Annales Univ. Sci. Budapest., Sect. Comp. 42 (2014) 199–208

DATA MINING OF EXTREME VALUE MODELLING EUROPEAN PRECIPITATION DATA

Csilla Hajas (Budapest, Hungary)

András Zempléni (Budapest, Hungary)

Dedicated to András Benczúr on the occasion of his 70th birthday

Communicated by László Lakatos

(Received June 1, 2014; accepted July 1, 2014)

Abstract. This paper shows how can the peaks over threshold model of gridded European precipitation data be combined with various data mining tools. The motivation is that even the 0.5 grade-grid of 63 years of the European Climate Assessment daily precipitation data is a massive data set, where there is little hope to find valuable results without reasonable preprocessing. This step is based on the peaks over threshold approach, which is a sound model for the extremes. We have applied a moving window methodology in order to catch the changes in the pattern of the high precipitations. Our results show that indeed there are spatially different tendencies observable.

1. Introduction

Detection of signs for climate changes is a very important and actual question. As mathematicians, we are not in the position of giving exact explanations

Key words and phrases: cluster analysis, daily precipitation data, generalized Pareto distribution, moving window, return level
2010 Mathematics Subject Classification: 62P12

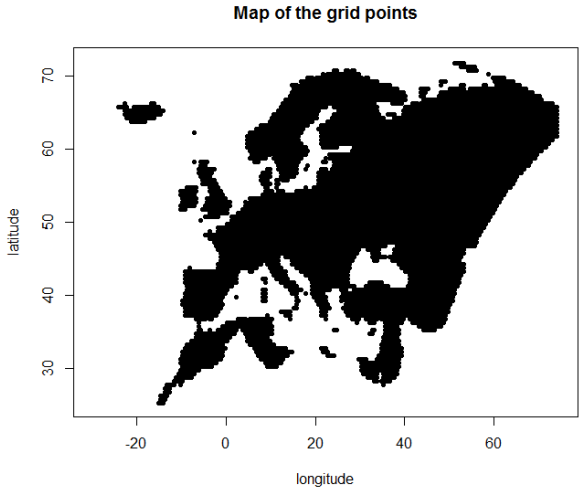

Figure 1. Map of the region covered by our data

for the changes, but we may try to reveal them. This revelation is not easy at all, as there are huge data sets of various quality available and it is difficult to get useful results out of them. Precipitation is a very important meteorological phenomenon, here in Europe we often feel the economic effect of its extremes – the last major flood affected most of Central Europe in 2013. The used observations are the 63 years of daily precipitation data of the European Climate Assessment (E-OBS, `http://www.ecad.eu`). We have worked with the data based on 0.5-grade grid points, available for Europe and Northern Africa. Figure 1 depicts the covered region. This gridded data base has been used extensively for climate analysis, see [4]. The quality has been evaluated in [5], and the results show that it may be considered reliable for most of Central Europe. However, especially in the African and Middle-Eastern region there are missing periods of various length, which have to be taken into account.

The used mathematical models are first the peaks over threshold model, then we apply various data mining tools for the sequence of estimated return levels. We give details of these models in Section 2. In Section 3 we show the applications of the models. Section 4 contains the conclusions.

2. Models

2.1. Peaks over threshold models

There are two widely investigated types of approaches to model extreme values. The classical models are based on the annual maxima of the data. The other, more recent approach focuses on all observations exceeding a high threshold. The latter class of models is called peaks over threshold (POT) models. In this paper we will concentrate on such models, because they allow for the use of more data which is important for us, as we investigate the time-dependence of the fitted distributions by the help of moving windows. Beguería et al. [2] have used the POT model to build a spatial pattern for extreme precipitation hazard. Our approach differs from this, as our intention is to determine temporal trends.

Threshold models have been introduced in the 1970s [1]. Under fairly general regularity conditions the threshold exceedances have an asymptotic distribution. To be more specific, let $\mathbf{X}_n = (X_1, ..., X_n)$ be a sequence of independent random variables with common unknown distribution function F. Then under fairly general conditions (which is true for all important continuous distributions)– for high thresholds u – the conditional excess distribution function converges:

$$F_u(z) = P(X_i - u \leq z | X_i > u) \overset{u \to \infty}{\longrightarrow} H(z),$$

where $H(z)$ is a distribution function over the nonnegative numbers z with parameters $\xi \in \mathbb{R}$ and $\sigma > 0$

$$H(z) = \begin{cases} 1 - \left(1 + \frac{\xi z}{\sigma}\right)^{-\frac{1}{\xi}} & \text{if } \xi \neq 0; \\ 1 - e^{-\frac{z}{\sigma}} & \text{if } \xi = 0. \end{cases} \tag{2.1}$$

The family defined in (2.1) is called generalized Pareto distribution (GPD). Depending on the parameter ξ, this distribution includes three types of distribution familes:

(I) $\xi > 0$: heavy tailed (ordinary Pareto) distribution;

(II) $\xi < 0$: short-tailed distribution;

(III) $\xi = 0$: exponential distribution.

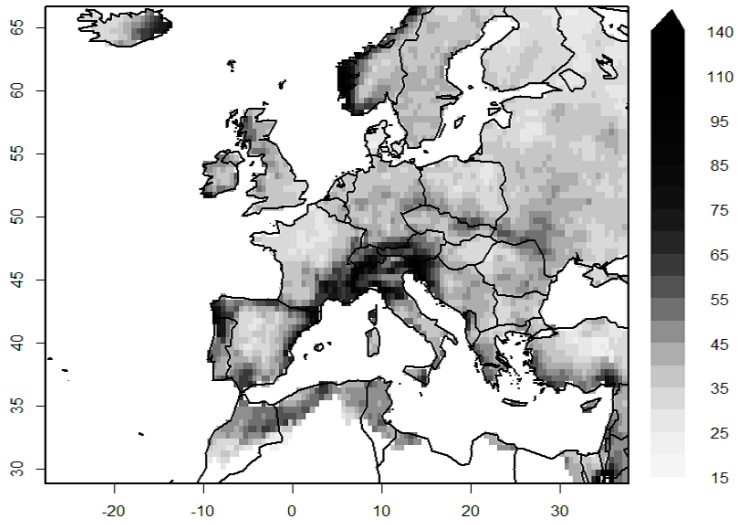

Figure 2. Estimated 10-years return levels for the daily precipitation in mm, based on the 63 years of observations

These distributions have proved to be a suitable model for precipitation data, see for example [3], where several reasonable families were compared and the Pareto distribution was clearly the best fit. In our case both ordinary and type II Pareto distributions appear, depending on the geographical properties of the regions. Rakonczai et al. [10] have investigated 5 grid points of the same dataset and found interesting tendencies both in its univariate and the bivariate properties.

The parameters of the GPD can be estimated by maximum likelihood, which has the usual optimality properties and asymptotic normality if $\xi > -0.5$, which is the case in almost all applications (including ours).

In meteorology return levels, which correspond to certain return periods – for instance 10, 20 or 50 years – are especially important. The q-quantile of the GPD can be given as

$$H^{-1}(q) = u + \frac{\sigma}{\xi}\left(\frac{P(X > u)}{q}\right)^{\xi - 1}$$

if $\xi \neq 0$. Otherwise it is simply $u + \sigma \log(P(X > u)) - \log(p)$. It is important to note that in general if we have n observations over the threshold in l years* and the return period of interest is m years, then the corresponding quantile is $q = 1 - \frac{1}{m}\frac{l}{n}$. For example if there are 1000 such observations in 100 years, then

*This means in average $\frac{n}{l}$ observations in a single year.

the 50 years return level is the 0.998-quantile of the distribution, which models the single exceedances. In our cases we shall estimate the 10-years return level.

2.2. Data mining

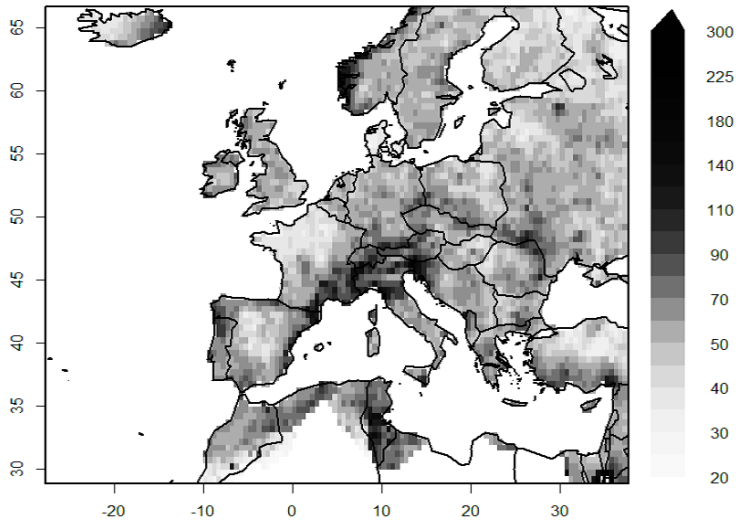

Figure 3. Maximum of the estimated 10-years return levels for the daily maximum precipitation in mm, based on the 57 moving windows

Our data is spatio-temporal, even after the parametric modeling by the GPD - as we apply moving windows in order to capture the time-dependence of the fitted parametric model. Spatio-temporal data mining has been developed in the last few years, see for example [9] or Chapter 10 in the book [6]. However, the methodologies are yet mostly ad hoc. For example if one wishes to find anomalies in a data set, the calculation of some anomaly scores are proposed, but there is no definite form given for these scores. We shall show that in our cases this approach does not identify any real anomalies.

Another, more classical data mining tool is clustering. Here we apply the k-means clustering, which is a traditional, simple and quick method even in our case of over 20000 data points in the 57 dimensional space. We shall see that a practical preprocessing makes this tool especially useful. In our case this approach ensured that we focused on the temporal aspect in the data mining (as the time series of estimated quantiles for the moving windows is analyzed), but the method turned out to be suitable for detecting areas sharing similar features – thus also the spatial aspects were included in the analysis.

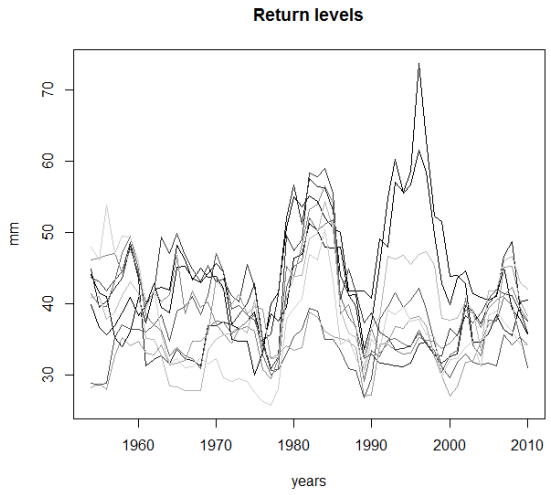

Figure 4. Time series of the 10-years return levels, based on the moving windows of length 7 years for 9 neighbouring grid points in Central Europe

3. Applications

Having shown the tools, now we are in position of actually carrying out the data analysis.

We have used 63 years of daily precipitation data of the 0.5 grade grid points, as shown on Figure 1. We have not investigated the time series, which may show some seasonality (see [10] for the analysis of Hungarian data) as we always use complete years in the analysis, so this effect does not have an influence. This is also true for the extremes, we are interested in.

First we show a map of the 10-years return levels, based on all observations (Figure 2). We see substantial differences in the expected amount of precipitation.

Our main aim is to detect if there are any significant changes in the time series. We have carried out an analysis, based on moving windows of 7 years (altogether there were 57 such windows, as we have repeated the analysis every year). In the paper of [7] the authors investigated a similar 2.5 grade gridded database of South America, by windows of 25-years. Their approach was based on the direct investigation of the parameters, including the quantiles of the fitted distribution. However, our grid consists of much more points, so we indeed need the data mining tools.

First we have fitted the POT model separately to the data of the moving windows (see Section 2.1). The threshold was chosen as the 95% quantile of the observations in the given window. This means that we have used different values for each site and time period. This choice allows the use of enough data (over 100) in each of the windows, while it is high enough in order the theory to hold.

For detecting possible anomalies, first we considered the maximum of the 57 estimated quantiles for all grid points. The result is shown in Figure 3. There are quite a few surprisingly high values in the Middle East or Africa, but not these time series are thought to be the most reliable. However, that for the whole Denmark we have lower estimated return levels than the values of its neighbours is interesting. These are those observations, which cannot be found easily by any data mining tools. If we calculate scores as the mean quadratic deviation between a point and its neighbours (excluding those points, where there is no observation) then we do not get any patterns, when the large deviations are plotted.

The time development of the 10-years return level, calculated by the estimated model parameters for some neighbouring grid points is shown in Figure 4. We see that there is similarity between these paths. We explored this further when clustering those sites, where there were no missing data. However, we were not so much interested in the differences in the precipitation itself, but more in the trends. So instead of the original estimates, we have used the relative value of the actual estimates, where the reference was calculated on the basis of the whole data set. These 57 dimensional data were clustered into 4 clusters by the k-means method, with the results shown in Figures 5 and 6.

As here all the values fluctuate around 1, there was no need to standardise the data. Figure 5 shows the time-development of the 4 cluster centers. We can clearly differentiate between 4 patterns (with respect to increasing darkness): the lightest one has one peak around 1990, the second one is two-peaked, the third one is decreasing after an early peak, and the last(black) one is increasing. It is interesting to investigate which areas belong to the clusters. The shades of Figure 6 correspond to those of Figure 5, so we may state that large parts of Central Europe belong to the black (increasing) cluster.

4. Conclusions

As a conclusion we can formulate that there are indeed interesting patterns observable in the gridded precipitation data we have analysed. Clustering and other spatial methods turned out to be useful methods for finding the places

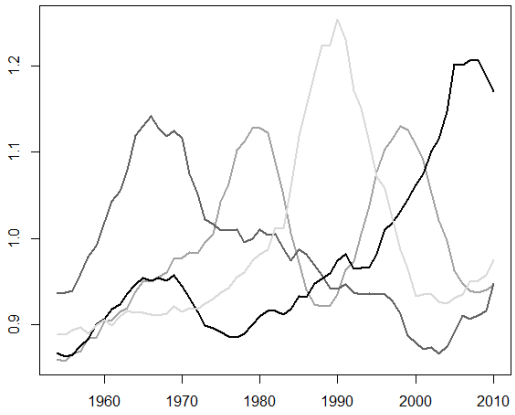

Figure 5. Cluster centers, relative to the estimation, based on the whole time span of 57 years

of special interest.

The POT model and the moving windows method provided reasonable amount of data: summarizing the most important characteristics of the observations at hand.

The findings: recently increased return levels, combined with similar observations on the increased dependence, see [10], between the sites show the danger of floods – something which has indeed been observed in summer 2013 over the Danube and the Elbe basin, so these tendencies are worth investigating further and in more detail.

References

[1] **Balkema, A.A. and L. de Haan,** Residual lifetime at great age, *Ann. Probab.*, **2** (1974), 792–804.

[2] **Beguería, S. and S. M. Vicente-Serrano,** Mapping the hazard of extreme rainfall by peaks over threshold extreme value analysis and spatial regression techniques, *Journal of Applied Meteorology and Climatology*, **45** (2006), 108–124.

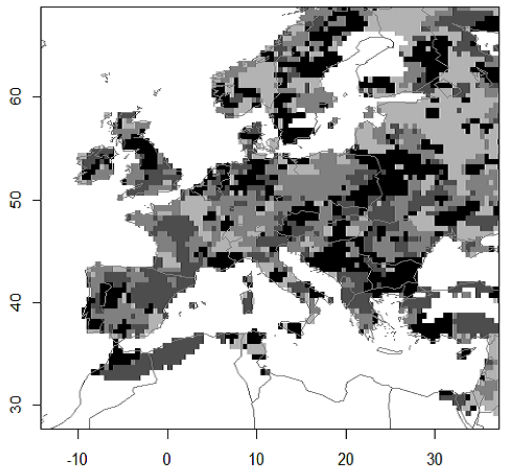

Figure 6. The clusters, where the darkness correspond to those of Figure 5

[3] **Dan'azumi, S., S. Shamsudin, and A. Aris,** Modeling the distribution of rainfall intensity using hourly data, *American Journal of Environmental Sciences*, **6** (2010), 238-243.

[4] **Haylock, M., N. Hofstra, T. Klein, M.G. Albert, E.J. Klok, P.D. Jones and M. New,** A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006, *Journal of Geophysical Research: Atmospheres (1984–2012)*, **113** (2008), Issue D20, DOI:10.1029/2008JD010201

[5] **Hofstra, N., M. Haylock, M. New, and P.D. Jones,** Testing E-OBS European high-resolution gridded data set of daily precipitation and surface temperature, *Journal of Geophysical Research: Atmospheres (1984–2012)*, **114** (2009), Issue D21, DOI:10.1029/2009JD011799

[6] **Giannotti, F. and D. Pedreschi (eds.),** *Mobility, Data Mining and Privacy,* Springer Verlag, 2008.

[7] **Khan, S., G. Kuhn, A.R. Ganguly, D.J. Erickson III and G. Ostrouchov,** Spatio-temporal variability of daily and weekly precipitation extremes in South America, *Water Resour. Res.*, **43** (2007), Issue 11, DOI:10.1029/2006WR005384

[8] **Pickands, J.** Statistical inference using extreme order statistics, *Annals of Statistics*, **3** (1975), 119–131.

[9] **Shekhar, S., M.R. Evans, J.M. Kang and P. Mohan,** Identifying patterns in spatial information: a survey of methods, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, **1** (3) (2011), 193–214.

[10] **Rakonczai, P., L. Varga, and A. Zempléni,** Applications of threshold models and the weighted bootstrap for Hungarian precipitation data, *Theoretical and Applied Climatology* (accepted) `http://arxiv.org/abs/1310.7918`

Csilla Hajas
Department of Information Systems
Eötvös Loránd University
Budapest, Hungary
`sila@inf.elte.hu`

András Zempléni
Department of Probability Theory and Statistics
Eötvös Loránd University
Budapest, Hungary
`zempleni@math.elte.hu`