Introduction

A typical research question of many empirical studies in the area of environmental processes is how to compare the differences among time series with regard to their corresponding extremal behavior. The study of extreme values is of particular interest in a climate change context. For example, in studies of regional variability of time series of daily mean temperature it is important to identify locations exhibiting similar behavior in terms of their corresponding distributions of return values for a pre-defined period of time (see e.g., Scotto et al. [11]). Further examples can be found in the analysis of regional variability of tide gauge records. Knowledge about extreme sea-level is essential for prediction of flooding risks, coastal management and design of coastal infrastructure systems; see Scotto et al. [10] and the references therein.

Objective

In this work, a testing procedure based on subsampling techniques for testing the equality of Generalized Pareto distributions (GPDs) related with the excesses of two stationary time series that may not be considered independent is developed. The Kolmogorov-Smirnov distance is measured as metrics between the two corresponding extremal distribution functions. The procedure introduced by Alonso and Maharaj [1] is closely followed. The performance of the testing procedure is illustrated through a simulation study and by means of a real data analysis concerning the daily mean temperatures recorded over the period 1990 and 2004 in 17 locations in Spain.

The GPD

The GPD plays a central role in extreme value theory (EVT) as the limiting distribution of the sample of excesses above a high threshold. This method is commonly referred to as the Peak-Over-Threshold (POT) (Davidson and Smith, [4]; Pickands, [36]). The POT methodology has been widely applied to several areas, such as hydrology, precipitation, environment and ocean engineering just to mention a few. The cumulative distribution function of the GPD is given as:

\[ F(x;k,\sigma) = \begin{cases} 1 - \left( \frac{x - \theta}{\sigma} \right)^{-\frac{1}{\gamma}}, & k \neq 0, \\ 1 - e^{-\frac{x}{\theta}}, & k = 0 \end{cases} \]

where \(k\) and \(\sigma\) are the shape and scale parameters, respectively. For \(k \geq 0\), \(x > 0\), while \(0 < x < -\theta/k\) provided \(k < 0\). A review of methods for estimation the parameters of the GPD is done by, e.g., de Zee Buremzde and Koe [5, 36], for the general theory of EVT see e.g., Coles [4] or Embrechts et al. [7].

Procedure

Let \((X_1)\) and \((Y_1)\) be two strictly stationary processes with underlying models \(F_X\) and \(F_Y\), respectively. Let \(X_n = \{X_1, \ldots, X_n\}\) be a vector of observations from \((X_1)\) and \((Y_1)\), respectively. Let \(G_X\) and \(G_Y\) be the GPDs for the excesses above a sufficiently high threshold of \((X_1)\) and \((Y_1)\), respectively. We focus on the following two-sided problem of testing

\[ H_0: G_X = G_Y, \quad H_1: G_X \neq G_Y \]

i.e., we want to test if the limiting distribution of the excesses is the same in both processes. To this extend, we consider a distance-based test statistic

\[ T_j = d(G_{X_j}, G_{Y_j}), \]

where \(G_{X_j}\) and \(G_{Y_j}\) are the estimated GPDs from the subsamples \(X_j\) and \(Y_j\), respectively. The normalization constant \(n_j = \sqrt{n(n + 1)(2n + 1)/6}\) is the GPD’s estimate from the sample \(X = \{X_1, \ldots, X_n\}\) and \(Y = \{Y_1, \ldots, Y_n\}\) be vectors of observations from \((X_1)\) and \((Y_1)\), respectively. Let \(G_X\) and \(G_Y\) be the GPDs for the excesses above a sufficiently high threshold of \((X_1)\) and \((Y_1)\), respectively. We focus on the following two-sided problem of testing

\[ H_0: G_X = G_Y, \quad H_1: G_X \neq G_Y \]

The analysis of the daily mean temperature collected at 17 locations in Spain clearly identifies a difference along the 15 years (see Figure 1 (left)). The corresponding daily exceedances are presented in Figure 1 (center). The box-plots of the exceedances above the \(T_{250}\) are contained in Figure 1 (right) for all the communities (Cantabria and Melilla are not considered). The exceedance temperatures are not independent as required by the classical EVT. The estimates of the extremal index \(\theta\) (see e.g. Coles [36]), for the exceedances obtained for \(k = 2\) are presented in Table 3. In this situation, \(\theta = 2\) observations above \(k\) belong to different clusters if there are, at least, two consecutive observations below \(k\). The table also contains the estimated \(\lambda\)'s, \(\hat{\lambda}_x\) and \(\hat{\lambda}_y\), of the GPD parameters, obtained by the Extremal Percentile Method (Castillo and Hadi [2]). The estimates of \(\theta\) correspond, roughly, to 2 or 3 very hot consecutive days. They show a reasonable “independence” structure of the highest temperatures. The estimates of \(\theta\) are all negative reflecting underlying light-tailed distributions with a finite upper bound.

Modeling the high temperatures

The mean daily temperatures exhibit the usual yearly seasonal variation, although there is no apparent trend along the 15 years (see Figure 2 (right)). The corresponding daily exceedances are presented in Figure 2 (center). The box-plots of the exceedances above the \(T_{250}\) are presented in Figure 2 (right) for all the communities (Cantabria and Melilla are not considered). The exceedance temperatures are not independent as required by the classical EVT. The estimates of the extremal index \(\theta\) (see e.g. Coles [36]), for the exceedances obtained for \(k = 2\) are presented in Table 3. In this situation, \(\theta = 2\) observations above \(k\) belong to different clusters if there are, at least, two consecutive observations below \(k\). The table also contains the estimated \(\lambda\)'s, \(\hat{\lambda}_x\) and \(\hat{\lambda}_y\), of the GPD parameters, obtained by the Extremal Percentile Method (Castillo and Hadi [2]). The estimates of \(\theta\) correspond, roughly, to 2 or 3 very hot consecutive days. They show a reasonable “independence” structure of the highest temperatures. The estimates of \(\theta\) are all negative reflecting underlying light-tailed distributions with a finite upper bound.

Comments and future work

The analysis of the daily mean temperature collected at 17 locations in Spain clearly identifies a difference between the four northeast communities on the shores of the Bay of Biscay and the remaining communities. A clear direction is also detected between AND, BIC, CLM, and MUR (the communities that exhibit the highest temperatures) from the remaining locations. The inclusion of covariates into the parameters of the GPD in a regression-like approach, allowing for trends or cycles in the upper tail, would be also relevant for classification purposes. This remains a topic for future research.

References