Chapter 4
Discussion and Conclusions

We have demonstrated on several data sets related to natural disasters of various nature that using logarithms of the original observations is more appropriate for fitting of heavy tails. By doing so, power-like tails (in particular those obeying the Pareto law with an arbitrary index) are transformed into exponential tails, and the corresponding GPD form parameter becomes non-positive. Zero value of the GPD form parameter corresponds to the exponential tail, whereas its negative values correspond to a distribution with a finite end point M_{max}. Tails heavier than any power-like tail are not frequently encountered in practice, so for the log-transformed data it is sufficient to consider GPDs with non-positive indexes. Thus, the peak-over-threshold distributions of log-sizes of events are best approximated by the GPD with a negative parameter (see Tables 4.1, 4.2). The density function of such distributions takes very small values at the approach of its final point M_{max}, which results in a "duck beak" shape, see Fig. 2.2. For instance, the limit behavior of probability density function of earthquake magnitudes taken from the Harvard catalog is best approximated by the following power law: $(M_{max} - x)^{-1-1/\xi} \cong (M_{max} - x)^{5.14}$. This fact explains in particular the origin of unstable statistical estimates of the parameter M_{max}: smalls changes in earthquake magnitudes can result in significant fluctuations of the corresponding estimates of M_{max}. In contrast, estimates of the integral parameter $Q_\tau(q)$ are typically stable and robust, as we have demonstrated above.

We would like to emphasize that a reliable estimation of quantiles of levels $q > 1 - 1/n$ can be obtained only with some additional assumptions on the behavior of the distribution's tail. Sometimes, such assumptions can be made on the basis of physical processes behind the studied phenomena. Here we have used for this purpose certain theorems of the extreme value theory (EVT). In our case, these EVT based assumptions boil down to assuming a regular behavior of the tail $1 - F(m)$ of the distribution of sizes of events in the vicinity of its rightmost point M_{max}. It should be noted that the assumptions regarding the asymptotic behavior of the distribution's tail cannot equally apply to all practical cases, and they should be supported by additional information for each particular studied phenomenon. In fact, the EVT suggests a statistical methodology for the extrapolation of quantiles

V. F. Pisarenko and M. V. Rodkin, *Statistical Analysis of Natural Disasters and Related Losses*, SpringerBriefs in Earth Sciences,
DOI: 10.1007/978-3-319-01454-8_4, © The Author(s) 2014

beyond the data range; whether such an extrapolation is justified should be thoroughly investigated in each particular case. In our view, the EVT provides us with the best statistical approach to this problem.

Application of the EVT to different extreme events data is reduced to fitting of the GPD to the tail of the corresponding distribution of event sizes or their logarithms. According to the EVT, the Generalized Pareto Distribution is the only possible limit distribution for the "peaks over threshold" events. GPD is a flexible two-parametric family of densities with well-known statistical properties. In certain cases however, even the GPD fails to reasonably approximate the distribution's tail. This may happen in a case when the Limit Theorem of the EVT is inapplicable to a particular data set, since the behavior of the sample's DF in the extreme range cannot be described by a single asymptotic function. For example, it may switch from a power-law like behavior for a certain range of values to an exponential one for the next range of values. In such cases, we have no well defined criteria to choose the value of the threshold for "the peaks over threshold" method, and the application of the exposed approach is not recommended.

Tables 4.1 and 4.2 summarize the main characteristics of the natural disasters analyzed above, together with the parameters of the corresponding fitted GPDs. The first column of Table 4.1 we indicates whether the log-transform was applied to the original values. The third column contains the estimates of the form parameter of the GPD. In two cases the form parameter is null, which corresponds to the exponential distribution (exponential distribution is the limit case of the GPD when $\xi \to 0$). In all the other presented cases, the form parameter estimates are negative, which indicates the finiteness of the corresponding distributions.

In the fourth column we give the p values which represent the probability to exceed the discrepancy between the observed and the fitted distributions, also known as, the Kolmogorov distance. We consider that if the p value is less than 0.1 one has grounds to reject the fitted curve). One can see that the GPD approximates reasonably well the extreme parts of the distribution's tail for all the considered catalogs of natural disasters. Only in one case (fatalities from floods in USA, 1995–2011) the p value is less than 0.4 which indicates a poor quality of fit. There are two cases (economic losses resulting from floods in USA) when the p value equals 0.90 which corresponds to a very close approximation.

As discussed above, the absolute value $|\xi|$ indicates the steepness of decrease of the extreme part of the distribution's tail. According to Tables 4.1 and 4.2, the steepest extreme tails are observed for the economic losses produced by floods and hurricanes, whereas the corresponding fatality and the injured/affected distributions have, as a rule, smaller parameter $|\xi|$, which corresponds to a slower decay of the tail. As was previously noted, the (unlimited) exponential distribution of $\log(x)$ corresponds to the (unlimited) Pareto distribution of x. This situation occurred once (the last row of Table 4.1) for the case of tornado related fatalities in USA. It is obvious that the maximum number of fatalities in any disaster is limited, however in that particular case a more accurate statistical approximation is observed for an unlimited model.

Table 4.1 Characteristics of disasters and parameters of fitted GPD law

	Lower threshold h, sample size n, intensity λ (1/year)	Form parameter ξ	Goodness-of-fit (p value)	Maximum observed effect	Quantile $Q_{0.95}(10)$
Seismic moment magnitude m_W, Harvard catalog 1976–2012	$m_W \geq 6.8$ $n = 324$ $\lambda = 8.80$	–0.163 ± 0.076	0.59	$m_W = 9.1$	9.13
Earthquake fatalities, Japan, 1900–2011 lg(x)	$h = 3$ persons $n = 44$ $\lambda = 0.339$	–0.260 ± 0.111	0.43	142,807 persons	58,000 persons
Injured in earthquakes, Japan, 1900–2011 lg(x)	$h = 3$ persons $n = 99$ $\lambda = 0.884$	–0.374 ± 0.063	0.69	103,733 persons	75,000 persons
USA, perished in floods, 1995–2011 x	$h = 3$ persons $n = 41$ $\lambda = 1.11$	0.0	0.22	35 persons	53 persons
Affected in floods, USA, 1995–2011 lg(x)	$h = 500$ persons $n = 52$ $\lambda = 3.06$	–0.182 ± 0.113	0.66	11,000,148 persons	17,700,000 persons
USA, estimated economic losses from floods, 1995–2011 in millions of $, lg($x$)	$h = 80$ $n = 32$ $\lambda = 1.88$	–0.486 ± 0.091	0.90	12,000	12,400
USA, perished in tornadoes, 1953–2012 lg(x)	$h = 20$ persons $n = 53$ $\lambda = 0.88$	0.0	0.75	1,200 persons	1,480 persons

Table 4.2 Characteristics of annual disasters and form parameter of fitted GPD-law

	Lower log-threshold h (10^h) sample size n	Form parameter ξ	Goodness-of-fit (p-value)	Maximum observed effect, lg(x) (x)	Quantile $Q_{0.95}(10)$ (10^Q)
Annual economic losses from floods in USA, in 10^9 $, 1940–2011 lg($x$)	h = 0.4 (2.5) n = 48	–0.354 ± 0.093	0.90	lg(x) = 1.71 (51.3)	1.66 (45.7)
Annual economic losses from hurricanes in USA, 1940–2010 in 10^6 $, lg($x$)	h = 1.5 (31.6) n = 64	–0.636 ± 0.045	0.48	lg(x) = 5.15 (141,000)	5.09 (123,000)

One can observe that in certain cases the quantile $Q_{0.95}(10)$ is less than the observed maximum event size, while in certain other cases it exceeds that value.. This is a result of an interplay between the parameters of the fitted GPD, namely intensity λ and time interval τ. It should also be remarked that such characteristics as economic losses resulting from natural disasters are strongly influenced by a rapid global development of the economic infrastructure and the population growth. Therefore, it is quite difficult to reliably forecast such characteristics for long time spans, say beyond 10–15 years. This remark should be kept in mind when one estimates quantiles of future losses.

Table 4.2 summarizes the results of the analysis of annualized data. The aggregation of event sizes over one year intervals represents in essence a linear filtration (smoothing) of the corresponding time series of sizes. That is why the tails of annualized distributions are as a rule less heavy compared to the tails of original distributions of marked point processes. This fact can explain higher values of the form parameter (in terms of its absolute value) of annualized distributions in Table 4.2 compared to the corresponding form parameters in Table 4.1. One exception is the case of the economic losses from floods, which can be explained by a very small sample size in this case: $n = 32$ (single event losses) and $n = 48$ (annualized losses). We remind that the theoretical maximum M_{max} of the GPD distribution with negative form parameter ξ is expressed as

$$M_{max} = h - \frac{s}{\xi},$$

and the lesser $|\xi|$ the larger M_{max} is.

One can also note, that the correlation between the high quantile $Q_{0.95}(10)$ and the maximum observed size is stronger for the annualized data, as it could be expected.

Table 4.3 Ratio of sum of 10 % largest effects to total sum

	Ratio of sum of 10 % largest effects to total sum (%)
Affected in floods, USA, 1995–2011	98
Earthquake fatalities, Japan, 1900–2011	98
Injured in earthquakes, Japan, 1900–2011	94
Annual economic losses from hurricanes in USA, 1940–2010	70
USA, estimated economic losses from floods, 1995–2011	68
USA, perished in tornadoes, 1953–2012	60
USA, perished in floods, 1995–2011	55
Annual economic losses from floods in USA, 1940–2011	40

We gave in Chap. 1 theoretical relations (1.3)–(1.4) connecting the sample maximum $M_{\max}^{(n)} = max(x_1, \ldots, x_n)$ with the total sum $S_n = x_1 + \cdots + x_n$. We can as well compare S_n with the sum of k largest observations. The ratio of such sums for the analyzed catalogs is presented on Figs. 3.27, 3.35, 3.42, 3.50, 3.59, 3.64, and 3.70. These ratios reflect in a more in detailed manner the contributions of the rightmost part of tail to the total sum. Let us consider for comparison one particular value on these curves, namely the ratio of 10 % of the largest observations to the total sum. One can say, that the higher this ratio, the more events are concentrated around the tail's extreme range. Table 4.3 presents a collection of such ratios for all the considered event catalogs. One can conclude that the highest concentration of events around the distribution's tail is observed for the data sets related to the number of individuals affected by floods (USA), to earthquake fatalities (Japan) and to the injured by earthquakes (Japan). For these cases, 10 % of the largest events are responsible for more than 95 % of the total loss. Intermediate values of the event concentration toward the tail's end (about 60–70 %) are observed for annualized economic losses from hurricanes (USA), economic losses from floods (USA) and fatalities from tornadoes (USA). Weak concentration (40–55 %) is observed for flood fatalities (USA) and annualized economic losses from floods (USA). It should be noted, that our concentration graphs are in essence an extended analog of the Pareto principle (or the 80-20 rule): "for many phenomena roughly 80 % of the effects come from 20 % of causes" (Italian economist Vilfredo Pareto observed in 1906 that 80 % of the land in Italy was owned by 20 % of population).