A Topological Clustering of Variables

Rafik Abdesselam

University of Lyon, Lumière Lyon 2, ERIC - COACTIS Laboratories
Department of Economics and Management, 69365 Lyon, France
(e-mail: rafik.abdesselam@univ-lyon2.fr)
(http://perso.univ-lyon2.fr/~rabdesse/fr/)

Abstract. The clustering of objects (individuals or variables) is one of the most used approaches to exploring multivariate data. The two most common unsupervised clustering strategies are hierarchical ascending clustering (HAC) and k-means partitioning used to identify groups of similar objects in a dataset to divide it into homogeneous groups.

The proposed topological clustering of variables, called TCV, studies an homogeneous set of variables defined on the same set of individuals, based on the notion of neighborhood graphs, some of these variables are more-or-less correlated or linked according to the type quantitative or qualitative of the variables. This topological data analysis approach can then be useful for dimension reduction and variable selection. Its a topological hierarchical clustering analysis of a set of variables which can be quantitative, qualitative or a mixture of both. It arranges variables into homogeneous groups according to their correlations or associations studied in a topological context of principal component analysis (PCA) or multiple correspondence analysis (MCA). The proposed TCV is adapted to the type of data considered, its principle is presented and illustrated using simple real datasets with quantitative, qualitative and mixed variables. The results of these illustrative examples are compared to those of other variables clustering approaches.

Keywords: Hierarchical clustering, proximity measure, neighborhood graph, adjacency matrix, multivariate quantitative, qualitative and mixed data analysis, dimension reduction.

1 Introduction

The objective of this article is to propose a new approach for classifying variables. This is a topological approach that is different from those that already exist and with which it is compared.

Besides classical and well know methods devoted to the clustering of objects, there are some approaches specifically devoted to the clustering of variables, the Varclus classification procedure [23] implemented in the SAS software, the ClustOfVar approach [9], the CVLC approach [29,28] for clustering variables arround latent components and the Clustatis approach [19], but as far as we know, none approach, is proposed in a topological context.

A clustering of variables can also be considered as a dimension reduction approach, like a factor analysis. The purpose of the classification of variables is to group together the variables strongly related to each other, that is to say to separate the variables into classes of variables. It will be possible to summarize each class of variables by a single quantitative synthetic variable.

The interest here is to understand the structures underlying the data, to constitute a summary of the information carried by the data or to detect redundancies, for example with a view to reducing number of variables in another process.

The objective of the clustering of variables is to obtain linked and redundant classes of variables. Specific algorithms have thus been developed for the clustering of variables. To create profiles from variables grouped in a questionnaire, we can achieve this using two main types of methods: non-hierarchical clustering such as K-means or dynamic clusters, and hierarchical clustering of the ascending or descending type.

Similarity measures play an important role in many areas of data analysis. The results of any operation involving structuring, clustering or classifying objects are strongly dependent on the proximity measure chosen.

Generally the variables are homogeneous in the sense that they revolve around a particular theme. Unlike the clustering of individuals, which is generally done from a single set of homogeneous variables relating to a single theme, the clustering of variables can process several sets of homogeneous variables from several different themes. The clusters of variables of the chosen partition can be considered as a selection of variables, each cluster of variables can then be synthesized separately using a factor analysis for example.

The TCV can be considered as a method of reduction of dimensions where each class of correlated variables of the partition can be represented by the synthesis variable of the variables of the class, or again, as a method of selection of variables where each class can be represented by the significant variables of the class.

The present study proposes a topological hierarchical clustering of variables, with no restriction on the type, quantitatives, qualitatives or a mixture both of them.

Several topological studies have been proposed in factorial analyses context, discrimination analysis [4], simple and multiple correspondence analyses [2] and principal component analysis [1] but none on clustering of variables.

Therefore, this paper focuses on unsupervised clustering of a set of variables of any type, quantitative, qualitative or a mixture of both. The eventual associations or correlations between the variables partly depends on the database being used and the results of the topological clustering of these variables can change according to the selected proximity measure. A proximity measure is a function which measures the similarity or dissimilarity between two objects or variables within a set.

This paper is organized as follows. In section 2, we briefly recall the basic notion of neighborhood graphs, we define and show how to construct an adjacency matrix associated with a proximity measure within the framework of the analysis of the correlation or association structure of a set of variables. Section 3 presents the principles of the TCV according to the three types of variables. It is illustrated in section 4 using simple examples on real data. The TCV results are compared according to the type of variables, with those of different known clustering of variables approaches. Finally, section 5 gives concluding remarks of this work.

2 Topological context

Topological data analysis is an approach based on the concept of the neighborhood graph. The basic idea is actually quite simple, for a given proximity measure for continuous or binary data and for a chosen topological structure, we can match a topological graph induced on the set of objects.

Consider a set $E = \{x^1, \cdots, x^j, \cdots x^p, y^{11}, \cdots, y^{1m_1}, \ldots, y^{q1}, \cdots, y^{qm_q}\}$ of a mixture variables, p quantitative variables $\{x^1, \cdots, x^j, \cdots x^p\}$ and q qualitative variables $\{y^1, \cdots, y^k, \cdots y^q\}$, where, $m = \sum_{k=1}^{q} m_k$ is the total number of modalities and m_k denotes the number of modalities of the variable y^k.

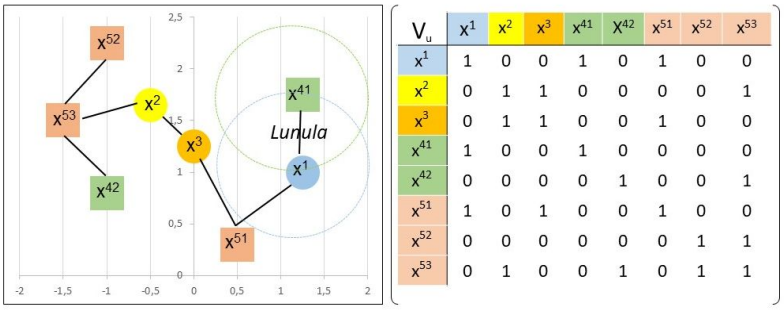
We can, by means of a proximity measure u, define a neighborhood relationship V_u to be a binary relationship on $E \times E$. There are many possibilities for building this neighborhood binary relationship.

Thus, for a given proximity measure u, we can build a neighborhood graph on E, where the vertices are the variables and the edges are defined by a property of the neighborhood relationship.

Many definitions are possible to build this binary neighborhood relationship. One can choose the Minimal Spanning Tree (MST) [15], the Gabriel Graph (GG) [21] or, as is the case here, the Relative Neighborhood Graph (RNG) [27].

For any proximity measure u listed in Table 9 given in the appendix, we construct the associated adjacency binary symmetric matrix V_u of order $p + m$, where, all pairs of neighboring variables in E satisfy the following RNG property:

$$V_u(x^k, x^l) = \begin{cases} 1 \text{ if } u(x^k, x^l) \leq \max[u(x^k, x^t), u(x^t, x^l)]; \\ \qquad\qquad \forall x^k,\ x^l,\ x^t \in E,\ x^t \neq x^k \ \ and \ \ x^t \neq x^l \\ 0 \text{ otherwise.} \end{cases}$$



V_u	x^1	x^2	x^3	x^{41}	x^{42}	x^{51}	x^{52}	x^{53}
x^1	1	0	0	1	0	1	0	0
x^2	0	1	1	0	0	0	0	1
x^3	0	1	1	0	0	1	0	0
x^{41}	1	0	0	1	0	0	0	0
x^{42}	0	0	0	0	1	0	0	1
x^{51}	1	0	1	0	0	1	0	0
x^{52}	0	0	0	0	0	0	1	1
x^{53}	0	1	0	0	1	0	1	1

Fig. 1. RNG structure - Euclidean proximity measure - Associated adjacency matrix

This means that if two variables x^k and x^l which verify the RNG property are connected by an edge, the vertices x^k and x^l are neighbors.

Figure 1 shows an example in $\mathbb{R}^2$ of a set of eight objects, three quantitative variables $\{x^1, x^2, x^3\}$ and five dummy variables $\{x^{41}, x^{42}, x^{51}, x^{52}, x^{53}\}$ of two qualitative variables $\{x^4, x^5\}$, which verify the RNG graph structure with the chosen proximity measure u, the Euclidean distance.

For example, for the first quantitative variable x^1 and the first modality of the first qualitative variable x^{41}, $V_u(x^1, x^{41}) = 1$, it means that on the geometrical plane, the hyper-Lunula (intersection between the two hyperspheres centered on the two variables x^1 and x^{41}) is empty.

For a given neighborhood property (MST, GG or RNG), each measure u generates a topological structure on the objects in E which are totally described by the adjacency binary matrix V_u.

2.1 Reference adjacency matrices

Three topological approaches are described according to the type of variables considered, quantitative or qualitative or a mixture of both.

2.2 Quantitative variables

We assume that we have at our disposal a set $\{x^j; j = 1, \cdots, p\}$ of p quantitative variables and n individuals-objects. The interest lies in whether there is a topological correlation between all the considered variables [1].

We construct the adjacency matrix denoted by $V_{u_\star}$, which corresponds to the correlation matrix. Thus, to examine the correlation structure between the variables, we look at the significance of their linear correlation coefficient. This adjacency matrix can be written as follows using the t-test or Student's t-test of the linear correlation coefficient ρ of Bravais-Pearson:

Definition 1. The reference adjacency matrix $V_{u_\star}$ associated to reference measure $u_\star$ is defined as:

$$V_{u_\star}(x^k, x^l) = \begin{cases} 1 \text{ if } \quad \text{p-value } = P[\mid T_{n-2} \mid > \text{t-value }] \leq \alpha \; ; \; \forall k, l = 1, p \\ 0 \text{ otherwise.} \end{cases}$$

Where p-value is the significance test of the correlation coefficient for the two-sided test of the null and alternative hypotheses, $H_0 : \rho(x^k \, , \, x^l) = 0$ vs. $H_1 : \rho(x^k \, , \, x^l) \neq 0$.

Let T_{n-2} be a t-distributed random variable of Student with $\nu = n - 2$ degrees of freedom. In this case, the null hypothesis is rejected with a p-value less or equal a chosen α significance level, for example $\alpha = 5\%$. Using linear correlation test, if the p-value be very small, it means that there is very small opportunity that null hypothesis is correct, and consequently we can reject it. Statistical significance in statistics is achieved when a p-value is less than a chosen significance level of α. The p-value is the probability of obtaining results which acknowledge that the null hypothesis is true.

2.3 Qualitative variables

We assume that we have at our disposal $\{y^k; k = 1, .., q\}$, a set of $q \geq 2$ qualitative variables and partitions of $n = \sum_{k=1}^{q} n_k$ individuals-objects into

m_k modalities-subgroups. The interest lies in whether there is a topological association between all these variables [4].
- $Y_k = Y_{(n,m_k)}$ the disjonctif table, data matrix associated to the m_k dummy variables of the qualitative variable y^k with n rows-objects and m_k columns-modalities, we check that $\forall_{i=1,n}$, $\Sigma_{k=1}^{m_k} y_i^k = 1$ and $\Sigma_{i=1}^{n} y_i^k = n_k$
- $Y_{(n,m)} = [\, Y_1 | Y_2 | \cdots | Y_q \,]$ the indicator matrix, juxtaposition of the q binary tables Y_k, with n rows-objects and $m = \sum_{k=1}^{q} m_k$ columns-modalities, we check that $\Sigma_{k=1}^{m_k} y_i^k = q$, $\forall_i$ and $\Sigma_{i=1}^{n} \Sigma_{k=1}^{m_k} y_i^k = nq$,
- $\mathcal{B}_{(m,m)} = {}^tY\, Y$ the symmetric Burt matrix of the two-way cross-tabulations of the q variables,

The dissimilarity matrice associated with a proximity measure is computed from data given by the Burt table $\mathcal{B}$. The attributes of any two points' modalities' y^k and y^l in $\{0,1\}^n$ of the proximity measures can be easily written and calculated from the Burt matrix.

A contingency table is one of the most common ways to summarize categorical data. Generally, interest lies in whether there is an association between the row variable and the column variable that produce the table; sometimes there is further interest in describing the strength of that association. The data can arise from several different sampling frameworks, and the interpretation of the hypothesis of no association depends on the framework. The question of interest is whether there is an association between the two variables.

In this case, we build the adjacency matrix $V_{u_\star}$, which corresponds best to the Burt table. Thus, to examine similarities between the modalities we examine the gap between each profile-modality and its average profile, that is, the gap to independence. This best adjacency matrix can be written as follows:

Definition 2. The reference adjacency matrix $V_{u_\star}$ associated to reference measure $u_\star$ is defined as:

$$V_{u_\star}(y^{kr}, y^{ls}) = \begin{cases} 1 \text{ if } \frac{\mathcal{B}_{kr\,ls}}{\mathcal{B}_{kr\,..}} \geq \frac{\mathcal{B}_{kr\,..}}{nq^2}; \;\; \forall k,l = 1,q \;;\; r = 1,m_k \; and \; s = 1,m_l \\ 0 \text{ otherwise.} \end{cases}$$

$\mathcal{B}_{kr\,ls} = \Sigma_{i=1}^{n} y_i^{kr} y_i^{ls}$, element of the Burt matrix that corresponds to the number of individuals who have the modality r of the variable k and the modality s of the variable l,
$\mathcal{B}_{kr\,..} = \Sigma_{l=1}^{q} \Sigma_{s=1}^{m_s} b_{kr\,ls}$ is the row margin of the modality r of the variable k,
$\frac{\mathcal{B}_{kr\,ls}}{\mathcal{B}_{kr\,..}}$ is the row profile of the modality r of the variable k,
$\frac{\mathcal{B}_{kr\,..}}{nq^2}$ is the average profile of the modality r of the variable k, nq^2 being the total number.

2.4 Mixed variables

In this case, the variables for clustering can be a mixture of both quantitative and qualitative variables.

Let $\{x^j; j = 1, \cdots, p\}$ and $\{y^k; k = 1, \cdots, q\}$ be two sets with p quantitative variables and q qualitative variables respectively, with partitions of

$n = \sum_{k=1}^{q} n_k$ individuals-objects into m_k modalities-subgroups which total $m = \sum_{k=1}^{q} m_k$ modalities. The interest lies in whether there is a topological dependency between all the mixed variables.

Simultaneous treatment of mixed data (quantitative and qualitative) cannot be achieved directly by conventional methods of data analysis. So, firstly we transform qualitative data into quantitative data [5]. This transformation is based on multivariate analysis of variance (MANOVA) and on the maximization of the mixed criterion, proposed in terms of correlation squares by Saporta [24] and geometrically in terms of square cosines of angles by Escofier [12]. Then secondly, we build the adjacency matrix $V_{u_\star}$, associated to reference proximity measure $u_\star$, from the correlation matrix of all variables, quantitative and transformed qualitative variables, according to the definition 1. Then secondly, we build the adjacency matrix $V_{u_\star}$, associated to reference proximity measure $u_\star$, from the correlation matrix of all variables, quantitative and transformed qualitative variables, according to Definition 1.

3 Topological clustering of variables - Selective review

Whatever the type of the set of variables considered, the binary and symmetric adjacency matrix build $V_{u_\star}$ is associated with an unknown reference proximity measure $u_\star$.

The robustness according to the α error risk chosen for the null hypothesis: no linear correlation in the case of quantitative variables, or the positive deviation from independence in the case of qualitative variables, can be studied by setting a minimum threshold in order to analyze the sensitivity of the results. Certainly the numerical results will change, but probably not their interpretation.

In order to describe the similarities between variables and to group them into homogeneous groups, we apply the notion of the thémascope or structural analysis of survey data [17], which is a methodological sequence of a clustering method on the principal components of a factorial analysis method. In this case here, it is a topological factorial analysis followed by a Hierarchical Ascendant Classification (HAC). For the topological factorial analysis method, we carry out the classical Multidimensional Scaling (MDS), namely factorial analysis on the similarity table [8], the reference adjacency matrix $V_{u_\star}$ associated with the proximity measure $u*$, the most appropriate measure for the considered data.

Definition 3. The Topological Clustering of Variables (TCV) consist to perform a HAC algorithm based on the Ward criterion[1] [30], on the significant components, of the topological multiple correspondence analysis (TMCA) if

[1]Aggregation based on the criterion of the loss of minimal inertia. Ward's method is a criterion applied in hierarchical cluster analysis; it is a general agglomerative hierarchical clustering procedure. With the square of the Euclidean distance, this criterion allows one to minimize the total within-cluster variance or, equivalently, maximize the between-cluster variance.

the variables are qualitative or of the topological principal component analysis (TPCA) if the variables are quantitative or a mixture of quantitative and qualitative variables.

The TCV hierarchical approach and its dendrogram are easily programmable from the PCA and HAC procedures of the SPAD, SAS or R software.

As for classical methods devoted to the clustering of observations, there are many methods devoted specifically to the clustering of variables, particularly quantitative ones. One of the most used is the Varclus procedure [23] of the SAS software, but we can also apply the ClustOfVar procedure [9] implemented in R, the CVLC procedure [29], clustering around latent variables or the Clustatis procedure [19].

In the case of the TCV of quantitative variables, it is considered that two positively correlated variables are related and that two negatively correlated variables are related, but remote, we will therefore take into account the sign of the correlation between variables. It should be noted that the Varclus procedure implemented in the SAS software, dedicated to the classification of variables, also includes this option. Varclus procedure is more precisely a Hierarchical Descending Classification (HDC).

4 Illustration on real data of simple examples

We illustrate the TCV approach in each of the three types of variables, quantitative, qualitative and mixed variables.

4.1 Case of a set of quantitative variables

The illustrative data table from [13] includes 38 French brands of bottled water described by 8 variables relating to the ion composition (mg/liter). The data comes from the information provided on the bottle labels. The objective is to group together these variables which form a homogeneous set of the ion contents of French brands of bottled water. Simple statistics of these variables are displayed in Table 1.

Table 1. Summary statistics of ion content of French brands of bottled water

Variable	Frequency	Mean	Standard Deviation (N)	Coefficient of variation (%)	Min	Max
CA - Calcium	38	104.184	114.40	109.81	1.00	528.00
MG - Magnesium	38	28.105	29.50	104.95	0.00	95.00
NA - Sodium	38	115.658	210.43	181.94	0.00	968.00
K - Potassium	38	15.079	28.18	186.89	0.00	130.00
SULP - Sulphates	38	119.237	289.83	243.07	1.00	1342.00
NO3 - Nitrates	38	1.842	2.64	143.06	0.00	12.00
HC03 - Carbonates	38	561.368	696.23	124.02	4.00	3380.00
CL - Chlorides	38	40.868	75.35	184.37	0.00	387.00

Figure 2 presents the adjacency matrix $V_{u_\star}$ associated to the proximity measure $u_\star$ adapted to the data considered, is build from the correlations matrix Table 6 given in Appendix, according to Definition 1.

The correlation circle of the two first TPCA factors gives an overview of groups of correlated and uncorrelated variables, an HAC according to Ward's criterion is then applied on the TPCA principal components.

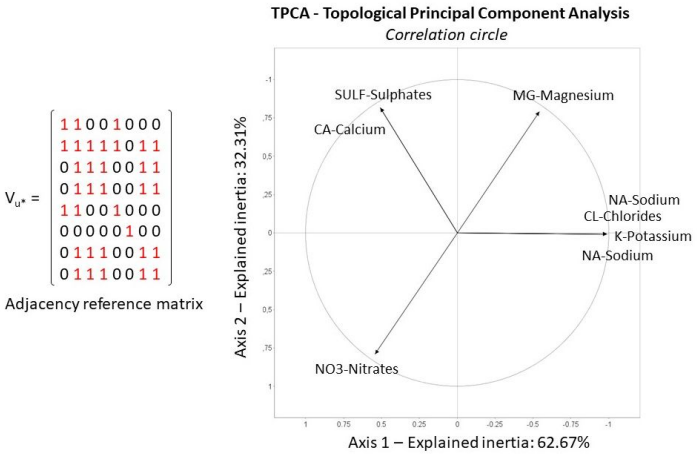


Fig. 2. Representation of the ion composition of French brands of bottled water

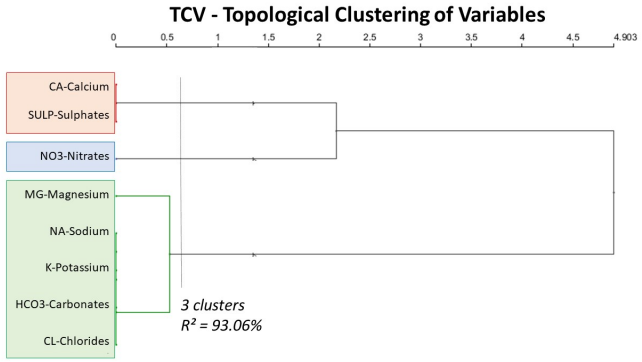


Fig. 3. TCV dendrogram of the ion composition of French brands of bottled water

The dendrogram cluster given in Figure 3 allows to visualize and identify the topological structure of the variables. The aggregation indices of TCV suggests a partition into 3 clusters of the eight variables.

The characterization of the classes by the variables, Table 2, shows with a risk of error less than or equal to 5%, that the first cluster composed of 2 variables, Calcium and Sulfates, are positively correlated and negatively correlated with the variables Sodium, Potassium, Carbonates and Chlorides. The Nitrates variable alone constitutes the second cluster, it is negatively correlated with the Magnesium variable. As for the third cluster, composed of 5 variables, only 4 variables Sodium, Potassium, Carbonates and Chlorides are positively

correlated with each other, the Magnesium variable does not significantly characterize this class.

Cluster Frequency (%)	Cluter 1 2 (25.00%)	Cluster 2 1 (12.50%)	Cluster 3 5 (62.50%)
Profile	CA-Calcium SULF-Sulphates	NO3-Nitrates	NA-Sodium K-Potassium HCO3-Carbonates CL-Chlorides
Anti-profile	NA-Sodium K-Potassium HCO3-Carbonates CL-Chlorides	MG-Magnesium	

Table 2. Characterization of clusters

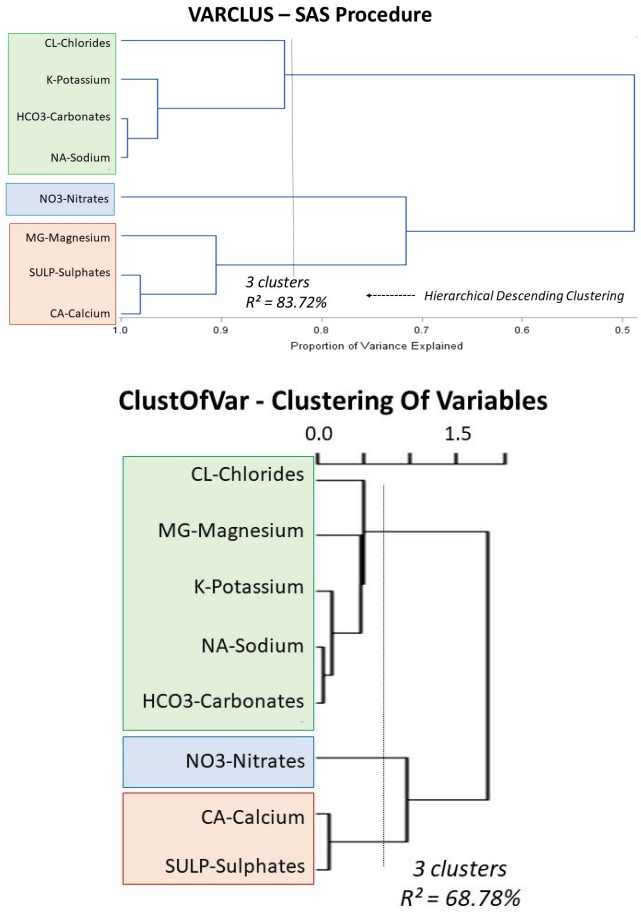


Fig. 4. Varclus and ClustOfVar dendrograms

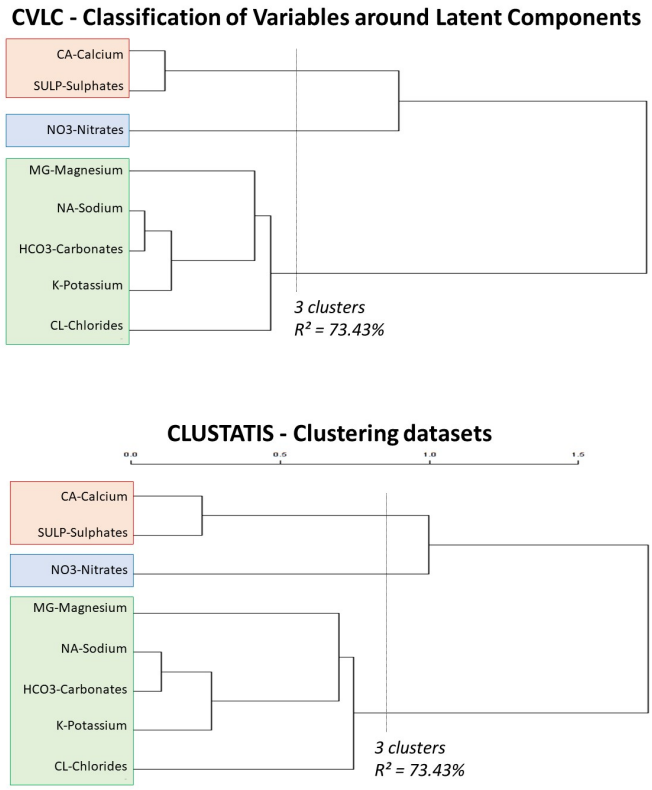


Fig. 5. CVLC and Clustatis dendrograms

From a dimension reduction or variable selection point of view, we can perform in each of the three clusters, a PCA of the variables that characterize it significantly, see Table 2. We can then keep only the first principal component of each of the three PCAs. We thus end up with three synthetic variables of the clusters.

For comparison, Figures 4 and 5 show dendrograms of other variables clustering approaches. Note that for a 3 cluster partition, the constitution of the clusters is the same except for the Varclus approach.

Table 3 presents the percentages of the total variance explained by the 3-cluster partition of the different approaches. The percentage of the TCV approach is much higher than the percentages of the other four approaches, so the TCV clusters are more homogeneous.

4.2 Case of a set of qualitative variables

To illustrate our approach from a set of qualitative variables, we consider a study on female entrepreneurship conducted in Dakar Senegal in 2014. The

Table 3. Comparison criteria

Clustering approach	TCV	Varclus	CVLC	Clustatis	ClustOfVar
R^2 : Variance explained (%)	**93.06**	83.72	73.43	73.43	68.78

data displayed in Table 4 have been collected from 153 female of Dakar region. The objective here is to provide a topological clustering of the demographic characteristics of the female entrepreneurs.

In Figure 6, we can see the adjacency matrix $V_{u_\star}$ associated to the best adapted proximity measure $u_\star$ to the considered data established from the profile Table 7 given in Appendix, according to Definition 2.

Table 4. Burt table - Female Entrepreneurship in Dakar - Senegal

Modality / Variable	Age			Marital status				Number of children			Level of study		
Under 25	22	0	0	18	2	1	1	13	3	6	3	1	18
25 to 50 years	0	80	0	16	9	21	34	14	11	55	58	5	17
Over 50	0	0	51	3	8	24	16	8	35	8	30	10	11
Single	18	16	3	37	0	0	0	20	3	14	9	1	27
Divorcee	2	9	8	0	19	0	0	3	10	6	13	5	1
Monogamous bride	1	21	24	0	0	46	0	7	21	18	26	5	15
Polygamous bride	1	34	16	0	0	0	51	5	15	31	43	5	3
No children	13	14	8	20	3	7	5	35	0	0	11	5	19
From 1 to 3 children	3	11	35	3	10	21	15	0	49	0	27	9	13
More than 3 children	6	55	8	14	6	18	31	0	0	69	53	2	14
Illiterate-Primary	3	58	30	9	13	26	43	11	27	53	91	0	0
Secondary	1	5	10	1	5	5	5	5	9	2	0	16	0
Higher	18	17	11	27	1	15	3	19	13	14	0	0	46

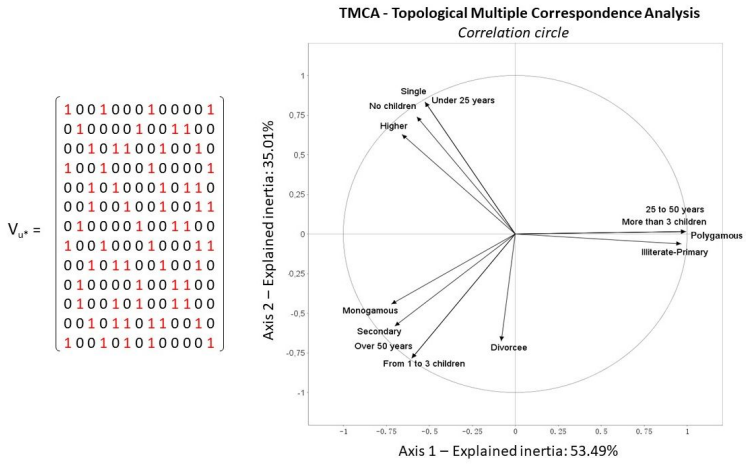


Fig. 6. TCV : The demographic characteristics of women entrepreneurs

The representation on the first principal plane of TMCA gives a first view of the linked groups of modalities, then, a Ward's HAC was performed on the TMCA principal components.

Figure 7 shows the dendrogram of the thirteen demographic characteristics of the female entrepreneurs. We choose according the dendrogram to cut this hierarchical tree into 3 clusters.

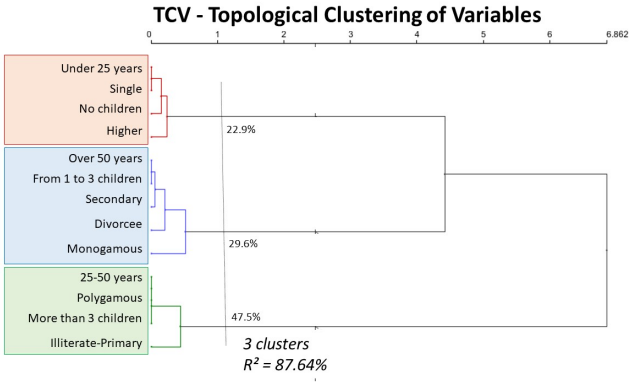


Fig. 7. TCV : Dendrogram of the demographic characteristics of women entrepreneurs

Figure 8 shows the hierarchical dendrogram obtained by the Corresp and Cluster procedures of SAS software, its percentage of total explained inertia (32.90%) is much lower than that of the TCV approach (87.64%) for a partition into 3 clusters.

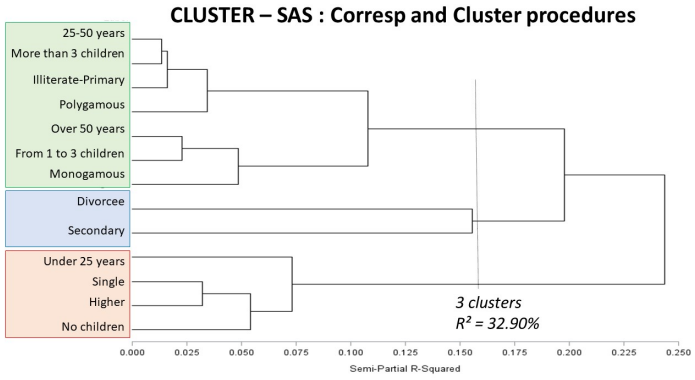


Fig. 8. CLUSTER - Tree Diagram of the demographic characteristics of women entrepreneurs

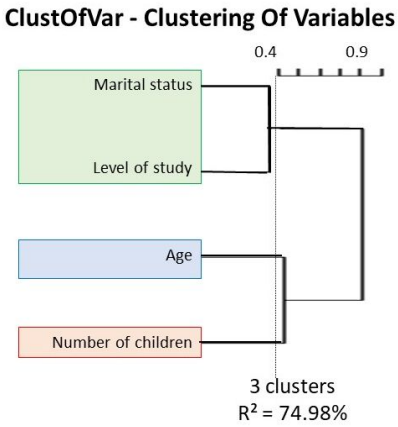


Fig. 9. ClusOfVar - Tree Diagram of the demographic characteristics of women entrepreneurs

Figure 9 is given as an indication and not for comparison, the ClusOfVar approach partitions the qualitative variables and not the modalities of the variables as is the case with the TCV and Cluster approaches.

4.3 Case of a set of mixed variables

In some real data situations, variables of a thematic are measured on different scales with at a mixture of quantitative and qualitative variables.

Table 5. Summary Statistics and frequency distributions

Quantitative variable	Frequency	Mean	Std Dev (N)	Minimum	Maximum
Urban Consumption	27	7.14	1.12	5.60	9.30
Cubic Capacity	27	1165.63	204.17	903.00	1597.00
Maximum Speed	27	154.26	21.94	115.000	200.00
Boot Volume	27	901.41	301.67	202.00	1200.00
Weight/Power	27	18.65	5.42	10.20	33.10
Length	27	3.62	0.07	3.40	3.70
Qualitative variable	Modality	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Horsepower	HP4	13	48.15	13	48.15
	HP5	5	18.52	18	66.67
	HP6	9	33.33	27	100.00
Brand Country Manufacturer	French	10	37.04	10	37.04
	Foreign	17	62.96	27	100.00
Price	Price1	10	37.04	10	37.04
	Price2	5	18.52	15	55.56
	Price3	8	29.63	23	85.19
	Price4	4	14.81	27	100.00

To illustrate this approach, we take the data published in [16], they cover a sample of 27 small cars of the Belgian market. We have a homogeneous theme of nine mixed variables of which six quantitative and three qualitative characteristics totaling nine modalities.

The objective here is to synthesize simultaneously in the sense of correlations all of these mixed characteristics. Table 5 summarizes the elementary statistics of the mixed variables.

Figure 10 gives the adjacency matrix $V_{u_\star}$ associated to the adapted proximity measure $u_\star$ for the considered data, build from the correlations matrix, see Table 8 given in Appendix, according to Definition 1. The correlation circle of the two first TPCA factors gives an overview of groups of correlated and uncorrelated quantitative and modalities of qualitative variables. An HAC according to Ward's criterion is then applied on the TPCA components represented by the dendrogram of the characteristics of small cars on the Belgian market presented in Figure 11.

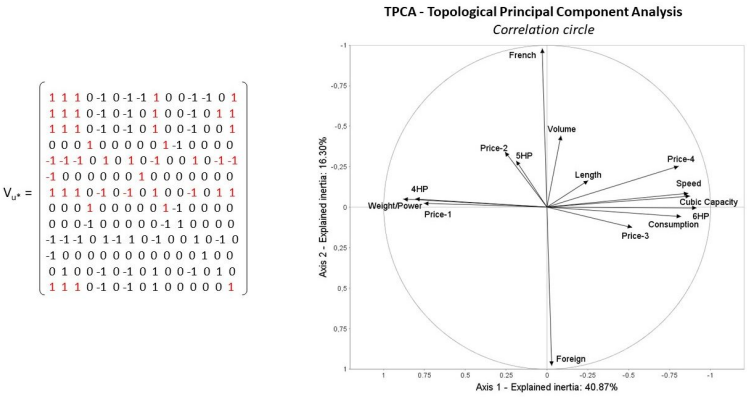


Fig. 10. Representation of the cars characteristics

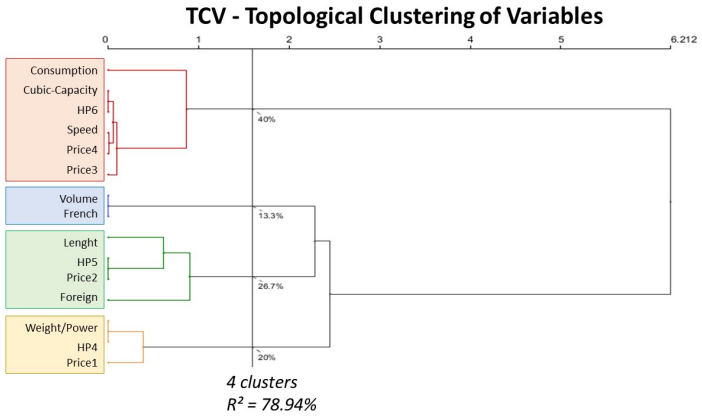


Fig. 11. TCV - Car characteristics dendrogram

The TCV percentage of total explained inertia is equal to 78.94% for the partition into 4 classes.

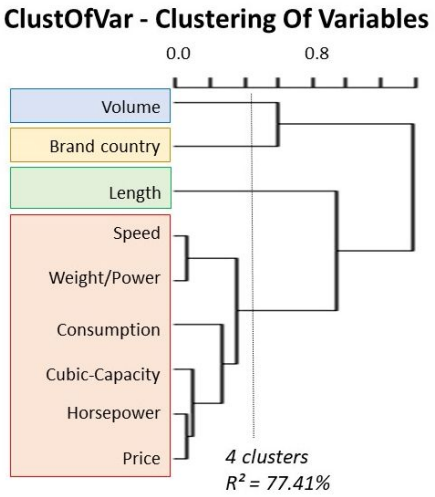


Fig. 12. ClusOfVar - Car characteristics dendrogram

Figure 12 presents, as an indication and not for comparison, the tree diagram of hierarchical clusters of the ClustOfVar approach; the latter considers the qualitative variables and not their modalities.

5 Conclusion

This paper proposes a new topological method of clustering of variables which enriches the methods of data analysis within the framework of the clustering of a set of quantitative or qualitative variables or a mixture of both. The results the proposed topological approach of classifying variables, based on the notion of neighborhood graph, are as well as good, or event better according to the R square than those of the existing methods. This approach is be easily implemented on SAS, SPAD or R software. Future work consists in extending this topological approach to other methods of data analysis, in particular in the context of evolutionary data analysis, both in the concept of a factorial analysis and that of a clustering as well individuals and variables.

6 Appendix

Table 6. Pearson correlation matrix (p-values)

Variable	CA	MG	NA	K	SULF	NO3	HCO3	CL
CA	1.0000							
MG	0.6672	1.0000						
	(< .0001)							
NA	0.0042	0.5649	1.0000					
	(0.9757)	(< .0001)						
K	0.1072	0.6703	0.8817	1.0000				
	(0.4358)	(< .0001)	(< .0001)					
SULF	0.8997	0.5629	-0.0957	-0.0546	1.0000			
	(< .0001)	(< .0001)	0.4872	0.6923				
NO3	-0.0473	-0.1756	-0.0830	-0.1529	-0.1288	1.0000		
	(0.7317)	(0.1998)	0.5469	0.2650	0.3486			
HCO3	0.1491	0.6583	0.9474	0.8866	-0.0573	-0.0541	1.0000	
	0.2774	(< .0001)	(< .0001)	(< .0001)	0.6776	0.6947		
CL	0.0578	0.52094	0.5646	0.7187	-0.0276	-0.1053	0.4794	1.0000
	0.6749	(< .0001)	(< .0001)	(< .0001)	0.8406	0.4443	(0.0002)	

Table 7. Row and Average profiles

Row-Profiles	Age			Marital status				Number of child			Level of study		
Under 25 years	0.25	0	0	0.205	0.023	0.011	0.011	0.148	0.034	0.068	0.034	0.011	0.205
25 to 50 years	0	0.25	0	0.050	0.028	0.066	0.106	0.044	0.034	0.172	0.181	0.016	0.053
Over 50 years	0	0	0.25	0.015	0.039	0.118	0.078	0.039	0.172	0.039	0.147	0.049	0.054
Single	0.122	0.108	0.020	0.25	0	0	0	0.135	0.020	0.095	0.061	0.007	0.182
Divorcee	0.026	0.118	0.105	0	0.25	0	0	0.040	0.132	0.079	0.171	0.066	0.013
Monogamous	0.005	0.114	0.130	0	0	0.25	0	0.038	0.114	0.098	0.141	0.027	0.082
Polygamous	0.005	0.167	0.078	0	0	0	0.25	0.025	0.074	0.152	0.211	0.025	0.015
No children	0.093	0.100	0.057	0.143	0.021	0.050	0.036	0.25	0	0	0.079	0.036	0.136
From 1 to 3 child	0.015	0.056	0.179	0.015	0.051	0.107	0.077	0	0.25	0	0.138	0.046	0.066
More than 3 child	0.022	0.199	0.029	0.051	0.022	0.065	0.112	0	0	0.25	0.192	0.007	0.051
Illiterate-Primary	0.008	0.159	0.082	0.025	0.036	0.071	0.118	0.030	0.074	0.146	0.25	0	0
Secondary	0.016	0.078	0.156	0.016	0.078	0.078	0.078	0.078	0.141	0.031	0	0.25	0
Superior	0.098	0.092	0.060	0.147	0.005	0.082	0.016	0.103	0.071	0.076	0	0	0.25
Average profile	0.036	0.131	0.083	0.061	0.031	0.075	0.083	0.057	0.080	0.113	0.149	0.026	0.075

Table 8. Pearson correlation matrix (p-values)

Variable	Consumption	Cubic capacity	Speed	Volume	Weight/Power	Length	HP4	HP5	HP6	French	Foreign	Price1	Price2	Price3	Price4	Variable
Consumption	1.0000															Consumption
Cubic Capacity	0.7966 (< .0001)	1.0000														Cubic capacity
Speed	0.78044 (< .0001)	0.83222 (< .0001)	1.0000													Speed
Volume	0.2946 (0.1358)	0.1125 (0.5766)	0.0220 (0.9134)	1.0000												Volume
Weight/ Power	-0.6825 (< .0001)	-0.7788 (< .0001)	-0.9376 (< .0001)	0.1020 0.6127	1.0000											Weight/Power
Length	0.1966 (0.3258)	0.2897 (0.1427)	0.1552 (0.4396)	-0.0734 (0.7161)	-0.0977 (0.6280)	1.0000										Length
HP4	-0.5481 (0.0031)	-0.8376 (< .0001)	-0.6602 (0.0002)	-0.1048 (0.6031)	0.6091 (0.0007)	-0.3058 (0.1208)	1.0000									HP4
HP5	-0.3819 (0.0494)	0.0611 (0.7621)	-0.1447 (0.4714)	-0.2655 (0.1807)	0.1083 (0.5909)	0.1650 (0.4109)	-0.4594 (0.0159)	1.0000								HP5
HP6	0.8957 (< .0001)	0.8375 (< .0001)	0.8190 (< .0001)	0.3298 (0.0930)	-0.7348 (< .0001)	0.1882 (0.3472)	-0.6814 (< .0001)	-0.3371 (0.0855)	1.0000							HP6
French	-0.0939 (0.6415)	0.0431 (0.8310)	-0.0100 (0.6198)	0.4093 (0.0340)	0.2039 (0.3077)	0.1488 (0.4589)	-0.1251 (0.5342)	0.2267 (0.2555)	-0.0542 (0.7882)	1.0000						French
Foreign	0.0939 (0.6415)	-0.0431 (0.8310)	0.0100 (0.6198)	-0.4093 (0.0340)	-0.2039 (0.3077)	-0.1488 (0.4589)	0.1251 (0.5342)	-0.2267 (0.2555)	0.0542 (0.7882)	-1.0000 (< .0001)	1.0000					Foreign
Price1	-0.4020 (0.0376)	-0.6767 (0.0001)	-0.6035 (0.0009)	-0.0110 (0.9568)	0.5475 (0.0031)	-0.4024 (0.0374)	0.7959 (< .0001)	-0.3656 (.0607)	-0.5423 (0.0035)	-0.1118 (0.5789)	0.1118 (0.5789)	1.0000				Price1
Price2	-0.3904 (0.0441)	-0.2901 (0.1422)	-0.2969 (0.1327)	-0.1015 (0.6145)	0.3086 (0.1173)	0.1254 (0.5330)	0.1131 (0.5744)	0.2636 (0.1839)	-0.3371 (0.0855)	0.2267 (0.2555)	-0.2267 (0.2555)	-0.3656 (0.0607)	1.0000			Price2
Price3	0.2465 0.2151	0.4807 0.0112	0.3251 0.0980	-0.0232 0.9086	-0.3901 0.0443	0.2313 0.2458	-0.6253 (0.0005)	0.3171 (0.1071)	0.4015 (0.0379)	-0.3297 (0.0931)	0.3297 (0.0931)	-0.4977 (0.0083)	-0.3093 (0.1164)	1.0000		Price3
Price4	0.6565 (0.0002)	0.6191 0.0006	0.7270 (< .0001)	0.1557 (0.4382)	-0.5803 (0.0015)	0.1126 (0.5760)	-0.4019 (0.0377)	-0.1988 (0.3202)	0.5898 (0.0012)	0.3278 (0.0950)	-0.3278 (0.0950)	-0.3199 (0.1039)	-0.1988 (0.3202)	-0.2706 (0.1722)	1.0000	Price4
Variable																

Table 9. Some proximity measures for continuous and binary data

Measure	**Distance and Dissimilarity for continuous data**
Euclidean	$u_{Euc}(x,y) = \sqrt{\sum_{j=1}^{p}(x_j - y_j)^2}$
Manhattan	$u_{Man}(x,y) = \sum_{j=1}^{p} \lvert x_j - y_j \rvert$
Minkowski	$u_{Min_\gamma}(x,y) = (\sum_{j=1}^{p} \lvert x_j - y_j \rvert^\gamma)^{\frac{1}{\gamma}}$
Tchebychev	$u_{Tch}(x,y) = \max_{1\leq j\leq p} \lvert x_j - y_j \rvert$
Normalized Euclidean	$u_{NE}(x,y) = \sqrt{\sum_{j=1}^{p} \frac{1}{\sigma_j^2}[(x_j - \overline{x}_j) - (y_j - \overline{y}_j)]^2}$
Cosine dissimilarity	$u_{Cos}(x,y) = 1 - \frac{\sum_{j=1}^{p} x_j y_j}{\sqrt{\sum_{j=1}^{p} x_j^2}\sqrt{\sum_{j=1}^{p} y_j^2}} = 1 - \frac{<x,y>}{\|x\|\|y\|}$
Canberra	$u_{Can}(x,y) = \sum_{j=1}^{p} \frac{\lvert x_j - y_j \rvert}{\lvert x_j \rvert + \lvert y_j \rvert}$
Pearson Correlation	$u_{Cor}(x,y) = 1 - \frac{(\sum_{j=1}^{p}(x_j - \overline{x})(y_j - \overline{y}))^2}{\sum_{j=1}^{p}(x_j - \overline{x})^2 \sum_{j=1}^{p}(y_j - \overline{y})^2} = 1 - \frac{(<x-\overline{x}, y-\overline{y}>)^2}{\|x-\overline{x}\|^2 \|y-\overline{y}\|^2}$
Squared Chord	$u_{Cho}(x,y) = \sum_{j=1}^{p}(\sqrt{x_j} - \sqrt{y_j})^2$
Doverlap measure	$u_{Dev}(x,y) = max(\sum_{j=1}^{p} x_j, \sum_{j=1}^{p} y_j) - \sum_{j=1}^{p} min(x_j, y_j)$
Weighted Euclidean	$u_{WEu}(x,y) = \sqrt{\sum_{j=1}^{p} \alpha_j (x_j - y_j)^2}$
Gower's Dissimilarity	$u_{Gow}(x,y) = \frac{1}{p} \sum_{j=1}^{p} \mid x_j - y_j \mid$
Shape Distance	$u_{Sha}(x,y) = \sqrt{\sum_{j=1}^{p}[(x_j - \overline{x}_j) - (y_j - \overline{y}_j)]^2}$
Size Distance	$u_{Siz}(x,y) = \mid \sum_{j=1}^{p}(x_j - y_j) \mid$
Lpower	$u_{Lpo_\gamma}(x,y) = \sum_{j=1}^{p} \lvert x_j - y_j \rvert^\gamma$

Where, p is the dimension of space, $x = (x_j)_{j=1,\ldots,p}$ and $y = (y_j)_{j=1,\ldots,p}$ two points in R^p, $\overline{x}_j$ the mean,σ_j the Standard deviation, $\alpha_j = \frac{1}{\sigma_j^2}$ and $\gamma > 0$.

Measure	**Similarity and Dissimilarity for binary data**	
Jaccard	$s_1 = \frac{a}{a+b+c}$	$u_1 = 1 - s_1$
Dice, Czekanowski	$s_2 = \frac{2a}{2a+b+c}$	$u_2 = 1 - s_2$
Kulczynski	$s_3 = \frac{1}{2}(\frac{a}{a+b} + \frac{a}{a+c})$	$u_3 = 1 - s_3$
Driver, Kroeber and Ochiai	$s_4 = \frac{a}{\sqrt{(a+b)(a+c)}}$	$u_4 = 1 - s_4$
Sokal and Sneath 2	$s_5 = \frac{a}{a+2(b+c)}$	$u_5 = 1 - s_5$
Braun-Blanquet	$s_6 = \frac{a}{max(a+b,a+c)}$	$u_6 = 1 - s_6$
Simpson	$s_7 = \frac{a}{min(a+b,a+c)}$	$u_7 = 1 - s_7$
Kendall, Sokal-Michener	$s_8 = \frac{a+d}{a+b+c+d}$	$u_8 = 1 - s_8$
Russell and Rao	$s_9 = \frac{a}{a+b+c+d}$	$u_9 = 1 - s_9$
Rogers and Tanimoto	$s_{10} = \frac{a+d}{a+2(b+c)+d}$	$u_{10} = 1 - s_{10}$
Pearson ϕ	$s_{11} = \frac{ad-bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$	$u_{11} = \frac{1-s_{11}}{2}$
Hamann	$s_{12} = \frac{a+d-b-c}{a+b+c+d}$	$u_{12} = \frac{1-s_{12}}{2}$
Sokal and Sneath 5	$s_{13} = \frac{ad}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$	$u_{13} = 1 - s_{13}$
Michael	$s_{14} = \frac{4(ad-bc)}{(a+d)^2+(b+c)^2}$	$u_{14} = \frac{1-s_{14}}{2}$
Baroni, Urbani and Buser	$s_{15} = \frac{a+\sqrt{ad}}{a+b+c+\sqrt{ad}}$	$u_{15} = 1 - s_{15}$
Yule Q	$s_{16} = \frac{ad-bc}{ad+bc}$	$u_{16} = \frac{1-s_{16}}{2}$
Yule Y	$s_{17} = \frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}$	$u_{18} = \frac{1-s_{18}}{2}$
Sokal and Sneath 4	$s_{18} = \frac{1}{4}(\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{d+b} + \frac{d}{d+c})$	$u_{18} = 1 - s_{18}$
Gower and Legendre	$s_{19} = \frac{a+d}{a+\frac{(b+c)}{2}+d}$	$u_{19} = 1 - s_{19}$
Sokal and Sneath 1	$s_{20} = \frac{2(a+d)}{2(a+d)+b+c}$	$u_{20} = 1 - s_{20}$

Where, $a = \mid X \cap Y \mid$ is the number of attributes common to both points x and y, $b = \mid X - Y \mid$ is the number of attributes present in x but not in y, $c = \mid Y - X \mid$ is the number of attributes present in y but not in x and $d = \mid \overline{X} \cap \overline{Y} \mid$ is the number of attributes in neither x or y and $\mid . \mid$ the cardinality of a set.

References

1. Abdesselam, R.: A Topological Principal Component Analysis. *International Journal of Data Science and Analysis*. Science Publishing Group, USA, Vol.7, Issue 2, pp.20-31, 2021.
2. Abdesselam, R.: Selection of proximity measures for a Topological Correspondence Analysis. *Data Analysis and Applications 3, Computational, Classification, Financial, Statistical and Stochastic Methods, Vol.5, Part 2. Classification Data Analysis and Methods*. Wiley,103–120, 2020.
3. Abdesselam, R.: A Topological Multiple Correspondence Analysis. *Journal of Mathematics and Statistical Science*, ISSN 2411-2518, Science Signpost Publishing Inc., USA, Vol.5, Issue 8, 175–192, 2019.
4. Abdesselam, R.: A Topological Discriminant Analysis. *Data Analysis and Applications 2, Utilization of Results in Europe and Other Topics*, Vol.3, Part 4. Wiley, 167–178, 2019.
5. Abdesselam, R.: Analyse en Composantes Principales Mixte. Classification : points de vue croisés, RNTI-C-2, *Revue des Nouvelles Technologies de l'Information* RNTI, Cépaduès Editions, 31-41, 2008.
6. Batagelj, V., Bren, M.: Comparing resemblance measures. *In Journal of classification*, 12, 73–90, 1995.
7. Benzécri, J.P.: L'Analyse des Données. Tome 1 : La Taxinomie. Tome 2 : L'analyse des correspondances, *"2ème édition Dunod, Paris*, 1976.
8. Caillez, F. and Pagès, J.P.: Introduction à l'Analyse des données. *S.M.A.S.H., Paris*, 1976.
9. Chavent M., Kuentz V., Liquet B., Saracco J., ClustOfVar: An R Package forthe Clustering of Variables, *Journal of Statistical Software*. Vol. 50, 1-16, 2012.
10. Cohen, J.: A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, Vol 20, 27–46, 1960.
11. Demsar, J.: Statistical comparisons of classifiers over multiple data sets. *The journal of Machine Learning Research*, Vol. 7, 1–30, 2006.
12. Escofier, B. : Traitement simultané de variables qualitatives et quantitatives en analyse factorielle. *Cahier de l'analyse des données*, Vol.4(2), 137-146, 1979.
13. Govaert, G.: Analyse des données. *Hermes*, 19–42, 2003.
14. Hubert, L. and Arabie, P.: Comparing partitions. *Journal of Classification*, 193-–218, 1985.
15. Kim, J.H. and Lee, S.: Tail bound for the minimal spanning tree of a complete graph. *In Statistics & Probability Letters*, 4, 64, 425–430, 2003.
16. Lambin, J.J.: La recherche marketing, Analyser - Mesurer - Prévoir. McGraw-Hill, 1990.
17. Lebart, L.: Stratégies du traitement des données d'enquêtes. *La Revue de MODULAD, 3, 21–29, 1989.*
18. *Lesot, M. J., Rifqi, M. and Benhadda, H.: Similarity measures for binary and numerical data: a survey.* In IJKESDP*, 1, 1, 63-84, 2009.*
19. *Llobell, F. and Qannari, E.M.: Clustering datasets by means of Clustatis with identification of atypical datasets. Application to sensometrics.* Food Quality and Preference*, Elsevier, 75, 97–104, 2019.*
20. *Mantel, N.: A technique of disease clustering and a generalized regression approach.* In Cancer Research*, 27, 209–220, 1967.*
21. *Park, J. C., Shin, H. and Choi, B. K.: Elliptic Gabriel graph for finding neighbors in a point set and its application to normal vector estimation.* In Computer-Aided Design Elsevier*, 38, 6, 619–626, 2006.*

22. *Rifqi, M., Detyniecki, M. and Bouchon-Meunier, B.: Discrimination power of measures of resemblance.* IFSA'03 Citeseer*, 2003.*
23. *SAS Institute Inc. SAS/STAT Software, the VARCLUS Procedure, URL http://support.sas.com/documentation/onlinedoc/stat/930/varclus.pdf.*
24. *Saporta, G.: Simultaneous treatment of quantitative and qualitative data.* In Attidela XXXV Riunone scientifica; Società Italiana di Statistica*, 63–72, 1990.*
25. *Schneider, J. W. and Borlund, P.: Matrix comparison, Part 1: Motivation and important issues for measuring the resemblance between proximity measures or ordination results.* In Journal of the American Society for Information Science and Technology*, 58, 11, 1586–1595, 2007.*
26. *Schneider, J. W. and Borlund, P.: Matrix comparison, Part 2: Measuring the resemblance between proximity measures or ordination results by use of the Mantel and Procrustes statistics.* In Journal of the American Society for Information Science and Technology*, 11, 58, 1596–1609, 2007.*
27. *Toussaint, G. T.: The relative neighbourhood graph of a finite planar set.* In Pattern recognition*, 12, 4, 261–268, 1980.*
28. *Vigneau, E., Qannari, E.M.: Classification of variables around latent components.* Communications in statistics Simulation and Computation*, 32(4), 1131-1150, 2003.*
29. *Vigneau, E., Qannari, E.M., Sahmer. K. and Ladiray, D.: Classification de variables autour de composantes latentes* Revue de statistique appliquée*, tome 54, no 1, 27-45, 2006.*
30. *Ward, J. R.: Hierarchical grouping to optimize an objective function.* In Journal of the American statistical association JSTOR*, 58, 301, 236–244, 1963.*
31. *Warrens, M. J.: Bounds of resemblance measures for binary (presence/absence) variables. In Journal of Classification,* Springer*, 25, 2, 195–208, 2008.*
32. *Zighed, D., Abdesselam, R., and Hadgu, A.: Topological comparisons of proximity measures.* In the 16th PAKDD 2012 Conference*. In P.-N. Tan et al., Eds. Part I, LNAI 7301, Springer-Verlag Berlin Heidelberg, 379–391, 2012.*