Statistical Theory of Extremes
Homepage: http://www.gathacognition.com/book/gcb14
http://dx.doi.org/10.21523/gcb1

Part 5
Complements

Chapter 17
Some Case Studies

J. Tiago de Oliveira
Academia das Ciências de Lisboa (Lisbon Academy of Sciences), Lisbon, Portugal.

Abstract

This chapter deals with the statistical analysis of data considered in the parts: univariate extremes data and multivariate extremes data related to stochastic processes and sequence of extremes connected to multivariate extremes data. The case studies of maximum wind speed data in Lisbon, flood discharges of the North Saskachevan River at Edmonton and flood discharges of the Fox River at Berlin and Wrightstown are analysed.

Published Online

23 June 2017

Keywords

Univariate extremes data;
Multivariate extremes data;
Multivariate extremes;
Exceedance probabilities.

Editor(s)

J.C. Tiago de Oliveira

17.1 Introduction

We will deal here with the statistical analysis of data considered in Parts 2 (univariate extremes data) and 3 (multivariate extremes data); the situation related to stochastic processes and sequence of extremes (still in its infancy) must, in general, be connected to multivariate extremes data and was, in part, sketched in the Chapters of Part 4 and Annex 6 and Chapter 16 (*GEA*).

We will begin by analysing some univariate data and then, presupposing the analysis technique, deal with bivariate and multivariate data. We will not of course detail the computation of the estimators of parameters, of quantiles and of exceedance probabilities because they follow directly from the formulae and examples given in Parts 2 and 3.

Originally published in 'Statistical Analysis of Extremes', 1997, 2016
http://dx.doi.org/10.21523/gcb1.1728

17.2 Maximum wind speed data in Lisbon

The data that follow correspond to maximum wind speeds in Lisbon between 1941 and 1970[1]. The data for exploratory and analytical study are contained in Table 17.1.

Table 17.1

Year	Km/h	Year	Km/h	Year	Km/h
1941	129	1951	96	1961	86
1942	117	1952	72	1962	91
1943	100	1953	98	1963	96
1944	100	1954	85	1964	89
1945	132	1955	124	1965	90
1946	94	1956	108	1966	89
1947	108	1957	102	1967	89
1948	113	1958	102	1968	84
1949	96	1959	112	1969	107
1950	113	1960	107	1970	111

a) Quick exploration of extremes data:
The calculation of the statistic

$$Q = \frac{x'_n - x'_{[n/2]+1}}{x'_{[n/2]+1} - x'_1} = \frac{x_{max} - x_{med}}{x_{med} - x_{min}},$$

as $x_{max} = 132,\ x_{med} = 100\,, x_{min} = 72$ gives

$$Q_{30} = 1.1428571.$$

As for $\theta = 0$ we have, for $n = 30$,

$$\beta_{30,0} = \frac{\log 30 + \log(\log 2)}{\log\log 30 - \log\log 2} = 1.9078381$$

and $\alpha_{30,0} = \frac{1}{\log\log 30} = .8169083\,,$
we see that

1. The author thanks the Portuguese Institute for Meteorology and Geophysics for having kindly supplied the data.

$\frac{Q_{30}-\beta_{30,0}}{\alpha_{30,0}} = -.9364343.$

Recall that $(Q_n - \beta_{n,0})/\alpha_{n,0}$ has asymptotically a reduced Gumbel distribution if the i.i.d. sample has the same distribution we see that $(Q_{30} - \beta_{30,0})/\alpha_{30,0}$ is between the bounds for shortest 95% - probability interval for the Gumbel distribution and therefore we can, initially, assume it as the approximate distribution of maxima.

For the plotting we must know $x'_{[.367n]+1} = x'_{12} = 96$ and $x'_{[.692n]+1} = x'_{21} = 108$ and so the plot in the statistical choice modification of Gumbel probability papers gives the graph (Figure 17.1) that follows, agreeing with the use of the Gumbel model.

b) Analytical choice of the model:
Once having seen a general trend for the interpretation of data as (approximately) described by the Gumbel distribution, let us confirm (or reject) it by analytic study.

The ML estimators of (λ, δ) are

$\hat{\lambda} = 94.71, \hat{\delta} = 12.49,$

and so the statistical choice statistic has the value

$\hat{V}_{30} = -9.63$.

Consequently, as $\hat{V}_n/\sqrt{2.09797 \times n}$ has the value,

$\hat{V}_{30}/\sqrt{2.09797 \times 30} = -1.21,$

which is in the acceptance region for the asymptotically normal distribution of $\hat{V}_n/\sqrt{n\,\hat{\sigma}_0^2}$, thus confirming the exploratory analysis. For the value of $\hat{V}_n/\sqrt{n\,\hat{\sigma}_0^2} = -1.21$, obviously, we do not need to go to the refined bounds which are, for $\alpha = .05$ and to the order $O(n^{-1})$, $b_{30} = -1.71$ and $a_{30} = 2.91$, confirming the decision.

It would be useful to divide the sample into two parts (about 2/3 and 1/3) and confirm the results.

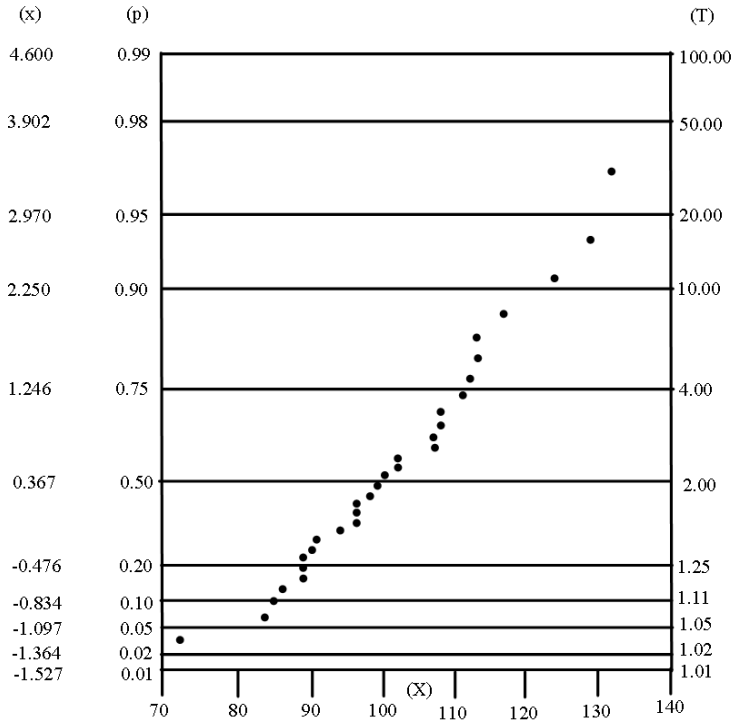

Figure 17.1. Plot of maximum wind speed at Lisbon (1941-70)

To a certain extent this was done in the papers by Tiago de Oliveira (1981) and Fransén and Tiago de Oliveira (1984): in the former, data for greatest ages of death for men and for women in Sweden are studied for the years 1905/1958, and in the latter corresponding data for a larger set of years (1905/1970); and the results are coherent.

Other cases have been considered in the paper by Fransén and Tiago de Oliveira (1984) and can be considered as exercises. We will only follow another case leading to a different decision — the choice of the Fréchet distribution.

It would also be useful to read Whitmore *et al*. (1987) where data for maximum wind speeds are analysed for Canada.

17.3 Flood discharges of the North Saskachevan River at Edmonton

In this case we consider not only the raw data but, as a consequence of the rejection of the Gumbel distribution as a possible model, their logarithms. The ordered raw data, for a period of 47 years in $1000\ \mathrm{ft}^3/\mathrm{s}$, is[2] (Table 17.2):

Table 17.2

i	x_i	i	x_i	i	x_i	i	x_i
1	19.885	13	30.380	25	40.400	37	61.740
2	20.940	14	31.500	26	42.250	38	65.440
3	21.820	15	32.600	27	44.020	39	65.597
4	23.700	16	32.680	28	44.730	40	66.000
5	24.88	17	34.400	29	44.900	41	74.100
6	25.460	18	35.347	30	46.300	42	75.800
7	25.760	19	35.700	31	50.330	43	84.100
8	26.720	20	38.100	32	51.442	44	106.600
9	27.500	21	39.020	33	57.220	45	109.700
10	28.100	22	39.200	34	58.700	46	121.970
11	28.600	23	40.000	35	58.800	47	185.560
12	30.200	24	40.400	36	61.200		

a) Quick exploration of extremes data:

We will avoid the graphical analysis, which is as before, but will only consider the statistic

$$Q_{47} = \frac{x_{max} - x_{med}}{x_{med} - x_{min}}$$

which as $x_{max} = 185.560$, $x_{med} = 40.400$ and $x_{min} = 19.885$ takes the value $Q_{47} = 7.0757982$; we have $\beta_{47,0} = 2.0317188$ and of $\alpha_{47,0} = .7417784$. So

$$\frac{Q_{47} - \beta_{47,0}}{\alpha_{47,0}} = 6.7999815\ ,$$

2. Data taken from van Montfort (1970).

thus rejecting the acceptance of the Gumbel distribution as an approximate model.

b) Analytical choice of the model:

The values of $(\hat{\lambda}, \hat{\delta})$ are $\hat{\lambda} = 38.151$ and $\hat{\delta} = 17.740$, thus giving $\hat{V}_{47} = 37.77531$ and then $\hat{V}_{47}/\sqrt{2.09797 \times 47} = 3.804$, outside the acceptance region of the asymptotically normal distribution of $\hat{V}_{47}/\sqrt{n\hat{\sigma}_0^2}$ at a level even smaller than 2%. But this value leads to the Fréchet distribution in the statistical choice procedure.

c) Analysis of the logarithms of raw data:

The value of $(Q_{47} - \beta_{47,0})/\alpha_{47,0}$ and of $\hat{V}_{47}/\sqrt{2.09797 \times 47}$ suggest the relevance of trying the Fréchet distribution. Denoting by $'x = \log x$ we get $'x_{max} = 5.2234$, $'x_{med} = 3.6988$ and $'x_{min} = 2.9900$ and $'Q_{47} = 2.1509594$ and so $('Q_{47} - \beta_{47,0})/\alpha_{47,0} = .1607495$, which leads us to accept the Gumbel distribution for $\log x$ and thus suggests the use of the Fréchet approximation.

Let us confirm this analytically. We have $'\hat{\lambda} = 3.5489$, $'\hat{\delta} = .3963$, $'\hat{V}_{47} = -1.95849$ and $'\hat{V}_{47}/\sqrt{2.09797 \times 47} = -1.197$, thus leading to the acceptance of Gumbel distribution for the logarithms of raw data and thus the Fréchet distribution for raw data, as seen in (b).

17.4 Flood discharges of the Fox River at Berlin and Wrightstown

In this analysis we will use flood data the Fox river at two points of its course: Berlin (upstream) and Wrightstown (downstream), both in Wisconsin, USA[3]. The set of data of $n = 33$ pairs of observations is given in Table 17.3.

We can, evidently, expect dependence between the yearly flood discharges at Berlin (B) and Wrightstown (W). The graphs on the probability paper show that the Gumbel distribution fits well to the observed marginal flood discharges. The ML estimators are $\hat{\lambda}_B = 3.21070, \hat{\delta}_B = 1.33880$, $\hat{\lambda}_W = 10.89643$ and $\hat{\delta}_W = 4.61371$ and the empirical Kolmogoroff-Smirnov statistics are $KS_B = .07462$ and $KS_W = .09957$ with $\sqrt{n}\, KS_B = .429$ and

3. Data taken from Gumbel and Mustafi (1967) .

$\sqrt{n}\,KS_W = .572$ accepting thus the Gumbel model for the margins. Evidently the better analysis is to use the $\widehat{V}_n$ statistics for each margin.

Table 17.3

Year	Berlin 1000 ft^3/s	Wrightstown 1000 ft^3/s	Year	Berlin 1000 ft^3/s	Wrightstown 1000 ft^3/s
1918	6.05	16.3	1935	4.34	11.1
1919	2.67	13.1	1936	4.34	6.3
1920	5.15	16.6	1937	3.26	13.5
1921	2.45	14.2	1938	6.19	18.0
1922	5.95	20.1	1939	4.91	18.2
1923	6.05	13.7	1940	4.72	17.5
1924	4.02	15.5	1941	3.54	16.6
1925	2.52	8.3	1942	2.74	19.8
1926	3.44	9.1	1943	5.08	21.3
1927	3.17	13.3	1944	2.29	10.8
1928	5.92	15.1	1945	3.46	15.8
1929	6.62	20.6	1946	6.90	21.3
1930	3.00	6.6	1947	3.16	11.0
1931	1.14	3.1	1948	4.54	10.3
1932	1.91	9.9	1949	2.00	6.4
1933	2.60	8.9	1950	4.63	10.9
1934	1.91	6.7			

The elements for our statistical decision are the values of the estimated correlation coefficient $\rho^* = .693$, the medians $\tilde{\mu}_B = 3.54$ and $\tilde{\mu}_W = 13.5$ and the number $N = 24$ of pairs (x_i, y_i) such that $x_i > \tilde{\mu}_B$ and $y_i > \tilde{\mu}_W$ (first quadrant) or $x_i < \tilde{\mu}_B$ and $y_i < \tilde{\mu}_W$ (third quadrant). This value of N is not coincident with the total of the first and third quadrants (22) given in Gumbel and Mustafi (1967) because the authors, instead of computing the medians directly from the sample, estimate the margin parameters and from them compute estimated medians. The test of independence, as $\sqrt{n}\,\rho^* = \sqrt{33} \times .693 = 3.981$ is larger than the values $\lambda_{0.05} = 1.64$ and $\lambda_{0.01} = 2.33$, leads to rejection of independence at level 1%. Consequently we will have to estimate θ.

Let us consider first the mixed model. As the larger value of ρ^* in the mixed case is 2/3, we have no point estimator of θ by this method (the best being evidently $\theta^* = 1$, allowing for sampling errors).

The relation $p^{**} = p_M(\theta^{**}) = 2^{-2+\theta^{**}/2} = \frac{N}{2n} = \frac{24}{2\times 33} = .36364$ gives $\theta^{**} = 1.081$ and thus is not solvable for θ^{**} as we should have $0 \le \theta^{**} \le 1$; also note that the maximum value of $p_M(\theta)$ is $p_M(1) = .35355 < \frac{24}{2\times 33} = .36364$. The mixed model must therefore be rejected.

Consider now the logistic model. As $0 \le \rho_L(\theta) \le 1$ the equation $\rho_L(\theta^*) = \theta^*(2-\theta^*) = .693$ can be solved giving $\theta^* = 446$. As $p_L(\theta) = 2^{-2^{1-\theta}}$, the equation $p^{**} = p_L(\theta^{**}) = \frac{N}{2n} = .36364$ gives $\theta^{**} = .4546$, which is close to the previous estimate, justifying to a certain extent the use of the logistic model.

Note that this is a margin (location-dispersion) parameter-free approach. As said, Gumbel and Mustafi (1967) used estimators of the margin parameters and computed the *estimated* reduced difference, comparing it, by the Kolmogorov-Smirnov test, with the logistic and confirming the suggestion to use the logistic model as an approximation.

Recall that estimation of the margin parameters involves as yet unsolved problems with the use of the Kolmogorov-Smimov test, although it gives a hint for the logistic model, as stated.

The 5% asymptotic confidence interval is given by

$|p_L(\theta) - p^{**}| \le 1.96\sqrt{p^{**}(.5-p^{**})}/\sqrt{n}$ which is, in our case,

$|p_L(\theta) - .36364| \le .07598$ or $.28766 \le p_L(\theta) \le .43962$ and so

$.154 \le \theta \le .754$, a very large (asymptotic) confidence interval.

Let us now consider the biextremal model. The equation $\rho_B(\theta^*) = .693$ gives $\theta^* = .55$. The solution from $p_B(\theta^{**}) = 2^{-2+\theta^{**}} = .36364$ is $.541$; the approximate confidence interval for the significance level 5% is $|p_B(\theta) - .36364| \le .07598$ is $.202 \le \theta \le .814$.

For the Gumbel model the equation $\rho_G(\theta^*) = .693$ gives $\theta^* = .723$ and the confidence interval from $p_G(\theta) = p_B(\theta)$ is also $.202 \leq \theta \leq .814$, which suggests that the Gumbel model is not appropriate to this problem, the solution of $\rho(\theta^*) = .6846$ being close to the upper bound of the confidence interval; in fact the model arises naturally in questions connected with failures. A simple check can be the estimation of θ from the use of Kendall's $\tau_G = \theta/(2-\theta)$.

The question of discriminating between models, as for the remaining ones (logistic, biextremal) in this example is, as yet, an unsolved problem. Here, though, the solution is simple.

As in the biextremal model we have for the reduced values $Y = X$ with probability θ and $Y > X$ with probability $1-\theta$, we can expect that, roughly, $n\,\theta$ observations (x_i, y_i) should be in a straight line $y = ax + b$ and rest should be such that $y > ax + b$. In our case, as $\theta^* = .541$ and $n = 33$ we should have roughly $.541 \times 33 \approx 18$ points in a straight line; the scatter diagram given in Gumbel and Mustafi (1967) shows immediately that this is not the case. Of the four models considered, the only one not yet rejected is, for the moment, the logistic. Also the plotting of *estimated* reduced differences on the logistic paper, given also in Gumbel and Mustafi (1967), speaks in favour of the logistic model.

The application of the strip method — see Posner *et al*. (1969) referred to in Chapter — to the logistic model as a flood model for the Fox River confirms its usefulness.

It should be remarked here that if we *expect* bivariate models to be absolutely continuous here, as in other cases, the logistic model seems to be a good candidate. Also, as shown in Tiago de Oliveira (1987) the discrepancy between the logistic and $(2, N)$ natural model is small and not easily seen for such small samples. The intrinsic estimation will also not give large discrepancies.

References

Fransén, A. and Tiago de Oliveira, J., 1984. Statistical choice of univariate extreme models, part II, *in Statistical Extremes and Applications*, Tiago de Oliveira, J, ed., 373-394, D. Reidel, Dordrecht.

Gumbel, E. J. and Mustafi, C. K., 1967. Some analytical properties of bivariate extremal distributions. *J. Amer. Statist. Assoc.*, 62, 569-588.

Posner, E. C., Rodemich, E. R., Ashlock, J. C. and Sandra L., 1 969. Application of an estimator of high efficiency in bivariate extreme value theory, *J. Amer. Statist. Assoc.*, 64 (328), 1403-1414.

Tiago de Oliveira, J., 1981. *Statistical choice of univariate extreme models, Statistical Distributions in Scientific work*, 6, G. P. Patil ed., 6, 367-387, D. Reidel, Dordrecht.

Tiago de Oliveira, J., 1987. Comparaison entre les modeles logistique et naturel pour les maxima et extensions, *C. R. Acad. Sc. Paris, t.* 305(1), 481-484.

Van Montfort, M.A.J., 1970. On testing that the distribution of extremes is of type I when type II is the alternative, *J. Hydrology*, 11, 421-427.

Whitmore, G. A., *et al*. 1987. Case studies in data analysis, nº 5. *Canad. J. Statist*., 15, (4), 311-337.

*****