A Comparison of Clustering Techniques for Meteorological Analysis

Ángel Arroyo, Verónica Tricio, Emilio Corchado and Álvaro Herrero

Abstract Present work proposes the application of several clustering techniques (*k*-means, SOM *k*-means, *k*-medoids, and agglomerative hierarchical) to analyze the climatological conditions in different places. To do so, real-life data from data acquisition stations in Spain are analyzed, provided by AEMET (Spanish Meteorological Agency). Some of the main meteorological variables daily acquired by these stations are studied in order to analyse the variability of the environmental conditions in the selected places. Additionally, it is intended to characterize the stations according to their location, which could be applied for any other station. A comprehensive analysis of four different clustering techniques is performed, giving interesting results for a meteorological analysis.

Keywords Clustering techniques • *K*-means • SOM *k*-means • *K*-medoids • Agglomerative hierarchical clustering • Meteorology

Á. Arroyo (✉) · Á. Herrero
Department of Civil Engineering, University of Burgos, Burgos, Spain
e-mail: aarroyop@ubu.es

Á. Herrero
e-mail: ahcosio@ubu.es

V. Tricio
Department of Physics, University of Burgos, Burgos, Spain
e-mail: vtricio@ubu.es

E. Corchado
Departamento de Informática y Automática, University of Salamanca, Salamanca, Spain
e-mail: escorchado@usal.es

Á. Herrero et al. (eds.), *10th International Conference on Soft Computing Models in Industrial and Environmental Applications*, Advances in Intelligent Systems and Computing 368, DOI 10.1007/978-3-319-19719-7_11

1 Introduction

As they are usually perceived as similar issues, it is necessary to distinguish between meteorology and climatology. On the one hand, meteorology consists in the study of the atmosphere, the scientific study of phenomena and physical processes occurring in the atmosphere, and atmospheric effects on the weather. Meteorologists then produce forecasts that are intended to predict weather conditions over the short term. On the other hand, climatology is the study of atmospheric changes, that define average climates and their change over time, due to both natural and anthropogenic climate variability. Climatology studies the same parameters as the meteorology, but its purpose is different; not because it seeks to make immediate forecast, but to study the long-term climatic characteristics. Climatology employs a long-term perspective, analyzing models that are designed to predict changes in weather patterns in the years to come. Present study focuses on the study of the meteorology in four places in Spain for a certain period of time.

In Spain, a network of stations for meteorological data acquisition can be found at [1]. These measurement stations acquire data continuously and these data are available for further study and analysis.

Clustering can be defined as the unsupervised classification of patterns into groups [2]. Hence, clustering (or grouping) techniques divide a given dataset into groups of similar objects, according to several different "similarity" measures. These sets of techniques have been previously applied to meteorological data [3, 4].

Differentiating from previous work, in present paper several clustering methods are applied to a detailed case study, where four places with different climates are selected with the more significant variables. Results are analyzed in two ways: the meteorology of the four places selected and the comparison of the clustering techniques to establish the strengths of each method. The main idea of present study is to analyse data describing meteorology from a case study associated to four places in Spain. Firstly, Principal Component Analysis [5] is applied to reduce the dimensionality of analyzed data and get an intuitive visualization of their internal structure. By doing so, we can determine an approximate number of clusters. In a second step, clustering techniques are applied to the original data set in order to find the best possible clustering of data. Four relevant hierarchical [6] and partitional [7] clustering techniques are applied, combined with the most widely-used distance measures. The number of clusters identified in the first step is applied in the second step as some of the techniques do need that figure.

The rest of this paper is organized as follows. Section 2 presents the techniques and methods that are applied. Section 3 details the real-life case study that is addressed in present work, while Sect. 4 describes the experiments and results. Finally, Sect. 5 sets out the main conclusions and future work.

2 Clustering Techniques and Methods

This study checks the performance of several clustering techniques when analyzing meteorological variables (described in Sect. 3), in order to study the behavior of the climatology in different locations.

In order to analyze data sets with meteorological information, several clustering methods [2, 8] have been applied. Clustering is one of the most important unsupervised learning problems [9]. It can be defined as the process of organizing objects into groups whose members are similar in some way. A cluster is a collection of objects which are similar to those in the cluster and are dissimilar to those belonging to other clusters. Clustering techniques can be divided, in general terms, into two categories: partitional and agglomerative. Partitional clustering algorithms divide the data set into a specified number of clusters trying to minimize certain criteria [10]. On the contrary, agglomerative clustering algorithms begin with each pattern in a distinct (singleton) cluster, and successively merges clusters together until a stopping criterion is satisfied [2].

Those methods and measure distances are described in this section.

2.1 Partitional Clustering

2.1.1 *k*-Means

The well-known *k*-means [11] is an algorithm for grouping data into a given number of clusters. Its application requires two input parameters: the number of clusters (k) and their initial centroids, which can be chosen by the user or obtained through some pre-processing. Each data element is assigned to the nearest group centroid, thereby obtaining the initial composition of the groups. Once these groups are obtained, the centroids are recalculated and a further reallocation is made. The process is repeated until the centroids do not change. Given the heavy reliance of this method on initial parameters, a good measure of the goodness of the grouping is simply the sum of the proximity Error Sums of Squares (SSE) that it attempts to minimize:

$$SSE = \sum_{j=1}^{k} \sum_{x \in G_j} \frac{p(x_i, c_j)}{n} \tag{1}$$

where p is the proximity function, k is the number of the groups, c_j *are* the centroids and n the number of rows. In the case of Euclidean distance, the expression is equivalent to the global mean square error.

2.1.2 SOM *k*-Means

Traditional Self Organizing Maps (SOM) [12] cannot provide with precise clustering results, while traditional *k*-means depends on the initial value and it is difficult to find the centroid of cluster [13].

To overcome the limitations of both methods, SOM *k*-means [12] is proposed. It combines SOM and *k*-means in the following way: when the SOM training finishes, the *k*-means algorithm is applied to refine the weights obtained by the SOM. When the SOM clustering finishes, *k*-means is also applied to refine the final result of clustering.

2.1.3 *k*-Medoids

The objective function of *k*-medoids (partitioning around medoids) algorithm is to partition a given dataset (X) into c clusters. The input and output arguments are the ones that *k*-means uses [11]. The main difference between the two methods consists in the way cluster centers are calculated; in *k*-medoids, the new cluster center is the nearest data point to the mean of the cluster points [14]. The algorithm generates random cluster centres, and not a partition matrix for initialization.

2.2 Agglomerative Hierarchical Clustering

Algorithms in this category generate a cluster tree also called dendrogram by using heuristic techniques. The most popular algorithms that use merging to generate the cluster tree are called agglomerative. There are many implementations of agglomerative hierarchical algorithms [15].

2.3 Measure Distances

The above mentioned clustering techniques, take into account distance in order to cluster the data. Different distance criteria are defined. The distance measures applied in present study are described in this subsection.

2.3.1 Euclidean Distance

This is the most common metric, where each centroid is the mean of the points in that cluster:

$$d(x-c) = (x-c)(x-c)' \tag{1}$$

where d is the distance from the point x to the centroid c.

2.3.2 Seuclidean Distance

In Standardized Euclidean metric (Seuclidean), each coordinate difference between rows in X is scaled by dividing it by the corresponding element of the standard deviation:

$$d(x-c) = (x-c)V^{-1}(x-c)' \tag{2}$$

where V is the n-by-n diagonal matrix.

2.3.3 Cityblock Distance

In this case, each centroid is the component-wise median of the points in that cluster.

$$d(x-c) = 1 - \sum_{j=1}^{p} |x_j - c_j|' \tag{3}$$

where the exponent P is a scalar positive value and j an observation in the vector X.

2.3.4 Cosine Distance

This is defined as one minus the cosine of the included angle between points (treated as vectors). Each centroid is the mean of the points in that cluster, after normalizing those points to unit Euclidean length:

$$d(x-c) = 1 - \frac{xc}{\sqrt{(xx')(cc')}}' \tag{4}$$

2.3.5 Correlation Distance

In this case, each centroid is the component-wise mean of the points in that cluster, after centering and normalizing those points to zero mean and unit standard deviation.

$$d(x-c) = 1 - \frac{(x-\bar{x})(c-\bar{c})}{\sqrt{(x-\bar{x})(c-\bar{c})}\sqrt{(c-\bar{c})(c-\bar{c})}} \tag{5}$$

2.3.6 Minkowski Metric

The Minkowski distance is a metric in a normed vector space which can be considered as a generalization of both the Euclidean distance and the Manhattan distance, as defined by:

$$d(x-c) = \sqrt[p]{\sum_{j=1}^{n} \left|x_{sj} - x_{tj}\right|^{p}} \tag{6}$$

where *p* is a scalar positive value of the exponent, *s* and *t* are the indexes of the rows of the vector *x* and *j* is index of the column of the same vector *x*.

3 Real-Life Case Study

This study is focused on the analysis of meteorological data recorded in four different places in Spain, which is a country with a noticeable climatic diversity. As it was described in Sect. 1, the data were provided by the Spanish Meteorological Agency (AEMET) [1, 16]. From the database of AEMET, the following four stations were selected for present analysis, based on their very different meteorology and sparse geographical location:

1. **Burgos Airport**. Geographical coordinates: 42°21’22”N; 03°37’17”W; 891 meters above sea level, moderate Continental climate. Labelled as **BU**.
2. **Santiago de Compostela Airport**. Geographical coordinates: 42°53’51”N; 08°24’38’W; 370 meters above sea level, Atlantic climate. Labelled as **SA**.
3. **Almería Airport**. Geographical coordinates: 36°50’47”N; 02°21’25”W; 21 meters above sea level, Mediterranean dry climate. Labelled as **AL**.
4. **Palma de Mallorca Port**. Geographical coordinates: 39°33’12”N; 02°37’31”E; 3 meters above sea level, typical Mediterranean climate. Labelled as **PM**.

From the timeline point of view, data are selected from years 2004 and 2005. Year 2003 was characterized by extreme values, particularly a heat wave during the month of August, in three of the analyzed locations. During years 2004 and 2005 normal values were again recorded for the analyzed variables. This fact, together with the intention of analyzing subsequent years on these studies, are the reasons for selecting years 2004 and 2005 in present study There are a total of 2,924 samples as data are collected on a daily basis (365 days for 2004 and 2005), that is

730 samples for each one of the 4 stations and one sample per day (daily average). The following parameters (six meteorological variables) are gathered:

1. Maximum absolute temperature: maximum temperature in the whole day (C°).
2. Minimum absolute temperature: minimum temperature in the whole day (C°).
3. Wind speed: air maximum gust recorded in the whole day (m/s).
4. Number of hours of sunshine in the day (hours).
5. Maximum absolute atmospheric pressure in tenths of hectopascal in the whole day (hPa).
6. Minimum absolute atmospheric pressure in tenths of hectopascal in the whole day (hPa).

4 Experiments and Results

As previously stated, Principal Component Analysis (PCA) [17] has been firstly applied to the dataset. In this step, the inner structure of the dataset is searched. In this case study, three main clusters of data are identified. This figure is used as an approximation to the number of clusters to be selected in the subsequent experiments. Once the initial approximate number of clusters is obtained, several clustering techniques are compared, namely: *k*-means, SOM *k*-means, *k*-medoids, and agglomerative hierarchical. For present study, the Matlab [18] implementations of such methods have been applied.

The results obtained by those techniques (after twenty valid runs for *k*-means, *k*-medoids and agglomerative hierarchical methods and ten valid runs for SOM *k*-means) are listed and described in this section: Tables 1, 2, 3 and 4 shows the parameter values of the applied technique and the allocation of data (according to the meteorological station they come from: BU, AL, SA, and PM) to the defined clusters (K). Additionally, execution time is also gathered to compare the different methods.

In Table 1, column k represents the number of clusters specified for the algorithm in advance, Distance is the measure distance applied (see Sect. 2), Time is the execution time (in seconds) and the Cluster Samples Allocation represents the percentage of samples from each one of the stations (BU, AL, SA and PA) that are allocated to each one the clusters; e.g. [100 0] represents 100 % of samples allocated to the first cluster and 0 % allocated to the second one.

From the data in Table 1, two main aspects can be highlighted. Firstly, the big difference between the meteorology of Burgos and the one of the other three locations, as well as the similar Mediterranean conditions in Almería and Palma de Mallorca. This can be seen in the following tendency; as the number of clusters is increased, the samples belonging to Burgos tend to remain together (in the same cluster), while the subdivision of samples in different clusters is more usual for the locations Almería and Palma de Mallorca. It is important to highlight that in all cases, samples from Mallorca and Almería are included in the same clusters. It is

Table 1 *K*-means clustering results

			Cluster samples allocation (%)			
k	Distance	Time (s)	BU	AL	SA	PM
2	Seuclidean	0.051601	[0 100]	[100 0]	[98 2]	[100 0]
2	Cityblock	0.045175	[100 0]	[0 100]	[9 91]	[0 100]
2	Cosine	0.045913	[63 37]	[40 60]	[67 33]	[44 56]
2	Correlation	0.631190	[29 71]	[69 31]	[41 59]	[75 25]
3	Seuclidean	0.039746	[100 0 0]	[0 100 0]	[0 0 100]	[0 100 0]
3	Cityblock	0.071559	[100 0 0]	[0 100 0]	[0 0 100]	[0 99 1]
3	Cosine	0.054590	[28 42 30]	[48 3 49]	[51 29 20]	[44 8 48]
3	Correlation	0.839933	[47 9 44]	[11 45 44]	[29 19 52]	[29 19 52]
4	Seuclidean	0.051418	[0 0 100 0]	[59 0 0 41]	[0 100 0 0]	[51 0 0 49]
4	Cityblock	0.080244	[52 0 0 48]	[0 100 0 0]	[0 0 100 0]	[0 100 0 0]
4	Cosine	0.058141	[18 28 39 15]	[34 27 2 38]	[45 27 21 7]	[38 20 5 37]
4	Correlation	0.084806	[39 5 40 15]	[24 44 7 26]	[31 15 18 36]	[21 50 6 23]
5	Seuclidean	0.060528	[0 49 0 51 0]	[0 0 59 0 41]	[100 0 0 0 0]	[0 0 51 0 49]
5	Cityblock	0.063433	[0 0 100 0 0]	[0 0 0 52 48]	[36 64 0 0 0]	[0 0 0 45 54]
5	Cosine	0.078143	[22 35 37 5 1]	[29 7 1 29 33]	[43 10 18 26 4]	[31 0 4 29 35]
5	Correlation	0.079106	[39 27 1 14 18]	[6 1 44 21 28]	[15 14 11 33 27]	[6 0 51 17 26]
6	Seuclidean	0.069633	[0 100 0 0 0 0]	[0 0 48 30 23 0]	[60 0 0 0 0 40]	[0 0 46 26 28 0]
6	Cityblock	0.081311	[31 0 6 19 27 17]	[2 34 25 12 1 27]	[10 5 21 24 14 27]	[2 44 18 10 0 25]
6	Cosine	0.104866	[34 34 17 10 3 1]	[5 1 15 20 26 33]	[11 14 22 28 22 3]	[0 3 9 26 29 33]
6	Correlation	0.082790	[16 3 23 1 31 26]	[18 33 9 35 3 1]	[29 19 18 8 11 14]	[12 35 6 43 4 0]

also worth mentioning the great influence of the measure distance applied. While ‘cosine’ and ‘correlation’ usually split samples from a location in different clusters, ‘seuclidean’ and ‘cityblock’ generally keep the samples from the same location in the same cluster. This is because ‘cosine’ and ‘correlation’ measures the difference in the angle between two vectors and not the difference in the magnitude between two vectors [10]. Finally, regarding the elapsed time in executing the *k*-means algorithms, it could be say that ‘Correlation’ provides the lowest response when k equals 1, 2, and 3. A similar response is obtained when applying the other measure distance when k equals 4, 5 and 6.

In Table 2, the results obtained by SOM *k*-means are shown. In this table, Type is the type of applied algorithm in the neurons initialization process (it can be sequential or batch). Additionally Err shows the total quantization error for the data set, according to the distance from any given data point to a cluster center weighted by that data point’s membership grade.

One of the first conclusions that can be drawn from Table 2 is that, as expected, the ‘seq’ algorithm is slower than the ‘batch’ one. Both are iterative algorithms, but the batch version is much faster in Matlab since matrix operations can be utilized efficiently [19]. In relation to the cluster sample allocation process, SOM *k*-means offer similar results to *k*-means (Table 1) when applying ‘seuclidean’ distance; this is because SOM k-means uses also a simple distance measure. The same patterns in the process of cluster samples allocation are repeated respect to *k*-means (Table 1).

In Table 3 the results of applying *k*-medoids to the original data set are shown.

By applying *k*-medoids (Table 3), the cluster sample allocation is similar to that obtained by *k*-means (Table 1) when the measure distance is ‘seuclidean’. However, it is quite different when ‘cosine’ and ‘correlation’ distances are applied. In these cases *k*-means makes a wider division of the samples into clusters. The main drawback of *k*-medoids respect to *k*-means (Table 1), is the computational cost; as can be seen, when the number of clusters is increased, the execution time is much bigger in the case of *k*-medoids.

Table 4 shows the very different results of applying agglomerative hierarchical clustering technique to the original dataset from those obtained in partitional methods, (Tables 1, 2 and 3).

The main difference between agglomerative hierarchical clustering and the other three methods is that the former allocates, with almost 100 % of accuracy, the samples to clusters according to the location of the stations. Increasing the number of cluster is not then useful in the case of agglomerative hierarchical clustering because most of the additional clusters are empty (no samples are allocated to them). Not a very reliable response is detected in the samples cluster allocation process when k equals 3, 4, 5, and 6 and when the measure distance selected is ‘seuclidean’, the total of the samples are allocated in the same cluster. Additionally, when k equals 3 and 6 and selected distance is ‘cityblock’, the samples of Santiago de Compostela, Almería and Palma de Mallorca are allocated in the same cluster. Agglomerative hierarchical clustering is able to distinguish with almost 100 % accuracy three groups of data: Burgos, Santiago de Compostela and Almería together with Palma de Mallorca, when (k equals 5 and 6) and for ‘euclidean’ or

Table 2 SOM *k*-means clustering results

				Cluster samples allocation (%)			
K	Type	Err	Time (s)	BU	AL	SA	PM
2	Seq	3.74	1.934958	[0 100]	[100 0]	[98 2]	[100 0]
2	Batch	3.73	0.025540	[0 100]	[100 0]	[98 2]	[100 0]
3	Seq	0.99	1.923084	[0 0 100]	[0 100 0]	[100 0 0]	[0 100 0]
3	Batch	0.99	0.028748	[0 0 100]	[0 100 0]	[100 0 0]	[0 100 0]
4	Seq	0.83	1.928557	[0 0 0 100]	[59 0 40 0]	[0 100 0 0]	[51 0 49 0]
4	Batch	0.87	0.032140	[0 0 51 49]	[0 100 0 0]	[100 0 0 0]	[0 100 0 0]
5	Seq	0.74	1.917720	[0 100 0 0 0 0]	[0 0 28 24 48 0]	[99 0 1 0 0 0]	[0 0 25 29 46 0]
5	Batch	0.77	0.034465	[0 49 51 0 0 0]	[57 0 0 0 43 0]	[0 0 0 99 0 0]	[69 0 0 0 31 0]
6	Seq	0.66	1.943915	[0 100 0 0 0 0]	[0 0 48 29 0 23]	[32 0 0 0 68 0]	[0 0 46 26 0 29]
6	Batch	0.66	0.037405	[0 0 0 100 0 0]	[48 0 30 0 23 0]	[0 40 0 0 0 60]	[46 0 26 0 28 0]

Table 3 *K*-medoids clustering results

			Cluster samples allocation (%)			
K	Distance	Time (s)	BU	AL	SA	PM
2	Euclidean	0.477961	[0 100]	[100 0]	[93 7]	[100 0]
2	Seuclidean	0.624595	[100 0]	[0 100]	[50 50]	[1 99]
2	Cosine	0.590097	[0 100]	[100 0]	[92 8]	[100 0]
2	Correlation	0.461034	[0 100]	[100 0]	[92 8]	[100 0]
3	Euclidean	0.835087	[0 100 0]	[100 0 0]	[0 0 100]	[100 0 0]
3	Seuclidean	1.092403	[0 0 100]	[50 50 0]	[15 70 15]	[50 50 0]
3	Cosine	0.623375	[0 100 0]	[0 0 100]	[100 0 0]	[0 0 100]
3	Correlation	0.633741	[0 0 100]	[100 0 0]	[0 100 0]	[100 0 0]
4	Euclidean	1.352807	[0 0 100 0]	[0 50 0 50]	[100 0 0 0]	[0 47 0 53]
4	Seuclidean	1.420630	[0 45 0 55]	[50 0 50 0]	[76 1 15 9]	[50 0 50 0]
4	Cosine	1.063674	[100 0 0 0]	[0 51 0 49]	[0 0 100 0]	[0 46 0 54]
4	Correlation	1.134566	[0 100 0 0]	[0 0 50 50]	[100 0 0 0]	[0 0 46 54]
5	Euclidean	1.293159	[0 0 57 43 0]	[0 50 0 0 50]	[100 0 0 0 0]	[0 46 0 0 54]
5	Seuclidean	1.676479	[0 45 0 55 0]	[48 0 39 0 13]	[15 1 24 1 60]	[47 0 41 0 11]
5	Cosine	1.309666	[0 0 0 44 56]	[0 51 49 0 0]	[100 0 0 0 0]	[0 46 54 0 0]
5	Correlation	1.674484	[0 100 0 0 0]	[31 0 0 25 44]	[0 0 100 0 0]	[26 0 0 31 43]
6	Euclidean	1.946891	[0 57 0 43 0 0]	[0 0 25 0 44 31]	[100 0 0 0 0 0]	[0 0 31 0 43 26]
6	Seuclidean	1.544820	[0 55 0 0 0 45]	[30 0 29 4 38 0]	[15 2 11 58 13 0]	[37 0 10 10 42 0]
6	Cosine	1.947368	[0 100 0 0 0 0]	[46 0 0 27 0 28]	[0 0 66 0 34 0]	[41 0 0 23 0 35]
6	Correlation	1.974048	[0 57 0 43 0 0]	[50 0 50 0 0 0]	[0 0 0 0 39 61]	[47 0 53 0 0 0]

Table 4 Agglomerative hierarchical clustering results

			Cluster samples allocation (%)			
K	Distance	Time (s)	BU	AL	SA	PM
2	Euclidean	31.217975	[0 100]	[100 0]	[100 0]	[100 0]
2	Seuclidean	31.722818	[0 100]	[0 100]	[0 100]	[0 100]
2	Cityblock	30.559558	[100 0]	[0 100]	[0 100]	[0 100]
2	minkowsky	30.697608	[0 100]	[100 0]	[100 0]	[100 0]
3	Euclidean	31.083268	[0 100 0]	[0 0 100]	[0 0 100]	[0 0 100]
3	Seuclidean	31.626975	[0 100 0]	[0 100 0]	[0 100 0]	[0 100 0]
3	Cityblock	31.734806	[0 0 100]	[0 100 0]	[0 100 0]	[0 100 0]
3	minkowsky	31.577912	[0 100 0]	[0 0 100]	[0 0 100]	[0 0 100]
4	Euclidean	31.163991	[0 0 0 100]	[0 100 0 0]	[0 100 0 0]	[0 100 0 0]
4	Seuclidean	30.652015	[0 100 0 0]	[0 100 0 0]	[0 100 0 0]	[0 100 0 0]
4	Cityblock	31.023600	[0 100 0 0]	[0 0 0 100]	[0 0 0 100]	[0 0 0 100]
4	minkowsky	30.651821	[0 0 0 100]	[0 100 0 0]	[0 100 0 0]	[0 100 0 0]
5	Euclidean	30.652530	[0 0 0 0 100]	[0 100 0 0 0]	[100 0 0 0 0]	[0 100 0 0 0]
5	Seuclidean	30.602942	[0 100 0 0 0]	[0 100 0 0 0]	[0 100 0 0 0]	[0 100 0 0 0]
5	Cityblock	30.619702	[0 0 0 100 0]	[0 100 0 0 0]	[0 100 0 0 0]	[0 100 0 0 0]
5	minkowsky	30.623085	[0 0 0 0 100]	[0 100 0 0 0]	[100 0 0 0 0]	[0 100 0 0 0]
6	Euclidean	30.643442	[0 0 0 0 0 100]	[0 100 0 0 0 0]	[0 1 99 0 0 0]	[0 100 0 0 0 0]
6	Seuclidean	30.694775	[0 100 0 0 0 0]	[0 100 0 0 0 0]	[0 100 0 0 0 0]	[0 100 0 0 0 0]
6	Cityblock	30.762574	[0 100 0 0 0 0]	[0 0 0 100 0 0]	[0 0 0 100 0 0]	[0 0 0 100 0 0]
6	minkowsky	30.737657	[0 0 0 0 0 100]	[0 100 0 0 0 0]	[0 1 99 0 0 0]	[0 100 0 0 0 0]

'minkowsky' distance. This is because 'euclidean' distance is a particular case of 'mikowski' distance. Another drawback is that this technique is highly computationally demanding, regardless of the number of selected clusters or the distance metric applied.

5 Conclusions and Future Work

Main conclusions derived from previously explained results (see Sect. 4) can be divided into two groups, firstly, those regarding the analysis of meteorological conditions in the analyzed case study. Secondly, those related to the behaviour of the different clustering techniques applied in the case study.

Talking about the meteorological conditions in the four selected places, a clear conclusion is the big difference between the climatology in Burgos and that in the other three places. Also, in Santiago de Compostela it is appreciated a different climatology from the other three places, but not as pronounced as in the case of Burgos. However, the climate in Palma de Mallorca and Almería are very similar between them, as none of the applied methods has been able to split those samples in different clusters.

Regarding the applied clustering techniques, it should be emphasized the different results offered by the hierarchical agglomerative method compared with the partitional methods, and also the differences between the application of different measures of distance. In many cases, the agglomerative hierarchical clustering do not show a reliable response, not being able to allocate samples from different places in different clusters.. *K*-means, SOM *k*-means and *k*-medoids attain similar results, and the selected measure distance selected is a key factor. *K*-means is the best method in terms of computational load. None of the techniques has been able to separate the samples of the four locations in four different clusters, as it was initially detected through PCA.

Future work will consists on expanding the time window to analyze the temporal evolution of meteorological data. It will also include the application of probabilistic methods and different evaluating techniques, in order to extend the comparison.

References

1. National Network of meteorological stations—Spanish Agency of Meteorology. http://www.aemet.es/es/eltiempo/observacion/ultimosdatos
2. Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review. ACM Comput Surv (CSUR) 31(3):264–323
3. Lu Y, Ma T, Yin C, Xie X, Tian W, Zhong S (2015) Implementation of the fuzzy C-means clustering algorithm in meteorological data. Int J Database Theory Appl 6:1–18
4. Tian W, Zheng Y, Yang R, Ji S, Wang J (2015) A survey on clustering based meteorological data mining. Int J Grid Distributed Comput 7:229–240

5. Hotelling H (1933) Analysis of a complex of statistical variables into principal components. J Educ Psychol 24:417–444
6. Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, Eccles MP, Cane J, Wood CE (2013) The behavior change technique taxonomy (v1) of 93 Hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 46(1):81–95
7. Aparna K, Nair MK (2015) Comprehensive study and analysis of partitional data clustering techniques. Int J Bus Anal (IJBAN) 2:23–38
8. Anil K (2010) J.: Data clustering: 50 years beyond K-means. Pattern Recogn Lett 31:651–666
9. Barlow H (1989) Unsupervised learning. Neural Comput 1:295–311
10. Jain AK, Maheswari S (2013) Survey of recent clustering techniques in data mining. J Curr Comput Sci Technol 3
11. Ding C, He X (2004) K-means clustering via principal component analysis. In: Proceedings of the twenty-first international conference on Machine learning, vol 29 (2004)
12. Kohonen T (1990) The self-organizing map. Proc IEEE 78:1464–1480
13. Napoleon D, Pavalakodi S (2011) A New method for dimensionality reduction using K means clustering algorithm for high dimensional data set. Int J Comput Appl 13:41–46
14. Park HS, Jun CH (2009) A simple and fast algorithm for K-medoids clustering. Expert Syst Appl 36:3336–3341
15. Day WHE, Edelsbrunner H (1984) Efficient algorithms for agglomerative hierarchical clustering methods. J Classif 1:7–24
16. Hotelling H (1933) Analysis of a complex of statistical variables into principal components. J Educ Psychol 24:498–520
17. Mathworks. http://es.mathworks.com/products/matlab/?refresh=true (2015)
18. Vesanto J, Himberg J, Alhoniemi E, Parhankangas J (1999) Self-organizing map in Matlab: the SOM toolbox. In: Proceedings of the Matlab DSP Conference, vol 99, pp 16–27