Boletín de Estadística e Investigación Operativa
Vol. 29, No. 2, Junio 2013, pp. 92-102

Estadística

Extreme-based Clustering of Environmental Time Series

Manuel G. Scotto
Departamento de Matemática
Universidade de Aveiro, Portugal
✉ mscotto@ua.pt

Susana M. Barbosa
Instituto Dom Luiz
Universidade de Lisboa, Portugal
sabarbosa@fc.ul.pt

Andrés M. Alonso
Departamento de Estadística
Instituto Flores de Lemus
Universidad Carlos III de Madrid, España
andres.alonso@uc3m.es

Abstract

This work provides an up-to-date review on clustering techniques to classify time series on the basis of their corresponding extremal properties with a bias towards describing the authors' ongoing work. Applications to clustering time series of sea-level and daily mean temperature are presented.

Keywords: Extreme value theory, Cluster analysis, Bayesian analysis, Return values.
AMS Subject classifications: 62H30, 62C10, 60G70.

1. Introduction

A topic of major current interest in the analysis of environmental processes is to investigate the occurrence of future rare events which can have catastrophic consequences for human activities, through their impacts on health and on the natural and constructed environments. It is widely believed that the frequency of events such as avalanches, heat waves, floods and so forth is increasing as a result of climatic and other changes, yet they are hard to predict and their effects

© 2013 SEIO

are poorly understood. Recent developments in the statistics of extremes (e.g., Chavez-Demoulin and Davison, 2012), optimal alarm systems (e.g., Baxevani et al., 2011) and in multivariate statistical techniques for discrimination, clustering and dimension reduction for time series (e.g., Jolliffe and Philipp, 2010), and the increasing availability of relevant high-quality data have the potential both to deepen our grasp of the underlying physical phenomena, and to aid on the construction of models and tools for forecasting the occurrence and impact of rare events. There are various issues involved in predicting such events, from long-term prediction to forecasting individual events with potentially catastrophic consequences.

In studies of regional sea-level or temperature variability, time series records are often analyzed individually for characterizing variability at each location. Marginal analysis, however, is in itself insufficient to come to an accurate description of regional variability. An alternative approach is to consider simultaneously the whole data set from a given region and characterize regional variability in terms of locations exhibiting similar behavior, through clustering techniques. Although clustering techniques have been popular for the analysis of non-time series environmental data, its extension to time series data are hindered by the serial dependence and high dimensionality of the observations. The reader is referred to Galeano and Peña (2000) for an overview of the early work in this area and to Caiado et al. (2012) and Scotto et al. (2010) for recent developments.

In this work, a time series clustering approach which combines Bayesian methodology, extreme value theory and classification techniques is presented for the analysis of the regional variability of sea-level and temperature extremes. In both cases time series records are clustered on the basis of their corresponding predictive distributions for 25-, 50- and 100-years return values.

The rest of the paper is laid out as follows: for completeness and reader's convenience, in Section 2 some background results on extreme value theory are given. The time series clustering procedure is explained in Section 3. An application to hourly tide gauge records from the North Atlantic, and to time series of daily mean temperature in Europe is presented in Section 4. Finally, in Section 5 some concluding remarks are given.

2. Extreme Value Framework

Extreme Value Theory (EVT in short) provides simple techniques for estimating probabilities of future extremal events given historical data. The celebrated *Fisher-Tippett Extreme Value Theorem* states that if the distribution of conveniently normalized partial maxima of an independent and identically distributed sequence of random variables with common (unknown) distribution F, converges to a non-degenerate limit distribution, say, G, then F is in the domain of attraction of G, and G must be the *Generalized Extreme Value* (GEV in short) distribution

$$G(x) \equiv G_{\xi}(x) := \exp\left\{ - \left[1 + \xi \left(\frac{x-\mu}{\sigma} \right) \right]^{-\frac{1}{\xi}} \right\}, \tag{2.1}$$

defined on $\{x : 1 + \xi(x-\mu)/\sigma > 0\}$ with location $\mu \in \mathbb{R}$, scale $\sigma > 0$ and shape parameter (also called tail index) $\xi \in \mathbb{R}$. The shape parameter determines the three extreme value types. Specifically, when ξ takes negative values, positives values or when $\xi = 0$, interpreted by taking the limit in (2.1) as $\xi \to 0$, the GEV distribution is the negative Weibull, the Fréchet or the Gumbel distribution, respectively. The Fréchet domain of attraction embraces heavy-tailed distributions with polynomially decaying tails. All d.f.'s belonging to Weibull domain of attraction are light-tailed with finite right endpoint. The intermediate case $\xi = 0$ is of particular interest since this class includes distribution functions with very different tails, ranging from moderately heavy (such as the lognormal distribution) to light (such as the Normal distribution) having finite right endpoint or not.

Therefore, separating statistical inference procedures according to the most suitable domain of attraction for the sampled distribution is an impending problem. A test for Gumbel domain against Fréchet or Weibull max-domain has received the general designation of statistical choice of extreme domains of attraction (see e.g., Marohn, 1998a,b; Wang et al., 1996; Fraga Alves and Gomes, 1996; Hasofer and Wang, 1992). In this respect, Hasofer and Wang's test may be pointed out as one of the most commonly used testing procedure.

A useful parameter of interest in many extreme value studies is the quantile x_p for a specified exceedance probability p, defined as

$$x_p := \left\{ \begin{array}{ll} \mu - \frac{\sigma}{\xi} \left\{ 1 - \left(-\log(1-p) \right)^{-\xi} \right\}, & \xi \neq 0 \\ \mu - \sigma \log\left\{ -\log(1-p) \right\}, & \xi = 0 \end{array} \right. , \tag{2.2}$$

where $G(x_p) = 1 - p$. Roughly speaking, x_p is the return value that is associated with the return period $1/p$ for small p, in units of, say, years, if the GEV corresponds to the annual maximum.

A typical application in EVT is the Annual Maxima method which consists in fitting the Generalized Extreme Value distribution to annual maxima. This approach is often viewed as advantageous since with the requirement of only a simplified data summary cumulates the reduction of dependence in the data. Nevertheless, it is sometimes difficult to detect seasonal cycles in the measurements that may be feasibly identified with blocks. This issue motivated the search for characterizations of extreme value behavior that enables modeling the observed data by other than just the block maxima and ultimately to make better use of the information given by the data. There are two well-known alternative characterizations. One is based on the behavior of the r-largest order statistics within a block of large size, for small values of r. This approach was extensively used by Guedes Soares and Scotto (2004) to obtain predictions of extreme values of significant wave heights collected on the northern North Sea and by Scotto et al. (2011a) and Scotto et al. (2010) for long-term predictions of hourly tide gauge records and daily mean temperatures.

The other approach is based on exceedances of a high (fixed or time-dependent) threshold. The rationale behind this procedure is that (i) if the threshold level is taken high enough, the distribution of the peak excesses is expected to be close to the Generalized Pareto distribution; (ii) observations in particular cluster peaks, belonging to different clusters are expected to be independent. Note that, while (i) avoids the choice of arbitrary models, (ii) ensures that the fundamental concept of return value is applicable. Applications of this method to the extrapolation of wave data are described in Jonathan and Ewans (2013), Mínguez et al. (2010), Menéndez et al. (2008), Méndez et al. (2006) and Caires and Sterl (2005).

While the application of extreme value theory is often based on likelihood approaches, Bayesian methods have the advantage of a substantially more flexible inference. The application of Bayesian extreme value analysis to wave and tide gauge records (e.g., Scotto et.al, 2011b; Scotto et al., 2010; Scotto and Guedes Soares, 2007; Egozcue et al., 2005) allows to retrieve the complete posterior density of the parameters and the return values, and therefore to quantify the degree of estimation uncertainty. Applications to daily mean temperature records can be found in Scotto et al. (2011a).

3. Time Series Clustering Procedure

In Scotto et al. (2010) a new approach for time series clustering based on long-term predictions of extreme values of sea-level records was introduced. We outline here the essential of the method. The implementation of the clustering procedure comprises three steps:

1. *firstly, the algorithm starts with the estimation of the posterior predictive distributions of the 25-, 50-, and 100-years return values for each time series.*

Since the GEV distribution in (2.1) admits no conjugate prior distributions a near-flat Normal multivariate distribution is adopted for $\theta := (\mu, \sigma, \xi)$ reflecting the absence of external information. The Metropolis-Hastings algorithm implemented in an appropriate Markov Chain Monte Carlo scheme is adopted to draw a sample $(\theta^{(1)}, \ldots, \theta^{(N)})$ from the posterior distribution of θ, being the initial values the maximum log-likelihood estimates obtained from the distribution of the r-largest order statistic model. From the Markov Chain sequence and (2.2) a sample from the posterior predictive distribution of the return values is generated.

2. *The second step consists in calculating the distance among pairs of distribution functions.*

   To this extent, an adequate metric between univariate distribution functions is required. The choice of this metric should reflect the final goal of the clustering procedure in the sense that the distance captures the discrepancies between predictive distributions of return values. In this case, the weighted L_2-Wasserstein distance between posterior predictive distributions is adopted. This means that the distance between two time series, say, $\mathbf{X}^{(i)}$ and $\mathbf{X}^{(j)}$ is defined as

$$D_{ij}^2 := \int_0^1 \left(F_{x_p^{(i)}}^{-1}(y|\boldsymbol{x}) \; - \; F_{x_p^{(j)}}^{-1}(y|\boldsymbol{x}) \right)^2 y(1-y)\, dy,$$

   where $F_{x_p^{(i)}}(\cdot|\boldsymbol{x})$ and $F_{x_p^{(j)}}(\cdot|\boldsymbol{x})$ denote the posterior predictive distribution functions of the return values of the ith and jth time series $x_p^{(i)}$ and $x_p^{(j)}$ respectively with $p = 1/m$, corresponding to a return period of m-years. Consistent estimators for similar distances have been proposed by Vilar-Fernández et al. (2010) and Alonso et al. (2006).

3. *Finally, a dendrogram based on the application of classical cluster techniques to the dissimilarity matrix is built.*

   That gives the different clusters formed by the predictive distributions for 25-, 50- and 100-years return values. In particular, agglomerative hierarchical methods with nearest distance (single linkage), furthest distance (complete linkage) and unweighted average distance (average linkage) are used as grouping criteria.

4. Applications

In this section, the clustering procedure described above is applied to hourly tide gauge records from the North Atlantic and to time series of daily mean temperature obtained from the European Climate Assessment (ECA) data set. A complete description of these results is provided in Scotto et al. (2011a) and Scotto et al. (2010), respectively.

4.1. Results for the North Atlantic Data Set

It is well-known that sea-level plays a fundamental role as an indicator of the state of the Earth's system. Oceanographic and atmospheric changes influence sea-level variability on a wide range of scales, ranging from local to global spatial scales and from minutes to millennia temporal scales. The height of the sea surface provides therefore an integrated record of changes of the Earth's climate. Furthermore, it assumes particular relevance in a climate change context since long-term changes in extreme sea-levels may be associated with regional climatic variability, and coastal regions may face an increased flooding risk, particularly for areas where the combination of local subsidence and global sea-level rise enhances the rate of relative sea-level change.

Hourly tide gauge records for the Northwest Atlantic are obtained from the Research Quality Data Set of the University of Hawaii Sea Level Center. Only continuous and very long records (> 40 years) are considered. A total of 14 stations with records extending from at least January 1911 to December 2005 are selected for the analysis. Information of the analyzed 14 tide gauge record is given in Table 1 below. The algorithm described in the previous section is applied to obtain clusters of the sea-level observations on the basis of 25-, 50- and 100-years return values. The results reveals that:

Tide gauge	Longitude (°E)	Latitude (°N)	Country
Atlantic City (AC)	-74.42	39.35	United States
Boston (BS)	-71.01	42.35	United States
Brest (BR)	-4.5	48.38	France
Charleston (CH)	-79.93	32.78	United States
Eastport (ES)	-66.98	44.90	United States
Fort Pulaski (FP)	-80.90	32.03	United States
Key West (KW)	-81.82	24.55	United States
Lewes (LW)	-75.12	38.78	United States
Nantucket (NN)	-70.10	41.28	United States
New London (NL)	-72.08	41.35	United States
Newport (NW)	-71.03	41.05	United States
New York (NY)	-74.02	40.70	United States
Portland (PR)	-70.25	43.65	United States
Wilmington (WL)	-77.95	34.23	United States

Table 1: Analyzed tide gauge records.

- for WL, AC, BR, ES, FP and PR, the posterior mean for ξ is negative. Moreover, the posterior probability of non-negative values for ξ is negligible clearly indicating that a bounded upper tail distribution may be a reasonable choice to fit the data sets corresponding to these locations. This is in contrast with the posterior mean of ξ in NW, for which the posterior probability of non-negative values for ξ is non-negligible, being approximately 0.43. Thus, distributions with light tails (bounded or unbounded) are also consistent with the data. Furthermore, for BS, NL, NY and KW the most plausible values of the tail index are those ranging from -0.10 to 0.15 (i.e., light tail distributions). Finally, CH, LW and NN distributions with light or moderately heavy distribution might be a suitable choice.

- The results of the clustering procedure show that the dendrograms based on average linkage exhibit more comprehensive clusters than the dendrograms based on single and complete linkage. However, the results from the three clustering strategies are consistent, showing a clear separation between stations at the highest latitudes (that is ES, BR, PR and BS), for which the return values are largest ($>$ 6m) and the remaining stations. Despite the large difference in longitude of the BR station, this tide gauge does not appear isolated in the dendrograms, indicating that latitude is the dominant influence on the values of the return levels. BR is closer in terms of the returns distribution function to ES, the highest latitude station in the US east coast data set. BS and PR (PR) form a distinct cluster joining at a slightly higher level, indicating the closeness of the returns distribution function for these two neighboring stations, though not as close as in the case of the BR cluster. For the lower latitude stations, the dendrograms display two distinct clusters, a cluster including the stations with lowest return values, $<$3.5m (NW, WL, NN and KW) and a group comprising LW, NY, CH, NL, AC, and FP, corresponding to the stations with intermediate return levels, between 3.5m and 6m. These two groups are spatially quite heterogeneous and do not follow criteria of geographical proximity. However, the differences in the return distribution functions for the locations in these two groups are considerably smaller than between the two northern clusters, since they are joined at a lower level than the two individual clusters of the northern stations (BR & ES and PR & BS).

4.2. Results for the ECA Data Set

Stations in western Europe with data from at least January 1901 to December 2007 were selected from the ECA blended data set. Furthermore, only time series for which the percentage of missing values is smaller than 2% were considered in the study. Details are displayed in Table 2.

The analysis of the results allows to conclude that:

- a common feature of the posterior distributions associated with the shape parameter, for all the 32 stations, is that the probability of non-negative

Station	Longitude (°E)	Latitude (°N)	Country
St.Petersburg (STP)	30.3	59.97	Russia
Stockholm (STO)	18.05	59.35	Sweden
Vestervig (VES)	8.32	56.77	Denmark
Koebenhavn (KOE)	12.53	55.68	Denmark
Hammer odde fyr (HOF)	14.78	55.3	Denmark
Hamburg (HAM)	10.25	53.48	Germany
Dublin (DUB)	353.68	53.35	Ireland
Bremen (BRE)	8.78	53.05	Germany
Berlin (BER)	13.3	52.45	Germany
Potsdam (POT)	13.07	52.38	Germany
De bilt (DEB)	5.18	52.1	Holland
Halle (HAL)	11.98	51.48	Germany
Leipzig (LEI)	12.23	51.43	Germany
Kyiv (KYI)	30.53	50.4	Ukraine
Frankfurt (FRE)	8.67	50.12	Germany
Bamberg (BAM)	10.88	49.88	Germany
Paris (PAR)	2.33	48.82	France
Stuttgart (STU)	9.22	48.72	Germany
Kremsmuenster (KRE)	14.13	48.05	Austria
Salzburg (SAL)	13	47.8	Austria
Hohenpeissenberg (HOH)	11.02	47.8	Germany
Zuerich (ZUE)	8.57	47.38	Switzerland
Saentis (SAE)	9.35	47.25	Switzerland
Graz (GRA)	15.45	47.08	Austria
Sonnblick (SON)	12.95	47.05	Austria
Geneve (GEN)	6.13	46.25	Switzerland
Ljubljana bezigrad (LJU)	14.52	46.05	Slovenia
Lugano (LUG)	8.97	46	Switzerland
Zagreb-gric (ZAG)	15.97	45.82	Croatia
Osijek (OSI)	18.63	45.53	Croatia
Bologna (BOL)	11.25	44.48	Italy
Lisboa (LIS)	350.85	38.72	Portugal

Table 2: Analyzed air temperature records.

values for ξ is negligible indicating that a bounded upper tail distribution may be a reasonable choice to fit the data sets corresponding to these locations. A bounded upper tail distribution is not only reasonable from the statistical point of view but also from a physical perspective, in the sense that thermodynamic considerations lead to a upper limit to the Earth's air temperature.

- The results obtained by the three clustering approaches are in general similar particularly for the complete and average linkage methods. The largest distance between stations distinguishes a cluster with SAE and SON from the remaining locations. These are the highest altitude stations exhibiting the lowest 25-years return values. The second largest distance discriminates mainly the northern stations with maritime climate or on the continental-maritime boundary (STO, VES, KOE, BRE and DUB) from the remaining stations in central and southern Europe. Within this large remaining cluster of European stations there is a further discrimination between roughly northern and southern stations, with a further distinction between the south-eastern stations and the more western (PAR, BOL and LIS) locations.

5. Conclusions

Clustering time series based on extreme values is a promising approach for the analysis of temporal variability within a spatial context, and specifically for linking future estimates and spatial distributions. This information is relevant in practical applications and complements and serves as input to models lacking the spatial resolution to identify such localized behavior. It is important to stress that by adopting a Bayesian approach, the inference is substantially more flexible. Moreover, by working with noninformative priors the Bayesian framework is essentially formal, but it leads to an inference in which parameter uncertainty is properly formalized and for which inferences on return values are naturally handled.

Acknowledgement

The authors are grateful to the Associate Editor Miguel López Díaz for inviting us to contribute to this issue of BEIO. This research was partially supported by FEDER funds through COMPETE – Operational Programme Factors of Competitiveness ("Programa Operacional Factores de Competitividade") and by Portuguese funds through the Center for Research and Development in Mathematics and Applications and the Portuguese Foundation for Science and Technology ("FCT–Fundação para a Ciência e a Tecnologia"), within project PEst-C/MAT/UI4106/2011 with COMPETE number FCOMP-01-0124-FEDER-022690. The third author also acknowledge the support by CICYT (Spain) grants SEJ2007-64500 and ECO2011-25706.

References

[1] Alonso, A.M., Berrendero, J.R., Hernández, A. and Justel, A. (2006). Time series clustering based on forecast densities. *Comput. Statist. Data Anal.*, **51**, 762-776.

[2] Baxevani, A., Wilson, R. and Scotto, M.G. (2011). Prediction of catastrophes in space over time. Submitted.

[3] Caiado, J., Maharaj, E.A. and D'urso, P. (2012). Time series clustering, in *Handbook of Cluster Analysis*, C. Henning, M. Meila, F. Murtagh, R. Rocci (eds.), CRC Press (to appear).

[4] Caires, S. and Sterl, A. (2005). 100-years return value estimates for ocean wind speed and sigfinicant wave height from the era-40 data. *J. Climate*, **18**, 1032–1048.

[5] Chavez-Demoulin, V. and Davison, A.C. (2012). Modelling time series extremes. *Revstat-Statistical Journal*, **10**, 109–133.

[6] Egozcue J.J., Pawlowsky-Glahn V and Ortego M.I. (2005). Wave-height hazard analysis in Eastern Coast of Spain - Bayesian approach using generalized Pareto distribution. *Adv. Geosci.*, **2**, 25–30.

[7] Fraga Alves, M.I. and Gomes, M.I. (1996). Statistical choice of extreme value domains of attraction - a comparative analysis. *Commun. Statist. Theory Meth.*, **25**, 789–811.

[8] Galeano, P. and Peña, D. (2000). Multivariate analysis in vector time series. *Resenhas*, **4**, 383–403.

[9] Guedes Soares, C. and Scotto, M.G. (2004). Application of the r-order statistics for long-term predictions of significant wave heights. *Coastal Eng.*, **51**, 387–394.

[10] Hasofer, A.M. and Wang, Z. (1992). A test for extreme value domain of attraction. *J. Amer. Statist. Soc.*, **87**, 171–177.

[11] Jolliffe, I.T. and Philipp, A. (2010). Some recent developments in cluster analysis. *Phys. Chem. Earth*, **35**, 309-315.

[12] Jonathan P. and Ewans, K. (2013). Statistical modelling of extreme ocean environments for marine design: a review. *Ocean Eng.*, **62**, 91-109.

[14] Marohn, F. (1998a). An adaptive test for Gumbel domain of attraction. *Scand. J. Statist.*, **25**, 311–324.

[14] Marohn, F. (1998b). Testing the Gumbel hypothesis via the POT-method. *Extremes*, **1**, 191–213.

[15] Méndez, F.J., Menéndez, M., Luceño, A. and Losada, I.J. (2006). Estimation of the long term variability of extreme significant wave height using a time-dependent POT model. *J. Geophys. Res.*, **111**, C07024.

[16] Menéndez, M., Méndez, F.J., Losada, I.J. and Graham, N.E. (2008). Variability of extreme wave heights in the northeast pacific ocean based on buoy measurements. *Geophys. Res. Lett.* **35**, L22607.

[17] Mínguez, R., Menéndez, M., Méndez, F.J. and Losada, I.J. (2010). Sensitivity analysis of time-dependent generalized extreme value models for ocean climate variables. *Adv. Water Resour.* **33**, 833-845.

[18] Scotto M.G. and Guedes Soares C. (2007). Bayesian inference for long-term prediction of significant wave height. *Coastal Eng.*, **54**, 393-400.

[19] Scotto, M.G., Alonso, A.M. and Barbosa, S.M. (2010). Clustering time series of sea levels: extreme value approach. *J. Waterway, Port, Coastal, and Ocean Engrg.*, **136**, 215-225.

[21] Scotto, M.G., Barbosa, S.M. and Alonso, A.M. (2011a). Extreme value and cluster analysis of European daily temperature series. *J. Appl. Statist.*, **38**, 2793-2804.

[21] Scotto, M.G., Barbosa, S.M. and Alonso, A.M. (2011b). Model-Based Clustering of Extreme Sea Level Heights. Chapter 14. En: *Sea Level Rise, Coastal Engineering, Shorelines and Tides*, Nova Science Publishers, New-York (EE.UU).

[22] Vilar-Fernández, J.A., Alonso, A.M. and Vilar-Fernández, J.M. (2010). Nonlinear time series clustering based on nonparametric forecast densities. *Comput. Statist. Data Anal.*, **54**, 2850–2865.

[23] Wang, J.Z., Cooke, P. and Li, S. (1996). Determination of domains of attraction based on a sequence of maxima. *Austral. J. Statist.*, **38**, 173–181.

About the authors

Manuel G. Scotto is presently Assistant Professor at the Departamento de Matemática of the Universidade de Aveiro (Portugal). He completed his PhD in Statistics in 2001. His research interests center in applied probability and sometimes cross the boundary into statistics. Current topics of research gravitate towards problems in integer-valued time series analysis, forecasting, classification, extreme value theory and in applied statistics.

Susana M. Barbosa is Senior Researcher at the Instituto Dom Luiz of the Universidade de Lisboa (Portugal). Her research interests focus on time series analysis of geophysical data, particularly sea-level and climate records.

Andrés M. Alonso is Associate Professor at the Departamento de Estadística of the Universidad Carlos III de Madrid. His research interests are centered on applied statistics and applied econometrics. He had done research on resampling methods (bootstrap, subsampling and jackknife), classification methods, time series modeling and computation.