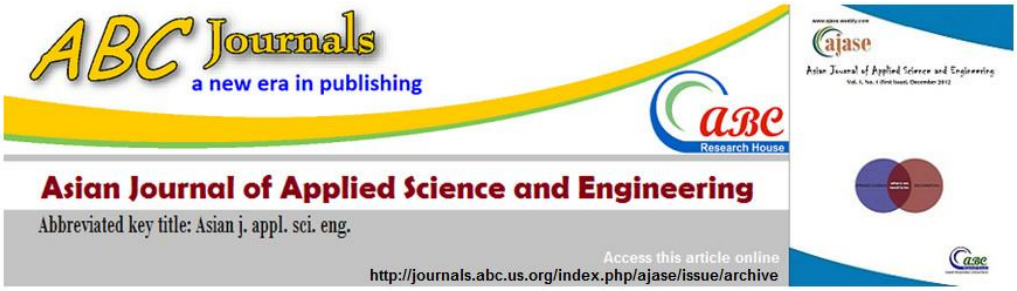

On Clustering Interval Data with Different Scales of Measures: Experimental Results

Áurea Sousa[1*], Helena Bacelar-Nicolau[2], Fernando C. Nicolau[3], Osvaldo Silva[4]

[1]Department of Mathematics and CEEpla, University of Azores, Ponta Delgada, **PORTUGAL**
[2]Laboratory of Statistics and Data Analysis, Faculty of Psychology, University of Lisbon, Lisboa, **PORTUGAL**
[3]Department of Mathematics, FCT, New University of Lisbon, Caparica, **PORTUGAL**
[4]Department of Mathematics and CICS.NOVA, University of Azores, Ponta Delgada, **PORTUGAL**

ARTICLE INFO

Received: Dec 9, 2014
Accepted: Feb 1, 2015
Published: Feb 26, 2015

***Corresponding Contact**
Email: **aurea@uac.pt**
Cell Phone: (351) 967864946

crossref doi Prefix 10.18034

ABSTRACT

Symbolic Data Analysis can be defined as the extension of standard data analysis to more complex data tables. We illustrate the application of the Ascendant Hierarchical Cluster Analysis (AHCA) to a symbolic data set (with a known structure) in the field of the automobile industry (car data set), in which objects are described by variables whose values are intervals of the real data set (interval variables). The AHCA of thirty-three car models, described by eight interval variables (with different scales of measure), was based on the standardized weighted generalized affinity coefficient, by the method of Wald and Wolfowitz. We applied three probabilistic aggregation criteria in the scope of the VL methodology (V for Validity, L for Linkage). Moreover, we compare the achieved results with those obtained by other authors, and with a priori partition into four clusters defined by the category (Utilitarian, Berlina, Sporting and Luxury) to which the car belong. We used the global statistics of levels (STAT) to evaluate the obtained partitions.

Keywords: Ascendant hierarchical cluster analysis, standardized weighted generalized affinity coefficient by the method of Wald and Wolfowitz, interval data, VL methodology

Source of Support: Nil, **No Conflict of Interest:** Declared.

How to Cite: Sousa Á, Bacelar-Nicolau H, Nicolau FC and Silva O. 2015. **On Clustering Interval Data with Different Scales of Measures: Experimental Results** *Asian Journal of Applied Science and Engineering,* 4, 17-25.

INTRODUCTION

The purpose of Cluster Analysis is to identify groups (clusters) of entities (data units/objects or variables), homogeneous and, preferably, well separated, on the basis of similarities or dissimilarities between these entities. There are two main classes of clustering methods:

hierarchic and non-hierarchic methods. The first ones return a nested sequence of partitions (hierarchical structure). On the other hand, the non-hierarchical methods seek to obtain a single partition of the input data into an appropriate number of clusters. The last ones usually produce clusters by (locally) optimizing an adequacy criterion. In this paper, we will focus on hierarchic agglomerative methods (Ascendant Hierarchical Cluster Analysis - AHCA).

With the advent of computers, it is possible to synthesize data in terms of their most relevant concepts, which may be described by different types of complex data (generalizations of classical data types), also known as symbolic or complex data. In a symbolic data table, rows correspond to data units (frequently, groups of individuals) and columns to variables. Each entry in the table 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, 2014; Doria et al., 2013). Some similarity and dissimilarity measures for the case of Symbolic data can be found, for instance, in Bock and Diday (2000).

The recording of interval data has become a more frequent practice with the recent advances in database technologies (Souza and De Carvalho, 2004). Some dissimilarity measures for interval data can be found in the literature (see f.i. Chavent and Lechevallier, 2002; Chavent et al., 2003; Souza and De Carvalho, 2004; De Carvalho et al., 2006a, 2006b), as well as some similarity measures which are capable of dealing with the particular case of interval data (e.g. Bacelar-Nicolau et al., 2009, 2010, 2014a, 2014b). In this paper, we face the problem of clustering data units described by variables whose values are intervals of the real data set (interval data), with different scales of measures. To show the usefulness of the standardized weighted generalized affinity coefficient by the method of Wald and Wolfowitz (Bacelar-Nicolau, 2000; Bacelar-Nicolau et al., 2009, 2010; Sousa et al., 2013a), for this type of data, a well-known interval data set (with a known structure) was considered. The AHCA was based on three probabilistic aggregation criteria (*AVL*, *AV1*, and *AVB*) included in a parametric family of methods in the context of the *VL* methodology (e.g. Nicolau, 1983; Bacelar-Nicolau, 1988; Nicolau and Bacelar-Nicolau, 1998; Lerman, 1972, 1981). In addition, we compare the achieved results with those obtained by other authors (e.g., De Carvalho et al., 2006a, 2006b; Souza et al., 2007), and with a priori partition. The validation of the obtained partitions is based on the global statistics of levels (STAT), as proposed by Lerman (1970, 1981) and Bacelar-Nicolau (1980, 1985).

The paper is organized as follows: the second section is related to the models of AHCA in the field of the Symbolic Data Analysis used in the present work. More emphasis is given to the weighted generalized affinity coefficient, and to the corresponding asymptotic standardized weighted generalized coefficient, under a permutational reference hypothesis based on the limit theorem of Wald and Wolfowitz. We present, in the third section, the best results obtained with the application of the AHCA to a data set (with a known structure) in the field of the automobile industry (car data set), in which objects (thirty-three car models) are described by eight variables whose values are intervals of the real line (interval variables). Finally, the fourth section contains some concluding remarks about the work and its results.

ASCENDANT HIERARCHICAL CLUSTER ANALYSIS OF SYMBOLIC DATA UNITS

The AHCA (agglomerative hierarchical methods) usually start with every single object in a single cluster (singleton). The algorithm successively merges the most similar clusters together until the entire set of elements to classify becomes one group. In this work, we use a similarity measure, called standardized weighted generalized affinity coefficient, by the method of Wald and Wolfowitz, which is described in the present section.

Weighted Generalized Affinity Coefficient for the case of Interval Data

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 (corresponding to pairs of columns or rows of a data matrix). A theoretical study of this coefficient and their asymptotic normal distributions may be found, e.g., in Bacelar-Nicolau (1980, 1988). Moreover, some simulation studies (Sousa, 2005; Sousa et al., 2013b) shown that the convergence of the affinity coefficient for the normal distribution is relatively fast (in general from sample sizes above 20). Afterwards, Bacelar-Nicolau extended that coefficient to different types of data, including complex and heterogeneous data (Bacelar-Nicolau, 2000, 2002; Bacelar-Nicolau et al., 2009, 2010, 2014a, 2014b). The so-called weighted generalized affinity coefficient, $a(k,k')$, between a pair of statistical data units, k and k' $(k,k' = 1,\dots,N)$, is an extension of the affinity coefficient for the case of symbolic data, and may be defined as follows:

$$a(k,k') = \sum_{j=1}^{p} \pi_j \; aff(k,k';j) = \sum_{j=1}^{p} \pi_j \sum_{\ell=1}^{m_j} \sqrt{\frac{x_{kj\ell}}{x_{kj\bullet}} \cdot \frac{x_{k'j\ell}}{x_{k'j\bullet}}} \quad (1)$$

where: $aff(k,k';j)$ is the generalized local affinity between k and k' over the *jth* variable, m_j is the number of modalities of the sub-table associated to the *jth* variable, $x_{kj\ell}$ is the number of individuals (in the unit k) which share category ℓ of variable Y_j, $x_{kj\bullet} = \sum_{\ell=1}^{m_j} x_{kj\ell}$, $x_{k'j\bullet} = \sum_{\ell=1}^{m_j} x_{k'j\ell}$ and π_j are weights such that $0 \le \pi_j \le 1$, $\sum \pi_j = 1$. Either the local affinities or the whole coefficient given by formula (1) assume values in the interval [0,1] (e.g. Bacelar-Nicolau, 2002; Bacelar-Nicolau et al. 2009).

The coefficient defined by formula (1) is suitable when mixed variables types are present in a database, often a large one, since the same coefficient can deal with different variables types (for details, see Bacelar-Nicolau et al., 2009, 2010). In the particular case of symbolic variables of interval type, Bacelar-Nicolau has defined the weighted generalized affinity coefficient, as is described below (for details, see Bacelar-Nicolau et al., 2009, 2010, 2014b).

Given N data units described by p interval variables, Y_j, with $j=1, \dots, p$, and a data matrix, as Table 1, where each cell (k, j) contains an interval $I_{kj} = [a_{kj}, b_{kj}]$ of the real data set, with k =1,…,N and j=1,…,p, the weighted generalized affinity coefficient between a pair of data units, k and $k'(k,k' = 1,\dots,N)$, can be expressed by:

$$a(k,k') = \sum_{j=1}^{p} \pi_j \cdot \frac{|I_{kj} \cap I_{k'j}|}{\sqrt{|I_{kj}| \cdot |I_{k'j}|}}, \quad (2)$$

where k and $k'(k,k' = 1,\dots,N)$ are a pair of data units, $|I_{kj}|$, $|I_{k'j}|$ and $|I_{kj} \cap I_{k'j}|$ represent, respectively, the ranges of the intervals I_{kj}, $I_{k'j}$ and $I_{kj} \cap I_{k'j}$. Bacelar-Nicolau called the coefficient defined by formula (2) a *generalized Ochiai coefficient for interval data*, which is associated with a 2×2 generalized contingency table that contains interval ranges instead of the usual cardinal numbers of any simple 2×2 contingency table.

Table I: Symbolic data table (interval data)

	Y_1	⋯	Y_j	⋯	Y_{1p}
1	I_{11}	⋯	I_{1j}	⋯	I_{1p}
⋮	⋮	⋮	⋮	⋮	⋮
K	I_{k1}	⋯	I_{kj}	⋯	I_{kp}
⋮	⋮	⋮	⋮	⋮	⋮
N	I_{N1}	⋯	I_{Nj}	⋯	I_{Np}

It was demonstrated that formula (2) arises as a particular case of formula (1) when we are dealing with variables of interval type (see e.g. Bacelar-Nicolau et al., 2009, 2010). Taking into account the decomposition of each interval into m_j elementary and disjoint intervals, $\{I_{j\ell}: \ell = 1, \cdots, m_j\}$, we obtain the following equalities:

$$a(k,k') = \sum_{j=1}^{p} \pi_j aff(k,k';j) = \sum_{j=1}^{p} \pi_j \cdot \sum_{\ell=1}^{m_j} \sqrt{\frac{x_{kj\ell}}{x_{kj\bullet}} \cdot \frac{x_{k'j\ell}}{x_{k'j\bullet}}} = \sum_{j=1}^{p} \pi_j \cdot \frac{|I_{kj} \cap I_{k'j}|}{\sqrt{|I_{kj}| \cdot |I_{k'j}|}},$$

with $x_{kj\ell} = |I_{kj} \cap I_{j\ell}|$, where $|\ |$ represents the interval range, $x_{kj\ell} = |I_{i\ell}|$ if $I_{kj} \cap I_{j\ell} = I_{j\ell}$, and $x_{kj\ell} = 0$, otherwise, π_j are weights such that $0 \leq \pi_j \leq 1$, $\sum \pi_j = 1$, $x_{kj\bullet} = \sum_{\ell=1}^{m_j} x_{kj\ell}$, $x_{k'j\bullet} = \sum_{\ell=1}^{m_j} x_{k'j\ell}$, and $|I_{kj}|$, $|I_{k'j}|$ and $|I_{kj} \cap I_{k'j}|$ are, respectively, the ranges of the intervals I_{kj}, $I_{k'j}$ and $I_{kj} \cap I_{k'j}$. Therefore, the weighted generalized affinity coefficient $a(k,k')$ between a pair $(I_{kj}, I_{k'j})$ of intervals $(k,k' = 1, \ldots, N)$, may be computed in two different ways, either by using the formula (1) or, alternatively, by using the formula (2).

Asymptotic Standardized Weighted Generalized Affinity Coefficient

The values of the proximity measures and of the clustering results are affected by the scales of the variables. Often, a standardization performed prior to the clustering process improves the performance of the clustering method (De Carvalho et al., 2006a).

Considering a permutational reference hypothesis based on the limit theorem of Wald and Wolfowitz (Fraser, 1975), the random variable associated to $aff(k,k';j)$ follows the asymptotic normal distribution, and an associated standardized coefficient, $a_{WW}(k,k')$, may be used, instead of $a(k,k')$ (see f.i. Bacelar-Nicolau, 1988; Bacelar-Nicolau et al., 2009, 2010, 2014a).

From the decomposition of each interval into a suitable number of elementary intervals, as referred above, the local standardized weighted generalized affinity coefficient by the method of Wald and Wolfowitz, $aff^*_{WW}(I_{kj}, I_{k'j}; j)$ or, more concisely, $aff^*_{WW}(k,k';j)$, between a pair $(I_{kj}, I_{k'j})$ of intervals $(k,k' = 1, \ldots, N)$, over the *jth* variable, is given by the formula:

$$aff^*_{ww}(k,k';j) = \sqrt{m_j - 1} \frac{\sum_{\ell=1}^{m_j} \sqrt{x_{kj\ell} \cdot x_{k'j\ell}} - \frac{1}{m_j} \sum_{\ell=1}^{m_j} \sqrt{x_{kj\ell}} \sum_{\ell=1}^{m_j} \sqrt{x_{k'j\ell}}}{\sqrt{\left[x_{kj\bullet} - \frac{1}{m_j}\left(\sum_{\ell=1}^{m_j} \sqrt{x_{kj\ell}}\right)^2\right]\left[x_{k'j\bullet} - \frac{1}{m_j}\left(\sum_{\ell=1}^{m_j} \sqrt{x_{k'j\ell}}\right)^2\right]}} \qquad (3)$$

where, the notations are the same as those in the previous subsection. Finally, the standardized weighted generalized affinity coefficient by the method of Wald and Wolfowitz is given by the formula:

$$a_{WW}(k,k') = \sum_{j=1}^{p} \pi_j \, aff^*_{WW}(k,k';j) \qquad (4)$$

Where π_j are weights such that $0 \leq \pi_j \leq 1$, $\sum \pi_j = 1$, and the local asymptotic normal affinity coefficient $aff^*_{WW}(k,k';j)$ also satisfies the main properties of a similarity coefficient (Bacelar-Nicolau et al., 2010). Furthermore, $a_{WW}(k,k')$ allows us to define a probabilistic coefficient in the scope of the *VL* methodology, 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).

Given the affinity similarity matrix, a data set can be classified through classical aggregation criteria or probabilistic ones (Bacelar-Nicolau et al., 2009, 2010, 2014b; Sousa et al., 2013a). In the present work, we used probabilistic aggregation criteria under the *VL*

probabilistic approach.

An important step in Cluster Analysis is to determine the best number of clusters. The values of validation indexes obtained from the values of the similarity (or dissimilarity) matrix between elements can be calculated, also in the case of symbolic data (Sousa, 2005; Sousa et al., 2010, 2013a, 2014). Here we use, as mentioned above, the global statistics of levels STAT (e. g. Lerman 1970, 1981; Bacelar-Nicolau, 1980, 1985; Sousa et al., 2014) as the validation index to find the obtained best partitions.

EXPERIMENTAL RESULTS: THE CAR DATA SET

The analyzed symbolic data matrix (see Table II) is referred in the literature of the symbolic data analysis (e. g. De Carvalho et al, 2006a, 2006b; Souza et al, 2007) and contains thirty-three car models (complex data units) described by eight interval variables (*Price*, *Engine Capacity*, *Top Speed*, *Acceleration*, *Step*, *Length*, *Width and Height*), two categorical non-ordered multi-valued variables (*Alimentation* and *Traction*) and one nominal (*Car Category*). This last variable, with the modalities *Utilitarian*, *Berlina*, *Sporting* and *Luxury*, reflects the *a priori* partition (indicated by the suffix attached to the car model denomination) into four groups according to the category (De Carvalho et al., 2006a, 2006b), as follows: ***Utilitarian:*** {1-Alfa 145/U; 5-Audi A3/U; 12-Punto/U; 13-Fiesta/U; 17-Lancia Y/U; 24-Nissan Micra/U; 25-Corsa/U; 28-Twingo/U; 29-Rover 25/U; 31-Skoda Fabia/U}; ***Berlina:*** {2-Alfa 156/B; 6-Audi A6/B; 8-BMW serie 3/B; 14-Focus/B; 21-Mercedes Classe C/B; 26-Vectra/B; 30-Rover 75/B; 32-Skoda Octavia/B}; ***Sporting:*** {4-Aston Martin/S; 11-Ferrari/S; 15-Honda NSK/S; 16-Lamborghini/S; 19-Maserati GT/S; 20-Mercedes SL/S; 27-Porsche/S}; ***Luxury:*** {3-Alfa 166/L; 7-Audi A8/L; 9-BMW serie 5/L; 10-BMW serie 7/L}.

Table II shows part of the symbolic data matrix. The complete data set is included in the SODAS (Symbolic Official Data Analysis System) software. In this work, the eight interval variables have been considered for the AHCA of the thirty-three car models based on the standardized weighted generalized affinity coefficient by the method of Wald and Wolfowitz (see the precedent section). The measure of comparison between elements has been combined with three probabilistic aggregation criteria, *AVL*, *AV1*, and *AVB* (Bacelar-Nicolau, 1988; Nicolau, 1983; Nicolau and Bacelar-Nicolau, 1998).

Table II: Symbolic data matrix - Car data set

Model	Price	Engine Capacity	Alimentation	Traction	. . .	Height	Category
Alfa 145	[27806, 33596]	[1370, 1910]	Gasoli, Diese	Anter	. . .	[143, 143]	Utilit
Alfa 156	[41593, 62291]	[1598, 2492]	Gasoli	Anter	. . .	[142, 142]	Berlina
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
Passat	[39676, 63455]	[1595, 2496]	Gasoli, Diese	Anter, Integ	. . .	[146, 146]	Luxo

Before calculating the coefficient given by formula (4), the domains of each variable were decomposed in a suitable number of elementary intervals (respectively, 65, 52, 46, 57, 30, 32, 24 and 12 elementary intervals, for the variables *Price*, *Engine Capacity*, *Top Speed*, *Acceleration*, *Step*, *Length*, *Width* and *Height*). Consequently, we obtained a new data matrix, subdivided into eight subtables (one for each variable), which contain a decomposition of the respective initial intervals into elementary intervals. Table III illustrates the decomposition corresponding to the variable *Price*. In that case, once sorted (in ascending order) the values corresponding to the lower and upper boundaries of the 33 intervals (corresponding to the car models), we considered the subintervals defined only by the distinct values.

Table III: Decomposition in elementary intervals – Variable *Price*

Car Model	Price	[16992, 18492]	⋯	[262500, 276792]	[276792, 389405]	⋯	[423000, 460000]
1/U	[27806, 33596]	0	⋯	0	0	⋯	0
2/B	[41593, 62291]	0	⋯	0	0	⋯	0
3/L	[64499, 88760]	0	⋯	0	0	⋯	0
4/S	[260500, 460000]	0	⋯	14292	112613	⋯	37000
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
9/L	[70292, 198792]	0	⋯	0	0	⋯	0
10/L	[104892, 276792]	0	⋯	14292	0	⋯	0
11/S	[240292, 391692]	0	⋯	14292	112613	⋯	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
31/U	[19519, 32686]	0	⋯	0	0	⋯	0
32/B	[27419, 48679]	0	⋯	0	0	⋯	0
33/L	[39676, 63455]	0	⋯	0	0	⋯	0

Note that in the column corresponding to the variable *Price* there are not initial intervals with identical lower and upper boundaries. Contrary, in the columns associated to other variables there are intervals in such conditions. We deal with this situation replacing these intervals by transformed intervals obtained from the first ones, by subtracting and adding 0.5, respectively to the lower and upper boundaries. Table IV exemplifies the procedure for the Variable *Height*.

Table IV: Decomposition in elementary intervals – Variable *Height*

Car Model	Height	Transformed intervals	[110.5, 111.5]	[123.5, 128.5]	⋯	[145.5, 146.5]	[147.5, 148.5]
1/U	[143, 143]	[142.5, 143.5]	0	0	⋯	0	0
2/B	[142, 142]	[141.5, 142.5]	0	0	⋯	0	0
3/L	[142, 142]	[141.5, 142.5]	0	0	⋯	0	0
4/S	[124, 132]	[123.5, 132.5]	0	5	⋯	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
9/L	[144, 144]	[143.5, 144.5]	0	0	⋯	0	0
10/L	[143, 143]	[142.5, 143.5]	0	0	⋯	0	0
11/S	[130, 130]	[129.5, 130.5]	0	0	⋯	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
31/U	[145, 145]	[144.5, 145.5]	0	0	⋯	0	0
32/B	[143, 143]	[142.5, 143.5]	0	0	⋯	0	0
33/L	[146, 146]	[145.5, 146.5)	0	0	⋯	1	0

According to the STAT index, the selected (best) partition is the partition into six clusters (STAT=14.7191), which was obtained at level 27 by the *AVL* method: C1:{1/U, 14/B, 26/B, 32/B, 2/B, 33/L, 5/U, 8/B, 30/B, 3/L, 18/L};
C2:{6/B, 21/B, 7/L, 9/L, 22/L, 10/L, 23/L, 20/S}; C3:{12/U, 31/U, 29/U, 24/U, 25/U, 17/U, 13/U, 28/U}; C4:{4/S, 11/S, 19/S, 27/S}; C5:{15/S}; C6:{16/S}. On the other hand, from the dendrogram provided by *AVL* method (see Figure 1), we can see four well-defined clusters. Table V contains the main clusters that we can see from the dendrograms provided by the applied methods (*AVL*, *AV1*, and *AVB*). In this table, we present for each cluster their individuals and respective *a priori* class labels.

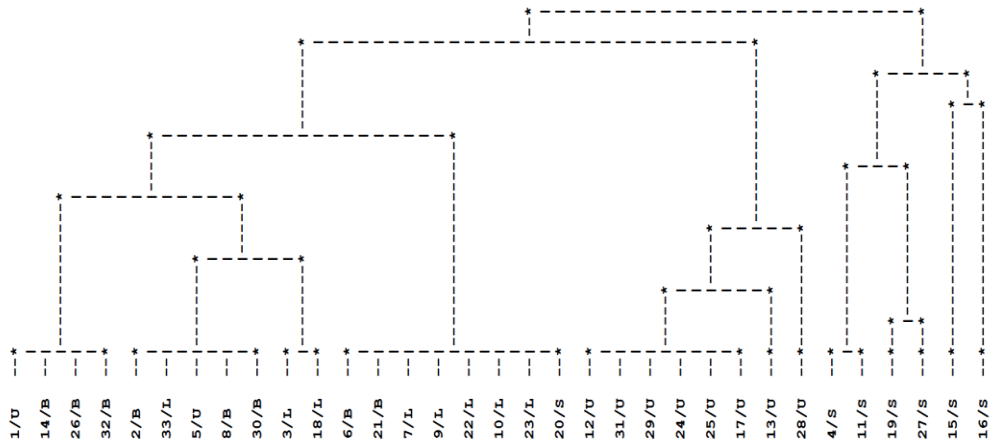

Figure 1: **Dendrogram obtained with AVL (levels 21 to 32)**

The four clusters provided by the AVL method are in accordance with the ones identified by other authors (e.g., De Carvalho et al., 2006a, 2006b; Souza et al., 2007), using different standardization methods, except with regard to the objects (car models) 20/S and 21/B (see Table V). In particular, the clusters {6/B, 21/B, 7/L, 9/L, 22/L, 10/L, 23/L, 20/S} and {12/U, 31/U, 29/U, 24/U, 25/U, 17/U, 13/U, 28/U} were found in all obtained dendrograms. Note that the cluster {12/U, 31/U, 29/U, 24/U, 25/U, 17/U, 13/U, 28/U}, containing most Utilitarian cars, also was identified by the referred authors. In addition, it appears that the partition into four clusters provided by AVL is compatible with the *a priori* partition defined by the variable *Car category*. This fact points out to the satisfactory performance of the standardized weighted generalized affinity coefficient, by the method of Wald and Wolfowitz.

Table V: Clustering results for the Car data set

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
$a_{ww}+AVL$	1/U, 14/B, 26/B, 32/B, 2/B, 33/L, 5/U, 8/B, 30/B, 3/L, 18/L	6/B, 21/B, 7/L, 9/L, 22/L, 10/L, 23/L, 20/S	12/U, 31/U, 29/U, 24/U, 25/U, 17/U, 13/U, 28/U	4/S, 11/S, 19/S, 27/S, 15/S, 16/S			
$a_{ww}+AV1$	26/B, 32/B, 2/B, 33/L, 5/U, 8/B, 30/B	6/B, 21/B, 7/L, 9/L, 22/L, 10/L, 23/L, 20/S	12/U, 31/U, 29/U, 24/U, 25/U, 17/U, 13/U, 28/U	4/S, 11/S, 19/S, 27/S	15/S , 16/S	1/U, 14/B	3/L, 18/L
$a_{ww}+AVB$	26/B, 32/B, 2/B, 33/L, 5/U, 8/B, 30/B, 3/L, 18/L	6/B, 21/B, 7/L, 9/L, 22/L, 10/L, 23/L, 20/S	12/U, 31/U, 29/U, 24/U, 25/U, 17/U, 13/U, 28/U	4/S, 11/S, 19/S, 27/S	15/S, 16/S	1/U, 14/B	
***Other authors* (standardization methods)**	1/U, 14/B, 26/B, 32/B, 2/B, 33/L, 5/U, 8/B, 30/B, 3/L, 18/L, 21/B	6/B, 7/L, 9/L, 22/L, 10/L, 23/L	12/U, 31/U, 29/U, 24/U, 25/U, 17/U, 13/U, 28/U	4/S, 11/S, 19/S, 27/S, 15/S, 16/S, 20/S			

CONCLUDING REMARKS

It is important to stress that the weighted generalized affinity coefficient, $a(k,k')$, between

a pair $(I_{kj}, I_{k'j})$ of intervals $(k, k' = 1, \ldots, N)$, may be computed from the formulae (1) or (2). The decomposition of each interval into elementary intervals is mainly interesting in the case of the standardized weighted generalized affinity coefficient, $a_{WW}(k, k')$, due to the associated reference hypothesis based on the limit theorem of Wald and Wolfowitz. The use of the similarity measure $a_{WW}(k, k')$, instead of the $a(k, k')$, allows us to work with comparable values of the similarity measure, which are realizations of random variables with the same distribution (asymptotically normal standard).

In this paper, we faced the problem of clustering interval data, with different scales of measures (case of the analyzed data set), in the scope of the *VL* methodology. The obtained results with the application of our methods to these data are congruent with the ones reported by other authors. Indeed, several applications of the used methodology to various data sets (with a known structure) have shown that the used methods are very promising, that is, they reproduce in a satisfactory way the proprieties of the data structures.

REFERENCES

Bacelar-Nicolau, H. (1980), *Contributions to the study of comparison coefficients in cluster analysis*. Ph.D Thesis (in Portuguese), Universidade de Lisboa, Portugal.

Bacelar-Nicolau, H. (1985), "The affinity coefficient in cluster analysis", *Methods of Operations Research*, vol. 53, M. J. Beckmann, K.-W. Gaede, K. Ritter, & H. Schneeweiss (Eds.), Verlag Anton Hain, Munchen, pp. 507-512.

Bacelar-Nicolau, H. (1987), "On the distribution equivalence in cluster analysis. In Devijver, P.A. & Kittler, J. (Eds.) *Pattern Recognition Theory and Applications*, NATO ASI Series, Series F: Computer and Systems Sciences, vol. 30, Springer - Verlag, New York, pp. 73-79.

Bacelar-Nicolau, H. (1988), "Two probabilistic models for classification of variables in frequency tables", In: Bock, H.-H. (Eds.), *Classification and related methods of data analysis*, Elsevier Sciences Publishers B.V., North Holland, pp. 181-186.

Bacelar-Nicolau, H. (2000), "The affinity coefficient", In: *Analysis of symbolic data: Exploratory methods for extracting statistical information from complex data*, H.-H. Bock & E. Diday (Eds.), Series: Studies in Classification, Data Analysis, and Knowledge Organization, Springer-Verlag, Berlin, pp. 160-165.

Bacelar-Nicolau, H. (2002), "On the generalised affinity coefficient for complex data", *Biocybernetics and Biomedical Engineering*, vol. 22, no. 1, pp. 31-42.

Bacelar-Nicolau, H., Nicolau, F.C., Sousa, Á., & Bacelar-Nicolau, L. (2009), "Measuring similarity of complex and heterogeneous data in clustering of large data sets", *Biocybernetics and Biomedical Engineering*, vol. 29, no. 2, pp. 9-18.

Bacelar-Nicolau, H., Nicolau, F.C., Sousa, Á., & Bacelar-Nicolau, L. (2010), "Clustering complex heterogeneous data using a probabilistic approach", *Proceedings of the Stochastic Modeling Techniques and Data Analysis International Conference* (SMTDA2010), pp. 85-93. Available from http://www.smtda.net/images/ SMTDA_2010_Proceedings_pp_1-356.pdf (Accessed: 05 February 2015).

Bacelar-Nicolau, H., Nicolau, F.C., Sousa, Á., & Bacelar-Nicolau, L. (2014a), "Clustering of variables with a three-way approach for health sciences", *Testing, Psychometrics, Methodology in Applied Psychology* (*TPM*), vol. 21, no. 4, pp. 435-447.

Bacelar-Nicolau, H., Nicolau, F.C., Sousa, Á., & Bacelar-Nicolau, L. (2014b), "On cluster analysis of complex and heterogeneous data", *Proceedings of the 3rd Stochastic Modeling Techniques and Data Analysis International Conference* (SMTDA2014), C. H. Skiadas (Eds.), 2014 ISAST, pp. 99-108. Available from: http://www.smtda.net/images/1_A-F_SMTDA2014_Proceedings_NEW.pdf (Accessed: 05 February 2015).

Bock, H.-H., & Diday, E. (Eds.) (2000), *Analysis of symbolic data: Exploratory methods for extracting statistical information from complex data*, Series: Studies in Classification, Data Analysis, and Knowledge Organization, Springer-Verlag, Berlin.

Chavent, M., & Lechevallier, Y. (2002), "Dynamical clustering algorithm of interval data: Optimization of an adequacy criterion based on Hausdorff distance", In *Classification, clustering, and data analysis*, K. Jajuga, A. Sokolowski, H.-H. Bock (Eds.). Springer-Verlag, Berlin, pp. 53-60.

Chavent, M., De Carvalho, F.A.T., Lechevallier, Y., & Verde, R. (2003), "Trois nouvelles méthods de classification automatique de données symboliques de type intervalle", *Revue de Statistique Applique*, vol. LI, no. 4, pp. 5-29.

De Carvalho, F.A.T., Brito, P., & Bock, H.-H. (2006a), "Dynamic clustering for interval data based on L_2 distance", *Computational Statistics*, vol. 21, no. 2**,** pp. 1-19.

De Carvalho, F.A.T., Souza, R.M.C.R. de, Chavent, M., & Lechevallier, Y. (2006b), "Adaptive Hausdorff distances and dynamic clustering of symbolic interval data", *Pattern Recognition Letters,* vol. 27, no. 3, pp. 167-179.

Diday, E., & Noirhomme-Fraiture, M. (Eds.) (2008) "Symbolic Data Analysis and the SODAS software", John Wiley & Sons, Chichester.

Doria, I., Sousa, Á., Bacelar-Nicolau, H., & Le Calvé, G. (2013), "Comparison of modal variables using multivariate analysis", In: João Lita da Silva, Frederico Caeiro, Isabel Natário and Carlos A. Braumann (Eds.), *Advances in regression, survival analysis, extreme values, Markov processes and other statistical applications*, Studies in Theoretical and Applied Statistics, Springer, Berlin, Heidelberg, pp. 363-370.

Lerman, I. C. (1970), "Sur l`analyse des données préalable à une classification automatique (Proposition d'une nouvelle mesure de similarité)", *Rev. Mathématiques et Sciences Humaines*, vol. 32, no. 8, pp. 5-15.

Lerman, I. C. (1972), *Étude distributionelle de statistiques de proximité entre structures algébriques finies du même type: Apllication à la classification automatique*, Cahiers du B.U.R.O., 19, Paris.

Lerman, I. C. (1981), *Classification et analyse ordinale des données*, Dunod, Paris.

Matusita, K. (1951), "On the theory of statistical decision functions", *Annals of the Institute of Statistical Mathemathics*, vol. 3, pp. 17-35.

Nicolau, F.C. (1983), "Cluster analysis and distribution function", *Methods of Operations Research*", vol. 45, pp. 431-433.

Nicolau, F.C., & Bacelar-Nicolau, H. (1998), "Some trends in the classification of variables", In: Hayashi, C., Ohsumi, N., Yajima, K., Tanaka, Y., Bock, H.-H., &Baba, Y. (Eds.), *Data Science, Classification, and Related Methods*. Springer-Verlag, pp. 89-98.

Sousa, Á. (2005), *Contributions to the VL methodology and validation indexes for data of complex nature*. Ph.D Thesis (in Portuguese), Universidade dos Açores, Portugal.

Sousa, Á., Nicolau, F.C., Bacelar-Nicolau, H., & Silva, O. (2010), "Weighted generalised affinity coefficient in cluster analysis of complex data of the interval type", *Biometrical Letters*, vol. 47, no. 1, pp. 45-56.

Sousa, Á., Nicolau, F.C., Bacelar-Nicolau, H., & Silva, O. (2013a), "Clustering of symbolic data based on affinity coefficient: Application to a real data set", *Biometrical Letters*, vol. 50, no. 1, pp. 27-38.

Sousa, Á., Silva, O., Bacelar-Nicolau, H., & Nicolau, F.C. (2013b), "Distribution of the affinity coefficient between variables based on the Monte Carlo simulation method", *Asian Journal of Applied Sciences*, vol. 1, no. 5, pp. 236-245. Available from: http://www.ajouronline.com/index.php?journal=AJAS&page=article&op=view&path%5B%5D=746&path%5B%5D=411 (Accessed: 05 February 2015).

Sousa, Áurea, Nicolau, F.C, Bacelar-Nicolau, H., & Silva, O. (2014). "Cluster analysis using affinity coefficient in order to identify religious beliefs profiles", *European Scientific Journal (ESJ)*, vol. 3 (Special edition), pp. 252 - 261. Available from: http://eujournal.org/index.php/esj/article/viewFile/2943/2772 (Accessed: 05 February 2015).

Souza, R.M.C.R., & De Carvalho, F.A.T. (2004), "Clustering of interval data based on City-Block distances", *Pattern Recognition Letters*, vol. 25, pp. 353-365.

Souza, R.M.C.R., De Carvalho, F.A.T., & Pizzato, D.F. (2007), "A partitioning method for mixed feature-type symbolic data using a squared Euclidean distance". IN FREKSA, C., KOHLHASE, M., & SCHILL, K. (Eds.) *KI 2006: Advances in artificial intelligence*. 29th Annual German Conference on AI, KI 2006, Bremen, Germany, June 14-17, 2006, Proceedings. Series: Lecture notes in computer science, vol. 4314, Springer-Verlag, Berlin Heidelberg, pp. 260-273.

-- 0 --