Journal of Applied Logic 24 (2017) 76–89

Contents lists available at ScienceDirect

Journal of Applied Logic

www.elsevier.com/locate/jal

Analysis of meteorological conditions in Spain by means of clustering techniques

Ángel Arroyo [a,*], Álvaro Herrero [a], Verónica Tricio [b], Emilio Corchado [c]

[a] Department of Civil Engineering, University of Burgos, Burgos, Spain
[b] Department of Physics, University of Burgos, Burgos, Spain
[c] Departamento de Informática y Automática, University of Salamanca, Salamanca, Spain

ARTICLE INFO

Article history:
Available online 15 November 2016

Keywords:
k-means
SOM *k*-means
k-medoids
Agglomerative hierarchical clustering
Cluster based on Gaussian Mixture Models
Clustering evaluation techniques
Meteorology

ABSTRACT

A comprehensive analysis of clustering techniques is presented in this paper through their application to data on meteorological conditions. Six partitional and hierarchical clustering techniques (*k*-means, *k*-medoids, SOM *k*-means, Agglomerative Hierarchical Clustering, and Clustering based on Gaussian Mixture Models) with different distance criteria, together with some clustering evaluation measures (Calinski–Harabasz, Davies–Bouldin, Gap and Silhouette criterion clustering evaluation object), present various analyses of the main climatic zones in Spain. Real-life data sets, recorded by AEMET (Spanish Meteorological Agency) at four of its weather stations, are analyzed in order to characterize the actual weather conditions at each location. The clustering techniques process the data on some of the main daily meteorological variables collected at these stations over six years between 2004 and 2010.

1. Introduction

Meteorology and climatology are different fields of study, although they may often be confused. Meteorology is the scientific study of atmospheric phenomena, physical processes in the atmosphere and atmospheric effects on the weather. Meteorologists then produce weather forecasts that predict short changes in the weather. In contrast, climatology is the study of atmospheric changes that define average climates and their long-term changes, due to both natural and anthropogenic variations in the climate. Climatological studies therefore share certain meteorological parameters, although climatology predicts long-term weather patterns through climatic models rather than making short-term forecasts. The present study focuses on the clustering analysis of meteorological data from four locations in Spain over a five-year period.

* Corresponding author.
E-mail addresses: aarroyop@ubu.es (Á. Arroyo), ahcosio@ubu.es (Á. Herrero), vtricio@ubu.es (V. Tricio), escorchado@usal.es (E. Corchado).

http://dx.doi.org/10.1016/j.jal.2016.11.026
1570-8683/

Clustering is a useful technique [29] in the study of meteorological phenomena and correct selection of the right clustering algorithm, a requisite for successful experiments. Clustering can be defined as the unsupervised classification of patterns into groups [14]. Hence, clustering (or grouping) techniques will divide a given dataset into groups of similar objects, according to various "similarity" measures. These sets of techniques have previously been applied to meteorological data. In [25], 24-hour mass air trajectories were analyzed at a location in Spain over a three-year period. Clustering techniques with spherical trigonometry were applied, together with the kernel regression method, for their calculation. A multivariate data cube was investigated in [32], to establish whether climate and vegetation classes coincided. To do so, unsupervised clustering techniques were applied and differences between clustering of climate variables versus vegetation variables were studied. In [11], two clustering techniques and a neural network were applied, in an analysis of air quality in Greece and the impact of weather circulation patterns on urban air quality over a period of five-years. Principal Components Analysis (PCA) and Cluster Analysis (CA), were applied in [26] over a 3-year period to analyze the mass concentrations of Sulfur Dioxide (SO_2) and Particulate Matter (PM10) in Oporto. Finally, a clustering method for the study of multidimensional non-stationary meteorological time series was presented in [15] and the results were compared with standard fuzzy clustering techniques for a dataset with temperatures in Europe over forty years. In this study, unlike previous works, cluster evaluation measures, together with partitional and hierarchical clustering techniques, based on different distance measures, are employed to categorize different climate zones in Spain.

So, with these promising techniques, we can analyze the study of both the similarities of the main regional climates and their differences. The climate in Spain is highly variable, mainly due to its position in southern Europe, diverse relief, and extensive coastline. The Iberian Peninsula is in a temperate zone where currents of warm air and cold air merge to create its unique meteorological conditions. This great variability of climatic zones means that Spain is a European country of special interest for a meteorological study of the sort proposed in this study. Its various climatic subtypes [8] are reflected in the data gathered from four points: a Mediterranean island; an interior location on the meseta of the Iberian Peninsula; a city on the southern coastline; and, a city to the north-west of the Iberian Peninsula. The network of weather stations for meteorological data acquisition are constantly recording continuous data streams that are publicly accessible for research and analysis [23]. Described in detail in Section 3, these stations represent points within each of the four main Spanish climatic zones (continental, Atlantic, dry Mediterranean and typical Mediterranean).

Unlike the time window in a previous work [3] by the authors, a wider time window is used for data analysis in this study, running between 2004 and 2010. Additionally, a larger and more comprehensive set of techniques analyses the meteorological the extensive time series of data from the four different climatic zones (see Section 3). Firstly, four cluster evaluation measures yielded an accurate estimation of the recommended number of clusters for the dataset. Secondly, various clustering techniques applied to the original data set allowed us to assign the best possible data clustering technique. Four relevant partitional [2] techniques, one hierarchical [21] technique, and four cluster evaluation measures [17] were applied, combined with the most widely-used distance measures. The results were analyzed in two ways: through a study of the meteorology at the four selected locations and through clustering technique comparisons to establish the advantages of each method.

The rest of this paper is organized as follows. Section 2 presents the clustering techniques, distance criteria, and the cluster evaluation measures applied. Section 3 describes the real-life case study and, Section 4, the experimental results. Finally, Section 5 sets out the main conclusions and future lines of work.

2. Clustering techniques

This study reports the performance of several clustering techniques analyzing time series of data on meteorological conditions (described in Section 3), studying the climatology of different locations. Several

clustering methods [14,1] and clustering evaluation techniques: [5,28,6,30] have been applied in our analysis of the data sets with meteorological information,.

Clustering, a key unsupervised learning problems [4], can be defined as the process of organizing objects into groups that in some way have similar members. A cluster is a collection of objects that 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 seeking to minimize certain criteria [13]. On the contrary, agglomerative clustering algorithms begin with each pattern in a distinct (singleton) cluster, and successively merge clusters together until a stopping criterion is satisfied [14].

The evaluation measures, the clustering techniques applied, and distance criteria are described in this section and their Matlab [19] implementations are applied in this study.

2.1. Cluster evaluation measures

Clustering validation evaluates the goodness of clustering results [17]. The two main categories of clustering validation are external and internal. The main difference is whether external information (for which *a priori* knowledge of the dataset is required) is used for clustering validation. Internal validation measures can be used to choose the best clustering algorithm, as can the optimal numbers of clusters, with no further information needed. The following four internal validation measures were all applied in the present work.

Calinski–Harabasz Index. The Calinski–Harabasz Index [5] evaluates cluster validity based on the between-cluster means and the within-clusters covariance matrix. It measures separation in relation to the maximum distance between cluster centers, and compactness, as the sum of distances between objects and their cluster center. As separate and compact clusters are desirable, the between-class is maximized and the within-class scatter matrix is minimized. The value of k that maximizes the Calinski–Harabasz index points to an estimation of the optimal number of clusters.

Silhouette Index. The Silhouette index [28] scores clustering performance, based on the pairwise difference of between-cluster and within-cluster distances. In addition, the optimal cluster number is determined by maximizing the value of this index. As the objective is to obtain clusters with minimum intra-cluster distance and maximum inter-cluster distance, high Silhouette Index values are desirable. Thus, the optimal partition is the partition with the highest Silhouette Index for parameter k.

Davies–Bouldin Index. Similar to the Calinski–Harabasz Index, the Davies–Bouldin Index [6] obtains clusters with the minimum intra-cluster distance and the maximum distance between cluster centroids. The minimum value of the index indicates a suitable dataset partition. The Davies–Bouldin Index [6] is calculated as follows: for each cluster, the similarities between each cluster C and all other clusters are computed, and the highest value is assigned to C in terms of its cluster similarity. The Davies–Bouldin Index may then be obtained by averaging all the cluster similarities; however, the smaller this index, the better the clustering result.

Gap Index. The Gap Index [30] uses the output of any clustering algorithm, comparing the change in within-cluster dispersion with that expected under an appropriate reference null distribution. The Gap Index is especially useful on well-separated clusters and when used with a uniform reference distribution in the principal component orientation.

2.2. Partitional clustering

***k*-means.** The well-known k-means [9] 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 there are no further changes in the centroids. 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 Sums of Squared Error (SSE) that it attempts to minimize, where $p()$ is the proximity function, k is the number of the groups, c_j are the centroids, and n the number of rows:

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

In the case of Euclidean distance, the expression is equivalent to the global mean square error.

SOM k-means. Self Organizing Maps (SOM) [16] cannot provide precise clustering results, while the k-means statistic depends on the initial value and has difficulty finding the cluster centroid [22].

SOM k-means [16] is proposed to overcome the limitations of both methods. 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.

k-medoids. The objective function of the k-medoids (partitioning around medoids) algorithm is to partition a given dataset (X) into c clusters. The input and output arguments are those used by k-means [9]. The main difference between both methods is in their way of calculating cluster centers; in k-medoids, the new cluster center is the nearest data point to the mean of the cluster points [24]. The algorithm generates random cluster centers, rather than a partition matrix for initialization.

Cluster based on Gaussian Mixture Model. A Gaussian Mixture Model (GMM) [27] is a parametric probability density function represented as a weighted sum of Gaussian component densities. From a number of samples or observations, GMM calculates the estimation of the parameters of each of the distributions and subpopulations making up the mixture. GMM parameters are estimated from training data using the iterative Expectation–Maximization (EM) algorithm [10].

The EM algorithm enables parameter estimation in probabilistic models with incomplete data. The algorithm is a natural generalization of maximum likelihood estimation to the incomplete data case. The EM algorithm aims to maximize the density function of the data based on the parameters for likelihood estimation.

2.3. Agglomerative hierarchical clustering

Hierarchical clustering algorithms are either top-down or bottom-up approaches. Bottom-up algorithms treat each sample as a singleton cluster at the outset and then successively merge (or agglomerate) pairs of clusters until all clusters are merged into a single cluster that contains all the documents. Bottom-up hierarchical clustering is therefore called Hierarchical Agglomerative Clustering (HAC) [18]. Algorithms in this category generate a cluster tree or dendrogram by using heuristic techniques. A dendrogram consists of many U-shaped lines that connect data points in a hierarchical tree. The height of each U represents the distance between the two connected data points. The most popular algorithms that use merging to generate the cluster tree are called agglomerative. There are many implementations of agglomerative hierarchical algorithms [7]. Additionally, it may be highlighted that dendrograms are only shown to explain the bad results offered by Agglomerative hierarchical clustering, but a deeper analysis of these dendrograms lies outside the scope of the present paper.

2.4. Distance criteria

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

Given an mx-by-n data matrix X, which is treated as mx (1-by-n) row vectors x_1, $x_2, ..., x_{mx}$, and my-by-n data matrix Y, which is treated as my (1-by-n) row vectors y_1, $y_2, ..., y_{my}$, the various distances between the vector x_s and y_t are defined as follows:

Euclidean distance. This is the most common metric, where each centroid is the mean of the points in its cluster:

$$d_{st}^2 = (x_s - y_t)(x_s - y_t)' \tag{2}$$

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

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

$$d_{st}^2 = (x_s - y_t)V^{-1}(x_s - y_t)' \tag{3}$$

where V is the n-by-n diagonal matrix the jth diagonal element of which is $S(j)^2$, where S is the vector of standard deviations.

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

$$d_{st} = \sum_{j=1}^{n} |x_{sj} - y_{tj}| \tag{4}$$

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

Cosine distance. This distance 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 unitary Euclidean lengths:

$$d_{st} = 1 - \frac{x_s y_t'}{\sqrt{(x_s x_s')(y_t y_t')}} \tag{5}$$

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 a zero mean and a unit standard deviation.

$$d_{st} = 1 - \frac{(x_s - \bar{x}_s)(y_t - \bar{y}_t)'}{\sqrt{(x_s - \bar{x}_s)(x_s - \bar{x}_s)'}\sqrt{(y_t - \bar{y}_t)(y_t - \bar{y}_t)'}} \tag{6}$$

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

$$d_{st} = \sqrt[p]{\sum_{j=1}^{n} |x_{sj} - x_{tj}|^p} \tag{7}$$

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

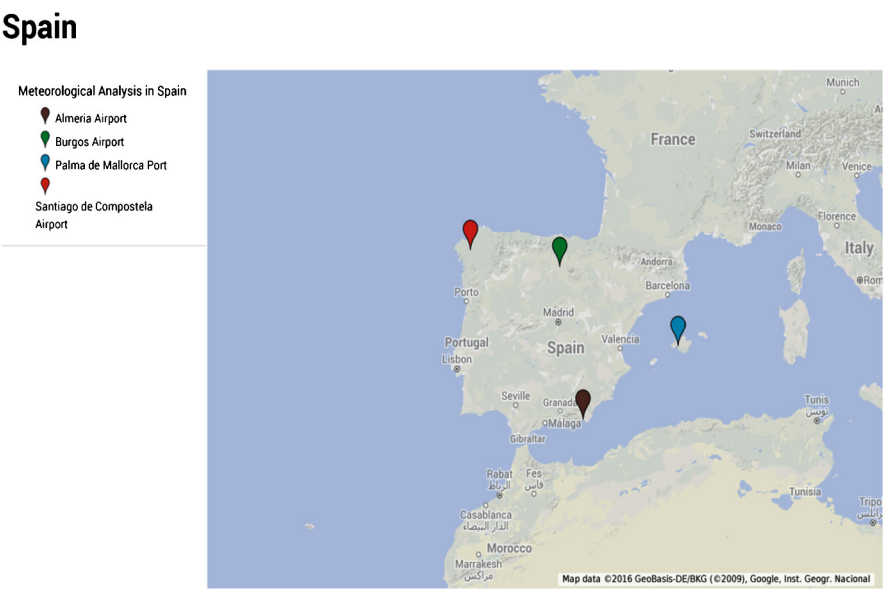

Fig. 1. Location of the four stations analyzed in Spain (source: Google Maps).

3. Real-life case study

This study presents an analysis of meteorological data recorded at four different points in Spain, a country with noticeable climatic variations. As mentioned in Section 1, the data under study were kindly supplied by the Spanish Meteorological Agency (AEMET) [23,20]. The following four stations were selected from the AEMET database, in view of their very different climatic conditions, typical of the four main climatic zones in Spain (see Fig. 1).

The following four data-acquisition stations supplied the data for this analysis:

1. **Burgos Airport** (Labeled as **BU**). Geographical coordinates: 42°21′22″N; 03°37′17″W; 891 meters above sea level, moderate continental climate. The continental climate is tempered by more rigorous climatic conditions from the Atlantic seaboard, with high diurnal and annual differences in temperature (frosty, icy winters below 0 °C) and generally low rainfall.
2. **Santiago de Compostela Airport** (Labeled as **SA**). Geographical coordinates: 42°53′51″N; 08°24′38′W; 370 meters above sea level, Atlantic climate. Rainfall is very abundant and usually at least 1,000 mm per month on average. Copious rains are well distributed throughout the year, with a peak in autumn-winter and a summer minimum, with more than 30 mm each month at low intensity. Under these conditions, the average relative humidity is high (80%).
3. **Almeria Airport** (Labeled as **AL**). Geographical coordinates: 36°50′47″N; 02°21′25″W; 21 meters above sea level, Mediterranean dry climate**.** The dry Mediterranean climate is given as a transition between the Mediterranean climate and the desert climate and is characterized by drought most of the year. The Mediterranean basin, where the typical Mediterranean weather patterns may be found, has a warmer climate than in the east (see Fig. 1) and less rainfall, ranging between 200 and 400 mm a month, concentrated around the colder seasons. Winter temperatures are hot, the summer is dry, with mild temperatures on the coast and very high (above 25 °C) temperatures in inland areas; the average maximum temperatures can exceed 45 °C in the event of heat waves.
4. **Palma de Mallorca Port** (Labeled as **PM**). Geographical coordinates: 39°33′12″N; 02°37′31″E; 3 meters above sea level, typical Mediterranean climate**.** This climate is characterized by hot and dry summers,

with average temperatures above 22 °C and humid and rainy winters with mild temperatures. In the colder months, there is more rain, and the warmest month is the driest month.

On a timeline, data were selected from 2004 up until 2010. Year 2003 was not included as it was characterized by extreme values, particularly a heat wave during the month of August, in three of the locations under analysis. So, together with non-availability of data since 2011, those reasons explain the selected time window in the present study. There are a total of 10,162 samples as data are collected on a daily basis (365 days for 7 years), that are about 2500 samples for each one of the 4 stations and one sample per day. Some data were omitted because of missing or corrupt data information. The main parameters in the study were the following six (daily average) meteorological variables:

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

4. Experiments and results

In a previous work, Principal Component Analysis (PCA) [12] was initially applied to the sample dataset with the aim of identifying its general inner structure. The PCA projection was used to gain an approximate idea of the number of clusters to be selected in the subsequent experiments. One characteristic of PCA is that clusters can be identified with the naked eye in a graphical representation, without any label or assignment of each sample to a certain group of data. These techniques are very useful to gain general knowledge about the structure of unlabeled datasets. In the dataset under analysis in this study, three main clusters of data were identified in the PCA projection. Subsequently, some cluster evaluation measures (see Section 2) were applied to obtain a recommended number of clusters. Once the initial approximate number of clusters was obtained, several clustering techniques were compared according to the estimated number of clusters.

The results obtained by those techniques are listed and described in this section: Tables 1–5 show the parameter values of the applied techniques and the allocation of data (by the meteorological station they come from: BU, AL, SA, and PM) to the defined number of clusters (k). Additionally, computing time is also shown for comparison with the different methods.

Table 1 shows the information on the cluster evaluation performed by applying the different measures with the Gaussian mixture distribution algorithm. In this table, column 'k' represents the optimum number of clusters selected by each one of the measures from the 'InspectedK' parameter (taking values from 2 to 6), 'Time' is the execution time (in seconds) and 'Criterion Values' corresponding to each proposed number of clusters in 'InspectedK', stored as a vector of numerical values. Each value of this vector is calculated according to the evaluation measure on cluster centroids, the number of points in each cluster, the sum of Squared Euclidean and the number of clusters. In Gap, 'Reference Distribution' is the reference data generation method, and 'LogW' is the natural logarithm of W based on the input data, stored as a vector of scalar values where W is the within-cluster dispersion computed using the distance measurement distance.

The output of the four measures applied (Calinski–Harabasz, Davies–Bouldin, Gap and Silhouette) was similar: the value of k obtained was three in the case of Calinski–Harabasz, Gap and Silhouette while Davies–Bouldin gave a value of two This result points to the usefulness of the k parameter, required as an input for the subsequent clustering techniques. The Gap evaluation measure was the slowest in terms of computing time.

Table 1
Cluster evaluation.

Cluster Evaluation Measure	K	Time (s)	Parameters
Calinski–Harabasz	3	2.73	Criterion Values: [38489.78 86717.77 63255.07 49307.77 45169.55]
Davies–Bouldin	2	2.29	Criterion Values: [0.34 0.39 0.93 1.32 1.55]
Gap	3	366.63	Criterion Values: [1.25 2.24 2.15 2.17 1.92] Reference Distribution: 'PCA'. LogW: [−7.88 −9.20 −9.29 −9.40 −9.48]
Silhouette	3	16.36	Criterion Values: [0.86 0.89 0.74 0.44 0.34]

Table 2
Initial k-means clustering results.

K	Distance	Time (s)	SumD	Cluster Samples Allocation (%)			
				BU	SA	AL	PM
2	Seuclidean	0.10	[4.46E−05 0.0003]	[100 0]	[4 96]	[0 100]	[0 100]
2	Cityblock	0.11	[0.74 2.37]	[100 0]	[11 89]	[0 100]	[0 100]
2	Cosine	0.10	[0.11 0.083]	[64 36]	[67 33]	[37 63]	[41 59]
2	Correlation	0.10	[0.090 0.089]	[69 31]	[71 29]	[32 68]	[21 79]
3	Seuclidean	0.13	[4.40E−05 3.06E−05 2.54E−05]	[0 100 0]	[0 0 100]	[100 0 0]	[100 0 0]
3	Cityblock	0.13	[0.47 0.90 0.53]	[0 0 100]	[100 0 0]	[1 99 0]	[1 99 0]
3	Cosine	0.18	[0.06 0.04 0.04]	[30 45 25]	[48 36 16]	[50 3 47]	[47 8 45]
3	Correlation	0.22	[0.041 0.053 0.046]	[9 48 43]	[22 36 42]	[49 19 32]	[59 11 30]

Table 2 shows the results obtained for the k-means with different distance criteria and the two suggested values for the k parameter (2 and 3). In this table, 'Distance' is the distance criterion applied (see Section 2) and 'SumD' is the within-cluster sums of point-to-centroid distances in the k-by-1 vector. 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. [1000] represents 2 clusters and 100% of samples allocated to the first cluster and 0% to the second one.

Two central points may be highlighted in Table 2. Firstly, the notable difference between the meteorology of Burgos and of the other three locations, as well as the similar Mediterranean conditions in Almeria and Palma de Mallorca. This can be seen in the following tendency: the samples belonging to Burgos tend to remain together (especially when applying 'Seuclidean' and 'Cityblock' distances), while the subdivision of samples in different clusters is more usual for the locations at Almeria and Palma de Mallorca. This is clearly shown in the 'Cosine' and 'Correlation' distance measures. In all cases, the highest percentage of samples from Mallorca and Almeria are included in the same clusters. Samples from Santiago de Compostela are split into more than one cluster, but only when k equals 2 and distance is 'Cosine' are these samples located in the same cluster as the samples from Burgos.

Although the recommended value for the k parameter is 2 and 3 (see Table 2), further experiments were conducted to see whether k-means is able to cluster data from Palma de Mallorca and Almeria in different clusters. Table 3 shows the clustering results obtained by increasing the value of k up to 6. These results are worth checking to see whether higher values of k lead to sample redistribution in the new clusters.

By raising the k parameter to 6, it can be seen how samples from Burgos all remain together allocated in the same cluster when applying 'Seuclidean' distance, and in 3 out of 4 cases when applying 'Cityblock'. Regarding the samples from Santiago de Compostela, in most cases they tend to gather in single cluster (different clusters than those for Burgos, Palma de Mallorca and Almeria). The samples from Almeria and Palma de Mallorca are distributed into the new clusters, but are mainly mixed up in the same cluster. No one cluster only gathers data from one of these stations, although there are empty clusters. At k values higher than 4, the samples from the two Mediterranean locations are still found together, which means that a value for the k parameter equal to 3 would be sufficient to obtain the best sample allocation

It is also worth mentioning the great influence of the distance criterion applied. While 'Cosine' and 'Correlation' distances usually split samples from the same location in different clusters, 'Seuclidean' and

Table 3
Additional k-means clustering results ($k = 4$, 5 and 6)

K	Distance	Time (s)	SumD	Cluster Samples Allocation (%)			
				BU	SA	AL	PM
4	Seuclidean	0.17	[3.06E−05 1.92E−05 2.53E−05 1.08E−05]	[100 0 0 0]	[0 0 100 0]	[0 64 0 36]	[0 57 0 43]
4	Cityblock	0.17	[0.53 0.43 0.47 0.27]	[100 0 0 0]	[0 0 100 0]	[0 52 0 48]	[0 55 1 44]
4	Cosine	0.24	[0.035 0.041 0.030 0.027]	[29 19 40 12]	[28 43 23 6]	[30 33 1 36]	[23 37 3 37]
4	Correlation	0.18	[0.033 0.026 0.023 0.031]	[38 6 16 40]	[32 14 34 20]	[22 41 29 8]	[22 53 21 4]
5	Seuclidean	0.26	[7.50E−05 3.06E−05 2.47E−05 9.79E−06 7.87E−37]	[0 100 0 0 0]	[0 0 100 0 0]	[67 0 0 33 0]	[71 0 0 29 0]
5	Cityblock	0.21	[0.23 0.53 0.20 0.48 0.18]	[0 100 0 0 0]	[0 0 0 100 0]	[44 0 29 0 27]	[42 0 34 0 24]
5	Cosine	0.29	[0.034 0.041 0 0.080 0.027]	[28 19 0 40 13	[28 43 0 23 6]	[30 33 0 1 36]	[23 37 0 3 37]
5	Correlation	0.30	[0.05 0 0 0.041 0.046]	[48 0 0 9 43]	[36 0 0 23 41]	[19 0 0 49 32]	[11 0 0 59 30]
6	Seuclidean	0.24	[2.70E−06 3.06E−05 2.44E−05 6.06E−06 4.42E−06 4.80E−06]	[0 100 0 0 0 0]	[0 0 100 0 0 0]	[15 0 0 40 27 18]	[18 0 0 39 28 15]
6	Cityblock	0.23	[0.069 0.90 0.091 0.096 0.47 0.072]	[22 0 33 26 0 19]	[0 0 0 0 100 0]	[0 99 0 0 1 0]	[0 99 0 0 1 0]
6	Cosine	0.26	[0.018 0.019 0.020 0.013 0.021 0.014]	[5 34 16 1 14 30]	[22 17 18 3 28 12]	[28 1 16 32 19 4]	[29 2 7 33 28 1]
6	Correlation	0.26	[0.018 0.009 0.017 0.013 0.016 0.013]	[18 25 30 0 19 8]	[25 15 9 4 25 22]	[26 1 2 33 14 24]	[25 0 1 46 9 19]

Table 4
SOM k-means clustering results.

K	Type	Err	Time (s)	Cluster Samples Allocation (%)			
				BU	SA	AL	PM
2	Seq	0.00037	6.64	[0 100]	[96 4]	[100 0]	[100 0]
2	Batch	0.00037	0.20	[100 0]	[4 96]	[0 100]	[0 100]
3	Seq	9.99E−05	7	[100 0 0]	[0 0 100]	[0 100 0]	[0 100 0]
3	Batch	9.99E−05	0.21	[0 0 100]	[100 0 0]	[0 100 0]	[0 100 0]

'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 [13]. Finally, regarding the computing time needed to run the k-means algorithms, it might be said that 'Seuclidean' and 'Cityblock' provide the shortest response time when k is greater than 2. The 'SumD' parameter has the lowest values for 'Seuclidean' distance; the clusters produced in this case are therefore more compact than those obtained by applying the other three distance measures.

In Table 4, the results obtained by SOM k-means are shown. In this table, 'Type' is the type of algorithm applied in the SOM neuron (sequential or batch) training process. In the traditional sequential training, samples are presented to the map one at a time, and the algorithm gradually moves the weight vector towards them. The batch algorithm is an online algorithm that aims to find a deterministic iterative procedure for the computation of points. Its speed is limited by its use of only the diagonal part of the Hessian matrix rather than the full matrix. In the batch training, the dataset is presented to the SOM as a whole, and the new weight vectors are weighted averages of the data vectors [31]. Additionally, 'Err' shows the total quantization error for the mapping, according to the distance from any given data point to a cluster center weighted by the membership grade of that data point.

One of the first conclusions that can be drawn from Table 4 is that SOM k-means is slower than k-means in both cases but especially for 'Seq' type. Regarding the cluster sample allocation, SOM k-means offers similar results to k-means (Table 2) when applying 'Seuclidean' distance; because SOM k-means also uses 'Euclidean' distance.

In Table 5 the results obtained by means of k-medoids to the original data set are shown.

By applying k-medoids, the cluster sample allocation is similar to the one obtained by k-means (Table 2), both 'Cosine' and 'Correlation' split the samples into more than one cluster, even for the samples of Burgos

Table 5
k-medoids clustering results.

K	Distance	Time (s)	SumD	Cluster Samples Allocation (%)			
				BU	SA	AL	PM
2	Euclidean	1.98	[1.51 0.52]	[100 0]	[90 100]	[0 100]	[0 100]
2	Seuclidean	0.31	[0.0003 4.67E−05]	[0 100]	[96 4]	[100 0]	[100 0]
2	Cosine	0.29	[0.074 0.12]	[33 67]	[29 71]	[59 41]	[57 43]
2	Correlation	0.68	[0.09 0.089]	[69 31]	[71 29]	[32 68]	[21 79]
3	Euclidean	0.31	[0.23 0.26 0.44]	[0 100 0]	[100 0 0]	[0 0 100]	[0 0 100]
3	Seuclidean	0.30	[3.13E−05 4.47E−05 2.55E−05]	[100 0 0]	[0 0 100]	[0 100 0]	[0 100 0]
3	Cosine	0.39	[0.052 0.040 0.062]	[26 44 30]	[18 31 51]	[49 3 48]	[47 6 47]
3	Correlation	0.39	[0.055 0.055 0.031]	[42 51 7]	[48 36 16]	[40 18 42]	[38 11 51]

Table 6
Cluster based on Gaussian Mixture Models.

K	Covariance	Nlogl	Time (s)	Cluster Samples Allocation (%)			
				BU	SA	AL	PM
2	Full	−5.38E+05	0.30	[40 60]	[39 61]	[49 51]	[61 39]
2	Diagonal	−5.37E+05	0.30	[100 0]	[4 96]	[0 100]	[0 100]
3	Full	−5.44E+05	0.61	[51 49 0]	[41 59 0]	[1 8 91]	[1 12 87]
3	Diagonal	−5.43E+05	0.31	[51 49 0]	[41 59 0]	[1 8 91]	[1 12 87]

Table 7
Agglomerative hierarchical clustering results.

K	Distance	Time (s)	Cluster Samples Allocation (%)			
			BU	SA	AL	PM
2	Minkowsky	1.28E+03	[100 0]	[0 100]	[0 100]	[0 100]
2	Cityblock	1.28E+03	[100 0]	[0 100]	[0 100]	[0 100]
2	Euclidean	1.34E+03	[100 0]	[0 100]	[0 100]	[0 100]
2	Seuclidean	1.32E+03	[0 100]	[0 100]	[0 100]	[0 100]
3	Minkowsky	1.27E+03	[0 0 100]	[0 100 0]	[0 100 0]	[0 100 0]
3	Cityblock	1.27E+03	[0 0 100]	[0 100 0]	[0 100 0]	[0 100 0]
3	Euclidean	1.27E+03	[0 0 100]	[0 100 0]	[0 100 0]	[0 100 0]
3	Seuclidean	1.27E+03	[0 100 0]	[0 100 0]	[0 100 0]	[0 100 0]

for both values of k. The samples from Almeria and Palma de Mallorca remain together in the same cluster in most cases. The 'SumD' parameter gets the best value for 'Seuclidean' distance, which means more compact clusters in these cases.

Table 6 shows the results obtained by applying the GMM with the EM algorithm. Means of the 'Covariance' parameter: diagonal if the covariance matrices are restricted to the 'Diagonal', and otherwise 'Full'. The 'Nlogl' parameter expresses the negative log-likelihood of the data.

When applying clustering based on GMM with the EM algorithm, the execution time was greater when the covariance was 'Full' than when the covariance was 'Diagonal' and when parameter k equaled 3, but the difference is not very relevant. Regarding the sample process allocation, the samples belonging to the location of Burgos were only grouped in one cluster when the covariance was diagonal and k equaled 2. In the other three cases, the samples from Burgos were distributed in more than one data cluster; quite a different result from those obtained in the above experiments (Tables 2 to 5). The samples from Almeria and Palma de Mallorca were assigned to different cluster in most cases. The samples from Santiago de Compostela were assigned to different clusters in all cases. The 'Nlogl' parameter gave similar results in all experiments.

Table 7 shows the very different results of applying the agglomerative hierarchical clustering technique to the original dataset in comparison with those obtained by partitional methods (Tables 1–6).

The main difference between agglomerative hierarchical clustering and the three previous methods is that the former allocates the samples to clusters according to the location of the stations, with an accuracy of

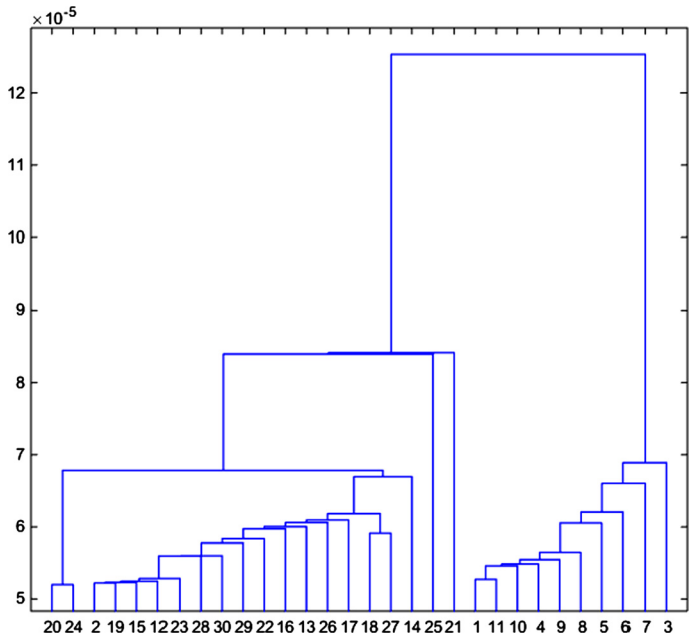

Fig. 2. Dendrogram with 30 leaf nodes ('Euclidean' distance criterion and 'Average' linkage method).

100%, in all cases. This result means that data from the same location are always allocated to the same cluster. In all cases, all the samples from Almeria Palma de Mallorca, and Santiago de Compostela are assigned to the same cluster. In one case, the samples from the four locations are assigned to the same cluster, which indicates the faulty performance of this technique in this case study. These results are quite different from those shown in Tables 2 to 6, where the partitional techniques were able to differentiate the samples from Santiago de Compostela from the samples from Almeria and Palma de Mallorca. These results are not consistent with the description of the case study as the samples from Burgos at least should be assigned to a different cluster than the samples from the other three locations, as is evident from the previous study applying Principal Component Analysis (PCA). Observing the dendrograms generated in Fig. 2 and Fig. 3, which are a visual complement of the agglomerative method, it can be seen that the samples from Burgos are clearly distinguishable alongside the samples from the other three locations. Another drawback is that this technique is highly demanding in terms of computing time, regardless of the number of selected clusters or the distance metric applied. The technique requires so much computer time, because it starts with individual samples and it generates groups from among them, which is not appropriate when the number of samples is so as high as in this case study.

As complementary information, Fig. 2 shows the dendrogram for agglomerative clustering with 'Euclidean' distance criterion and 'Average' linkage method for computing the distance between clusters. Number of leaf nodes: 30. This value is high enough to understand the subdivision process performed and to see the dendrogram output clearly in graphical form.

The samples are distributed in leafs as follows. Samples from Burgos are all grouped in leaf 1. Most samples from Almeria are grouped in leaf 2 with a few samples in leafs 12, 13, 14, 15 and 16. Most samples from Santiago de Compostela are grouped in leaf 2 with a few samples in leafs 18, 19, 20, 21, 22, 23, 24, 25, 26 and 27. Samples from Palma de Mallorca are located in leaf 2 and a few samples in leafs 15, 28, 29 and 30. As shown in Tables 2 to 7, the samples from Burgos were grouped together and were separated from the samples of the other three places. The fact that most of the samples were grouped in only two leafs (1 and 2) and that many of the leafs are empty may also be highlighted. These results underline the conclusions that may be drawn from the results shown in Table 7. On the one hand, the high levels of compaction in samples from Burgos are noteworthy, all of which are grouped on one leaf. On the other hand, the clear

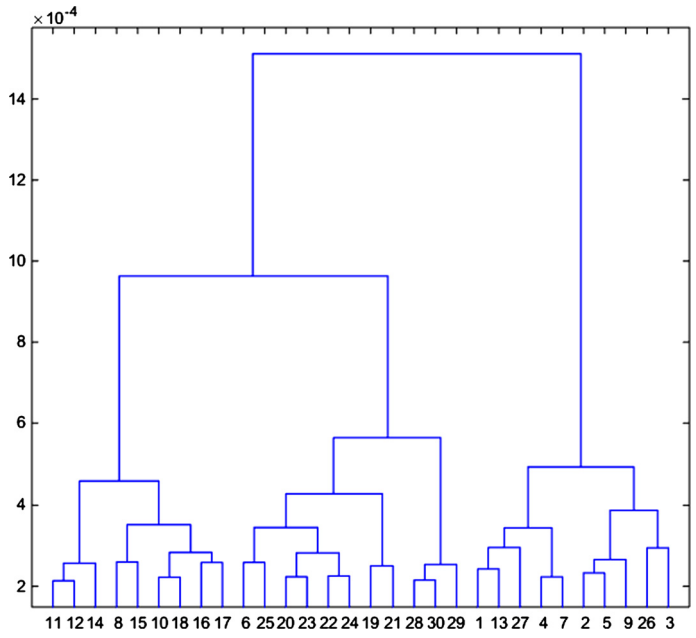

Fig. 3. Dendrogram with 30 leaf nodes ('Euclidean' distance criterion and 'Complete' linkage method).

differentiation of climatic conditions proper to Burgos from those of the other three locations may be seen, as in the initial division, the samples from Burgos were separated from the samples of the other three locations.

Fig. 3 shows the dendrogram (30 leaf nodes) for agglomerative clustering with the 'Euclidean' distance criterion and the 'Complete' linkage method for computing the distance between clusters.

In the dendrogram shown in Fig. 3, samples from Burgos are grouped in leafs 1, 2, 3, 4, 5, 7, 9, 13, 15, 25 and 26, while samples from Almeria are grouped in leafs 8, 10, 11, 12, 14, 15, 16, 17 and 18. Samples from Santiago de Compostela are grouped in leafs 6, 10, 19, 30, 28, 25, 24, 23, 22, 21, 20 and 19 while samples from Palma de Mallorca are grouped in leafs 8, 10, 11, 12, 14, 16, 17, 18 and 19. All the samples from Burgos are located in the large cluster to the right of the dendrogram, although the samples are split into more leaves than in the dendrogram with the 'Average' linkage method (Fig. 2). Samples from the other three locations are also distributed in more leafs than in the previous result (Fig. 2). In the dendrogram shown in Fig. 3, there are no empty leafs. In this type of dendrogram that applies the 'Complete' criterion linkage, having all the leafs with any samples may represent an easier way than in Fig. 2 of visualizing the formation of large data groups with the naked eye.

5. Conclusions and future work

The main conclusions derived from the previously explained results (see Section 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 performance of the different clustering techniques, criteria, and measures applied to the case study.

Climatological conditions are not analyzed in the present work for the following reasons: the time period under study is not long enough to consider the climates of the four locations in the study. Hence, conclusions regarding a change in the weather over the period under analysis cannot be drawn and both the available information and the results lend no support to long-term forecasting of climate conditions. Regarding the meteorological conditions at the four selected locations over the time period of the study, the notable difference between the meteorology in Burgos and at the other three locations may be highlighted. A different

meteorology from the other three sites may also be appreciated at Santiago de Compostela, but not as pronounced as in the case of Burgos. However, the similarity of the mean daily meteorological data from Palma de Mallorca and Almeria are very similar and none of the methods were capable of splitting those samples into different clusters. The samples from the two places with different Mediterranean climates (Almeria and Palma de Mallorca) tended to remain together in the same clusters.

The appropriateness of applying the cluster evaluation measures may be highlighted as a first step, in relation to the behavior of the clustering techniques. The results of the four main measures raises the question of why three of them suggest the same value for k; important for the selection of the k parameter considered in subsequent experiments. In a general comparison of the clustering techniques, with the selected distance criterion as a key factor, k-means, k-medoids and SOM k-means attain similar results. Moreover, it may be concluded that k-means is the best technique in terms of computational load for the data under analysis. Analyzing the distance measures applied, 'Euclidean' distances are usually the most reliable, while 'Cosine' and 'Correlation' distances have a tendency to split the samples from the same location into more than one cluster and not always in the most reliable way. GMM generates results that are similar to those obtained by the three previously mentioned techniques, although in some cases some inconsistent results have been obtained, handing out samples in different clusters that should be together. The different results of the hierarchical agglomerative technique should also be emphasized when compared with the partitional clustering techniques. In many cases, the agglomerative hierarchical clustering technique showed no reliable response, and failed to allocate samples from different locations in different clusters. Unlike the PCA technique used in the previous study, no single technique was able to sort the samples from the four locations into separate clusters.

The clustering techniques analyzed in this case study are useful to validate the meteorology of the four selected locations, corresponding to the four different climatic zones in Spain. By applying these techniques to the case study, it is easier to know which samples correspond to which climatic zone, which is much more difficult than applying dimensionality reduction. Another advantage is the possibility of analyzing the difference between the four climates and of inspecting the different measures returned by the techniques. The grade of compactness of each cluster may also be inspected. All of these finding indicate the variability of each climate and the differences between zones.

As future work, a hybrid intelligent system combining dimensionality reduction and clustering techniques is proposed, to demonstrate the complementarity of these two paradigms for the analysis of high-dimensionality climate data sets from various regions in the Iberian peninsula.

References

[1] J.K. Anil, Data clustering: 50 years beyond K-means, Pattern Recognit. Lett. 31 (2010) 651–666.
[2] K. Aparna, M.K. Nair, Comprehensive study and analysis of partitional data clustering techniques, IJBAN 2 (2015) 23–38. The Behavior Change Technique Taxonomy (v1) of 93 Hierarchically Clust. (2015).
[3] Á. Arroyo, V. Tricio, E. Corchado, Á. Herrero, A Comparison of Clustering Techniques for Meteorological Analysis, Advances in Intelligent Systems and Computing, vol. 368, Springer, 2015, 117130.
[4] H. Barlow, Unsupervised learning, Neural Comput. 1 (1989) 295–311.
[5] T. Caliński, J. Harabasz, A dendrite method for cluster analysis, Commun. Stat., Theory Methods 3 (1974) 1–27.
[6] D.L. Davies, D.W. Bouldin, A cluster separation measure, IEEE Trans. Pattern Anal. Mach. Intell. (1979) 224–227.
[7] W.H.E. Day, H. Edelsbrunner, Efficient algorithms for agglomerative hierarchical clustering methods, J. Classif. 1 (1) (2015) 7–24.
[8] M. de Castro, J. Martín-Vilde, S. Alonso, The climate of Spain: past, present and scenarios for the 21st century, in: A Preliminary General Assessment of the Impacts in Spain due to the Effects of Climate Change, vol. 162, Spanish Ministry of Environment, Madrid, 2005, p. 162.
[9] C. Ding, X. He, K-means clustering via principal component analysis, in: Proceedings of the Twenty-First International Conference on Machine Learning, 2004, p. 29.
[10] C.B. Do, S. Batzoglou, What is the expectation maximization algorithm?, Nat. Biotechnol. 26 (2008) 897–900.
[11] I. Hokenko, On clustering of non-stationary meteorological time series, Dyn. Atmos. Ocean. 49 (2) (2010) 164–187.
[12] H. Hotelling, Analysis of a complex of statistical variables into principal components, J. Educ. Psychol. 24 (1933) 498–520.
[13] A.K. Jain, S. Maheswari, Survey of recent clustering techniques in data mining, J. Curr. Comput. Sci. Technol. 3 (2013).
[14] A.K. Jain, M.N. Murty, P.J. Flynn, Data clustering: a review, CSUR 31 (3) (1999).

[15] P. Kassomenos, S. Vardoulakis, R. Borge, J. Lumbreras, C. Papaloukas, S. Karakitsios, Comparison of statistical clustering techniques for the classification of modelled atmospheric trajectories, Theor. Appl. Climatol. 102 (2010) 1–12.
[16] T. Kohonen, The self-organizing map, Proc. IEEE 78 (1990) 1464–1480.
[17] Y. Liu, Z. Li, H. Xiong, X. Gao, J. Wu, Understanding of Internal Clustering Validation Measures, IEEE, 2010.
[18] C.D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge University Press, ISBN 0521865719, 2008.
[19] MATLAB, MathWorks, http://mathworks.com/products/matlab/, 2016 (accessed 25.02.2016).
[20] Spanish Meteorology Agency – AEMET, Government of Spain, 2016, http://www.aemet.es/, 2016 (accessed 25.02.2016).
[21] S. Michie, M. Richardson, M. Johnston, C. Abraham, J. Francis, W. Hardeman, M.P. Eccles, J. Cane, C.E. Wood, 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) (2013).
[22] D. Napoleon, S. Pavalakodi, A new method for dimensionality reduction using k-means clustering algorithm for high dimensional data set, Int. J. Comput. Appl. 13 (2011) 41–46.
[23] National Network of meteorological stations, Spanish Agency of Meteorology, http://www.aemet.es/es/eltiempo/observacion/ultimosdatos, 2016 (accessed 25.02.2016).
[24] H.S. Park, C.H. Jun, A simple and fast algorithm for K-medoids clustering, Expert Syst. Appl. 36 (2) (2009) 3336–3341.
[25] I.A. Pérez, M.L. Sanchez, M.A. Garcia, N. Pardo, Analysis of air mass trajectories in the northern plateau of the Iberian Peninsula, J. Atmos. Sol.-Terr. Phys. 134 (2015) 9–21.
[26] J.C.M. Pires, S.I.V. Sousa, M.C. Pereira, M.C.M. Alvim-Ferraz, F.G. Martins, Management of air quality monitoring using principal component and cluster analysis—Part I: SO 2 and PM 10, Atmos. Environ. 42 (2008) 1249–1260.
[27] D. Reynolds, Gaussian mixture models, in: Encyclopedia of Biometrics, Springer, 2009, pp. 659–663.
[28] P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65.
[29] W. Tian, Y. Zheng, R. Yang, S. Ji, J. Wang, Research on clustering based meteorological data mining methods, Adv. Sci. Technol. Lett. IST 79 (2014) 106–112.
[30] R. Tibshirani, G. Walther, T. Hastie, Estimating the number of clusters in a data set via the gap statistic, J. R. Stat. Soc., Ser. B, Stat. Methodol. 63 (2001) 411–423.
[31] J. Vesanto, E. Alhoniemi, J. Parhankangas, Self-organizing map in Matlab: the SOM toolbox, in: Proceedings of the Matlab DSP Conference, 2000, pp. 35–40.
[32] J. Zscheischle, M.D. Mahecha, S. Harmeling, Climate classifications: the value of unsupervised clustering, ICCS 9 (2012) 897–906.