RESEARCH ARTICLE

CLUSTERING AN INTERVAL DATA SET – ARE THE MAIN PARTITIONS SIMILAR TO A PRIORI PARTITION?

Áurea Sousa, Helena Bacelar-Nicolau, Fernando C. Nicolau and Osvaldo Silva
1Department of Mathematics and CEEplA, University of Azores, Ponta Delgada, Portugal
2Faculty of Psychology and ISAMB, University of Lisbon, Lisbon, Portugal
3Department of Mathematics, FCT, New University of Lisbon, Caparica, Portugal
4Department of Mathematics and CICS.NOVA/UAc, University of Azores, Ponta Delgada, Portugal

ABSTRACT

In this paper we compare the best partitions of data units (cities) obtained from different algorithms of Ascendant Hierarchical Cluster Analysis (AHCA) of a well-known data set of the literature on symbolic data analysis (“city temperature interval data set”) with a priori partition of cities given by a panel of human observers. The AHCA was based on the weighted generalised affinity, $a(k, k')$, with equal weights, and on the probabilistic coefficient, $a_{GW}(k, k')$, associated with the asymptotic standardized weighted generalised affinity coefficient by the method of Wald and Wolfowitz. These similarity coefficients between elements were combined with three aggregation criteria, one classical, Single Linkage (SL), and the other ones probabilistic, $AV1$ and $AV2$, the last ones in the scope of the VL methodology. The evaluation of the partitions in order to find the partitioning that best fits the underlying data was carried out using some validation measures based on the similarity matrices. In general, global satisfactory results have been obtained using our methods, being the best partitions quite close (or even coinciding) with the a priori partition provided by the panel of human observers.

INTRODUCTION

With the computational progress and the increased use of large data sets that on occasion require to be aggregated into smaller more manageable data sizes we need more complex data tables sometimes called “symbolic data tables”, where rows correspond to data units (frequently, groups of individuals, considered as second-order objects) and columns to variables. In a table of this nature, each entry can contain just one value or several values, such as subsets of categories, intervals of the real data set, or frequency distributions (Bock and Diday, 2000; Bacelar-Nicolau, 2000, 2002; Diday and Noirhomme-Fraiture, 2008; Bacelar-Nicolau et al., 2009, 2010, 2014a, 2014b; Sousa et al., 2010, 2013a, 2013b, 2014; 2015). In particular, if each cell of a symbolic data table contains an interval we deal with interval variables. In fact, the interval-valued data arise in several situations such as recording monthly interval temperatures at meteorological stations (for instance, considering a town $w$, temperature ($w$) $= [6, 12]$ in January means that the temperature of the town $w$ varied in the interval $[6, 12]$ during the month of January), daily interval stock prices, etc. (De Carvalho et al., 2012). The symbolic data tables can also describe heterogeneous data and the values in their cells may be weighted and connected by logical rules and taxonomies (Bock and Diday, 2000). An extension of standard data analysis methods (exploratory, graphical representations, cluster analysis, factorial analysis,…) to symbolic data tables is required.

Cluster analysis frequently appears in the literature under different names in different contexts, such as for example unsupervised learning in pattern recognition, and taxonomy in biological sciences. The clustering aims at identifying and extract significant groups of elements to classify in the underlying data so that (based on a certain clustering criterion) the elements in a cluster are more similar to each other than the elements in different clusters. Different types of algorithms to cluster analysis (e.g., partitional clustering, hierarchical clustering, density-based clustering, grid-based clustering) have been developed (Jain et al., 1999; Lattin et al., 2003). Hierarchical clustering proceeds successively by either merging smaller clusters into larger ones (agglomerative methods), or by splitting larger clusters (divisive methods) (Halkidi et al.,...
Agglomerative methods usually start with each element to be classified in its own separate cluster. At each stage of the process, the most similar clusters (according to the selected aggregation criteria) are joined until only one cluster containing all elements remains. Given a set of $n$ elements to classify, the divisive methods start with all elements in a single cluster and proceed dividing one cluster into two at each step until $n$ clusters of size 1 remain. As is referred in Lattin et al. (2003) “Some methods are neither agglomerative nor divisive (e.g., various approaches that use least squares to fit certain tree structures)”. This paper is focused on agglomerative methods in the context of cluster analysis and on a symbolic data table where each cell contains an interval of the real axis. Some dissimilarity measures for interval data have been reported in the literature (e.g., Chavent and Lechevallier, 2002; Chavent et al., 2003; Souza and De Carvalho, 2004; De Carvalho et al., 2006a, 2006b, 2007), as well as some similarity measures which allow us to deal with this type of data (e.g., Guru et al., 2004; Bacelar-Nicolau et al., 2009, 2010, 2014a, 2014b; Sousa et al., 2010, 2013a, 2015).

A hierarchical algorithm allows us to obtain a tree of clusters, called dendrogram, which shows how the clusters are related. By cutting the dendrogram at an appropriate level, a partition of the elements to classify into disjoint groups is obtained. An important question is related to the number of clusters (How many clusters?). In fact, the evaluation of clustering results in order to find the partitioning that best fits the underlying data plays a very important role in cluster analysis (Halkidi et al., 2001).

Section 2 is devoted to the methods used to carry out the AHCa of cities, and to the measures of validation used to evaluate the obtained partitions. Section 3 is concerned to the main results obtained from the AHCa of the city temperature interval data set (issued from the symbolic data literature), and their comparison with apriori partition of cities given by a panel of human observers. The paper ends with some concluding remarks about the developed work.

**Methodological framework**

From the affinity coefficient between two discrete probability distributions proposed by Matusita (1951) as a similarity measure for comparing two distribution laws of the same type, Bacelar-Nicolau (e.g., 1980, 1988) introduced the affinity coefficient, as a similarity coefficient between pairs of variables or of subjects in cluster analysis context. After that, Bacelar-Nicolau extended that coefficient to different types of data, and the so-called weighted generalized affinity coefficient, $a(k, k')$, between a pair of statistical data units, $k$ and $k'$ ($k, k' = 1, \ldots, N$), is an extension of the affinity coefficient for the case of symbolic or complex data, which is able to deal with heterogeneous data (Bacelar-Nicolau, 2000, 2002; Bacelar-Nicolau et al., 2009, 2010, 2014a, 2014b).

Let $E = \{1, \ldots, N\}$ be a set of $N$ data units described by $p$ interval variables, $Y_1, \ldots, Y_p$, which values are intervals of the real data set (f. i., the entry ($k,j$), corresponding to the data unit $k$ ($k=1, \ldots, N$) and to the variable $Y_j$ ($j=1, \ldots, p$) of the data table, contains an interval $I_{kj} = [a_{kj}, b_{kj}]$). In this case, the weighted generalized affinity coefficient, between a pair of statistical data units, $k$ and $k'$ ($k, k' = 1, \ldots, N$), is defined in the following way:

$$a(k, k') = \sum_{j=1}^{p} \pi_j \frac{|I_{kj} \cap I_{kj'}|}{\sqrt{|I_{kj}| |I_{kj'}|}},$$

where $\pi_j$ are weights such that $0 \leq \pi_j \leq 1$, $\sum \pi_j = 1$, and corresponds to a generalized Ochiai coefficient for interval data, associated with a $2 \times 2$ generalized contingency table which entries contain interval ranges instead of the usual cardinal numbers of any simple $2 \times 2$ contingency table (Bacelar-Nicolau et al., 2009, 2010, 2014b; Sousa et al., 2015). The formula (1) is a particular case of the general formula of the weighted generalized affinity coefficient when we are dealing with variables of interval type. In fact, the weighted generalized affinity coefficient between a pair of intervals may be computed in two different ways, either by using the general formula of the weighted generalized affinity coefficient considering the decomposition of the initial intervals into $m_j$ elementary and disjoint intervals and working with the ranges of the elementary intervals; or, alternatively, by using the formula (1), without the decomposition of the initial intervals (for details, see Bacelar-Nicolau et al., 2009, 2010, 2014b).

Assuming a permutational reference hypothesis based on the limit theorem of Wald and Wolfowitz (Fraser, 1975), the random variable associated with $a(I_{kj}, I_{kj'})$ has an asymptotic normal distribution. Two of the coefficients related with the $a(k, k')$ coefficient, are the asymptotic standardized weighted generalized affinity coefficient, $a_{ww}(k, k')$, by the Wald and Wolfowitz method (see Bacelar-Nicolau, 1988; Bacelar-Nicolau et al., 2009, 2010, 2014a; Sousa et al., 2013a, 2015), and the associated probabilistic coefficient, $a_{ww}(k, k')$, in the scope of the $VL$ methodology ($V$ for Validity, $L$ for Linkage), in the line started by Lerman (1970, 1972, 1981) and developed by Bacelar-Nicolau (e.g., 1980, 1985, 1987, 1988) and Nicolau (e.g., 1983, 1998). This last coefficient validates the affinity coefficient between two data units $k$, $k'$ in a probabilistic scale (e.g., Bacelar-Nicolau, 1988, 2000; Bacelar-Nicolau et al., 2010; Lerman, 1972, 1981; Nicolau and Bacelar-Nicolau, 1998). The (hierarchical and non-hierarchical) clustering methods included in the $VL$-family are based on the cumulative distribution function of basic similarity coefficients (Bacelar-Nicolau, 1980, 1988; Nicolau, 1983; Nicolau and Bacelar-Nicolau, 1998).

Here, we used the weighted generalized affinity coefficient, $a(k, k')$, with equal weights ($\pi_j = 1/p$), and the probabilistic coefficient, $a_{ww}(k, k')$, associated with the asymptotic standardized weighted generalized affinity coefficient by the method of Wald and Wolfowitz. In the case of the data set under analysis (“city temperature interval data set”) the best clustering results were provided by the probabilistic coefficient, $a_{ww}(k, k')$, as a consequence of standardizing the affinity values, and of using the corresponding probabilistic scale values. Thus, in the next section, a special emphasis will be given to the results provided.
by the $\alpha_{WW}(k, k')$ coefficient. A brief reference to the results obtained from the $a(k, k')$ coefficient will be added. In order to compute these coefficients (similarity measures), we consider a previous decomposition of each interval of the original symbolic data table into a suitable number $m$ of elementary and disjoint intervals, $\{I_i; l = 1, \ldots, m\}$ (Bacelar-Nicolau et al., 2009, 2010, 2014b). In this paper, the measures of comparison between elements were combined with three aggregation criteria, one classical, Single Linkage (SL) or nearest neighbour method, and two probabilistic, AV1 and AVB, the last ones in the scope of the VL methodology, that use probabilistic notions for the definition of the comparative functions (e.g., Lerman, 1972, 1981, 2000; Nicolau, 1983; Bacelar-Nicolau, 1998; Nicolau and Bacelar-Nicolau, 1998).

An important step in cluster analysis is to determine the best number of clusters. In an optimal clustering scheme, the elements of each cluster should be as close to each other elements belonging to their cluster as possible (compactness), and the clusters should be widely spaced (isolation or separation). Therefore, is useful to use, for a cluster of elements, measures of its heterogeneity or lack of cohesion, and of its isolation or separation, from the rest of the data. These measures can be combined to provide measures of the adequacy of the partitions (Gordon, 1999). In fact, a general approach to finding the best partition involves defining a measure of the adequacy of a partition and seeking a partition of elements which optimizes that measure (Gordon, 1999).

Some measures of the heterogeneity of a cluster are defined in the literature (e.g., Gordon, 1999). Hennig (2005) refers to some different approaches that address different aspects of the validation problem, namely, use of external information (information that has not been used to generate the clustering), significance tests for clustering structure, comparison of different clustering structures on the same data set, validation indexes, stability assessment, and visual inspection. A global approach for evaluating the quality of clustering results provided from different clustering algorithms using the relevant information available (e.g., the stability, isolation and homogeneity of the clusters) was presented in Silva et al. (2012).

The city temperature interval data set was given to a panel of human observers for classification. The a priori partition of the cities given by the observers contains four clusters (Guru et al., 2004), which descriptions and corresponding latitudes are shown in Table 2:

Cluster 1: \{C2, C3, C4, C5, C6, C8, C11, C12, C15, C17, C19, C22, C23, C29, C31\};
Cluster 2: \{C0, C1, C7, C9, C10, C13, C14, C16, C20, C21, C24, C25, C26, C27, C28, C30, C33, C34, C35, C36\};
Cluster 3: \{C18\};
Cluster 4: \{C32\}.

We used the methodological framework, described in Sousa et al. (2014), in order to evaluate the obtained partitions according to measures of validation (adapted for the case of similarity measures) based on the values of the proximity matrix between elements (internal validation measures). Thus, the values of the statistics of levels $\text{STAT}$ and $\text{DIF}$ (Bacelar-Nicolau, 1980; Lerman, 1970, 1981), $P(2\text{mod}, \Sigma)$, and $\gamma$ (Goodman and Kruskal (1954)) indexes (for the partitions into three, four, and five clusters), were calculated. The “best” cluster is one that presents the largest values of $\text{STAT}$, $\text{DIF}$, and $\gamma$, and the smallest value of the $P(2\text{mod}, \Sigma)$. Furthermore, the values of the Sil* index based on the Silhouette plots (Rousseeuw, 1987) and of the U statistics (Mann and Whitney, 1947), namely the global U index ($U_0$) and the local U index ($U_i$) were calculated for the clusters of the most significant partition (according to the previous indexes), as described in Sousa et al. (2014), and for the a priori partition. The formulae of the $\text{STAT}$, $\text{DIF}$, $P(2\text{mod}, \Sigma)$, and $\gamma$ indexes, the last two ones adapted for the case of similarity measures, can be found in Sousa et al. (2013b). In the case of a cluster-L* we have $U_i=0$ and in the case of a ball cluster we have $U_i=0$ (Gordon, 1999). The best partition is compared with the a priori partition (external information) into four clusters given by a panel of human observers.

### AHCA of the city temperature interval data set

In this example, we consider the data set given in Guru et al. (2004) concerned to the minimum and the maximum monthly temperatures of 37 cities in degrees centigrade (city temperature interval data set) during a determined year. Table 1 shows a part of this data matrix.

<table>
<thead>
<tr>
<th>Pattern no.</th>
<th>Cities</th>
<th>Jan</th>
<th>Feb</th>
<th>Mar</th>
<th>Apr</th>
<th>May</th>
<th>Jun</th>
<th>Jul</th>
<th>Aug</th>
<th>Sep</th>
<th>Oct</th>
<th>Nov</th>
<th>Dec</th>
</tr>
</thead>
<tbody>
<tr>
<td>C0</td>
<td>Amsterdam</td>
<td>[4, 4]</td>
<td>[5, 3]</td>
<td>[2, 12]</td>
<td>\cdots</td>
<td>[5, 15]</td>
<td>[1, 10]</td>
<td>[-1, 4]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C1</td>
<td>Athens</td>
<td>[6, 12]</td>
<td>[6, 12]</td>
<td>[8, 16]</td>
<td>\cdots</td>
<td>[16, 23]</td>
<td>[11, 18]</td>
<td>[8, 14]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C2</td>
<td>Bahrain</td>
<td>[13, 19]</td>
<td>[14, 19]</td>
<td>[17, 23]</td>
<td>\cdots</td>
<td>[24, 31]</td>
<td>[20, 26]</td>
<td>[15, 21]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C3</td>
<td>Bombay</td>
<td>[19, 28]</td>
<td>[19, 28]</td>
<td>[22, 30]</td>
<td>\cdots</td>
<td>[24, 32]</td>
<td>[23, 32]</td>
<td>[20, 30]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\cdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C33</td>
<td>Tokyo</td>
<td>[0, 9]</td>
<td>[0, 10]</td>
<td>[3, 13]</td>
<td>\cdots</td>
<td>[13, 21]</td>
<td>[8, 16]</td>
<td>[2, 12]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C34</td>
<td>Toronto</td>
<td>[-8, -1]</td>
<td>[-8, -1]</td>
<td>[-4, 4]</td>
<td>\cdots</td>
<td>[6, 14]</td>
<td>[-1, 17]</td>
<td>[-5, 1]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C35</td>
<td>Vienna</td>
<td>[-2, 1]</td>
<td>[-1, 3]</td>
<td>[1, 8]</td>
<td>\cdots</td>
<td>[7, 13]</td>
<td>[2, 7]</td>
<td>[1, 3]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C36</td>
<td>Zurich</td>
<td>[-11, 9]</td>
<td>[-8, 15]</td>
<td>[-7, 18]</td>
<td>\cdots</td>
<td>[3, 22]</td>
<td>[0, 19]</td>
<td>[-11, 8]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The cities belonging to cluster 1 are located at latitudes between 0° and 40° and the cities included in the cluster 2 are mainly located at latitudes between 40° and 60°. The human observers have classified some cities (C1- Athens, C13- Lisbon, C27- San Francisco, C28- Seoul and C33- Tokyo), which are closer to the sea coast and are located at latitudes between 0 and 40, in the cluster 2, because those cities have low
temperature which is similar to that of the cities which are
located at latitudes between 40 and 60. "The cities nearer to sea
coast bear relatively low temperature because of the cool
breeze from the sea coast and also due to high humidity present
in the atmosphere" (Guru et al., 2004). Mauritius (the only
island in this data set) was included in a cluster with only one
element (singleton) and Tehran in another singleton due to its
irregular temperature. The main results of the hierarchical
cluster analysis of the 37 cities are presented in the remaining
section.

Table 2. Description of a priori partition – city temperature
interval data set

<table>
<thead>
<tr>
<th>Cluster 1</th>
<th>Latitude</th>
<th>Cluster 2</th>
<th>Latitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>C2-Bahrain</td>
<td>26°13’N</td>
<td>C0-Amsterdam</td>
<td>52°22’N</td>
</tr>
<tr>
<td>C3-Bombay</td>
<td>19°09’N</td>
<td>C1-Athens</td>
<td>37°58’N</td>
</tr>
<tr>
<td>C4-Cairo</td>
<td>30°32’N</td>
<td>C7-Copenhagen</td>
<td>55°41’N</td>
</tr>
<tr>
<td>C5-Calcutta</td>
<td>22°34’N</td>
<td>C9-Frankfurt</td>
<td>50°07’N</td>
</tr>
<tr>
<td>C6-Colombo</td>
<td>6°56’N</td>
<td>C10-Genova</td>
<td>46°12’N</td>
</tr>
<tr>
<td>C8-Dubai</td>
<td>25°15’N</td>
<td>C13-London</td>
<td>38°43’N</td>
</tr>
<tr>
<td>C11- Hong Kong</td>
<td>22°17’N</td>
<td>C14-London</td>
<td>51°30’N</td>
</tr>
<tr>
<td>C12- Kuala Lumpur</td>
<td>3°8’N</td>
<td>C16-Madrid</td>
<td>40°24’N</td>
</tr>
<tr>
<td>C15- Madras</td>
<td>13°05’N</td>
<td>C20-Monaco</td>
<td>55°45’N</td>
</tr>
<tr>
<td>C17-Manila</td>
<td>14°35’N</td>
<td>C21-Munich</td>
<td>48°08’N</td>
</tr>
<tr>
<td>C19-Mexico</td>
<td>19°26’N</td>
<td>C24-New York</td>
<td>42°54’N</td>
</tr>
<tr>
<td>C22-Nairobi</td>
<td>1°17’S</td>
<td>C25-Paris</td>
<td>48°51’N</td>
</tr>
<tr>
<td>C23-New Delhi</td>
<td>28°37’N</td>
<td>C26-Rome</td>
<td>41°54’N</td>
</tr>
<tr>
<td>C29 - Singapore</td>
<td>1°17’N</td>
<td>C27-San Francisco</td>
<td>37°47’N</td>
</tr>
<tr>
<td>C31-Sydney</td>
<td>33°52’S</td>
<td>C28-Seoul</td>
<td>37°34’N</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>C30-Stockholm</td>
<td>59°20’N</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>35°41’N</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>C33-Tokyo</td>
<td>35°41’N</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>43°42’N</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>C35-Vienna</td>
<td>48°13’N</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>47°22’N</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Cluster 3</td>
<td>Latitude</td>
<td>Cluster 4</td>
<td>Latitude</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>C18-Mauritius</td>
<td>20°10’S</td>
<td>C32-Tehran</td>
<td>53°42’N</td>
</tr>
</tbody>
</table>

Before calculating the values of the α(k, k’) and αuv(k, k’)
coefficients, the domains Duv of each variable Yv, i = 1, . . . , 12 for
the set E = {1, . . . , 37} of n = 37 objects (cities) were decomposed
into a suitable number of element and disjoint intervals.
For instance, the observed (interval-type) values of V1 (January)
are V1(e) = { [4, 4]; [6, 12]; [13, 19]; [19, 28]; [8, 20]; [13, 27];
[22, 30]; [2, 2]; [13, 23]; [10, 9]; [3, 5]; [13, 17]; [22, 31]; [8, 13];
[2, 6]; [20, 30]; [1, 9]; [21, 27]; [22, 28]; [6, 22]; [13, 6];
[6, 1]; [12, 25]; [6, 21]; [2, 4]; [1, 7]; [4, 11]; [6, 13]; [0, 7];
[3, 30]; [9, -5]; [20, 30]; [0, 5]; [0, 9]; [8, 1]; [1, 2]; [11, 9]}. Thus, the domain of variable V1 is the interval Dv1 = [13, 31]. Let u0 = 13, u1 = 11, u2 = 10, u3 = 9, u4 = 8, u5 = 6; u6 = 5;
u7 = 4, u8 = 3, u9 = 2, u10 = 1, u11 = 1, u12 = 2, u13 = 4, u14 = 5,
u15 = 6, u16 = 7, u17 = 8, u18 = 9, u19 = 10, u20 = 11, u21 = 12,
u22 = 13, u23 = 17, u24 = 19, u25 = 20, u26 = 21, u27 = 23, u28 = 25,
u29 = 27, u30 = 28, u31 = 30 be the 31 distinct values corresponding to
the lower and upper boundaries of the observed intervals of
V1(E) that are sorted in ascending order. The interval Dv1 is
decomposed into 33 element and disjoint intervals, [u0,..., u31] based on the 33 distinct values, u0, u1, . . . , u31, as follows:
[-13, -11]; [-11, -10]; [-10, -9]; [-9, -8]; [-8, -6]; [-6, -5]; [-5, -4];
[-4, -3]; [-3, -2]; [-2, 1]; [0, 1]; [1, 2]; [2, 4]; [4, 5];
[5, 6]; [6, 7]; [7, 8]; [8, 9]; [9, 11]; [11, 12]; [12, 13]; [13, 17];
[17, 19]; [19, 20]; [20, 21]; [21, 22]; [22, 23]; [23, 25]; [25, 27];
[27, 28]; [28, 30]; [30, 31] (see Table 1). For example,

Therefore, proceeding in this way for all other variables, we
obtained a new data matrix, subdivided into 12 subtables (one for
each variable), which contain a decomposition of the
respective initial intervals into elementary intervals.

Table 3. Decomposition into 33 elementary intervals – Variable Y1
(January)

<table>
<thead>
<tr>
<th>Cluster 1</th>
<th>[13, 14]</th>
<th>[11, 13]</th>
<th>[12, 14]</th>
<th>[11, -1]</th>
<th>[12, 1]</th>
<th>[1, 2]</th>
<th>[2, 4]</th>
<th>[28, 30]</th>
<th>[30, 31]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amsterdam</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Athens</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Bahrain</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Bombay</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Tokyo</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Toronto</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Vienna</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Zurich</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 4 contains the partitions into three, four, and five clusters
provided by the probabilistic similarity coefficient, αuv(k, k’),
combined with the three aggregation criteria (SL, AV1, and
AVB).

In the case of the probabilistic similarity coefficient,
αuv(k, k’), the partition into four clusters of the dendrogram
obtained from the SL method (see Table 4 and Figure 1) is
identical to that provided by the panel of human observers
(a priori partition). This partition was also obtained by Guru
et al. (2004), using a similarity measure for estimating the
degree of similarity among patterns (described by
interval type data) in terms of multivalued data, and an
unconventional agglomerative clustering technique, by
introducing the concept of mutual similarity value (Guru et al.,
2004). The partition into four clusters provided by the AV1
and AVB methods (see Table 4 and Figure 2) it is not the same
that the a priori partition (others authors (e.g., De Carvalho, 2007)
also have obtained partitions into four clusters that were
different to that a priori partition). It can be seen that the
partition into three clusters provided by the αuv(k, k’)
coefficient combined with the three applied aggregation
criteria is quite close to the a priori partition given by
the panel of human observers, excepting in what concerns
the location of city 18, which in the classification given by
the panel of human observers is a cluster with only one element
(singleton) [see Figures 1 and 2]. This is also the most significant
partition (the best partition), according to all applied
validation indexes, as is shown in Table 5, due to the maximum
values of STAT (18.9912), DIFF (2.3429 in the case of AV1
and AVB), and γ (0.8524), and to the minimum value of
P(I2mod, Σ).

Cluster 1: {C0, C1, C7, C9, C10, C13, C14, C16, C20, C21, C24,
C25, C26, C27, C28, C30, C33, C34, C35, C36};
Cluster 2: {C32};
Cluster 3: {C2, C3, C8, C4, C5, C6, C11, C12, C15, C17, C18, C19,
C22, C23, C29, C31}.
That partition was also the best partition according to the STAT (18.6694), DIF (0.4982), and $P(I_2^{\text{mod, } \Sigma})$ (0.1282) indexes obtained from the $\alpha(k, k')$ coefficient combined with the AV1 and AVB methods.

Table 4. Partitions into three, four and five clusters - $\alpha_{WW}(k, k')$

<table>
<thead>
<tr>
<th>Number of clusters</th>
<th>Methods</th>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
<th>Cluster 4</th>
<th>Cluster 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>$\alpha_{WW}(k, k') + SL$</td>
<td>C2, C3, C4, C5, C6, C8, C11, C12, C15, C14, C16, C20, C21, C24, C25,C10, C13, C19, C22, C23, C29, C31</td>
<td>C0, C7, C9, C10, C13, C14, C16, C20, C21, C24, C25, C26, C27, C28, C30, C33, C34, C35, C36</td>
<td>C0, C7, C9, C10, C13, C14, C16, C20, C21, C24, C25, C26, C27, C28, C30, C33, C34, C35, C36</td>
<td>C0, C7, C9, C10, C13, C14, C16, C20, C21, C24, C25, C26, C27, C28, C30, C33, C34, C35, C36</td>
<td>C0, C7, C9, C10, C13, C14, C16, C20, C21, C24, C25, C26, C27, C28, C30, C33, C34, C35, C36</td>
</tr>
<tr>
<td>4</td>
<td>$\alpha_{WW}(k, k') + AV1$</td>
<td>C0, C7, C10, C14, C20, C21, C25, C30, C34, C35</td>
<td>C1, C9, C13, C16, C24, C26, C27, C28, C33, C36</td>
<td>C0, C7, C9, C10, C13, C14, C16, C20, C21, C24, C25, C26, C27, C28, C30, C33, C34, C35, C36</td>
<td>C0, C7, C9, C10, C13, C14, C16, C20, C21, C24, C25, C26, C27, C28, C30, C33, C34, C35, C36</td>
<td>C0, C7, C9, C10, C13, C14, C16, C20, C21, C24, C25, C26, C27, C28, C30, C33, C34, C35, C36</td>
</tr>
<tr>
<td>5</td>
<td>$\alpha_{WW}(k, k') + AV1$</td>
<td>C0, C7, C10, C14, C20, C21, C25, C30, C34, C35</td>
<td>C3, C5, C6, C12, C15, C17, C18</td>
<td>C0, C7, C9, C10, C13, C14, C16, C20, C21, C24, C25, C26, C27, C28, C30, C33, C34, C35, C36</td>
<td>C0, C7, C9, C10, C13, C14, C16, C20, C21, C24, C25, C26, C27, C28, C30, C33, C34, C35, C36</td>
<td>C0, C7, C9, C10, C13, C14, C16, C20, C21, C24, C25, C26, C27, C28, C30, C33, C34, C35, C36</td>
</tr>
</tbody>
</table>

Table 5. Values of validation measures for the partitions into three, four, and five clusters

<table>
<thead>
<tr>
<th></th>
<th>STAT</th>
<th>DIF</th>
<th>$P(I_2^{\text{mod, } \Sigma})$</th>
<th>$\gamma$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SL</td>
<td>AV1/AVB</td>
<td>SL AV1/AVB</td>
<td>SL AV1/AVB</td>
</tr>
<tr>
<td>5 Clusters</td>
<td>18.5947</td>
<td>15.7538</td>
<td>0.6431</td>
<td>0.7185</td>
</tr>
<tr>
<td>4 Clusters</td>
<td>18.3947</td>
<td>16.6483</td>
<td>-0.2</td>
<td>0.8945</td>
</tr>
<tr>
<td>3 Clusters</td>
<td>18.9912</td>
<td>18.9912</td>
<td>0.5965</td>
<td>2.3429</td>
</tr>
</tbody>
</table>

The partition into four clusters provided by the $\alpha_{WW}(k, k')$ coefficient combined with the SL method is identical to the $a$ priori partition (the same is not verified in the case of the application of the $\alpha(k, k')$ coefficient combined with the SL method).

Figure 1. Dendrogram obtained by the probabilistic coefficient, $\alpha_{WW}(k, k')$, + SL method (last levels)

Figure 2. Dendrogram obtained by the probabilistic coefficient, $\alpha_{WW}(k, k')$, + AV1 method (last levels)
The values of $\text{Sil}^*$ and of $U$ statistics ($U_i/U_G$) for the cluster 2 belonging to the partition in to three clusters and to the a priori partition are, respectively, 0.5270 and 569/4782. Moreover, the cluster 1 belonging to the partition into three clusters has a higher value of $\text{Sil}^*$, and a lower value of $U$ statistics ($U_i/U_G$) compared to the corresponding ones of the cluster 1 of the partition provided by the human observers ($\text{Sil}^*=0.4917$; $U_i/U_G = 423/2598$ versus $\text{Sil}^*=0.4897$; $U_i/U_G = 474/2687$). According with the $U$ statistics of Mann and Whitney (Gordon, 1999), the clusters concerning to the partitions into three and to the a priori partition into four clusters, with more than one element, are neither ball clusters nor I*-clusters, however they are dense and well separated clusters, because they present relatively high values of the $\text{Sil}^*$ index. The "best" cluster is one that presents the largest value of $\text{Sil}^*$ and the smallest value of the $U$ statistics ($U_i$ and $U_G$). Thus, the best partition (into three clusters), according to the applied validation measures, is also slightly better than the partition into four clusters provided by the human panel.

Conclusion

The city temperature interval data set, as well as other experiments with different real and artificial interval data sets, have shown the usefulness of the weighted generalised affinity, $a(k, k')$, and of two related coefficients, namely, the asymptotic standardized weighted generalized affinity coefficient by the method of Wald and Wolofowitz, $a_{WW}(k, k')$, and the associated probabilistic coefficient, $a_{WW}(k, k')$. The use of the probabilistic coefficient, $a_{WW}(k, k')$, instead of the coefficient $a(k, k')$, allows us to work with comparable similarity values in a probabilistic scale. Moreover, the used validation measures are helpful in the selection of the best partitions of the elements to be classified.

Global satisfactory results were obtained using our approach, and one of the obtained partitions is in complete accordance with the partition into four clusters provided by the panel of human observers (a priori partition), although it is not the best one, according to the applied validation measures. The validation measures point to the partition into three clusters as the best partition.

REFERENCES


