Cluster Analysis of the Iberian Peninsula Temperatures Time Series

João C. S. Dias

Abstract

This work focuses on clustering maximum daily temperature time series of the Iberian Peninsula from 1995-1999. The clustering is applied to a set of 19 maximum daily temperature records defined according to the NUTS2 division. Two clustering techniques are used and compared: one using autocorrelation-based distances and another approach using quantile-based autocovariance distances. Moreover, different subsampling methodologies are considered and compared. The application of the clustering algorithms resulted mainly in: the choice of the better subsample length; the finding that the subsampling cases give dendrograms with higher cophenetic correlation coefficients than the no subsampling case; the choice of average-linkage as the best linkage method; and, the partition of the Iberian Peninsula in regions according to the locations' data: all methodologies agree in the partition when 2 clusters are considered.

Keywords: Clustering, autocorrelation function, quantile autocovariance function, subsampling

1 Introduction

Since the late 19th century, the global surface mean temperature has increased by 0.7 °C [Trenberth et al 2007]. Moreover, during the 20th century, there has been an increase in global and regional average temperatures, particularly in the last 50 years, being, nowadays, the analysis of its causes a lively topic of current research. Nevertheless, changes of mean values are not the only way in which changes on climate are manifested, the occurrence of extreme events have also the potential of contributing to shape climate change. Extreme temperature patterns are consistent with the global warming [Trenberth et al 2007], although different behaviours and intensities have been reported for different regions around the world. In Europe, for instance, a number of studies have reported significant changes in maximum and minimum temperatures during the 20th century, as well as increases in the mean temperature, mainly in Western Europe, west of the meridian 20°E [see Dias 2018 and references therein]. In terms of the Iberian Peninsula, average series of maximum and minimum temperature increased at a rate of about 0.5 °C/decade [see Dias 2018 and references therein].

Several methods for defining and characterising extreme events have been proposed in the literature, including the analysis of the statistical behaviour of the tail of a weather element's probability distribution or the use of extremal indices [see Dias 2018 and references therein]. Extreme value theory (EVT) can be applied in several ways to the analysis of extremes in meteorology. One is the so-called Peak-over-Threshold (POT) approach along with the Generalized Pareto Distribution (GPD). In fact, the GPD plays a major role in extreme value statistics, being considered the distribution of sample excesses above a high threshold [see Dias 2018 and references therein]. In the last two decades, extensive research has been done on characterising the GPD and on deriving probabilistic and statistical results. This approach, however, has more often been used for precipitation extremes [see Dias 2018 and references therein]. The EVT approach is generally used by engineers to assess the intensity of extreme meteorological events which must be endured by industrial facilities and buildings. An example of this is the work by Acero et al. 2014, who assessed these extremal events as long-period return levels (RLs), as corresponding to rare occurrences. The POT method has also been extensively applied in many other fields, besides extreme temperatures and precipitation, including hidrology, finance and insurance, environment, and extreme waves and waves parameters (see Dias 2018 and references therein). Alonso et al

2014 introduced a subsampling-based testing procedure for the comparison of exceedance‘s distributions of stationary time series. The performance of the testing procedure was illustrated by the authors through an application to a set of data of daily maximum temperatures in the 17 autonomous communities of Spain, for the period 1990-2004. Several other studies on climate extremes use high-resolution climate models (RCMs) entering in consideration with improved physics regarding clouds, land surface, boundary layer diffusion and convection process, which have shown to simulate regional climate. There are some advantages of the use of RCMs, namely the inclusion of regional effects as the topography, the coastlines, the mountains, the water bodies and the vegetation on the local climate [see Dias 2018 and references therein].

Comparing differences among time series with respect to their corresponding extremal behaviour is a topic of major current interest in many empirical studies in the area of the environment. For instance, in studies of regional variability of daily mean temperature time series, it is important to identify locations with analogous behaviour, using namely clustering, in terms of their corresponding predictive distributions of return values for a certain period of time [Scotto et al 2011]. The analysis of tide gauge records in a regional context is another example, being essential for predictions of flooding risks, coastal management and design of coastal infrastructure systems, through the identification of extremes in sea levels [see Scotto et al 2010 and references there in]. One popular approach to group time series, in some (pure) statistical studies is model-based clustering where the clustering is performed under the assumption that the time series have a specific generating model (see Dias 2018 and references therein). Commonly, these authors assume ARIMA-type generating mechanisms, although alternative models have been considered, such as GARCH, Markov chains and dynamic regression models [see Dias 2018 and references therein]. Other possibility is to focus on a feature-based approach, in which the raw observations are replaced by a reduced number of features describing the temporal structure of the time series, and thus dissimilarity is assessed in terms of such features. Moreover, in the time domain several studies have considered dissimilarity measures based on comparing estimations of simple or partial autocorrelation functions [Caiado et al 2006, D‘Urso and Maharaj 2009]. It is well documented that autocorrelations exhibit the ability to discriminate between processes, such as different classes of stationary time series (see Monte Carlo experiments in Caiado et al 2006). However, autocorrelation-based methods present also weaknesses, such as the lack of robustness to outliers or the presence of heavy tails or being unable to detect tail dependence. To tackle some of these limitations De Luca and Zuccolotto 2011 introduced a measure based on the tail dependence coefficient to group time series showing association between extreme low values, while D‘Urso et al 2013 consider two fuzzy clustering procedures making use of GARCH models. More recently, Lafuente-Rego and Vilar 2016 proposed to measure dissimilarity between pairs of time series by comparing quantile autocovariance functions (see also, e.g., Linton and Whang 2007, Lee and Rao 2012). In Lafuente-Rego and Vilar’s work, the authors use a statistic defined in terms of covariances of indicators functions, instead of considering the usual autocovariance and autocorrelation functions. The advantage of such method is being not dependent on moment conditions.

Alonso and Maharaj 2006 introduced a procedure based on subsampling for testing the hypothesis of equality of the generating processes of two stationary time series which can or can not be assume to be independently generated. Their approach combines the use of a distance between autocorrelation functions and a subsampling-based procedure to check the referred hypothesis test. The advantage of such procedure stems from the fact that no parametric fitting model is required.

The present work aims at analysing and clustering time series of maximum daily temperature collected at the Iberian Peninsula from 1995 to 1999. To this end, two methods for clustering time series are considered. Firstly, the sample autocorrelation functions of the time series are calculated in order to obtain the dissimilarities among time series, with the aim of classify them according with the null hypothesis of pairwise equality of the time series underlying generating mechanism. This is done through pairwise comparison of the time series. As a second approach, the pairwise comparison of quantile autocovariance functions is done, as proposed in a recent paper by Lafuente-Rego and Vilar 2016. Several subsampling methodologies based on the paper by

Alonso and Maharaj 2006, namely by using two different subsample lengths or by using a no subsampling-based procedure, were used when considering sample autocorrelation functions. Due to the length of computational time, no subsampling methodologies were used in the case of quantile autocovariance functions. Overall, 19 weather stations, distributed according to the NUTS2 for the continental Iberian Peninsula (including Balearic Islands), are included in this study. The time series data was collected from the site of the project European Climate Assessment & Dataset (ECAD) 2017. The software used for the calculations was R [R version 3.4.2 2017].

The rest of this paper is organised as follows. The time series clustering approaches are described in Section 2. The results of the cluster analysis are presented in Section 3, jointly with the summary of the data set used. Finally, concluding remarks are given in Section 4.

2 Methodology

2.1 Autocorrelation-based Distances

2.1.1 No Subsampling Case

In this section, we borrow the ideas of LaFuente-Rego and Vilar 2016 to cluster the set of time series according with the distance between sample autocorrelation instead of sample quantile autocovariance. The entries of the distance matrix are estimated according to

$$\hat{d}_{i,j} = \sum_{k=1}^{m} (\widehat{\rho}_{X^{(i)},k} - \widehat{\rho}_{X^{(j)},k})^2, \tag{1}$$

where $\widehat{\rho}_{X^{(i)},k}$ and $\widehat{\rho}_{X^{(j)},k}$ are the k^{th} sample autocorrelation of the time series $X_t^{(i)}$ and $X_t^{(j)}$ $(i,j$=1,...,N, N=19 in this work), respectively. $d_{i,j}$ is the squared Euclidean distance between the autocorrelation vectors $(\widehat{\rho}_{X^{(i)},1},...,\widehat{\rho}_{X^{(i)},m})$ and $(\widehat{\rho}_{X^{(j)},1},...,\widehat{\rho}_{X^{(j)},m})$, where *m* is the maximum lag considered (in this work m=10 was considered, which is justified by the autocorrelation structure of the time series; see discussion in Section 3).

2.1.2 Subsampling Case

In this case, the clustering of the time series is done through the comparison of each pair of stationary time series, according to the hypothesis test

$$H_0 : P_{X^{(i)}} = P_{X^{(j)}}$$
$$H_1 : P_{X^{(i)}} \neq P_{X^{(j)}}$$

where $P_{X^{(i)}}$ and $P_{X^{(j)}}$ represent the underlying generating models of $X^{(i)}$ and $X^{(j)}$, respectively. The underlying idea is to test if the generating process is the same in both series, $X^{(i)}$ and $X^{(j)}$. Recall that the complete probabilistic structure of a time series is determined by the joint distributions of the set of time series. In linear processes, and particularly in Gaussian linear processes, the characteristics of these joint distributions may be described in terms of the process mean $\mu_X = E[X_t]$ and the process autocovariance function $\gamma_{X,k} = Cov(X_t, X_{t+k})$, where k is a non-negative integer. Here, it is assumed that $\mu_{X^{(i)}} = \mu_{X^{(j)}}$, and a test statistic is defined as a function of the estimated autocorrelations of $X^{(i)}$ and $X^{(j)}$. In fact, after the removal of the trend and seasonal component of the original time series we obtain the time series of the residuals (X_t), which can be considered stationary and with zero mean. Thus, the proposed statistic, based on Alonso and Maharaj 2006, is

$$T \equiv T_{i,j} = n \sum_{k=1}^{m} (\widehat{\rho}_{X^{(i)},k} - \widehat{\rho}_{X^{(j)},k})^2, \tag{2}$$

where $\widehat{\rho}_{X^{(i)},k}$ and $\widehat{\rho}_{X^{(j)},k}$ are the k^{th} autocorrelations of $X^{(i)}$ and $X^{(j)}$, respectively, n is the sample size and m is the maximum lag considered (in this work m=10). In the work of Alonso and Maharaj 2006, the authors provide the theoretical basis for the subsampling method using the statistic T. It is important to stress that T represents the squared Euclidean distance between the autocorrelation vectors $(\widehat{\rho}_{X^{(i)},1}, ..., \widehat{\rho}_{X^{(i)},m})$ and $(\widehat{\rho}_{X^{(j)},1}, ..., \widehat{\rho}_{X^{(j)},m})$ multiplied by the normalizing constant (n).

In this method, the value of the proposed statistics T is compared with the sampling distribution obtained by subsampling as follows. Let $X_{i^{\cdot}}^{(i)} = (X_{i^{\cdot}}^{(i)}, X_{i^{\cdot}+1}^{(i)}, ..., X_{i^{\cdot}+l-1}^{(i)})$ and $X_{j^{\cdot}}^{(j)} = (X_{j^{\cdot}}^{(j)}, X_{j^{\cdot}+1}^{(j)}, ..., X_{j^{\cdot}+l-1}^{(j)})$ be the subsample of l consecutive observations of $X^{(i)}$ and $X^{(j)}$, for $i, j = 1, ..., N$ and $i^{\cdot}, j^{\cdot} = 1, ..., n - l + 1$, with the number of time series N and the length of each time series n. The $(i^{\cdot}, j^{\cdot})^{th}$ value of the subsampling statistic is obtained from

$$T_{i,j}^{(i^{\cdot},j^{\cdot})} = l \sum_{k=1}^{m} (\widehat{\rho}_{X_{i^{\cdot}}^{(i)},k} - \widehat{\rho}_{X_{j^{\cdot}}^{(j)},k})^2, \tag{3}$$

where $\widehat{\rho}_{X_{i^{\cdot}}^{(i)},k}$ and $\widehat{\rho}_{X_{j^{\cdot}}^{(j)},k}$ are the k^{th} autocorrelations of the subsample $X_{i^{\cdot}}^{(i)}$ and $X_{j^{\cdot}}^{(j)}$, each of length l. The sampling distribution function is then estimated through

$$\widehat{G}_{i,j}(x) = \frac{1}{(n-l+1)^2} \sum_{i^{\cdot}=1}^{n-l+1} \sum_{j^{\cdot}=1}^{n-l+1} I(T_{i,j}^{(i^{\cdot},j^{\cdot})} \leqslant x), \tag{4}$$

where $I(\cdot)$ is the indicator function. The distance between $X_t^{(i)}$ and $X_t^{(j)}$ is then obtained from the quantity $1 - p_{i,j}$, where $p_{i,j}$ represents the *p*-value obtained as

$$1 - p_{i,j} = \widehat{G}_{i,j}(T_{i,j}). \tag{5}$$

By repeating the same scheme for each pair of time series (i, j=1, ... , N), the distance matrix is obtained with entries $1 - p_{i,j}$ (i.e. $d_{i,j} = 1 - p_{i,j}$). Note that a *p*-value varies between 0 and 1, where a value closer to 0 provides less (statistical) evidence supporting H_0. On the other hand, a *p*-value closer to 1 does not allow the rejection of H_0 at a relatively high value of significance. As $d_{i,j} = 1 - p_{i,j}$ then $d_{i,j}$ quantifies the probability of having a value on the T statistics that is lower than the observed T value, thus supporting the hypothesis H_0. Therefore, $d_{i,j}$ turns out to be closer to 1 when $p_{i,j}$ is closer to 0, case when H_0 is rejected at relatively low significance values.

Different values of the subsample length l may be used. According to Alonso and Maharaj 2006, better results, i.e. with lower estimated size of the test (near 5%), are obtained for longer l, like l=128 or l=256 for a n =512. In the sequel, subsample lengths of 256 and 512 were considered, and their results compared, because our sample length (n=1826) is much larger than theirs, being 512 about one fourth of our sample length.

2.2 Quantile-based Autocovariance Distances

In this case, the clustering of the time series was done according to Lafuente-Rego and Vilar 2016, with the quantile-based autocovariance distance, instead of sample autocorrelation-base distance. The quantile autocovariance function of the sample X_t is defined as

$$\gamma_{X,l}(\tau,\tau^{`}) = Cov(I(X_t \leqslant q_\tau), I(X_{t+l} \leqslant q_{\tau^{`}})) = P(X_t \leqslant q_\tau, X_{t+l} \leqslant q_{\tau^{`}}) - \tau\tau^{`}, \tag{6}$$

where $\alpha = \{\tau, \tau^{`}\}$ are the percentile quantiles used, such that $0 \leqslant \alpha \leqslant 1$, q is the quantile value according to α, l is the time lag and $I(\cdot)$ is the indicator function. In practice, (6) is estimated with the values of the time series $X_1, ..., X_n$, based on the expression

$$\widehat{\gamma}_{X,l}(\tau,\tau^{`}) = \frac{1}{n-l}\sum_{t=1}^{n-l} I(X_t \leqslant \widehat{q}_\tau) I(X_{t+l} \leqslant \widehat{q}_{\tau^{`}}) - \tau\tau^{`}, \tag{7}$$

where $\widehat{q}_\alpha$, $\alpha = \{\tau, \tau^{`}\}$ are the empirical quantiles. The distance matrix is estimated with entries

$$\hat{d}_{ij} = \sum_{k=1}^{m}\sum_{i^{`}=1}^{r}\sum_{j^{`}=1}^{r}(\widehat{\gamma}_{X^{(i)},k}(\tau_{i^{`}},\tau_{j^{`}}) - \widehat{\gamma}_{X^{(j)},k}(\tau_{i^{`}},\tau_{j^{`}}))^2, \tag{8}$$

where m is the maximum lag considered and $\tau_{j^{`}}$ are the percentage quantiles $j^{`} = \{1, ..., r\}$. In concordance with LaFuente-Rego and Vilar 2016, we considered m=1 and three quantile levels (r=3) such that $\tau = \{0.1, 0.5, 0.9\}$.

The methodology presented in this subsection does not apply any subsampling technique. In fact, due to the huge computational resources, in terms of time and/or number of processors required, no subsampling technique was used here.

For each distance matrix, hierarchical agglomerative clustering algorithms were applied, i.e. the average-linkage method, the complete-linkage method and Ward‘s method. Each generated dendrogram is evaluated by three different methods, in an attempt to obtain the optimal cut of the dendrogram. The first two methods, are obtained from the paper of Montero and Vilar 2014, as described in the package TSClust for R: cluster.evaluation and loo1nn.cv. The third method is the Mojena‘s method [Faria et al 2012], which provides the optimal number of clusters for each dendrogram. The parameter k used was 1.25 according to Faria et al 2012. Moreover, dendrogram‘s goodness-of-fit is evaluated through the cophenetic index [Sneath and Sokal 1973].

3 Results and Discussion

3.1 The Iberian Peninsula Temperatures Data Set

This work aims at assessing the degree of similarity between time series representing maximum daily temperature in the Iberian Peninsula, through cluster analysis techniques. To this extent, weather stations in the continental Peninsula and nearby islands (Balearic Islands) are selected. From the whole set of stations, one station per NUTS2 (Iberian Peninsula) is selected, corresponding to the main city of this geographical division.

In order to reduce the impact of missing values in the posterior analysis, only time series for which the percentage of missing values is smaller than 15% are considered in the study, which results in 19 stations being selected. These stations are uniformly distributed having in consideration the climate zones of Iberian Peninsula (classification adapted from the Iberian Climate Atlas 2011): Mediterranean climate (Beja, Lisboa, Porto, Badajoz, Madrid, Valencia, Barcelona, Sevilla, Valladolid, Santiago de Compostella, Palma de Mallorca, Toledo), oceanic climate (Pamplona, Santander, Logroño, Vitoria, Oviedo), and semi-arid climate (Zaragoza, Murcia).

The period of time spans from 1st January 1995 to 31st December 1999 (5 years), since for more recent time spans arise larger quantity of missing values in some stations. The entries with missing values were filled as in Dias 2018.

Generally, to these clustering methodologies should be applied methods for the stabilisation of the mean and the variance in the original time series, in order to obtain a time series with the desirable characteristics (weakly stationarity). In this work, our time series present trend and periodicity, and then, the first step was the removal of these deterministic components. This decomposition was done in R with its function stl (package stats), which decomposes a time series into seasonal, trend and irregular (residuals) components using Loess decomposition.

As mentioned above the purpose of this work is two-fold: first, the clustering is based on the differences among the simple autocorrelation function of the different time series of the residuals (see Section 2). For such functions (not presented here), as a general rule, the sample autocorrelation decreases appreciably, becoming near the upper value of the confidence interval, in the first 10 lags or less. Valencia and Murcia are included in this group, although their sample autocorrelation function goes below the upper bound at about lag 5 instead of at lag 10. However, there are exceptions to the general rule: Beja, Lisboa, Badajoz, Madrid, Sevilla, Valladolid and Toledo have a bump after the first 10 lags. The second purpose of this work is to carry out a clustering proceeding based on quantile autocovariance functions (see Section 2 for details). For such functions, generally, no clear splitting of the locations is easily found, except, possibly, for the pair of quantile levels 0.5-0.5 (figure not presented here).

3.2 Results

3.2.1 Autocorrelation-based Distances: No Subsampling Case

The 19 maximum daily temperature time series are clustered according to the autocorrelation-based distances on the residuals. Dendrograms were produced with the complete-, average-linkage and Ward's methods. As expected, the dendrograms for the three linkage methods are very similar, due to the absence of outliers in the data. For the three linkage methods, the cophenetic correlation coefficient are very similar (within in the range 0.66-0.67), indicating that the clustering is reasonable fit.

Three methods were considered to cut the dendrogram and produce the clusters (see Section 2). The optimal number of clusters are not stable in relation to the methods: the optimal number of clusters ranges from 2 to 5, depending on the method. The Mojena‘s method produces the minimum optimal number of clusters (2 to 3). For further discussion, a solution with 3 clusters is considered for simplicity. Cluster 1 (C1) consists of Lisboa, Valladolid, Beja, Badajoz, Madrid, Toledo, Sevilla and Santiago de Compostella. Cluster 2 (C2) incorporates only Valencia and Murcia, while cluster 3 (C3) contains Palma de Mallorca, Zaragoza, Pamplona, Vitoria, Santander, Oviedo, Porto, Barcelona and Logroño. Generically, C3 corresponds to the northeast of the Peninsula, which have correspondence to the oceanic climate, except for Porto, Zaragoza and Barcelona, while C2 corresponds to two nearby cities located in the Mediterranean. C1 has the remaining locations and is dominated by the Mediterranean climate. Note that Porto is in a different cluster from that of nearby locations, which may be explained by the existence of many missing values in the original data set.

The clustering method considers the residuals, i.e. the maximum daily temperature discounted for the trend and the seasonal components, and then reflects the behaviour of the ACF of the difference from the referred components. More precisely, it reflects its first 10 autocorrelations. Comparing the ACF of these 3 clusters, it is noticed that the cluster C1 incorporates the time series for which the ACF is higher than the ACF of the time series grouped in clusters C2 and C3 (in the first 10 lags). On the other hand, C2 corresponds to the time series where the sample ACF tends to have the lowest values. Finally, C3 is the cluster where the ACF of its time series is in between the ACF profiles of C1 and C2.

3.2.2 Autocorrelation-based Distances: Subsampling Case

In the subsampling case, two subsample lengths were used: l=256 and l=512. The corresponding dendrograms were produced for three linkage methods were used for both subsample lengths: average-linkage, complete-linkage and Ward‘s method.

The dendrograms produced have some dissimilarities among linkage methods (mainly Ward‘s method), but also, between subsample lengths l=256 and l=512. There are differences in the clusters, mainly in the splittings after C1 and for the Ward‘s method, C1 is the first cluster to be originated, while for average-linkage and complete-linkage methods, C2 is the first cluster to appear. However, in the initial divisions of the dendrograms (i.e. the first three clusters formed), there are similarities.

For the three hierarchical clustering methods used, and for l=256, the cophenetic correlation coefficients are in the range 0.88-0.93, while for l=512, they are in the range 0.83-0.90, which are numbers near 1. Then, these dendrograms may be viewed as a good summary of the data. Also, with the subsample length of l=256, better results are obtained in terms of cophenetic correlation coefficients, and then, in terms of the quality of the dendrograms as a good summary of the data. These cophenetic correlation coefficients are larger than the ones obtained without subsampling.

The dendrograms were cut following the optimal number of clusters, varying from 2 to 6 clusters for all cutting methods used. For the Mojena‘s method, for all subsample lengths and linkage methods, the optimal number of clusters is equal to 3. Therefore, for further discussion, a solution with 3 clusters is considered for simplicity. For this division, the clusters formed by the three linkage methods and subsample lengths are similar. This splitting in 3 clusters is equal to the one obtained for no subsampling case. The properties of the ACF according to the clusters is the same as in the no subsampling case. Furthermore, the clusters C2 and C3 are nearer among them and more distant from the cluster C1 in the no subsampling case (all linkage methods). C3 and C1 are close and more distant from C2 in the subsampling case, for the average- and complete-linkage. For the Ward‘s method (subsampling case), the former statement (for the no subsampling case) is the one valid.

3.2.3 Quantile-based Autocovariance Distances: No Subsampling Case

In the case of quantile-based autocovariance distances, without subsampling, the 19 weather stations were classified according to the estimated quantile autocovariance of the residuals considering the time lag 1 and quantile levels of 0.1, 0.5, 0.9. The dendrograms were produced for three linkage methods: average-linkage, complete-linkage and Ward‘s method. Note that, the dendrograms obtained show some differences. E.g., for a solution with 3 clusters, the average- and the complete-linkage methods agree in the partition, whereas Ward‘s method provides a different partition, in which the group Palma de Mallorca, Oviedo and Murcia belongs to another cluster. However, the partitions with only 2 clusters are identical.

The cophenetic correlation coefficients for these dendrograms are approximately equal to 0.73 (for all linkage methods), and thus implying a reasonable fit. Three methods were applied to cut the dendrogram and produce the clusters, obtaining an optimal number of clusters between 2 and 7. For further discussion and for simplicity, a solution with 2 clusters is considered, according with Mojena‘s method. Cluster A (CA) consists of Lisboa, Valladolid, Beja, Badajoz, Madrid, Toledo, Sevilla and Santiago de Compostella. Cluster B (CB) contains Palma de Mallorca, Zaragoza, Pamplona, Vitoria, Santander, Oviedo, Porto, Barcelona, Logroño, Valencia and Murcia. Roughly, CA corresponds to the Mediterranean climate, while CB corresponds to the oceanic climate, however, Porto, Barcelona, Zaragoza, Valencia, Murcia, Palma de Mallorca and Sevilla are misclassified. This splitting in 2 clusters has similarities to that of 3 clusters obtained with autocorrelation functions, since Cluster A (quantile-based autocovariance) is equivalent to Cluster 1 (autocorrelation) and Cluster B (quantile-based autocovariance) is equal to the aggregation of Cluster 2 and Cluster 3 (autocorrelation). However, a more in-depth analysis of the clustering results reveals that they are very different from the ones obtained for autocorrelation functions. As an example, a solution with 3 clusters in the former case is different from the solution with the same number of clusters in the latter case: e.g., there are no cluster grouping only Valencia and Murcia (the previous cluster C2, from the autocorrelation part of this work) in the case of quantile-based autocovariance distances (all linkage methods). However, for average- and complete-linkage, these two locations belong to two different clusters.

The separation between clusters is not so evident, when we look at the plot of the quantile autocovariance functions (not presented here), as is for the case of autocorrelation-based distances. However, for the pairs of quantile levels 0.5-0.5 and 0.9-0.9, the locations of CA have larger quantile autocovariance than the locations of CB, although there is no gap of quantile autocovariance between the two clusters. Since the separation is not evident, principal component analysis was done for the case of quantile-based autocovariance distances. Then, considering all combinations of pairs of the used quantile levels (0.1, 0.5 and 0.9), it is obtained a cumulative variance of 88.7% with only the first principal component and 94.2% with the first two principal components. The graphic of the first two principal components (not presented here) shows a clear splitting between the locations of CA and the ones of CB, existing a gap between them in the first principal components. The first principal component is more associated with the pair of quantile levels 0.5:0.5, and then more associated with dependence on the median, while the second principal component is more associated with the pairs 0.1:0.1 and 0.9:0.9, which means that has a higher dependence on the extremes.

4 Conclusions

In this work, we have, for the analysis of 19 time series of maximum daily temperatures collected in the Iberian Peninsula, compared several clustering procedures, namely, autocorrelation-based distances (with different subsampling methodologies), and quantile-based autocovariance distances.

For the subsampling approach, the subsample length of 256 achieved higher cophenetic correlation coefficients of the resulting dendrograms, being them close to one. Thus, the dendrograms obtained for this subsample length provide a better summary of the data when compared with the subsample length 512.

Comparing subsampling with no subsampling, much higher cophenetic correlation coefficients are obtained for the subsampling approach. For the case of no subsampling, these coefficients are lower and far apart from one. However, both approaches provide fairly similar clusters.

The linkage method which provides the highest cophenetic correlation coefficients and gives dendrograms that better summarise the data is the average-linkage. In contrast, however, when using quantile autocovariance functions to obtain the dendrogram the differences among cophenetic correlation coefficients for the average-, the complete- and the Ward‘s linkage method are totally negligible.

For autocorrelation-based distances, all linkage methods and subsampling approaches agree in a partition of the weather stations in 3 clusters (see above). When using quantile-based autocovariance functions to obtain the dendrograms, for the no subsampling case, the partition obtained is in 2 clusters (see above). This splitting in 2 clusters is well shown through a principal component analysis, since only with the most important principal component, which accumulates 88.7% of the variance, it is possible to envisage such partition. The most important factor in the first principal component is the pair of quantile levels 0.5-0.5, with a contribution of 38.2%.

Finally, all considered methodologies agree when splitting the weather stations in 2 clusters, — being the partition obtained the one above mentioned. However, when more than 2 clusters are considered, the methodologies using the autocorrelation-based distances provide results different from the ones obtained using quantile-based autocovariance distances. The full dendrograms obtained for each one of these approaches are different (e.g, there are no cluster grouping only Valencia and Murcia in the case of quantile-based autocovariance distances, as in the case of autocorrelation-based distances). This difference in the partition may be due to the differences between methodologies to calculate the distance between time series.

Acknowledgments

The author wish to acknowledge the fruitful discussion with Prof. Manuel G. Scotto and Prof. Sónia C. A. Gouveia in the elaboration of this work.

References

Acero FJ, García JA, Gallego MC, Parey S, Dacunha-Castelle D. 2014. Trends in summer extreme temperatures over the Iberian Peninsula using nonurban station data. *Journal of Geophysical Research: Atmospheres* **119**: 39-53.

Alonso AM, Maharaj EA. 2006. Comparison of time series using subsampling. *Computational Statistics & Data Analysis* **50**: 2589-2599.

Alonso AM, de Zea Bermudez P, Scotto MG. 2014. Comparing generalized Pareto models fitted to extreme observations: an application to the largest temperature in Spain. *Stochastic Environmental Research and Risk Assessment* **28**: 1221-1233.

Caiado J, Crato N, Peña D. 2006. A periodogram-based metric for time series classification. *Computational Statistics & Data Analysis* **50**: 2668-2684.

De Luca G, Zuccolotto P. 2011. A tail dependence-based dissimilarity measure for financial time series clustering. *Advances in Data Analysis and Classification* **5**: 323-340.

Dias JCS. 2018. Cluster Analysis of the Iberian Peninsula Temperatures Time Series. Master Thesis. Instituto Superior Técnico.

D'Urso P, Maharaj EA. 2009. Autocorrelation-based fuzzy clustering of time series. *Fuzzy Sets and Systems* **160**: 3565-3589.

D'Urso P, Capelli C, Lallo DD, Massari R. 2013. Clustering of financial time series. *Physica A* **392**: 2114-2129.

European Climate Assessment & Dataset (ECAD). 2017. http://www.ecad.eu/. Date: Aug. 2017.

Faria PN, Cecon PR, da Silva AR, Finger FL, e Silva DD, Cruz CD, Sávio FL. 2012. Métodos de agrupamento em estudo de divergência genética de pimentas (Clustering methods in a study of genetic diversity of peppers). *Horticultura Brasileira* **30**: 428-432.

Iberian Climate Atlas. 2011. Eds: Agencia Estatal de Metereología, Instituto de Metereologia de Portugal. Closas-Orcoyen, Madrid.

Lafuente-Rego B, Vilar JA. 2016. Clustering of time series using quantile autocovariances. *Advances in Data Analysis and Classification* **10**: 391-415.

Lee J, Rao S. 2012. The quantile spectral density and comparison based tests for nonlinear time series. Unpublished manuscript. Department of Statistics, Texas A&M University, College Station, arXiv:1112.2759v2.

Linton O, Whang YJ. 2007. The quantilogram: with an application to evaluating directional predictability. *Journal of Econometrics* **141**: 250-282.

Montero P, Vilar JA. 2014. TSclust: Time series clustering utilities. https://CRAN.R-project.org/ package=TSclust. R package version 1.2.4.

Scotto MG, Alonso AM, Barbosa SM. 2010. Clustering time series of sea levels: extreme value approach. *Journal of Waterway, Port, Coastal, and Ocean Engineering* **136**: 215-225.

Scotto MG, Barbosa SM, Alonso AM. 2011. Extreme value and cluster analysis of European daily temperature series. *Journal of Applied Statistics* **38**: 2793-2804.

Sneath PHA, Sokal RR. 1973. Numerical Taxonomy: The Principles and Practice of Numerical Classification. Freeman, San Francisco, USA, p. 278 ff.

Trenberth KE, Jones PD, Ambejnje P, Bojariu R, Easterling D, Klein Tank A, Parker D, Rahimzadeh F, Renwick JA, Rusticucci M, Soden B, Zhai P. 2007. Observations: Surface and Atmospheric Climate Change. In The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt KB, Tignor M, Miller HL (eds). Cambridge University Press: Cambridge, UK. 1987.