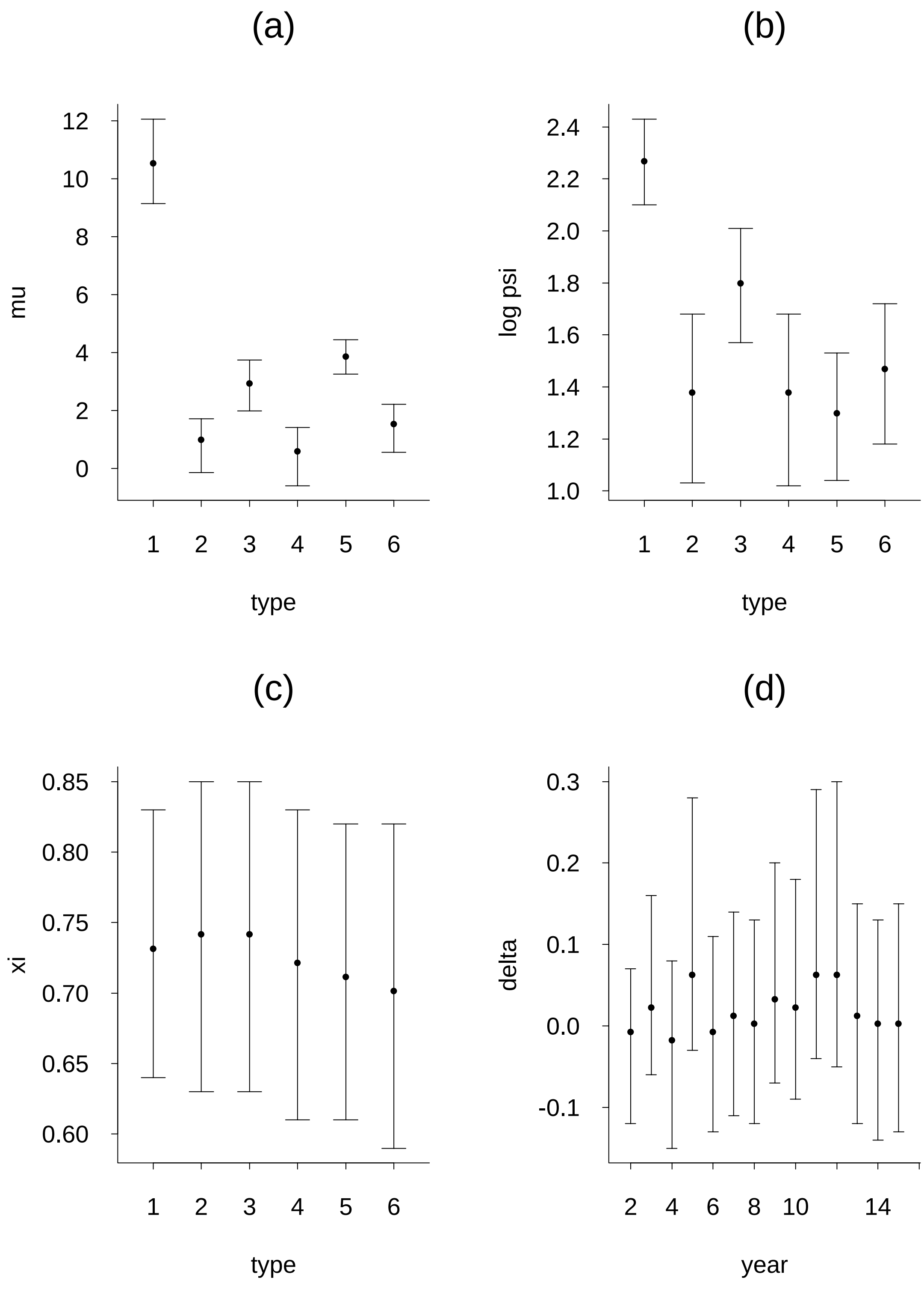

Fig. 5: Boxplots of posterior distributions of Level II variables. For each variables, the black dot represents the median of the posterior distributions and the horizontal bars the first and third quartiles.

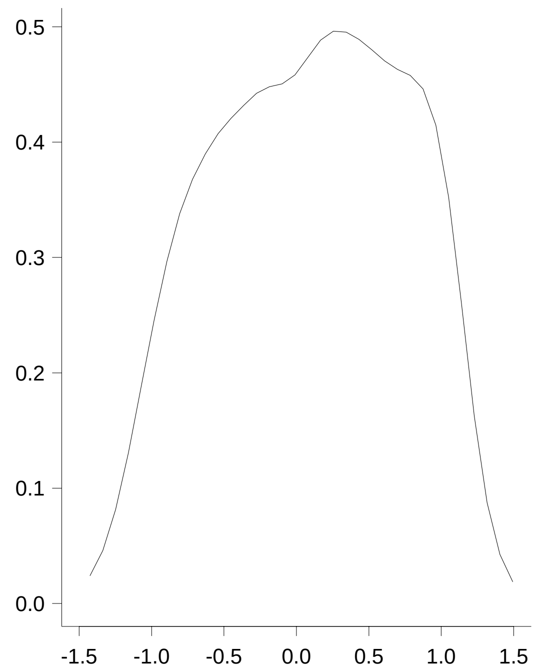

Fig. 6: Posterior density of ρ

E. Hierarchical model, $u = 2.5$, no year effects, outliers omitted.

F. Hierarchical model with year effects, $u = 2.5$, outliers left in the data.

G. Hierarchical model with year effects, $u = 2.5$, outliers omitted.

Based on the close proximity of curves D and F, there is little evidence that year effects are influencing our conclusions. However, there is a big gap between B and D (or F) implying that allowing for type of claim does have a marked effect on our risk assessment. As already remarked, it appears that if type effects are ignored, then the variability of scale between different types of claims has a masking effect causing the estimate of ξ to be increased, resulting in a more pessimistic risk assesment. The relationship between curves D and F with the outlier-removed versions E and G is similar to the relationship between A and C: remove the outliers, and the assessed risk goes down, but at the cost of ignoring "total losses" completely.

If we decide not to exclude the outliers, then the overall recommendation would be to use curve F on the basis that it includes the type-of-claim effects which do appear to be important, and that even though we are less sure about the significance of the year effects, it does appear to be desirable to take these into account having seen some evidence of their importance.

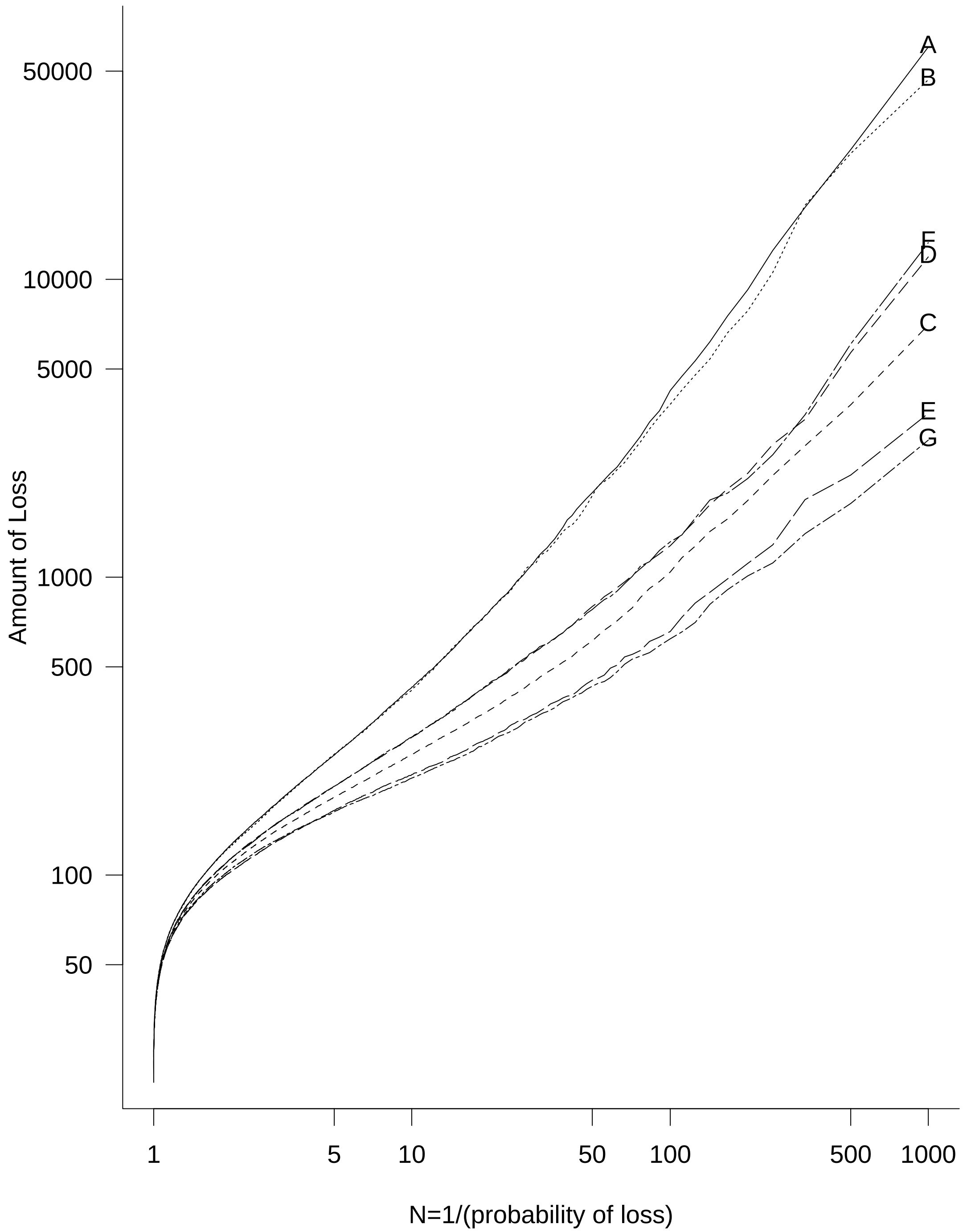

Fig. 7: Predictive loss curves under seven different models. See text for details of curves A–G.

Fig. 8 shows curve F again plotted for 1-year losses, and the same curve recomputed for 5-year and 10-year losses. One thing to note here is that the three curves do not obey a simple probability-scaling relationship. For example, the 1-year loss sustained with probability $\frac{1}{100}$ (1310) is quite different from the 10-year loss sustained with probability $\frac{1}{10}$ (2820). In conventional terminology these would both be called the "100-year loss" and, in the absence of trend or strong serial dependence, considered equivalent. The difference is explained by the fact that we are considering, not the fitted distribution with estimated parameters, but a predictive distribution in which the effect of unknown parameters is accounted for by integration over the posterior distribution. This departs not only from conventional actuarial practice, but also from more traditional methods of looking at extreme values, such as the "POT" method and its variants (Davison and Smith 1990).

A final comment is that a "loss curve" defined with respect to the sum of all losses over a fixed period of time is by no means the only way, or even necessarily the most sensible way, of assessing the impact of future large insurance claims on a company. It would be possible to extend the methodology in this paper to compute the predictive distribution of loss under a variety of assumptions about, for example, future reinsurance or investment strategies by the company. Ultimately this could be used to compare the effect of different strategies for managing large insurance claims.

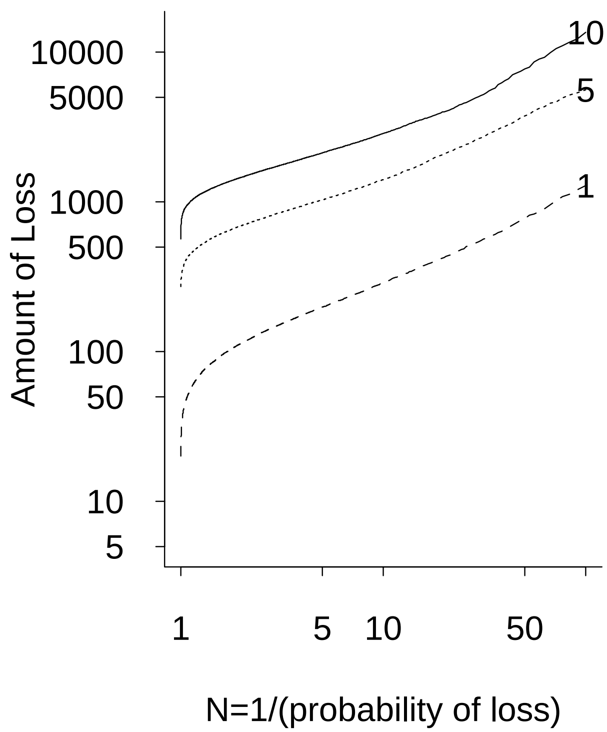

Fig. 8: Predictive loss curve using hierarchical model (curve F in Fig. 7) recomputed with respect to losses incurred over a 10-year (top curve), 5-year (middle) and 1-year (bottom) time horizon.

6 Conclusions

The methods described in this paper represent a completely new approach to the assessment of actuarial risks. Models from extreme value theory, previously employed in such contexts as the assessment of extreme floods or extreme air pollution episodes, are combined with modern statistical methodology based on Bayesian inference, hierarchical models, and Monte Carlo sampling. This allows us to compute a "loss curve" which represents our best estimate of the probability of a loss of a given size over a given time span, taking uncertainty of the model parameters into account.

The results confirm that the distribution of claims is very long-tailed, though when the type-of-claim effect is taken into account, the tail is less extreme ($\xi \approx 0.7$) than when it is not ($\xi \approx 0.9$). This has a substantial effect of the loss curve, which suggests the possibility that if we had even more ancillary information about the claims, we might come up with even more precise assessments of the individual distributions which could result in a further reduction of the assessed risks. When the analysis is further extended to include year effects, we see that there is an apparent peak in the frequency of claims around years 11 and 12, which is probably not a spurious effect, because there was a known change of policy towards insurance at about this time. However, there is no evidence of any long-term trends, and including the year effects makes little difference to the estimated loss curve. An intriguing possibility for future research is to extend the year effects to incorporate other factors, such as environmental variables. This could be a way of exploring the possible connection between patterns of insurance claims and long-term environmental effects such as global warming.

Finally, the whole analysis is heavily influenced by the two very large claims that correspond to "total losses", though the evidence in the data is that these are no more extreme than would be expected of the largest two observations, given the long-tailed nature of the entire distribution. Thus it is argued that there is no *statistical* reason to treat these as separate from the rest of the distribution, though there may of course be other reasons why the insurance industry would find it unreasonable to aggregate these with the partial losses into a single overall distribution.

References

Davison, A.C. & Smith, R.L. (1990), Models for exceedances over high thresholds (with discussion). *J.R. Statist. Soc. B* **52**, 393-442.

Pickands, J. (1975), Statistical inference using extreme order statistics. *Ann. Statist.* **3**, 119-131.

Smith, A.F.M. and Roberts, G.O. (1993), Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J.R. Statist. Soc. B* **55**, 3-23.

Smith, R.L. (1989), Extreme value analysis of environmental time series: An application to trend detection in ground-level ozone (with discussion). *Statistical Science* **4**, 367-393.

Smith, R.L. and Shively, T.S. (1995), A point process approach to modeling trends in tropospheric ozone. *Atmospheric Environment* **29**, 3489-3499.