Clustering Seasonal Time Series Using Extreme Value Analysis: An Application to Spanish Temperature Time Series

Elizabeth A. Maharaj[a], Andrés M. Alonso[b], Pierpaolo D'Urso[c]

[a]*Department of Econometrics and Business Statistics, Monash University, Australia*
[b]*Department of Statistics and Instituto Flores de Lemus, Universidad Carlos III de Madrid, Spain*
[c]*Dipartimento di Scienze Sociali ed Economiche, Sapienza - Universitá di Roma, Italy*

Abstract

A challenging aspect of grouping together regional temperature time series is that some regions have similar summer temperatures but different winter temperatures and vice versa. We explore this by applying cluster analysis to regional temperature time series in Spain using as features the parameter estimates of location, scale and shape, obtained from fitting the generalised extreme value (GEV) distribution to the block maxima and block minima of the series. Using this approach, our findings reveal that the identified clusters can be meaningfully interpreted and are well validated. The motivation for using this approach is that each time series is represented by just three easily extracted features. If features were to be extracted as a result of conventional time series modelling, they are likely to be impacted upon by the uncertainty of model selection. This is not the case with GEV modelling. Furthermore, GEV modelling enables long term projections of the maxima and minima which cannot otherwise be achieved from conventional time series modelling. For comparison purposes, we also explore clustering the block maxima and block minima of the times series. In addition, we explore the performance of this approach using simulated data.

Keywords:
Block Maxima, Generalised Extreme Value Analysis, Clustering, Returns.

1. Introduction

The analyses of extreme temperatures and sea levels are important tasks in this era of awareness of the effects of climate change. Many authors have used extreme value analysis to study sea level extremes (e.g., Tsimpis and Blackman [21], Unikrishnan et al. [22], Mendez et al. [15], Scotto et al. [19]) and temperature extremes (e.g., Scotto et al. [20], Alonso et al. [1]). In particular, Scotto et al. [19] combined a Bayesian Analysis of extreme sea levels to estimate predictive distributions, with hierarchical cluster analysis to distinguish groups of North Atlantic sea locations. Scotto et al. [20] applied the same methodology to European daily temperature series to group together similar locations, while Alonso et al. [1] compared Generalised Pareto models fitted to extreme temperature observations.

Accepted at Communications in Statistics - Case Studies and Data Analysis
https://doi.org/10.1080/23737484.2016.1179140

Because of the variability of climatic conditions across the regions of Spain, it is known that some regions experience similar summer and similar winter conditions while others experience similar summer conditions but different winter conditions, or different summer conditions but similar winter conditions. Hence, we are particularly interested in examining and modelling the series of annual maximum and minimum temperatures from the various regions to identify regions with similar climatic conditions. To this end, we use the estimates of the GEV distributions fitted to the series of annual maxima and annual minima as clustering features. As well, we cluster the actual maxima and minima and compare the groupings to those obtained from clustering the GEV estimates.

While most of these previous studies using clustering methods focussed on grouping together locations based on predictive distributions, the focus of our study is grouping the temperature series across the available record. We consider daily temperature time series from the provinces in Spain over a 15-year period. We use the GEV parameter estimates based on the series of the r largest and r smallest values of the blocks. In addition, we also use the block maxima and minima as clustering and classification features. We cluster the series using conventional non-hierarchical methods, namely, k-mean and k-medoids and we use a classification method, namely, the k-nearest neighbour (k-NN) algorithm to validate our cluster solutions.

We also conduct a simulation study to evaluate the performance of non-hierarchical cluster analysis using the GEV features where, specifically we emulate the scenario where we might have regions with similar summer maxima but with varying degrees of different winter minima. Here we obtain GEV estimates from the series of block maxima and minima as well as from the series of the r largest and r smallest values. For comparison purposes, we also perform the cluster analysis using the actual block maxima and minima.

The rest of the paper is organised as follows. We briefly describe the generalised extreme value (GEV) distribution and the cluster and classification procedures in Section 2, while in Section 3, we describe the daily times series under consideration, report and discuss the cluster solutions. In Section 4, we describe and report the results of the simulation study and conclude in Section 5.

2. Methods

2.1. Extreme Value Analysis

Extreme value analysis has relevance in areas such as flood frequency analysis, environmental science, insurance and finance (Reiss and Thomas [16]). In particular, the modelling of extreme values takes two forms; the method of maxima over fixed intervals, and the method of exceedance over high thresholds. In what follows, we focus on the method of maxima over fixed intervals by making use of the generalized extreme value (GEV) distribution. We also consider fitting the GEV distribution to the r-largest values over fixed intervals.

The generalised extreme value (GEV) distribution is a family of continuous probability distributions developed within extreme value theory to combine the Gumbel,

Fréchet and Weibull families, also known as Type I, II and III extreme value distributions, respectively. As a result of the extreme value theorem, the GEV distribution is the limiting distribution of normalised maxima of a sequence of independent and identically distributed random variables. Hence, the GEV distribution is used as an approximation to model the maxima of a long finite sequences of random variables. The GEV distribution has the following form:

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

defined on $\{x : 1+\xi(\frac{x-\mu}{\sigma}) > 0\}$ where $-\infty < \mu < \infty$, $\sigma > 0$, and $-\infty < \xi < \infty$. The three parameters μ, σ and ξ are the location, scale and shape parameters, respectively. The shape parameter determines the three extreme value types. When $\xi < 0$, $\xi > 0$ or $\xi = 0$, the GEV distribution is the negative Weibull, Fréchet and Gumbel distribution, respectively. This is assumed to be the case by taking the limit of Equation 1 as $\xi \to 0$. For maxima of m years, the log-likelihood function for the annual maxima is given by

$$\begin{aligned}\ell(\mu,\sigma,\xi) &= -m\log(\sigma) - (1+1/\xi)\sum_{i=1}^{m}\log\left[1+\xi\left(\tfrac{x_i-\mu}{\sigma}\right)\right] \\ &\quad -\sum_{i=1}^{m}\left[1+\xi\left(\tfrac{x_i-\mu}{\sigma}\right)\right]^{-1/\xi},\end{aligned} \tag{2}$$

provided $1+\xi\left(\frac{x_i-\mu}{\sigma}\right) > 0$ for $i = 1, 2, \ldots, m$. The expression in Equation 2 is valid for $\xi \neq 0$. For $\xi = 0$, the log-likelihood function for the annual maxima is given by

$$\ell(\mu,\sigma) = -m\log(\sigma) - \sum_{i=1}^{m}\left(\frac{x_i-\mu}{\sigma}\right) - \sum_{i=1}^{m}\exp\left[-\left(\frac{x_i-\mu}{\sigma}\right)\right]. \tag{3}$$

The above log-likelihood expression presents a difficulty in extreme value analysis when the number of extreme events is small. This is particularly severe when the method of maxima over fixed intervals is used. As mentioned in Coles [5], a possible solution is to consider the r-largest values over fixed intervals.

The number of largest values per year, r, should be chosen carefully since small values of it will produce likelihood estimators with high variance, whereas large values of r will produce biased estimates. In practice, r is selected as large as possible, subject to adequate model diagnostics. The validity of the models can be checked through the application of graphical methods, in particular, the probability plot, the quantile plot and the return level plot. For further details, see Coles [5] and the references therein.

The implications of a fitted extreme value model are usually made with reference to extreme quantiles. By inversion of the GEV distribution function, the quantile, x_p, for a specified exceedance probability p is

$$x_p = \mu - \frac{\sigma}{\xi}\left[1-(-\log(1-p)^{-\xi})\right] \text{ for } \xi \neq 0, \tag{4}$$

and

$$x_p = \mu - \sigma \log[-\log(1-p)] \text{ for } \xi = 0. \tag{5}$$

x_p is referred to as the return level associate with a return period $1/p$. It is expected to be exceeded by the annual maximum in any particular year with probability p.

While in most applications of extreme value theory, attention is focussed on the extreme quantiles of the GEV distribution, our main focus is on fitting GEV distributions to the series of block maxima or to the series of the r largest values per block of our times series, estimating the parameters and using these parameter estimates as the features for clustering and classification. In the application and simulation study, we will also fit the GEV distribution to the block minima or to the r smallest values per block. Note that the GEV distribution is fitted to the block minima by applying Equation 2 or Equation 3 to its negative values. Likewise, the GEV distribution is fitted to the r smallest values per block by applying the relevant equation (Coles [5] to its negative values.

2.2. Clustering Methods

The k-means clustering method generates a specific number of disjoint, non-hierarchical clusters. It is well suited to generating spherical clusters. The k-means method is numerical, unsupervised, non-deterministic and iterative.

Given an initial set of randomly assigned k centroids, $m_1^{(1)}, m_1^{(1)}, \cdots, m_1^{(k)}$, the method assigns each observation of a data set to the cluster with the closest centroid by determining the distance from the observation to each cluster centroid. It then determines a new mean for each cluster to be the centroid of the observations. If the data point is closest to its own cluster's centroid, it remains where it is, whereas if the data point is not closest to its own cluster's centroid, it is assigned to the cluster with the closest centroid. The process is repeated until a complete pass through of all the observations results in no observations moving from one cluster to another. At this point the clusters are stable and the clustering process ends. The choice of the initial partition can greatly affect the final clusters that result in terms of inter-cluster and intra-cluster distances and cohesion. Measures such as the Euclidean, Manhattan and Minkoski distances can be used.

The k-medoids method is related to the k-means method in that both attempt to minimise the distance between the observations assigned to a cluster and the point designated as the cluster centroid. However, the k-medoids method chooses actual observations which are referred to as medoids, as centres. This method is more robust to noise and outliers compared to the k-means method. Refer to Hastie et al. [12] for more details of these clustering methods.

For both of these methods, a useful tool for determining an appropriate number of clusters, k, is the silhouette method (Rousseeuw [17]). However, there are many other methods that are also available (e.g., refer to Charrad et al. [4]). We also explore the use of these methods to determine the appropriate number of clusters.

It should be noted that in the application, the observations referred to above are the extracted features of the time series. in particular, we use two sets of features:

(i) annual maximum and minimum temperatures; and (ii) estimates of the three GEV parameters fitted to the r largest (smallest) values of the summer (winter) season. That is, in the first set we have 15 variables related to summer temperatures and 15 variables related to winter temperatures, while in the second set we have three variables related to GEV distribution fitted to largest temperatures and three variables related to GEV distribution fitted to smallest temperatures. The algorithm could be summarize in the following step:

1. Given a set of n time series: $\mathcal{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$, extract the selected set of features related to summer and winter seasons, that is, set (i) or (ii).
2. Obtain the cluster solution using the extracted features as inputs.

In this last step, we use the clustering procedures for only summer related variables, for only winter related variables and for both. This allows us to illustrate the differences in the climatic conditions across regions.

2.3. Classification Method

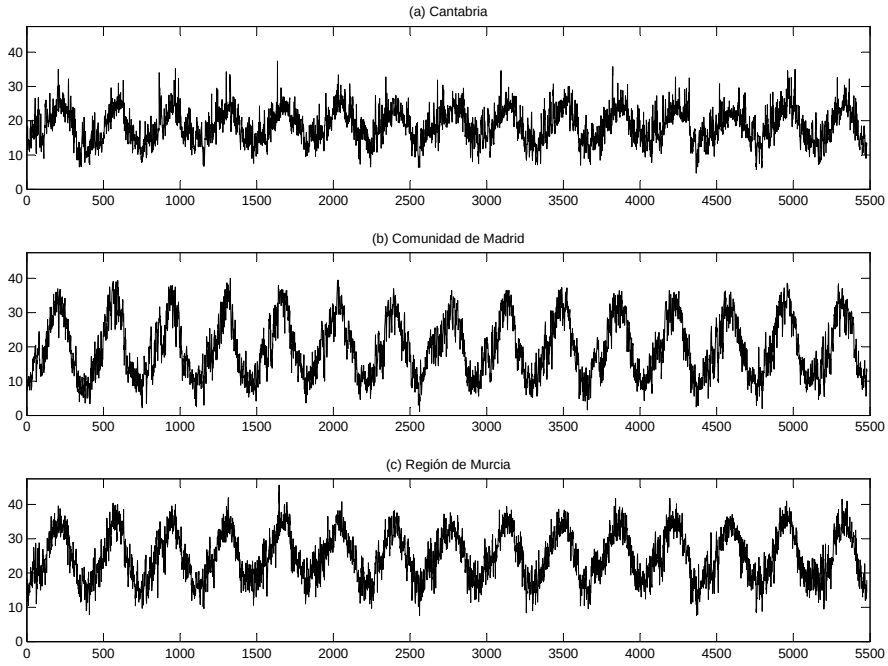
The k-nearest neighbours classification algorithm, k-NN, is a non-parametric method that predicts class memberships of observations based on the k closest training examples in the feature space. k-NN is a type of instance-based learning, or lazy learning procedure where the function is only approximated locally and all computation is deferred until classification (Altman [2]). An instance-based learning procedure does not use the training data points for generalisation during the learning phase. The k-NN is a simple algorithm that stores all available cases based on a similarity measure such as the Euclidean distance. An observation is assigned to the class most common amongst its k nearest neighbours, where k is a positive integer and is typically small. If $k = 1$, then the observation is simply assigned to the class of that single nearest neighbour. In general, a large k value is probably more precise since it reduces the overall noise. Cross-validation is one way of determining a good k value by using an independent dataset to validate the k value. Historically, the optimal k for most datasets has been between 3-10. Refer to Hastie et al. [12] for more details.

It should be noted that in the application, k-NN is used as a validation technique since we do not have an *a priori* class memberships of the observations. We will use the results of the clustering algorithms as if they were the "true" class of the observations.

3. Application

We consider the time series of daily maximum temperatures (in degrees Celsius, ^oC) observed in fifty provinces and two autonomous cities of Spain from 1990 to 2004. The temperatures exhibit the usual annual seasonal variation although there are no apparent trends across the years. As examples, the temperatures observed in *Cantabria* (a province in northern Spain with a humid oceanic climate), *Madrid* (a central province with continental Mediterranean climate) and *Murcia* (a province in south east Spain with Mediterranean climate) are plotted in Figure 1. A labeled provincial map of Spain

Figure 1: Daily maximum temperatures observed (1990 -2004) in (a) *Cantabria*; (b) *Madrid* and (c) *Región de Murcia*.

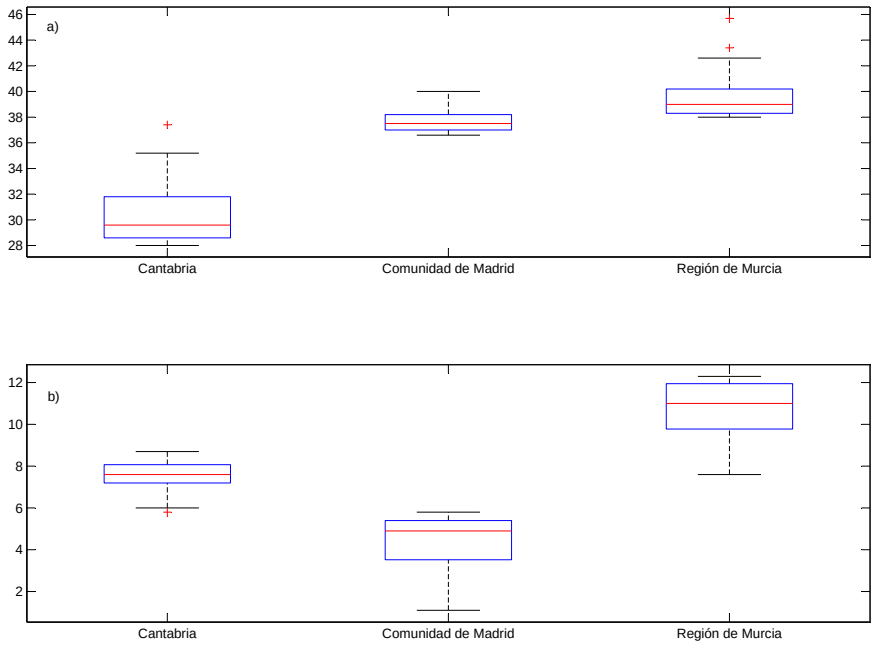


can be found at [23]. Other studies concerning the extremes in the Iberian Peninsula can be found in Alonso et al. [1], Fernández-Montes and Rodrigo [6], Furió and Meneu [7], Brunet et al. [3] and García-Herrera et al. [8].

The rest of the regions exhibit similar seasonal patterns to those in Figure 1. Figure 2(a) shows the values above the 95% percentile during the summer period and Figure 2(b) shows the values below the 5% percentile during the winter period. For simplicity, for all considered years, we define the period from June 21th until September 20th as summer, and the period from December 21th until March 20th as winter. Although it is clear that summer/winter occurs in the same period for the three regions, there are differences in the temperature extreme values. For instance, *Murcia* and *Madrid* have hotter summers compared to *Cantabria* while *Madrid* has cooler winters compared to *Cantabria* and *Murcia*.

First, we present the cluster solutions obtained by the k-means procedure using the annual maxima and minima as input features, that is, the 15 annual maxima during summer periods, and the 15 annual minima during the winter periods. Here and in what follows Euclidean distances are calculated from each observation to each cluster centroid. We perform three separate cluster analyses: (i) using only summer related variables; (ii) using only winter related variables and (iii) using both sets of variables. Notice that the first two analyses correspond to one block and the third analysis correspond to two blocks. The average silhouette statistics suggests using $k = 2$ for the first two analyses, and $k = 2$ or 3 for the third analysis. Using the R-Package, NbClust (Charrad et al. [4]) we explored using up to 30 different measures to determine the

Figure 2: (a) Box-plot of the exceedances (1990-2004) above the 95% percentile during summer period; (b) Box-plot of the exceedances below the 5% percentile during winter period.



optimal number of clusters for each of the three scenarios and we found that the majority of methods suggest that the best number of clusters is 2. The two maps in Figure 3 represent the obtained cluster solutions for $k = 2$ using only the summer or winter related variables and two maps in Figure 4 represent the obtained cluster solutions for $k = 2$ and $k = 3$ using both sets of variables.

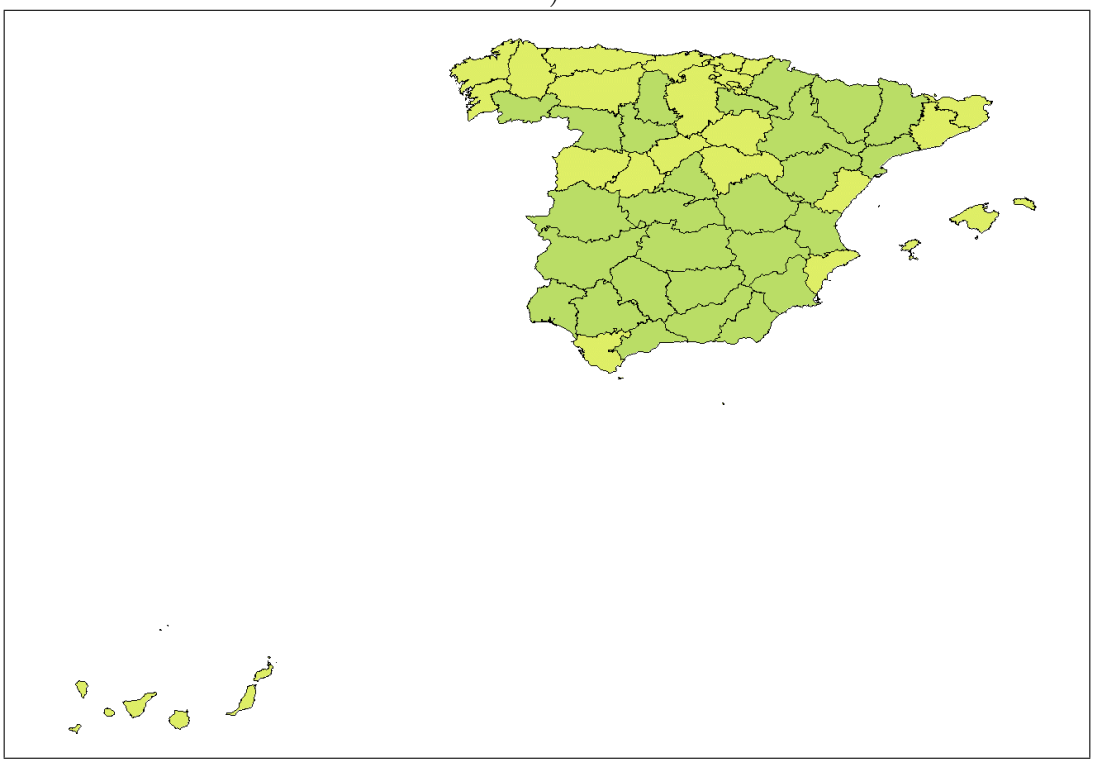
The map in Figure 3(a) corresponds to a two-cluster solution based on annual maxima. The means of the maximum temperatures in the two clusters are 34.68 and 39.14, that is, there is a cluster with mild summers (for instance, *Cantabria* belongs to this cluster) and a cluster with hot summers (for instance, (*Madrid* and *Murcia* belong to this cluster).

The map in Figure 3(b) corresponds to a two-cluster solution based on annual minima. The means of the minimum temperatures in the two clusters are 2.89 and 10.61, that is, there is a cluster with cool winters (for instance, *Madrid* belongs to this cluster) and a cluster with mild winters (for instance, (*Cantabria* and *Murcia* belong to this cluster). It is clear from these two maps that the two-cluster solutions differ in many regions. The similarity index between these two-cluster solutions is 0.5712 and this suggests a moderate disagreement in the classification based on summer only and winter only data.

The map in Figure 4(a) corresponds to a two-cluster solution based on annual minima and maxima. It is clear that this map coincides with the previous map, that is, the variables related to minimum temperature dominate the cluster solution. The means of the maximum (minimum) temperatures in the two clusters are 37.19 (2.89) and

Figure 3: (a) Two clusters based on summer variables; (b) Two clusters based on winter variables.

a)



b)

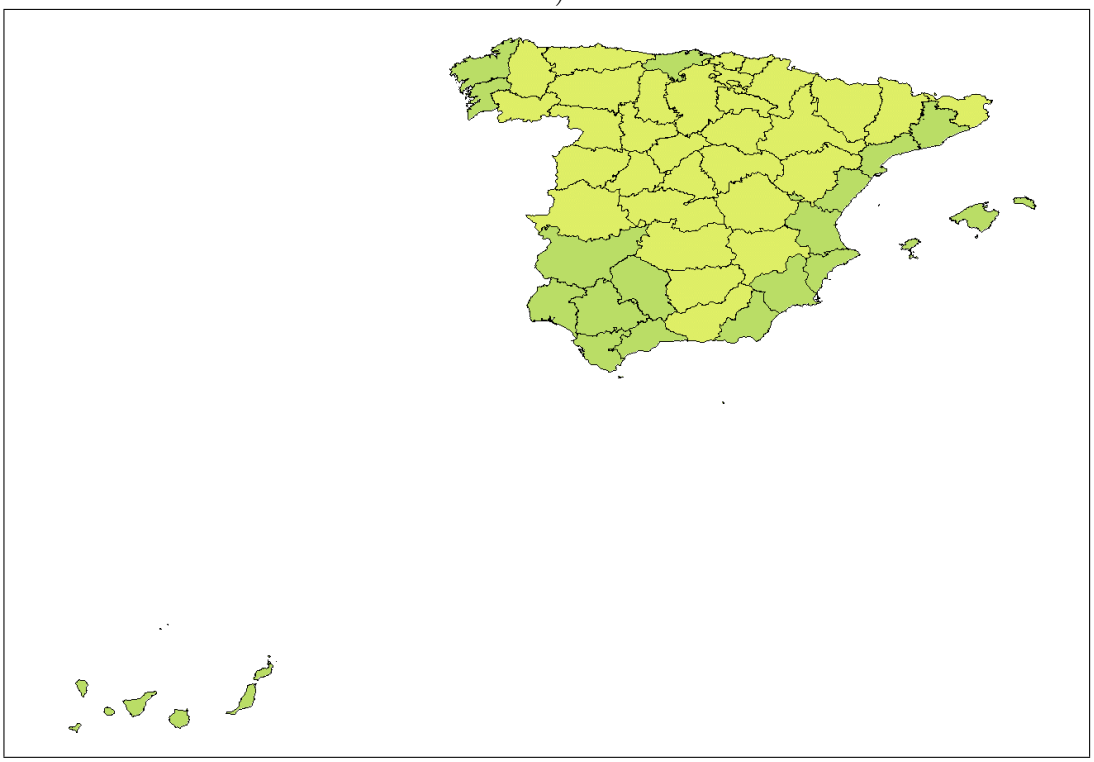
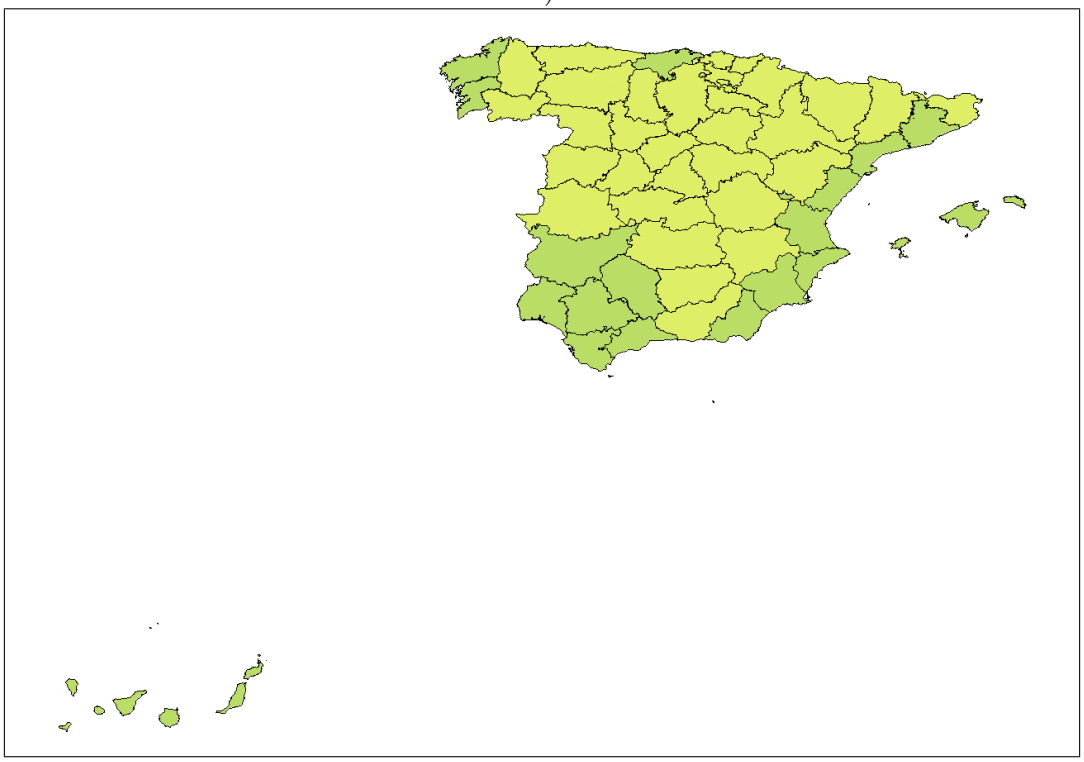
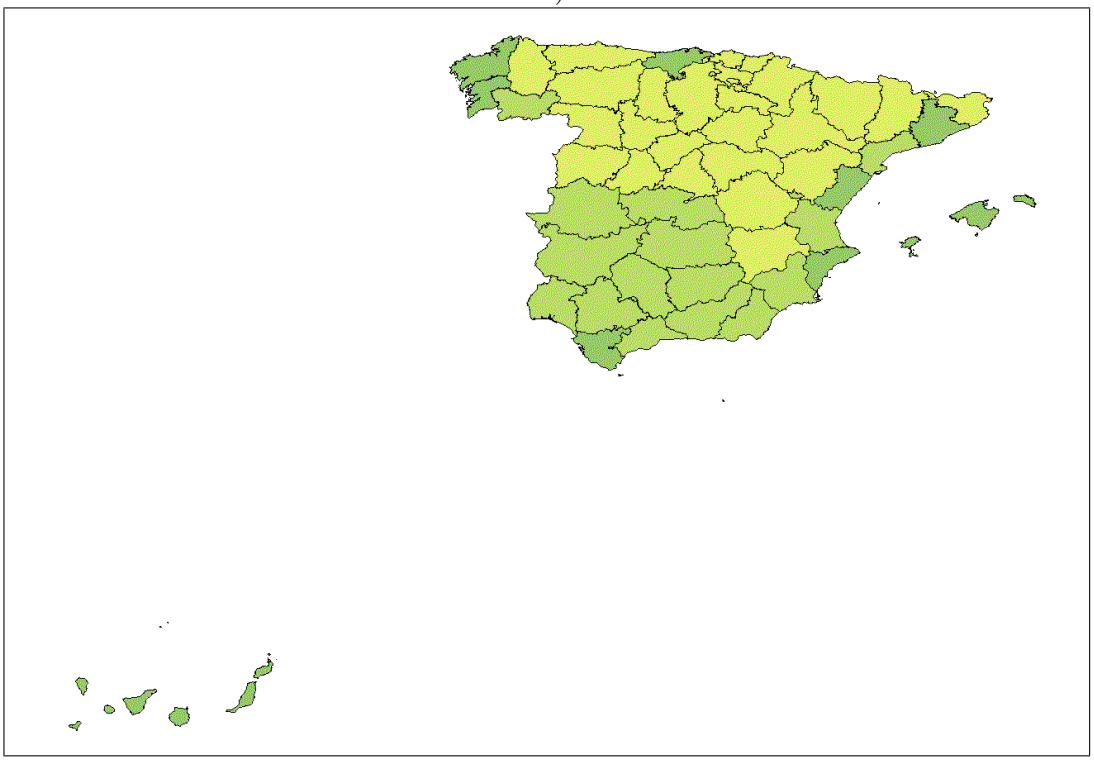


Figure 4: (a) Two clusters based on summer and winter variables; (b) Three clusters based on summer and winter variables.

a)



b)



36.70 (10.70). Notice that the means of the maximum temperatures of these clusters are close, the difference is less than 0.5 degrees Celsius, and this lends support to the fact that these clusters do not take into account the variables related to maximum temperature. For this additional reason, in Figure 4(b) we consider a three-cluster solution based on annual minima and maxima. For comparison purposes, the colour is related to the mean of the minimum temperatures. The means of the maximum (minimum) temperatures in the three clusters are 36.51 (2.24), 40.11 (8.30) and 34.11 (11.15). The temperatures of the first cluster correspond to hot summers and cooler winters (for instance, *Madrid* belongs to this cluster); the temperatures of the second cluster correspond to hotter summers and cool winters (for instance, *Murcia* belongs to this cluster) and temperatures of the third cluster correspond to mild summers and winters (for instance, *Cantabria* belongs to this cluster). The similarity indexes between the three-cluster solution and the previous two-cluster solutions are 0.6276 and 0.7067, respectively.

In Tables 1 to 4, we report the estimates of the GEV distribution fitted to the series of four annual highest (lowest) temperatures from 1990 to 2004 in each of the fifty Spanish provinces and the two autonomous cities, together with their standard errors. Since we have at most 15 maxima (or minima), we use the r-largest (smallest) values approach where we select r=4, hence estimating the GEV parameters from 60 highest (lowest) temperatures each time. Note that in order to extract the r-largest (smallest) observations within each season, we follow the procedure adopted by Guedes Soares and Scotto [10], that is, first we obtain the maximum (minimum) of the season and we exclude a week of observations around this maximum (minimum); then we obtain the second largest (smaller) value among the non-excluded observations. This exclusion guaranties that the first and second largest observations could be considered as independent. For the next largest (smallest) values we proceed in a similar way.

Figures 5 to 10 show four diagnostic plots from fitting GEV distributions to the series of four annual highest and lowest temperatures from 1990 to 2004 for each of three regions, namely, Cantabria, Madrid and Murcia. If the GEV fit is a reasonable estimated model for the corresponding population distribution, the points of the probability and quantile plots should lie close to the diagonal. The probability and quantile plots contain the same information but on different scales. The return level curve should be asymptotic to a finite level if the estimate of the shape paramater is negative. If the shape estimate is close to zero, the return level curve should be approxaimately linear. Finally, the curve of the density plot should be more or less consistent with the histogram of the data. From each set of these four plots, we observe that with the exception of the GEV fits to the Madrid highest temperatures and the Cantabria lowest temperatures, the other fits appear to be quite reasonable. GEV fits to the highest and lowest temperatures for the other regions also displayed mixed results.

Given the GEV estimates, we can now proceed as in the previous analysis. First, we obtain two-cluster solutions based on the GEV estimates for the summer period and for the winter period, then we obtain the two- and three-cluster solutions for combined sets of GEV estimates for the summer and winter periods. The four maps in Figures 11

Table 1: GEV estimates with standard errors for highest and lowest temperatures in the Spanish provinces: 1990-2004 - Part I.

	GEV estimates					
	Highest temperatures			**Lowest temperatures**		
Province	**Location**	**Scale**	**Shape**	**Location**	**Scale**	**Shape**
Alava	35.90	1.75	-0.57	2.52	1.82	-0.60
	0.39	0.15	0.10	0.45	0.17	0.13
Albacete	38.36	1.41	-0.57	4.73	1.82	-0.62
	0.32	0.12	0.09	0.42	0.17	0.10
Alicante	35.62	1.56	-0.15	11.71	1.60	-0.29
	0.34	0.16	0.06	0.38	0.16	0.12
Almeria	38.00	1.61	-0.56	13.00	1.44	-0.18
	0.36	0.14	0.10	0.34	0.19	0.15
Avila	34.50	1.54	-0.43	0.65	2.34	-0.48
	0.34	0.12	0.08	0.54	0.20	0.10
Badajoz	41.39	1.72	-0.42	8.61	1.89	-0.43
	0.38	0.14	0.08	0.45	0.18	0.12
Islas Baleares	34.59	1.50	-0.19	10.87	1.49	-0.34
	0.33	0.16	0.11	0.36	0.15	0.13
Barcelona	32.28	1.66	-0.17	8.34	2.01	-0.33
	0.37	0.18	0.09	0.47	0.19	0.10
Burgos	36.14	1.72	-0.61	1.02	1.56	-0.65
	0.38	0.14	0.08	0.39	0.15	0.13
Cáceres	40.11	1.27	-0.63	6.33	2.25	-0.52
	0.29	0.11	0.10	0.52	0.18	0.09
Cádiz	34.58	1.82	-0.43	12.43	1.44	-0.37
	0.40	0.15	0.09	0.33	0.13	0.09
Castellon	34.58	1.82	-0.43	10.02	1.70	-0.33
	0.40	0.15	0.09	0.41	0.18	0.13
Ceuta	31.70	1.97	-0.12	12.93	1.11	-0.20
	0.44	0.24	0.12	0.26	0.13	0.12

Table 2: GEV estimates with standard errors for highest and lowest temperatures in the Spanish provinces: 1990-2004 - Part II.

	GEV estimates					
	Highest temperatures			**Lowest temperatures**		
Province	**Location**	**Scale**	**Shape**	**Location**	**Scale**	**Shape**
Cordoba	42.81	1.75	-0.33	9.89	1.74	-0.42
	0.39	0.16	0.10	0.40	0.14	0.08
Coruña	28.68	2.11	-0.03	9.27	1.48	-0.39
	0.47	0.29	0.12	0.35	0.13	0.11
Ciudad Real	40.07	1.46	-0.37	5.06	2.32	-0.44
	0.32	0.11	0.07	0.54	0.20	0.10
Cuenca	36.63	1.38	-0.46	3.13	1.90	-0.58
	0.30	0.11	0.08	0.45	0.16	0.10
Gerona	35.56	2.18	-0.23	6.96	2.12	-0.27
	0.49	0.22	0.10	0.50	0.22	0.11
Granada	39.66	1.47	-0.51	6.73	1.78	-0.55
	0.33	0.13	0.10	0.42	0.16	0.12
Guadalajara	35.22	1.40	-0.44	1.13	2.53	-0.48
	0.31	0.11	0.07	0.59	0.21	0.09
Guipuzcoa	32.86	2.92	-0.40	3.51	1.67	-0.80
	0.66	0.25	0.10	0.42	0.19	0.14
Huelva	39.74	1.95	-0.36	11.68	1.75	-0.35
	0.44	0.17	0.09	0.41	0.16	0.11
Huesca	38.12	1.19	-0.56	3.30	3.36	-0.14
	0.27	0.11	0.11	0.77	0.41	0.12
Jaen	39.04	1.46	-0.49	6.36	1.85	-0.48
	0.32	0.12	0.09	0.43	0.16	0.10
Leon	34.27	1.42	-0.63	0.80	2.05	-0.42
	0.32	0.12	0.09	0.48	0.18	0.10
La Rioja	38.32	1.54	-0.57	3.40	2.00	-0.39
	0.36	0.14	0.12	0.46	0.17	0.09

Table 3: GEV estimates with standard errors for highest and lowest temperatures in the Spanish provinces: 1990-2004 - Part III.

	GEV estimates					
	Highest temperatures			**Lowest temperatures**		
Province	**Location**	**Scale**	**Shape**	**Location**	**Scale**	**Shape**
Lugo	33.99	2.80	-0.14	5.18	2.12	-0.11
	0.63	0.33	0.12	0.48	0.27	0.11
Lerida	38.25	1.14	-0.71	2.89	3.10	-0.21
	0.26	0.11	0.10	0.72	0.33	0.10
Madrid	37.68	1.34	-0.52	4.16	1.76	-0.48
	0.30	0.11	0.09	0.42	0.16	0.12
Malaga	38.90	1.99	-0.31	12.53	1.30	-0.32
	0.44	0.16	0.07	0.30	0.12	0.10
Melilla	35.55	2.29	-0.22	12.53	1.34	-0.21
	0.50	0.23	0.08	0.31	0.15	0.12
Murcia	39.71	1.77	-0.15	10.42	1.76	-0.51
	0.39	0.19	0.09	0.43	0.16	0.13
Orense	38.82	1.94	-0.34	6.15	2.15	-0.50
	0.44	0.18	0.11	0.50	0.18	0.10
Oviedo	30.96	2.29	-0.32	5.56	1.90	-0.41
	0.52	0.22	0.11	0.45	0.17	0.12
Palencia	36.95	1.47	-0.66	2.14	2.15	-0.41
	0.33	0.14	0.10	0.50	0.20	0.11
Las Palmas	32.09	2.11	-0.07	18.35	1.07	-0.25
	0.47	0.27	0.13	0.24	0.11	0.09
Pamplona	37.57	1.55	-0.49	2.35	1.79	-0.71
	0.34	0.12	0.08	0.47	0.20	0.16
Pontevedra	34.30	2.08	-0.41	8.52	1.57	-0.24
	0.48	0.18	0.11	0.36	0.16	0.09
Salamanca	36.60	1.39	-0.42	2.56	2.38	-0.43
	0.31	0.11	0.08	0.58	0.24	0.15

Table 4: GEV estimates with standard errors for highest and lowest temperatures in the Spanish provinces: 1990-2004 - Part IV.

	GEV estimates					
	Highest temperatures			**Lowest temperatures**		
Province	**Location**	**Scale**	**Shape**	**Location**	**Scale**	**Shape**
Santa Cruz de Tenerife	34.13	2.07	-0.15	18.00	0.97	-0.35
	0.46	0.23	0.10	0.23	0.09	0.10
Cantabria	31.14	2.64	-0.22	7.76	1.35	-0.63
	0.61	0.30	0.13	0.33	0.13	0.12
Segovia	35.88	1.42	-0.43	0.90	2.19	-0.64
	0.32	0.12	0.09	0.52	0.22	0.12
Sevilla	42.28	1.76	-0.28	11.59	1.61	-0.32
	0.39	0.16	0.09	0.37	0.15	0.09
Soria	35.26	1.31	-0.56	1.16	1.95	-0.45
	0.29	0.11	0.09	0.45	0.16	0.09
Tarragona	37.19	1.33	-0.32	8.20	2.29	-0.30
	0.30	0.12	0.09	0.54	0.24	0.12
Teruel	36.83	1.25	-0.52	2.36	2.80	-0.28
	0.28	0.10	0.08	0.65	0.27	0.09
Toledo	40.42	1.28	-0.60	5.72	2.63	-0.19
	0.28	0.11	0.09	0.63	0.32	0.15
Valladolid	37.68	1.39	-0.50	2.44	2.00	-0.43
	0.31	0.11	0.07	0.46	0.18	0.10
Valencia	36.79	2.10	-0.24	10.75	1.72	-0.22
	0.47	0.20	0.09	0.40	0.20	0.13
Vizcaya	35.62	2.57	-0.33	6.88	1.83	-0.33
	0.56	0.20	0.06	0.43	0.37	0.24
Zamora	37.33	1.61	-0.35	2.56	2.25	-0.40
	0.35	0.13	0.08	0.54	0.23	0.13
Zaragoza	38.82	1.58	-0.34	4.01	2.44	-0.38
	0.35	0.13	0.08	0.57	0.23	0.11

Figure 5: Diagnostic plots of the GEV fit to highest temperatures in Cantabria

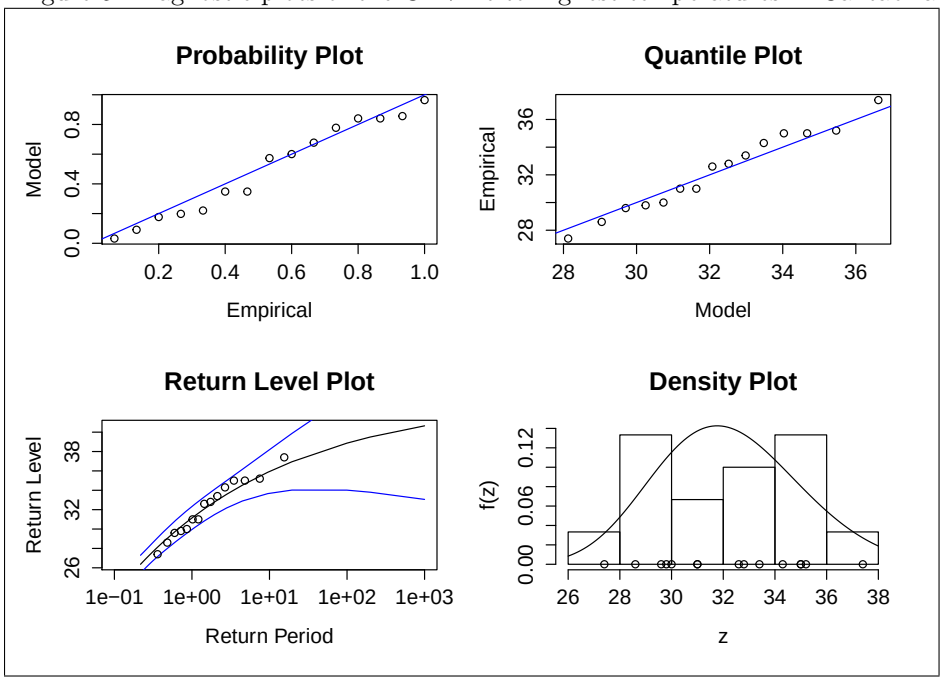


Figure 6: Diagnostic plots of the GEV fit to lowest temperatures in Cantabria

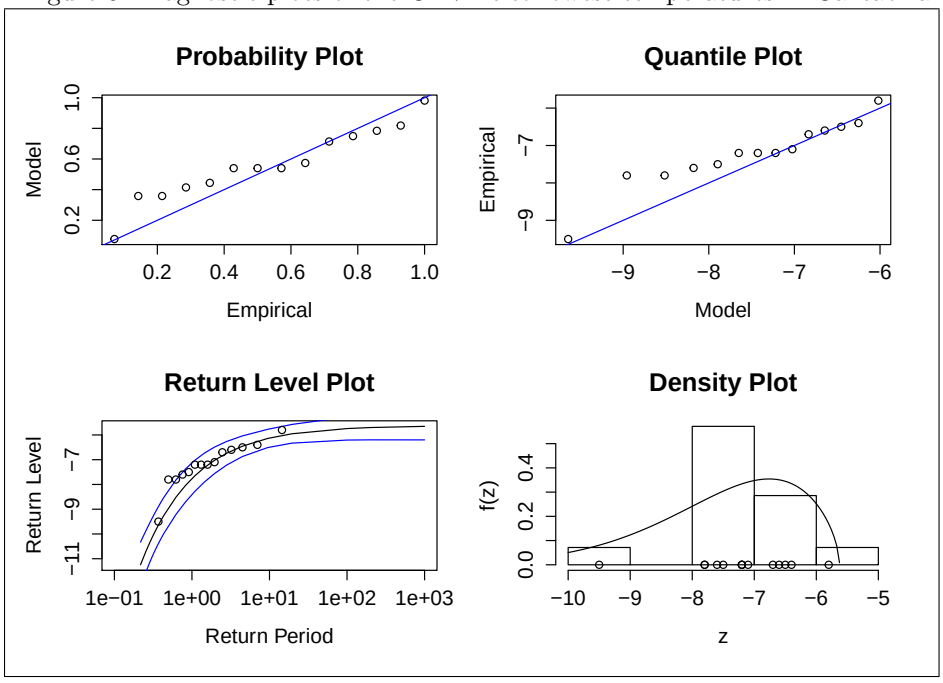


Figure 7: Diagnostic plots of the GEV fit to highest temperatures in Madrid

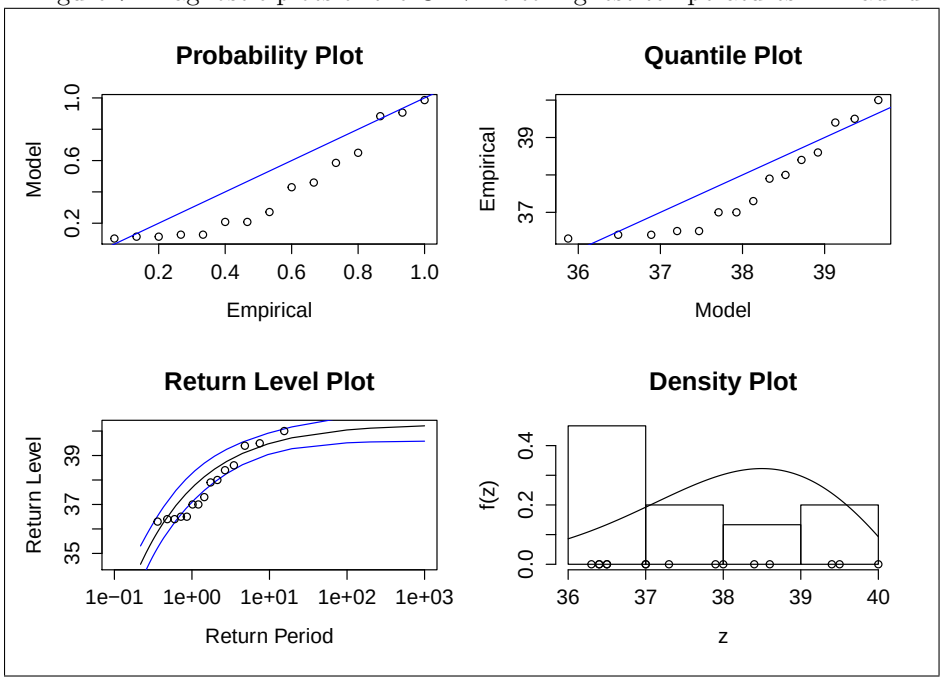


Figure 8: Diagnostic plots of the GEV fit to lowest temperatures in Madrid

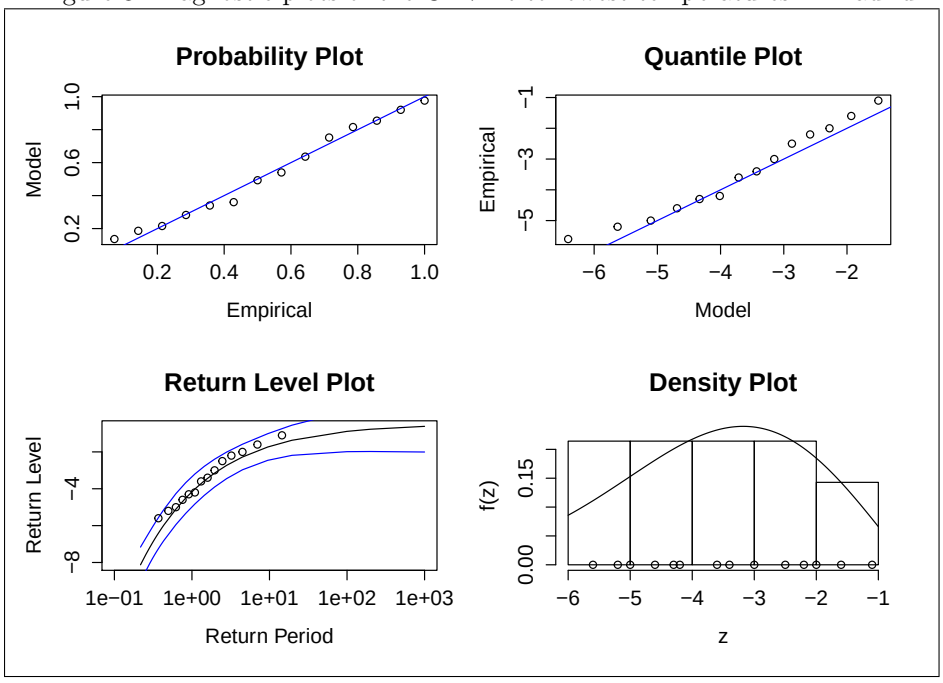


Figure 9: Diagnostic plots of the GEV fit to highest temperatures in Murcia

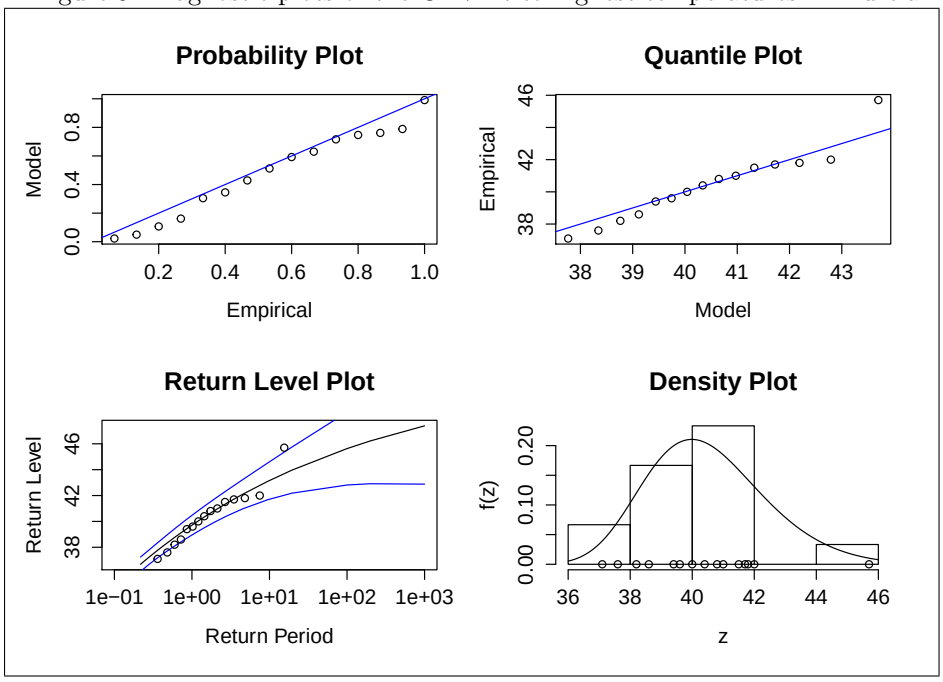


Figure 10: Diagnostic plots of the GEV fit to lowest temperatures in Murcia

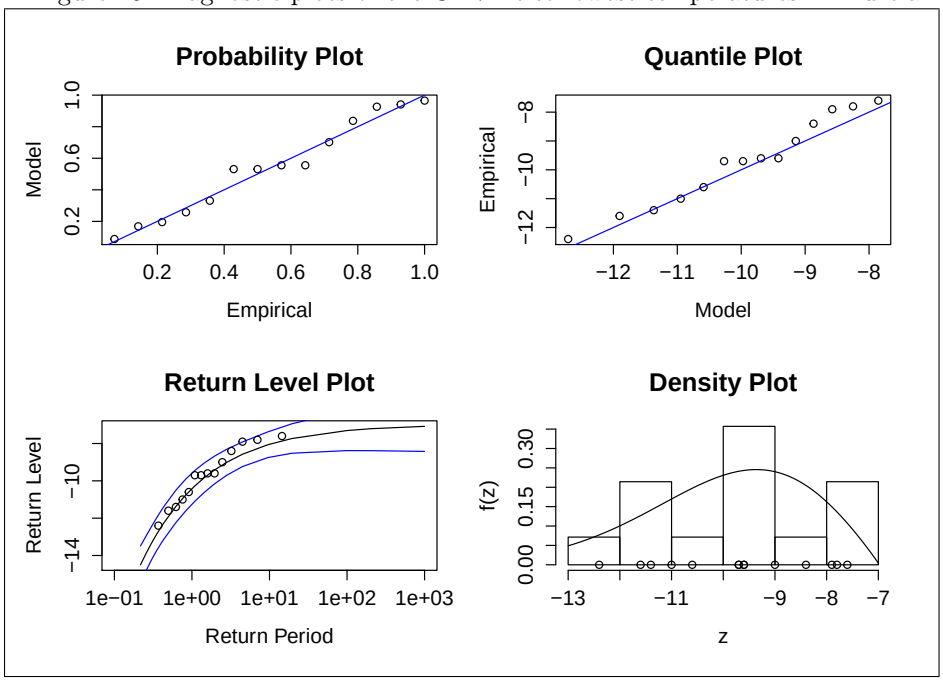
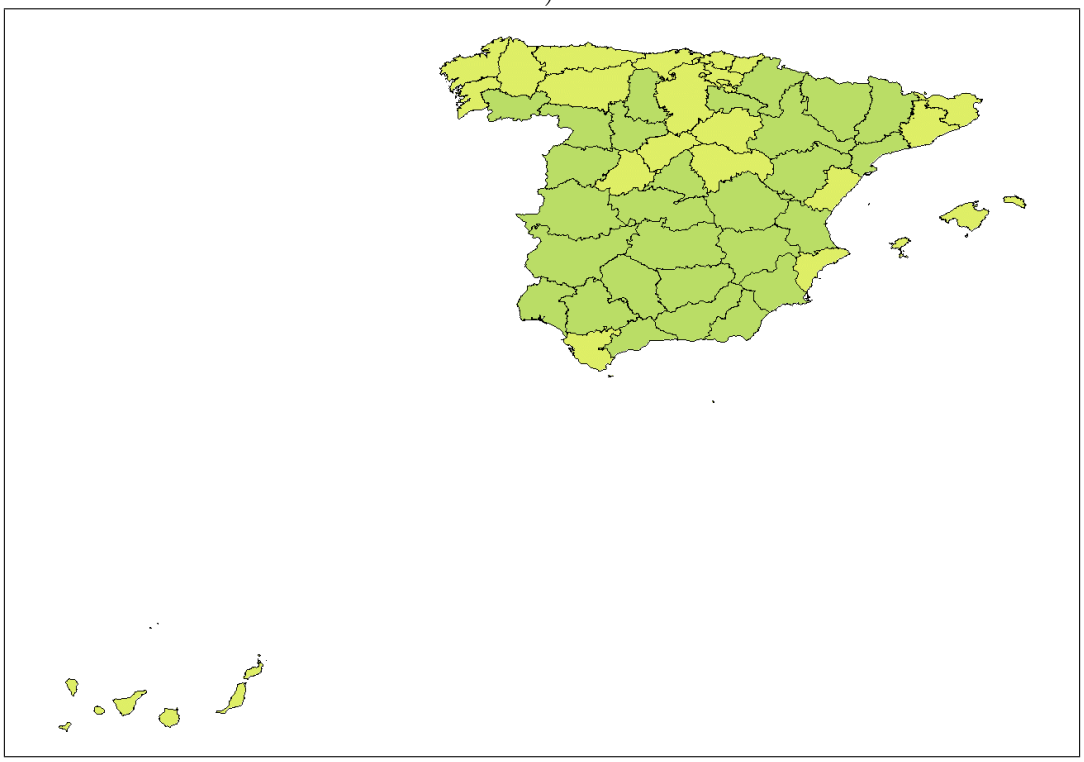


Figure 11: (a) Two clusters based on GEV estimates for highest temperatures; (b) Two clusters based on GEV estimates for lowest temperatures

a)



b)

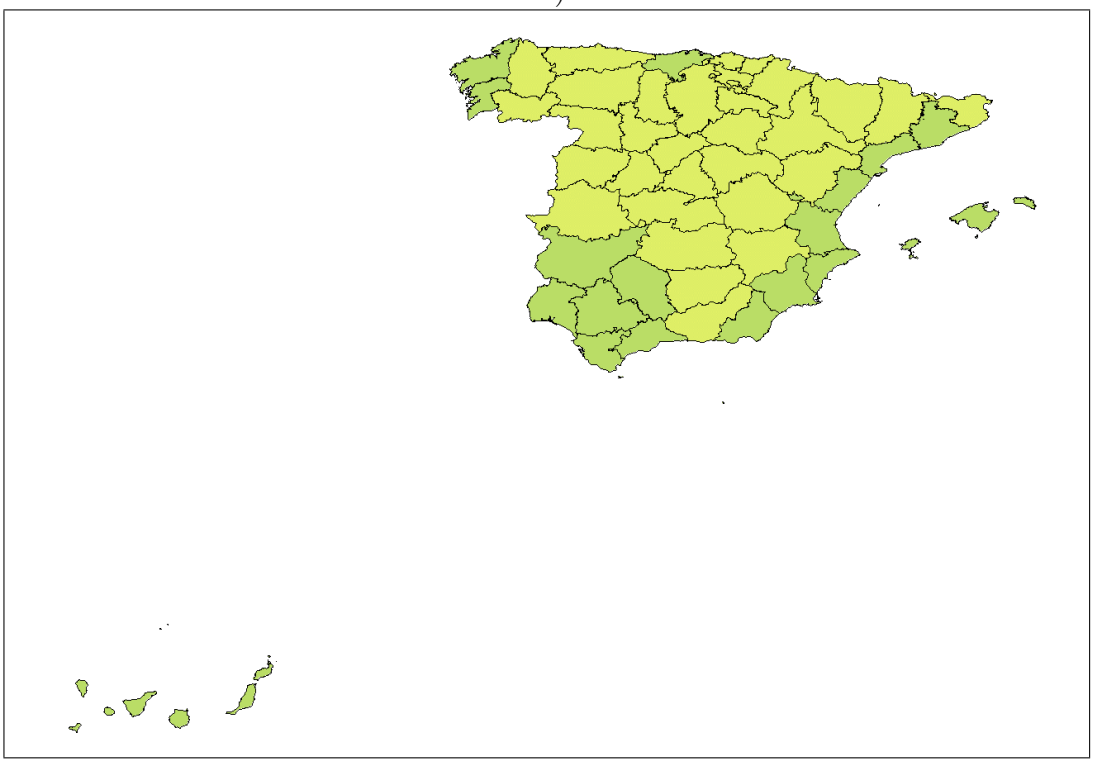
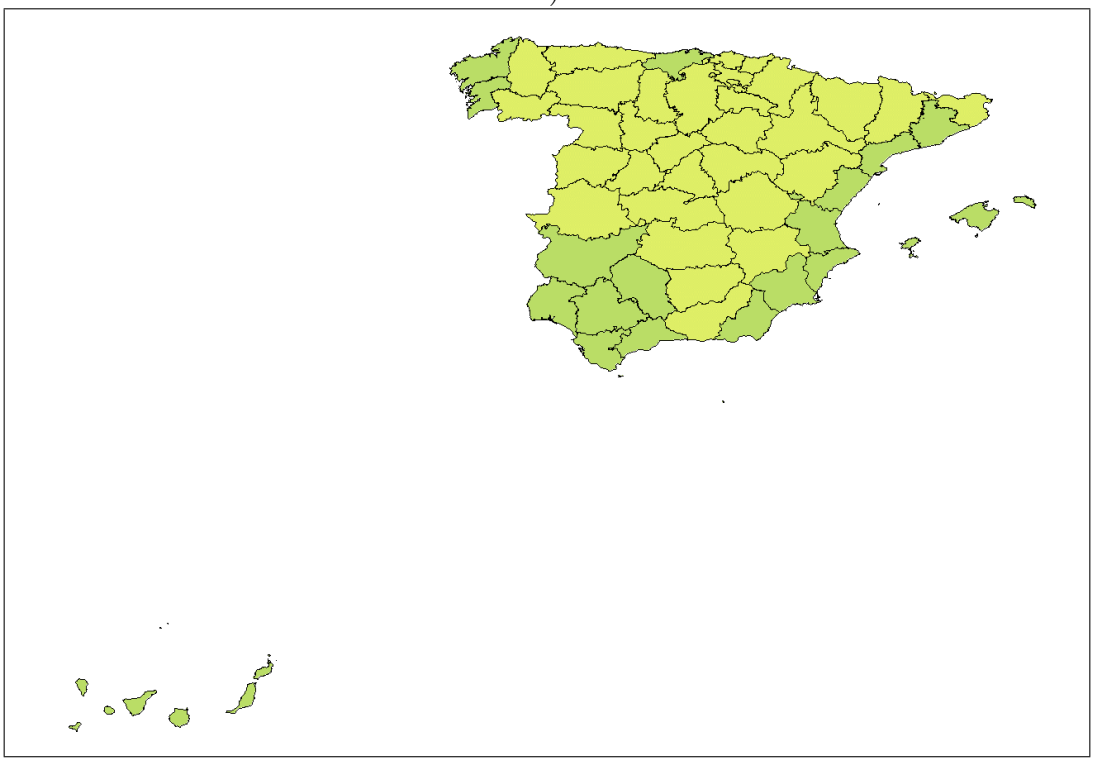
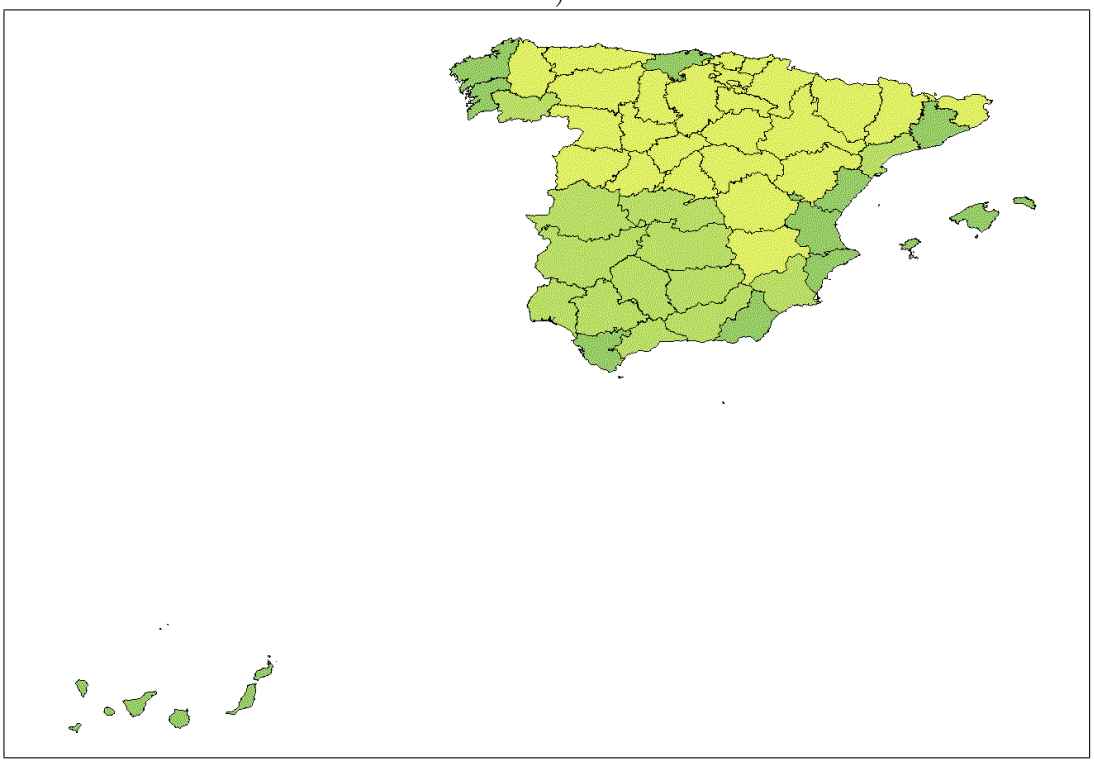


Figure 12: (a) Two clusters based on GEV estimates for highest and lowest temperatures; (b) Three clusters based on GEV estimates for highest and lowest temperatures.

a)



b)



and 12 represent the obtained cluster solutions.

The map in Figure 11(a) corresponds to a two-cluster solution based on GEV estimates for annual highest temperatures. Note that here and in what follows the GEV estimates of location, scale and shape are the clustering variables and Euclidean distances are calculated from each observation to each cluster centroid. The means of the maximum temperatures in the two clusters are 34.59 and 39.06, which are fairly close to the values obtained using the annual maxima. Only one region, *Salamanca*, differs in its cluster allocation. As expected, the similarity index attains a high value, namely, 0.9807.

The maps in Figures 11(b) and 12(a) correspond to a two-cluster solution based on GEV estimates for annual lowest temperatures and a two-cluster solution based on GEV estimates for annual highest and lowest temperatures, respectively. As in the previous analysis, both maps coincide. Moreover, they coincide with the maps in Figures 3(b) and 4(a). The coincidences observed in the maps are expected results since maps in Figures 3-4 and in Figures 11-12 are based on similar information, that is, GEV estimates are obtained using the annual highest and lowest temperatures.

The map in Figure 12(b) corresponds to a three-cluster solution based on GEV estimates for annual highest and lowest temperatures. The means of the maximum (minimum) temperatures in the three clusters are 36.51 (2.24), 40.41 (7.88) and 34.68 (11.13) and are close to the values obtained using the annual maxima and minima. In particular, the first clusters coincide in both approaches. Only two Mediterranean regions, *Almeria* and *Valencia*, change from the second to the third cluster. The similarity index between the three-cluster solutions (Figures 4(b) and 12(b)) is 0.9505, hence, reflecting the likeness or consistency of both approaches.

We also took into account the variability of estimates by clustering the 95% upper and lower confidence limits (six or twelve clustering variables) and found that the clustering results were identical to the clustering results obtained from just the estimates (three or six clustering variables). Hence, the uncertainty associated with the GEV estimates has not impacted on the clustering results.

Note that when the k-medoids clustering method was applied to the GEV estimates and to the block maxima and minima, similar cluster solutions were obtained.

3.1. Validation of Cluster Solutions

To check on the validity of the k-means 3-cluster solutions using the GEV features, and the maximum and minimum features, we ran the k-NN classification algorithm with one to five neighbours for the 52 regions with the groups designated according to the cluster solutions. The hold-out-one cross-validation method was used to evaluate the quality of the classification. The algorithm was run 52 times and the mean classification error obtained for each nearest neighbour.

From the results in Table 11, it can be observed that k-NN classification with three nearest neighbours produces the best result with a 4% classification error for the 3-cluster solution with GEV features. This is an indication that the 3-cluster solution using the k-means procedure with GEV features is reasonably well validated. However,

Table 5: k-Nearest Neighbours Classification

	Classification errors	
Number of Neighbours	**GEV**	**Max and Min**
1	8%	8%
2	8%	8%
3	4%	2%
4	8%	6%
5	6%	2%

the k-NN classification with the maxima and minima as features which resulted in a minimum classification error rate of 2% with three and five nearest neighbours, indicates that these features produce a higher quality of cluster separation.

3.2. Return Levels

One of the advantages of using the GEV features instead of the maxima and minima for clustering is that, we can interpret the cluster solutions using the N-year return levels (extreme quantiles), that is, the values that can be exceeded once every N-years. For instance, we use the expressions in Equations 4 or 5 for 25, 50 and 100 years in order to gain some insight into the three obtained clusters. These returns together with their 95% confidence intervals are presented in Table 6 and they confirm and complement our previous interpretation, that is, (1) the first cluster corresponds to regions having hot summers with temperatures that could be greater than 39^oC in periods of 25 or more years, and having the coolest winters with temperatures below 0^oC in periods of 25 or more years; (2) the second cluster corresponds to regions having the hottest summers with temperatures that could be greater than 43^oC in periods of 25 or more years, and cool winters with temperatures above 5^oC even in periods of 100 years; (3) the third cluster corresponds to regions having mild summers and mild winters. Refer to page 56 of Coles [5] for the formulae to obtain the standard error of the returns. Note that in all cases, the 95% confidence intervals of the returns are within reasonable ranges.

Table 6: Means of the 25, 50 and 100 years returns levels with 95% confidence intervals for the three clusters based on GEV estimates

Cluster		**25 yr**	**95% CI**		**50 yr**	**95% CI**		**100 yr**	**95% CI**	
1	sum	**39.12**	38.33	39.91	**39.40**	38.52	40.28	**39.61**	38.63	40.60
	win	**-0.63**	-1.41	0.15	**-1.01**	-1.94	-0.08	**-1.31**	-2.40	-0.23
2	sum	**43.08**	42.33	43.83	**43.41**	42.55	44.27	**43.67**	42.68	44.66
	win	**4.87**	4.15	5.59	**4.52**	3.68	5.35	**4.25**	3.29	5.20
3	sum	**38.37**	37.30	39.44	**39.04**	37.60	40.48	**39.63**	37.75	41.51
	win	**8.76**	8.03	9.48	**8.39**	7.56	9.23	**8.10**	7.13	9.07

Cluster analysis was also applied to the 25, 50 and 100-year return series for summer maxima and winter minima together and a comparison was made between each set of

3-cluster solutions and those obtained from 3-cluster solution of the GEV estimates. Table 7 shows the similarity indexes.

Table 7: Cluster Indexes comparing clusters from original data and 25, 50 and 100 year returns

	25-yr	50-yr	100-yr
1990-2004	88%	65%	62%
25-yr		73%	73%
50-yr			98%

The clusters obtained from the original data and from the 25-year return series are reasonably compatible. However, they become less compatible the further out the projections go. The clusters from the 50 and 100-year returns are quite compatible. It should be noted that returns are nonlinear transformation of the GEV parameters, hence, a full agreement is not expected since the closer the respective parameter estimates in a particular cluster are, the closer the returns but the reverse is not necessarily true. In Table 8, we present the results of a k-NN classification using the estimated returns as explanatory variables and the k-means 3-cluster solution using the GEV features as "true" class for the observations. The misclassification rates are small and particularly when three nearest neighbours are used.

4. Simulation Study

Following Safadi and Pena [18], we consider a dynamic factor model

$$y_t = Cf_t + e_t, \tag{6}$$

$$f_t = \sum_{i=1}^{p} \rho_i f_{t-1} + w_t, \tag{7}$$

where y_t is a $q\times 1$ vector of time series, C is a $q\times k$ matrix of factor loadings, $e_t \sim \mathcal{N}(0, \Gamma)$ where Γ is a $q \times q$ diagonal matrix. Each factor f_t is represented by a $k \times 1$ vector which follows a multivariate autoregressive distribution where the AR matrices ρ_i are diagonal matrices with ρ_i= diag$(\rho_{i1}, \rho_{i2}, \cdots, \rho_{ik})$, $i = 1, 2, \cdots, p$ and $\rho_{1j}, \rho_{2j}, \cdots, \rho_{pj}$,

Table 8: k-Nearest Neighbours Classification using the returns values as explanatory variables

Classification errors				
Number of Neighbours	**25-yr**	**50-yr**	**100-yr**	**(25-yr, 50-yr, 100-yr)**
1	12%	15%	23%	15%
2	12%	15%	23%	15%
3	8%	10%	12%	8%
4	10%	8%	13%	10%
5	8%	10%	12%	12%

$j = 1, 2, \cdots, k$ satisfy the stationary conditions and $w_t \sim \mathcal{N}(0, I_k)$. I_k is the identity matrix, and e_t and w_t are independent for all t and s.

We introduce seasonality into this dynamic factor model by adding a harmonic component to each factor in Equation 7 as follows:

$$f_{t,k} = \sum_{i=1}^{p} \rho_{i,k} f_{t-1,k} + A_k \Big(\sin \frac{2\pi t}{s} \Big) + B_k \Big(\cos \frac{2\pi t}{s} \Big) + w_t, \tag{8}$$

where s is the length of the cycle. $A_k = R_k \cos \theta_k$ and $B_k = -R_k \sin \theta_k$. For each factor, $f_{t,k}$, R_k is the amplitude or height of the cycle peaks and θ_k is the phase or the location of the peaks relative to time zero. Each factor can have different autoregressive dynamics and different seasonal dynamics, i.e., different amplitudes and phases.

We try to emulate the situation of regions experiencing similar summer maximum temperatures with different winter minimum temperatures as is the case in the application in Section 3. In order to simulate this case, we use one factor $f_{t,1}$ as described in Equation 8, and another factor $f_{t,2}$ with time variables A_2 and B_2 as follows:

$$A_2(t) = \begin{cases} A_1 & \text{for } t \text{ such that } 2j\pi \leq \frac{2\rho t}{s} \leq (2j+1)\pi \text{ for some } j \in \mathbb{N}, \\ A_1^* & \text{otherwise} \end{cases} \tag{9}$$

and

$$B_2(t) = \begin{cases} B_1 & \text{for } t \text{ such that } 2j\pi \leq \frac{2\rho t}{s} \leq (2j+1)\pi \text{ for some } j \in \mathbb{N}, \\ B_1^* & \text{otherwise} \end{cases} \tag{10}$$

where $A_1 = R_1 \cos \theta_1$ and $B_1 = -R_1 \sin \theta_1$, and $A_1^* = R_2 \cos \theta_k$ and $B_1^* = -R_2 \sin \theta_1$, $R_2 = r^* R_1$. These differences appear when t is such that $2j\pi \leqslant \frac{2\rho t}{s} \leqslant (2j+1)\pi$ for some $j \in \mathbb{N}$.

We emulate daily-type data with $s = 366$ over both 10 and 20 years, setting the amplitude for one group of five series to be $R_1 = 10$ and for a second group of five series to be $R_2 = r^* R_1$, with r^* between 0.1 to 0.9. Figures 13 to 15 show sections of a pairs of series generated for each scenario, viz., $R_1 = 10$ with each of $r^* = 0.9$, 0.5 and 0.1, i.e., $R_2 = 9$, 5 and 1. We expect that the greater the difference between R and $r^* R$, the greater the separation of the groups.

In order to estimate the GEV parameters for both the "summer" and "winter" seasons for each year, we take two blocks per year and hence obtain maxima and minima for each year. We fit the GEV distribution to the block maxima (r=1) as well as to the four largest per block (r=4). These steps are repeated for the minima. For both r=1 and r=4, we therefore have six GEV parameters, three estimated from the block maxima and three from the block minima.

Tables 9 to 12 show the results for the three scenarios ($R_1 = 10$ with $R_2 = 9$, 5, 1) for 10 and 20 years over 100 simulations, for the combined summer maxima and winter minima, summer maxima only, and winter maxima only using the estimated GEV parameters for one ($r = 1$) and four ($r = 4$) largest and smallest values as well as block maxima and/or minima as inputs into the clustering and classification methods. The clustering methods are evaluated by determining the similarity index used in Gavrilov

Figure 13: Time Series with the same peak heights, R_1=10 but different valley depths, R_1=10, R_2=9

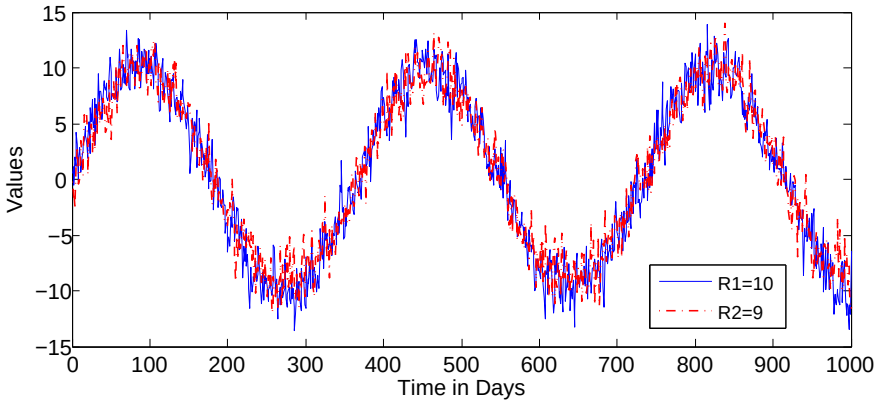


Figure 14: Time Series with the same peak heights, R_1=10 but different valley depths, R_1=10, R_2=5

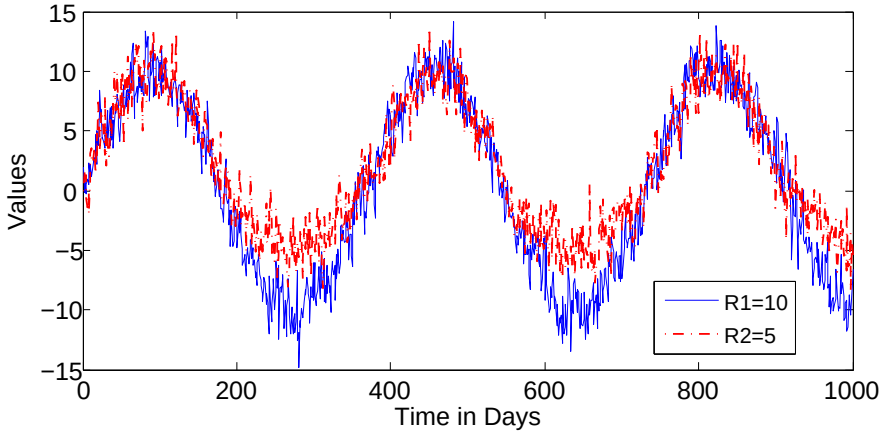
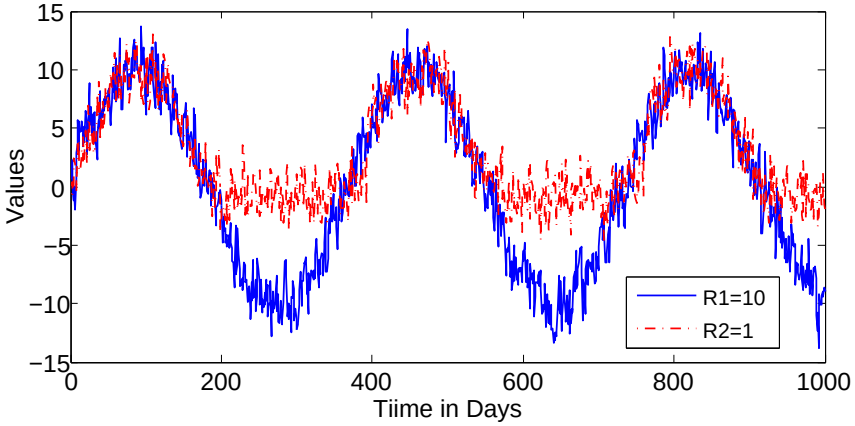


Figure 15: Time Series with the same peak heights, R_1=10 but different valley depths, R_1=10, R_2=1



et al. [9]. The index compares two different cluster solutions, $C = (C_1, C_2, \ldots, C_k)$ and $C' = (C'_1, C'_2, \ldots, C'_{k'})$ using the following formulae:

$$Sim(C_i, C'_j) = 2\frac{\#(C_i \bigcap C'_j)}{\#(C_i) + \#(C_i)},$$

and

$$Sim(C, C') = k^{-1} \sum_{i=1}^{k} \max_{1 \leq j \leq k'} Sim(C_i, C'_j),$$

where k and k' are the number of clusters in C and C', respectively. This index lies between zero and one and the closer it is to one, the better the performance. A useful property of this similarity index is that it allows for the comparison of cluster solutions with different numbers of cluster, i.e., k could be different to k'. Note that other comparable index, namely, the Adjusted Rand Index (Hubert and Arabie [13] produces similar results.

Hold-out-one cross-validation is used to evaluate the performance of the k-NN classification method. To this end, the proportion of correct classifications is obtained.

For the clustering methods, we make the following observations from Table 9:

- When R_1 =10 and R_2 = 9:
  - For both the 10-year and 20-year series, when the GEV features are inputs, the k-means method almost always performs better than the k-medoids method for each of the three scenarios. This is also the case when the maxima and minima, maxima only, and minima only, are the clustering features in the combined summer/winter, summer only and winter only scenarios, respectively.
  - For the winter scenario, i.e., when only minima are used as inputs, these methods perform slightly better than when both maxima and minima are the input features in the combined summer/winter scenario. This is also the case for the winter scenario when GEV features estimated from the minima (r=1) and from the four smallest values (r=4) are inputs compared to when GEV features estimated from the maxima and minima (r=1), and the estimated GEV features from the four largest and four smallest values (r=4) are inputs for the combined summer/winter scenario.
  - For the combined summer/winter scenario the performance of clustering methods is better when GEV features are estimated from the four largest and four smallest values (r=4) than when GEV features estimated from the maxima and minima (r=1), and than when the combined maxima and minima are features. The same observations are made for the winter only scenario, bearing in mind that only the minima are used, and the GEV features are estimated from the minima (r=1) and the four smallest values (r=4).

Table 9: Time series with similar summer-type peaks but different winter-type valleys: Similarity Index for Clustering Methods

		10 years, T=3660			20 years, T=7320		
R_2		**gev r=1**	**gev r=4**	**max/min**	**gev r=1**	**gev r=4**	**max/min**
				combined summer winter			
9	**k-means**	0.76	0.97	0.91	0.99	0.99	0.96
	k-medoids	0.74	0.94	0.81	0.96	1.00	0.88
5	**k-means**	1.00	1.00	1.00	1.00	1.00	1.00
	k-medoids	1.00	1.00	1.00	1.00	1.00	1.00
1	**k-means**	1.00	1.00	1.00	1.00	1.00	1.00
	k-medoids	1.00	1.00	1.00	1.00	1.00	1.00
				summer only			
9	**k-means**	0.62	0.59	0.60	0.60	0.59	0.60
	k-medoids	0.62	0.59	0.58	0.60	0.60	0.58
5	**k-means**	0.60	0.59	0.61	0.60	0.61	0.60
	k-medoids	0.58	0.61	0.57	0.60	0.61	0.58
1	**k-means**	0.61	0.60	0.62	0.60	0.60	0.61
	k-medoids	0.61	0.59	0.59	0.59	0.58	0.58
				winter only			
9	**k-means**	0.84	0.99	0.95	0.99	1.00	0.99
	k-medoids	0.83	0.96	0.84	0.97	1.00	0.93
5	**k-means**	1.00	1.00	1.00	1.00	1.00	1.00
	k-medoids	1.00	1.00	0.99	1.00	1.00	1.00
1	**k-means**	1.00	1.00	1.00	1.00	1.00	1.00
	k-medoids	1.00	1.00	1.00	1.00	1.00	1.00

  - For the combined summer/winter and winter only scenarios, the performance of the methods is always better for the 20-year series when comparing the same input features.

- For the combined summer/winter and winter only scenarios, when the winter valleys are further apart ($R_2 = 5$ and $R_2 = 1$), both methods are observed to perfectly differentiate between the two groups of time series.

- For the summer only scenario, for all three feature sets, maxima, GEV based on maxima (r=1) and GEV based on four largest values (r=4), the clustering performance of these methods is very poor for both the 10 and 20 year series. This is to be expected since the summer peaks are at similar heights.

Hence, it is clear from these observations, the winter features contribute most to the differentiation between the two groups of time series.

For the k-NN classification method (Tables 10 to 12), similar observations as for the clustering methods are made for most of the time. In all cases, the k-NN method appears to well validate the cluster separation for the combined summer/winter and winter only scenarios.

Table 10: Time series with similar summer-type peaks but different winter-type valleys (Combined summer and winter): Percentage of Correct Classifications using k-NN

		10 years, T=3660			**20 years, T=7320**		
R_2		**gev r=1**	**gev r=4**	**max/min**	**gev r=1**	**gev r=4**	**max/min**
9	**k-nn (1)**	0.85	0.98	0.85	0.99	1.00	0.93
	k-nn (2)	0.94	0.99	0.95	1.00	1.00	0.99
	k-nn (3)	0.81	0.97	0.87	0.99	1.00	0.95
	k-nn (4)	0.91	0.99	0.94	1.00	1.00	0.98
	k-nn (5)	0.73	0.97	0.85	0.98	1.00	0.94
5	**k-nn (1)**	1.00	1.00	1.00	1.00	1.00	1.00
	k-nn (2)	1.00	1.00	1.00	1.00	1.00	1.00
	k-nn (3)	1.00	1.00	1.00	1.00	1.00	1.00
	k-nn (4)	1.00	1.00	1.00	1.00	1.00	1.00
	k-nn (5)	1.00	1.00	1.00	1.00	1.00	1.00
1	**k-nn (1)**	1.00	1.00	1.00	1.00	1.00	1.00
	k-nn (2)	1.00	1.00	1.00	1.00	1.00	1.00
	k-nn (3)	1.00	1.00	1.00	1.00	1.00	1.00
	k-nn (4)	1.00	1.00	1.00	1.00	1.00	1.00
	k-nn (5)	1.00	1.00	1.00	1.00	1.00	1.00

Table 11: Time series with similar summer-type peaks but different winter-type valleys (Summer only): Percentage of Correct Classifications using k-NN

		10 years, T=3660			**20 years, T=7320**		
R_2		**gev r=1**	**gev r=4**	**max**	**gev r=1**	**gev r=4**	**max**
9	**k-nn (1)**	0.44	0.43	0.45	0.43	0.41	0.43
	k-nn (2)	0.74	0.70	0.71	0.72	0.70	0.71
	k-nn (3)	0.43	0.37	0.39	0.42	0.39	0.38
	k-nn (4)	0.66	0.63	0.64	0.65	0.64	0.63
	k-nn (5)	0.38	0.34	0.36	0.37	0.34	0.36
5	**k-nn (1)**	0.43	0.43	0.46	0.46	0.47	0.43
	k-nn (2)	0.71	0.71	0.74	0.73	0.74	0.72
	k-nn (3)	0.36	0.41	0.41	0.42	0.41	0.39
	k-nn (4)	0.62	0.65	0.63	0.66	0.65	0.63
	k-nn (5)	0.32	0.36	0.36	0.39	0.35	0.34
1	**k-nn (1)**	0.48	0.44	0.43	0.44	0.45	0.45
	k-nn (2)	0.74	0.72	0.71	0.72	0.71	0.71
	k-nn (3)	0.43	0.39	0.40	0.39	0.41	0.40
	k-nn (4)	0.65	0.65	0.66	0.64	0.64	0.62
	k-nn (5)	0.39	0.39	0.37	0.35	0.33	0.35

Table 12: Time series with similar summer-type peaks but different winter-type valleys (Winter only): Percentage of Correct Classifications using k-NN

		10 years, T=3660			**20 years, T=7320**		
R_2		**gev r=1**	**gev r=4**	**min**	**gev r=1**	**gev r=4**	**min**
9	**k-nn (1)**	0.92	0.98	0.90	0.99	1.00	0.96
	k-nn (2)	0.98	0.99	0.97	1.00	1.00	0.99
	k-nn (3)	0.89	0.98	0.92	0.99	1.00	0.97
	k-nn (4)	0.95	0.99	0.96	0.99	1.00	0.99
	k-nn (5)	0.83	0.98	0.91	0.98	1.00	0.97
5	**k-nn (1)**	1.00	1.00	1.00	1.00	1.00	1.00
	k-nn (2)	1.00	1.00	1.00	1.00	1.00	1.00
	k-nn (3)	1.00	1.00	1.00	1.00	1.00	1.00
	k-nn (4)	1.00	1.00	1.00	1.00	1.00	1.00
	k-nn (5)	1.00	1.00	1.00	1.00	1.00	1.00
1	**k-nn (1)**	1.00	1.00	1.00	1.00	1.00	1.00
	k-nn (2)	1.00	1.00	1.00	1.00	1.00	1.00
	k-nn (3)	1.00	1.00	1.00	1.00	1.00	1.00
	k-nn (4)	1.00	1.00	1.00	1.00	1.00	1.00
	k-nn (5)	1.00	1.00	1.00	1.00	1.00	1.00

5. Concluding Remarks and Future Directions

The clustering solutions in the application when the GEV features are used as clustering features can be meaningfully interpreted. This is also the case when the block maxima and block minima are used as clustering features. Validation of the cluster solutions in the application reveal that the block maxima and minima could be better separation features than the GEV estimates. However, an advantage of using GEV estimates as clustering features is that return level statements can be made about long-term maxima and minima which, in the case of applying this to real temperature time series may have policy implications in dealing with long-term extreme temperatures. Of course, other environmental variables (solar radiation, atmospheric ocean processes, pollutant gases concentration, etc.) can be analyzed with the proposed procedure.

It is clear from the simulation studies, the GEV estimates of location, scale and shape as well as the block maxima and minima are good separation features for temperature-type time series in general.

The future directions that we will be embarking on, in analysing real time series extremes is, (1) extend GEV fitted to extremes with trend, (2) consider the complete returns function for all quantiles leading to the clustering of functional data, and (3) examining the fuzzy behaviour of the series and incorporating their spatial features as an added source of information.

References

[1] Alonso, A.M., De Zea Bermudez, P. and Scotto, M.G. (2014), Comparing generalized Pareto models ftted to extreme observations: an application to the largest temperatures in Spain. *Stoch Environ Res Risk Assess.* **28**, 1221–1233.

[2] Altman, N. S. (1992), An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician* **46**3, 175–185.

[3] Brunet, M., Jones, P.D., Sigró, J., Saladié, O., Aguilar, E., Moberg, A., Della-Marta, P.M., Lister, D., Walther, A., López, D. (2007), Temporal and spatial temperature variability and change over Spain during 1850-2005. *J. Geophys. Res.* **112**, D12117.

[4] Charrad, M., Ghazzali, N., Boiteau, V., Niknafs A. (2014), NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software,* **61** 6, 1-36. URL http://www.jstatsoft.org/v61/i06/.

[5] Coles, S.(2001), An introduction to statistical modeling of extreme values. *Springer-Verlag: London.*

[6] Fernández-Montes, S., Rodrigo, F.S. (2012), Trends in seasonal indices of daily temperature extremes in the Iberian Peninsula, 1929–2005. *Int. J. Climatol.* **32**, 2320–2332.

[7] Furió, D., Meneu, V. (2011), Analysis of extreme temperatures for four sites across Peninsular Spain. *Theor. Appl. Climatol.* **104**, 83–99.

[8] García-Herrera, R., Díaz, J., Trigo, R.M., Hernández, E. (2005), Extreme summer temperatures in Iberia: health impacts and associated synoptic conditions. *Ann. Geophys.* **23**, 239–251.

[9] Gavrilov, M., Anguelov, D., Indyk, P., Motwani. R. (2000), Mining The Stock Market: Which Measure Is Best?. *KDD 2000, Boston, MA USA* ACM 2000 1-58113-233-6/00/08.

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

[11] Hagan, M.T., Demuth, H.B. and Beale, M. (1996), Neural Network Design *PWS Publishing,Boston, MA* .

[12] Hastie, T., Tibshirani, R. and Friedman. J. (2009), The Elements of Statistical Learning Data Mining, Inference, and Prediction, Second Edition *Springer Series in Statistics.*

[13] Hubert, L. and Arabie, F. (1985). Comparing partitions. *Journal of Classification,***2** 1, 193–218. doi:10.1007/BF01908075.

[14] Krishnapuram,R., Joshi, A., Nasraoui, O. and Yi, L. (2001), Low-complexity fuzzy relational clustering algorithms for web mining.. *IEEE Transactions on Fuzzy Systems* **9**4, 595–607.

[15] Mendez, F. J., Menéndez, M., Luceño, A. and Losada, I. J. (1997), Extreme sea-level distribution and return periods in the Aegean and Ionian seas. *J. Atmos. Ocean. Technol.* **24**, 894–911.

[16] Reiss, R-D. and Thomas, M. (2000), Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and Other Fields 3rd Edition *Birkhauser, Basel, Boston, Berlin.*

[17] Rousseeuw P.J.(1987) A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65.

[18] Safadi, T. and Peña, D. (2008), Bayesian analysis of dynamic factor models: an application to air pollution and mortality in Sao Paulo, Brazil. *Environmetrics* **19**, 582–601.

[19] Scotto, M.G., Barbosa, S.M. and Alonso, A.M. (2010), Clustering Time Series of Sea Levels: Extreme Value Approach. *J. Waterway, Port, Coastal, Ocean Eng.* **136**, 2793–2804.

[20] Scotto, M.G., Barbosa, S.M. and Alonso, A.M. (2011), Extreme value and cluster analysis of European daily temperature series. *Journal of Applied Statistics* **38, 12**, 215–225.

[21] Tsimplis, M. N., and Blackman, D. L. (1997), Extreme sea-level distribution and return periods in the Aegean and Ionian seas. *Estuarine Coastal Shelf Sci.* **44**, 79–89.

[22] Unnikrishnan, A. S., Sundar, D., and Blackman, D. L. (2004), Analysis of extreme sea level along the east coast of India. *J. Geophys. Res.,* **109**, C06023.

[23] Wikipedia Map of Spanish Provinces, http://en.wikipedia.org/wiki/File:Provinces_of_Spain.svg.