Extreme-value modelling: A preliminary analysis of monthly precipitation at Havana

Jesper Ryden

Recibido: octubre del 2005
Aprobado: diciembre del 2005

Resumen / *Abstract*

El análisis estadístico de valores extremos es de suma importancia en muchos campos de la ingeniería. Este trabajo es una breve introducción a la modelación con distribución de valores extremos. Consiste en la presentación de dos métodos básicos comúnmente empleados en el modelado de valores extremos para analizar el problema típico de estimación de valores de retorno. Como ilustración de los métodos, se investiga un conjunto de datos de la precipitación mensual en La Habana.

Palabras clave: análisis estadístico,valores extremos, modelación, distribución de valores extremos, estimación de valores de retorno, precipitación

Statistical analysis of extreme values is of importance in many fields of engineering. This paper will *serve as a brief introduction to modelling with extreme-value distributions. A presentation of two common, basic methods is given to analyse the standard problem in extreme-value modelling of estimation of return values. As an illustration of the methods, a data set of monthly precipitation at Havana is investigated.*

Key words: statistical analysis of extreme values, modelling, extreme-value distributions, extreme-value modelling, estimation of return values, precipitacion

INTRODUCTION

Knowledge of extreme values is important in many sciences and eventually for decision making in society, for example regarding questions related to energy production and consumption. How likely are certain unusual, perhaps undesired, events to happen . that is, how often? How high values of temperatures, wind speeds or other quantities with implications for energy supply or demand could be expected? In statistical analysis of extreme observations, the main interest lies in studying the tails of suitable probability distributions; this in contrast to standard statistical analysis, where the central parts of distributions are in focus. Several textbooks on extremevalue theory exist, for example by Castillo *et al.*[1], Coles,[2] the classical text by Gumbel[3] or chapters in the book on risk analysis by Rychlik and Ryden.[4]

In this paper, we discuss in the next section how extreme-value modelling may enter, directly or indirectly, in problems related to energy. That section is followed by a review of some 1 fundamental theoretical results and their implications. Thereafter, we will present commonly used methods how to estimate design values of randomly varying quantities and exemplify with a data set of monthly precipitation at Havana, Cuba.

APPLICATIONS TO ENERGY

Statistical extreme-value analysis has been used extensively in the environmental and meteorological sciences to analyse wind speeds, wave heights, temperatures and alike. Data are then often given as sequences of observations at some time scale: hourly, monthly, etc. The relation to problems related to energy supply and demand is obvious; knowledge of the (extreme) weather conditions assist in making predictions of energy demand. Moreover, in the development of technologies for renewable energy sources, mathematical and statistical modelling of solar radiation as well as of wind data is important (cf. the SWERA project, http://swera.unep.net).

In a broader perspective, consequences of the problem of global warming on energy consumption are studied. Harasawa,[5] presents a prediction that an increase of 1 ^{0}C in summer yields an increase in energy demand in Japan of 5 000 MW. Extreme weather may cause severe disturbances in national power systems; for example, violent storms hit France in December 1999 and caused damages to electricity transmission and distribution networks (see Le Du *et al.*[6]).

Important practical engineering applications of extreme-value theory are found for design purposes: to compute so-called design values and return periods. As will be motivated in the next section, extreme-value models are often particularly well describing situations where maxima of independent observations have been created. For example, Arvastson and Wollerstrand,[7] found an extreme-value distribution to describe well the distribution of daily domestic hot-water peak values.

Finally, statistical extreme-value analysis is not always suitable to model all more or less possible notions and quantities related to energy and society. For instance, the total yearly energy demand of a nation is often governed deterministically to a great extent by energy balances; hence, random mechanisms are not the major factor in determining such numbers.

EXTREME-VALUE MODELS

Assume that we have observations that are statistically independent and identically distributed. The aim is to study the distribution of the maximum of, say, n variables. Asymptotically, i.e. increasing the number of variables, it has been shown by the pioneer researchers in extreme-value 2 theory that the distribution of the maximum converges to one of three types of extreme-value distributions. One of these is the Gumbel distribution (also called double exponential, or simply extreme-value distribution), which in many applications is a natural model. The distribution function for a Gumbel distribution is given by

$$F(x) = \exp(e^{-(x.b)/a)}, \ - x < x < \alpha$$

where:
a: Scale parameter.
b: Location parameter.

In 1936, von Mises combined the three types of distributions in one, more general class of distributions, the so-called Generalized Extreme Value distribution (GEV) in which three parameters appear: in addition to scale and location parameter, a shape parameter is introduced. (The Gumbel distribution can be seen as a GEV distribution with shape parameter equal to zero.)

In practical work with data, a crucial issue is to decide whether an extreme-value distribution is well suited. Recall from above that these distributions arise as limits asymptotically, i.e. when the number of observations increases. Often the number of observations available is not that large. In practice, diagnostic tools are used, for instance plotting in different kinds of papers.Once a model distribution has been chosen, there exist methods for estimation of parameters in the distribution: by standard techniques (like maximum likelihood, ML) or specially developed (like method of moments).

CALCULATION OF RETURN VALUES

In this section, we will present two standard methods to compute return values, that is, values that have a return period of, say 100 years. Somewhat loosely speaking, the 100-return level is the level exceeded on average once in 100 years Data are assumed to be available at some time scale (hourly, monthly, weekly).

Method of annual maxima

This is the traditional method. Consider a fixed period of time (usually one year). The method is as follows: find each annual maximum and create the series of maxima, to which one of the three classical types of extreme-value distributions is fitted (often the Gumbel distribution). This method has its limitations, in particular, since possible dependence is not taken into account in the model. (This is often the case with seasonal loads.) Moreover, other large values during the year are ignored, for example the second and third highest largest maxima.

Peaks Over Threshold (POT)

The POT method makes use of a mathematical result for conditional distributions: exceedances over a high threshold can be shown to belong to a Generalized Pareto Distribution (GPD), provided the tail of the parent distribution belongs to the class of the extreme-value distributions.

As a special case of GPD is found the exponential distribution. If the GPD distributed exceedances of the level occur according to a Poisson process, the maxima belong to a GEV.

A crucial issue when using POT methods is the choice of the threshold. With a level chosen too high, there are not enough exceedances over the threshold to obtain good estimates of the parameters and their variances become high. On the other hand, with a too low threshold, the GPD might not be a good fit to the exceedances and hence a bias is introduced in the estimates. There is no simple rule how to choose the level. In practice, diagnostic plots are used; criteria minimizing a mean-square error expression are also found in the literature.

EXAMPLE: MONTHLY TOTAL PRECIPITATION IN HAVANA

In this section we illustrate the methods discussed above, making use of monthly total precipitation in Havana, from January 1949 to December 1970. Changes in amount and pattern of rainfall lead to impact on power generation, management and implementation of dams. Design issues may therefore be of interest for energy engineering community.

A time series of the observations is shown in figure 1 (left); in all there are 246 observations (hence, some missing values). The distribution over the year is interesting and presented in figure 1, right panel. There are seasonal effects, for instance, normally no huge amounts of precipitation are found during the first four months of a year.

The data originate from the National Oceanic and Atmospheric Administration (NOAA) and are compiled from information recieved at the National Climatic Data Center, Asheville, NC, USA. The data set can be found at the internet.

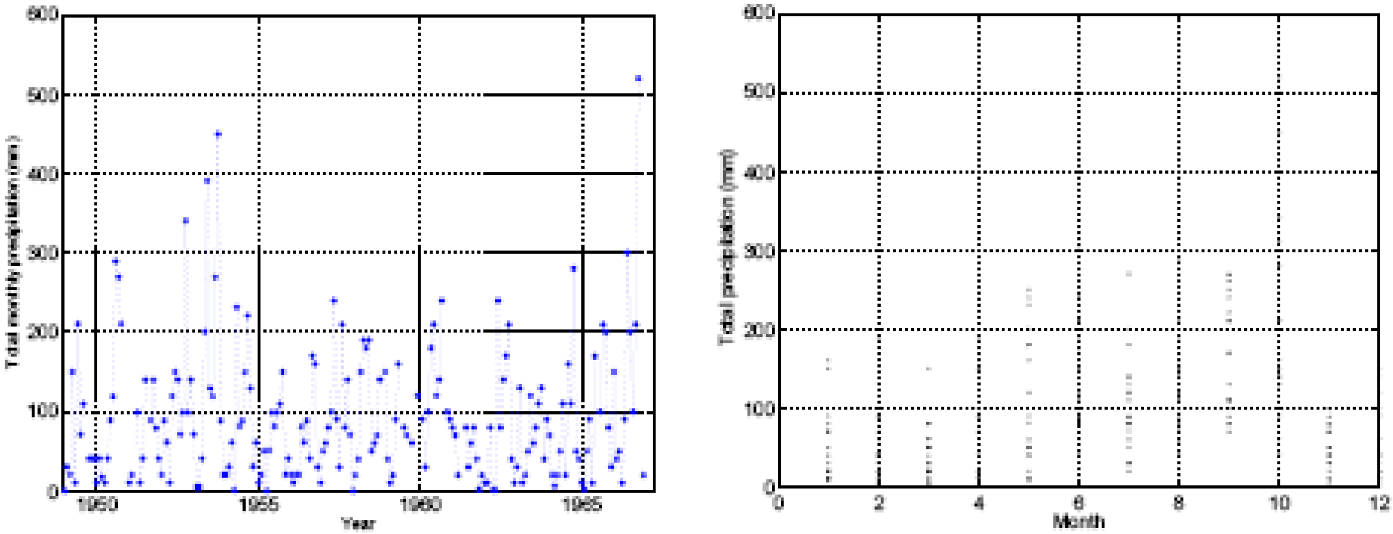
Annual maxima

The number of observed yearly maxima (22) is rather low, one cannot be sure whether the asymptotic results for extreme-value models are valid. In figure 2, the yearly maxima are plotted in a so-called probability paper for the Gumbel distribution. If the observations are quite close to a fitted straight line, a Gumbel distribution is assumed; this seems to be the case here. Furthermore, fitting of a GEV yields a shape parameter equal to 0,010 and a statistical 4 Month Total precipitation (mm) test implies that at the 0.05 confidence level, the hypothesis of shape parameter equal to zero cannot be rejected. Hence, the assumption of a Gumbel distribution is made.

The parameters in the distribution are found by maximum likelihood estimation, implemented in statistical software like the Matlab toolbox WAFO or the open-source statistical software R. By Matlab, we find the estimates of a and b to be *a*. = 80, *b*. = 201 and the 100-year return value of monthly total precipitation x_{100} is

$$x_{100} = b^* \cdot \; - a^* \cdot \ln(- \ln(1 - 1/100)) = 567 \text{ [mm]}.$$

Assuming asymptotic normal distribution for the parameter estimates, the standard error of x_{100} is 69 (mm) and an approximate 95% confidence interval is then [432, 701]. A Monte-Carlo based 95 % confidence interval, based on normal distributed parameters, is



Monthly total precipitation in Havana, Jan. 1949 - Dec., 1970. Left: Time series. Right: Data plotted monthly.

1

given by [421, 724].Note that these confidence intervals are quite wide.

In a careful study, the model needs to be valid: maxima from one year to another should be independent and identically distributed. If changes occur over the years, for example due to global warming, the model might not be valid anymore. Furthermore, convergence to the extreme-value distributionis in many instances slow, in this case we have only 22 observations.

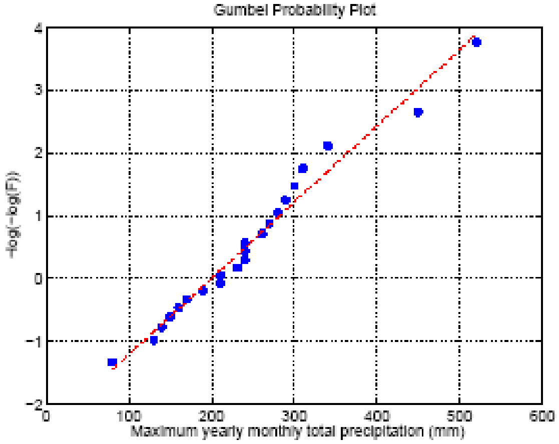
POT method

There exists no simple way to choose a suitable threshold. In practice, several diagnostic plots are used. Based on such plots and a recommendation made by Hasofer[8] that for the number of tail observations k (exceeding the threshold), $k \approx 1, 5\sqrt{n}$ where n is the total number of observations, a threshold about 210 mm is found. Point estimates given by Matlab[9] are as follows for some choices of thresholds:

Threshold (mm) 200 210 220 230
100-year return value (mm) 579 559 566 604.

The confidence intervals for the return values computed by the POT method are even wider in comparison to the method of annual maxima; for example, for the threshold 210 mm, a 95 % Monte Carlo confidence interval is given by [322, 1233].

SUMMARY



22 Yearly maxima plotted in Gumbel probability paper.

2

In this paper, we have presented some basic methods to analyse extreme values, given sequences of observations. The analysis performed is made simple to illustrate the basic ideas of the methods; the obtained values themselves are less interesting than the statistical reasoning. More refined procedures have been developed in the specialised literature, often adapted to the specific kind of data (wind speeds, wave heights). For real design purposes, several (additional) issues must be taken into account, for example the performance of the estimators; model validation (e.g. Gumbel or GEV?); the independence assumption (seasonal aspects, cf. figure 1, right panel); dependence on possible covariates. Bayesian modelling of the uncertainties of the design values might be an alternative in the case of few observations. Occasionally, the original observations are transformed before performing the extreme-value analysis to achieve 6 faster convergence to an extreme-value distribution; this is often made in wind engineering.

USEFUL LINKS

The statistical software R: http://www.r-project.org/
The Matlab toolbox WAFO: http://www.maths.lth.se/matstat/wafo/
National Climatic Data Center: http://www5.ncdc.noaa.gov/pubs/publications.html
IEA Energy Statistics: http://www.iea.org/Textbase/stats/index.asp

REFERENCES

1. **Castillo, E.** *et al.*: *Extreme Value and Related Models with Applications in Engineering and Science*, Wiley, 2004.
2. **Coles, S.:** *An Introduction to Statistical Odeling of Extreme Values*, Springer-Verlag, 2001.
3. **Gumbel, E. J:** *Statistics of Extremes*, Columbia University Press, New York. 1958 (Republication by Dover Publications 2004).
4. **Rychlik, I. and J. Ryden:** *Introduction to Probabilistic Modelling in Risk Analysis*, In preparation for Springer-Verlag, 2006.
5. **Harasawa, H.:** *Key Vulnerabilities and Critical Levels of Impact in East and South East Asia*, National Institute for Environmental Studies, Japan, 2005.
6. **Le Du, M.** *et al.*: *The French Power Network Facing the 1999,* Storms. Proceedings from Power Systems and Communications Infrastructures for the Future, Beijing, China, 2002.
7. **Arvastson, L. and J. Wollerstrand:** On Sizing of Domestic Hot Water Heaters of Instantaneous Type. 6th, International Symposium on Automation of District Heating Systems, Reykjavik, Island, 1997.
8. **Hasofer, A. M:** *Parametric Estimation of Failure Probabilities*, In Mathematical Models for Structural Reliability Analysis, eds. Casciati and Roberts, Boca Raton, Florida, CRC Press, 1996.
9. **Brodtkorb, P. A.,** *et al.*: *WAFO . a Matlab Toolbox for Analysis of Random Waves and Loads. Proceedings of the 10th International Offshore and Polar Engineering Conference, Seattle*, Vol. 3, pp. 343-350, 2000.

AUTHOR

Jesper Ryden
Applied Mathematics, School of Technology and Society, Malmö University
e-mail:jesper.ryden@ts.mah.se