A Topological Clustering on Evolutionary Data

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 objective of this paper is to propose a topological approach of clustering in evolutionary data analysis. We are interested in clustering resulting from exploratory methods of joint analysis of several data tables, methods applied more particularly to temporal data.

The clustering is one of the most widely used approaches to exploring multidimensional data. The two 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 approach, called Topological Clustering on Evolutionary Data (TCED), is based on the notion of neighborhood graphs in an evolutionary data context. It makes it possible to simultaneously explore several tables of data collected at different times on the same individual-rows, the variables possibly being different according to the tables considered. The columns-variables of each table are more-or-less correlated or linked according to whether the variable type. It analyzes in each table the structure of the correlations or associations observed between the variables according to their quantitative, qualitative type or a mixture of both.

The proposed TCED approach is presented and illustrated here using a real dataset with quantitative variables. Its results are compared with those resulting from the unsupervised clustering on evolutionary data analysis method - Multiple Factorial Analysis (MFA).

Keywords: Evolutionary data cluster, proximity measure, neighborhood graph, adjacency matrix, hierarchical clustering, clustering index.

1 Introduction

The aim of cluster analysis is to group objects into homogeneous classes, it is an example of unsupervised learning as we don't know how many groups will be formed. The other alternative that is supervised learning is also called pattern recognition, in this case, the number of groups is known.

The objective of this article is to propose a topological approach of data analysis applied to data tables crossing the same individuals with possibly different variables, quantitative, qualitative or mixed.

The proposed approach, called Topological Clustering on Evolutionary Data (TCED) is different from those that already exist, in particular the clustering on the results of the Multiple Factorial Analysis (MFA) [9–11] with which it is compared, or also on the results of the Structuring Tables with Three Indices of the Statistic (STATIS) method [17,19] or the Double Principal Component Analysis (DPCA) method [6].

There are topological approaches specifically devoted to the clustering [1,2,21] but as far as we know, none of these approaches has been proposed to analyze several data tables simultaneously. We can also cite the evolutionary data clustering approach proposed in [4] but not in a topological context.

The choice of proximity measure among the many existing measures plays an important role in multidimensional data analysis [5,18,28]. It has a strong impact on the results of any operation of structuring, grouping or classification of objects.

This study proposes an evolutionary topological classification of individuals, generally over time, regardless of the type of variables considered: quantitative, qualitative or a mixture of both.

The structure of correlation or dependence of the quantitative or qualitative variables of each evolutionary or temporal data table, depends on the considered data. Results may change depending on the proximity measure chosen for each data table. A proximity measure is a function that measures the similarity or dissimilarity between two objects or variables within a set.

This document is organized as follows. In section 2, we briefly recall the basic notion of neighborhood graphs, we define and show how to construct adjacency matrices associated with proximity measures within the framework of the analysis of the correlation structure of a set of evolving data tables, and we present the principle of the TCED approach. This is illustrated in section 3 using an example based on real evolutionnary data. The results of the TCED are compared with those of the classification applied to the results of the MFA. Finally, section 4 presents concluding remarks on this work.

2 Topological and evolutionary data contexts

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.

Topological analysis on evolutionary data consists of simultaneously analyzing several data tables $(X_t)_{t=1,T}$ collected at different times on the same individuals, the variables can be the same or different according to the tables.

We consider at time t, $E_t = \{x^1, \cdots, x^j, \cdots, x^{p_t}\}$ a set of p_t quantitative variables of the data table X_t. We can see in [2] cases of qualitative or even mixed variables.

We can, by means of a proximity measure u_t, define a neighborhood relationship, V_{u_t}, to be a binary relationship based on $E_t \times E_t$. There are many possibilities for building this neighborhood binary relationship.

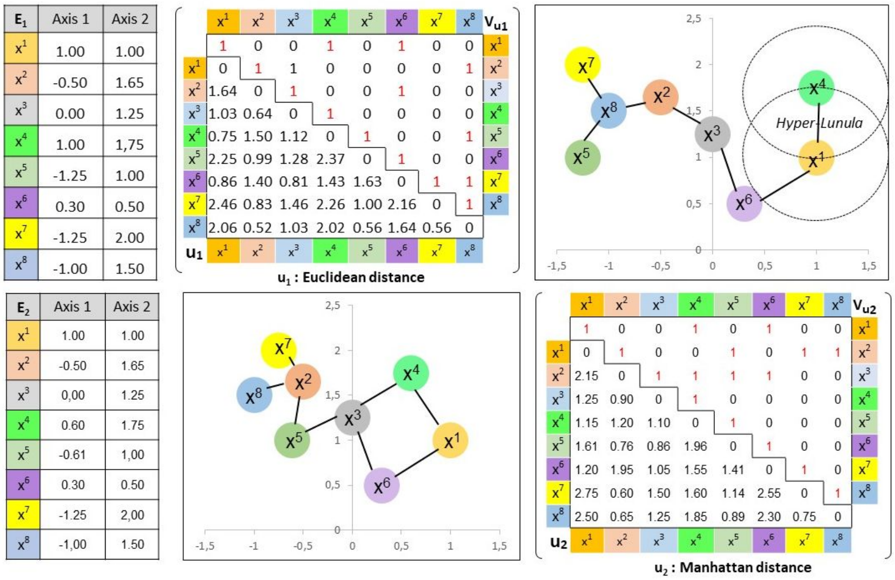
Thus, for a given proximity measure u_t, we can build a neighborhood graph on E_t, where the vertices are the variables and the edges are defined by a property of the neighborhood relationship.

Many definitions are possible to build this binary neighborhood relationship. One can choose the Minimal Spanning Tree (MST) [14], the Gabriel Graph (GG) [22] or, as is the case here, the Relative Neighborhood Graph (RNG) [25].

Given a set E_t of p_t variables of the data table X_t and a proximity measure u_t, for continuous or binary data, we can construct the associated adjacency binary symmetric matrix V_{u_t} of order p_t, where, all pairs of neighboring variables in E_t satisfy the following RNG property:

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

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.



E_1	Axis 1	Axis 2
x^1	1.00	1.00
x^2	-0.50	1.65
x^3	0.00	1.25
x^4	1.00	1,75
x^5	-1.25	1.00
x^6	0.30	0.50
x^7	-1.25	2.00
x^8	-1.00	1.50

	x^1	x^2	x^3	x^4	x^5	x^6	x^7	x^8	V_{u1}
	1	0	0	1	0	1	0	0	x^1
x^1	0	1	1	0	0	0	0	1	x^2
x^2	1.64	0	1	0	0	1	0	0	x^3
x^3	1.03	0.64	0	1	0	0	0	0	x^4
x^4	0.75	1.50	1.12	0	1	0	0	1	x^5
x^5	2.25	0.99	1.28	2.37	0	1	0	0	x^6
x^6	0.86	1.40	0.81	1.43	1.63	0	1	1	x^7
x^7	2.46	0.83	1.46	2.26	1.00	2.16	0	1	x^8
x^8	2.06	0.52	1.03	2.02	0.56	1.64	0.56	0	
u_1	x^1	x^2	x^3	x^4	x^5	x^6	x^7	x^8	

u_1 : Euclidean distance

E_2	Axis 1	Axis 2
x^1	1.00	1.00
x^2	-0.50	1.65
x^3	0,00	1.25
x^4	0.60	1.75
x^5	-0.61	1,00
x^6	0.30	0.50
x^7	-1.25	2,00
x^8	-1,00	1.50

	x^1	x^2	x^3	x^4	x^5	x^6	x^7	x^8	V_{u2}
	1	0	0	1	0	1	0	0	x^1
x^1	0	1	0	0	1	0	1	1	x^2
x^2	2.15	0	1	1	1	1	0	0	x^3
x^3	1.25	0.90	0	1	0	0	0	0	x^4
x^4	1.15	1.20	1.10	0	1	0	0	0	x^5
x^5	1.61	0.76	0.86	1.96	0	1	0	0	x^6
x^6	1.20	1.95	1.05	1.55	1.41	0	1	0	x^7
x^7	2.75	0.60	1.50	1.60	1.14	2.55	0	1	x^8
x^8	2.50	0.65	1.25	1.85	0.89	2.30	0.75	0	
u_2	x^1	x^2	x^3	x^4	x^5	x^6	x^7	x^8	

u_2 : Manhattan distance

Fig. 1. RNG - Euclidean and Manhattan distances - Adjacency matrices

Figure 1 shows a simple example in $\mathbb{R}^2$ of two variable sets E_1 and E_2 of the same eight quantitative variables, which verify the structure of the RNG graph with the Euclidean distance $u_1(x^k\ ,\ x^l) = \sqrt{\sum_{j=1}^{2}(x_j^k - x_j^l)^2}$ for the data table X_1 at time $t = 1$ and Manhattan distance $u_2(x^k, x^l) = \sum_{j=1}^{p} |x_j^k - x_j^l|$ for the data table X_2 at time $t = 2$, as well as the associated binary adjacency matrices V_{u_1} and V_{u_2}.

For example, at time $t = 1$, we can see that for the first and the fourth variables, $V_{u_1}(x^1, x^4) = 1$, it means that on the geometrical plane, the hyper-Lunula (intersection between the two hyperspheres centered on the two variables x^1 and x^4) is empty.

This generates a topological structure based on the objects in E_t which are completely described by the adjacency binary matrix V_{u_t}.

For a given neighborhood property (MST, GG or RNG), each measure u_t generates a topological structure on the objects in E_t which are totally described by the adjacency binary matrix V_{u_t}.

2.1 Reference adjacency matrices

The objective is initially, to analyze in a topological and evolutionary way the correlation structures of the variables [2] of the data tables considered, then to establish on this analysis, a clustering of the evolutionary individuals.

At time t, we construct the reference adjacency matrix noted $V_{u_\star t}$, in the case of quantitative variables, from the correlation matrix of data table X_t. The expressions of the suitable adjacency reference matrices in the case of qualitative variables or mixed variables are given in [2,3].

To examine the correlation structure between the variables of data table X_t, 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. For quantitative variables, the reference adjacency matrix $V_{u_t^\star}$ associated to reference measure $u_t^\star$ is defined as:

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

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

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.

Whatever the type of variable set being considered, the built reference adjacency matrix $V_{u_{\star t}}$ is associated with an unknown reference proximity measure $u_{\star t}$.

The robustness depends on the α error risk chosen for the null hypothesis: no linear correlation in the case of quantitative variables, or 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.

2.2 Evolutionary data analysis and clustering & Notations

We assume that we have at our disposal T tables of evolutionary data X_t with the same n rows-individuals and p_t different columns-variables or the same ones measured at different times $t, t = 1, \cdots, T$. We will use the following notations:

- $X_{t_{(n,p_t)}}$ is the data matrix with n individuals and p_t variables at time t,

- $X_{(n,p)} = [X_1|\cdots|X_t|\cdots|X_T]$ is the global data matrix with n rows-individuals and $p = \Sigma_{t=1}^{t=T} p_t$ columns-variables, concatenation in columns of T data tables X_t.

- $V_{u_t^\star(p_t)}$ is the symmetric adjacency matrix of order p_t, associated with the reference measure $u_{\star t}$ which best structures the correlations of the variables of the data table X_t,

- $V_{u^\star(p)} = Diag[V_{u_t^\star}]_{t=1,T}$ is the global diagonal adjacency matrix of order p, associated with the global data matrix X,

- $\widehat{X}_{(n,p)} = XV_{u^\star}$ is the projected data matrix with n individuals and p variables,

- M_p is the matrix of distances of order p in the space of individuals,

- $D_n = \frac{1}{n}I_n$ is the diagonal matrix of weights of order n in the space of variables.

We first analyze, in a topological way, the correlation structure of the variables using a Topological PCA, which consists of carrying out the standardized PCA [7,16] triplet $(\,\widehat{X}\,,\,M_p\,,\,D_n\,)$ of the projected data matrix $\widehat{X} = XV_{u^\star}$ and, for comparison, the MFA method.

We then proceed with a clustering of individuals based on the significant principal components of the previous topological PCA.

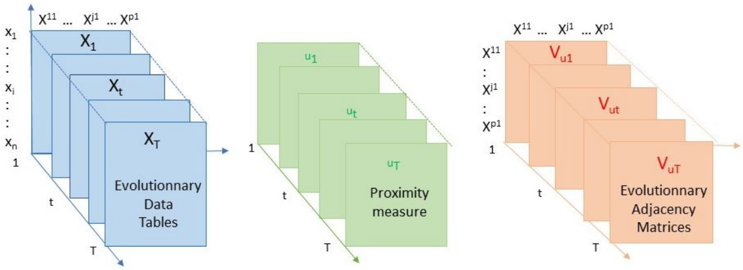


Fig. 2. Evolutionary data tables - Associated adjacency matrices

Definition 2. TCED consist to perform a HAC based on to the criterion, on the significant factors of the standardized topological PCA of the triplet $(\widehat{X}, M_p, D_n)$.

2.3 Measures for comparing clusterings

The following subsection gives the definitions of some indices generally used to compare clusterings. A partition of n individuals defines a qualitative variable whose categories are the classes of the partition. $\mathcal{C}_1$ and $\mathcal{C}_2$ are two partitions of the same n objects with the same number r of clusters. A comparison of two partitions is obtained by constructing the contingency table $N = (n_{ij})_{i,j=1,r}$ crossing these two variables, a $r \times r$ matrix whose $ij-th$ entry equals the number of elements in the intersection of the clusters $\mathcal{C}_1$ and $\mathcal{C}_2$. We thus count the pairs of individuals who remain or do not remain in the same clusters among the $C_n^2 = \frac{n(n-1)}{2}$ pairs of individuals.

A very intuitional approach to comparing clusterings is counting pairs of objects that are "classified" in the same way in both clusterings, i.e. pairs of objects that are in the same cluster (in different clusters, respectively) under both clusterings.

We use indices and measures based on counting pairs to compare the two clusterings with the same number of classes, the proposed TCED and the HAC-MFA, the HAC performed on the results of the MFA [10,11].

In order to measure their concordance, various popular indexes [27] can be calculated such as Rand, Jaccard, Fowlkes and Mallows, Adjusted Rand index, Cohen's kappa, etc.. Thus, can we consider that the configurations of these two clusterings are similar?

(I) R SQUARED (R^2)

The homogeneity of a cluster, its internal consistency, can be analyzed from the variance or the inertia of the objects that compose it. The more the objects are concentrated around their center of gravity, the more the cluster is homogeneous (low inertia within classes: I_{Within}). A common descriptive statistic in cluster analysis is the R^2 that measures the overall proportion of variance explained by the cluster means.

The ratio of variances or inertia explained by classes ; $R^2 = \frac{I_{Between}}{I_{Total}}$ must be as high as possible, its measures the quality of the clustering, its value should be as close as possible to one without too many classes. The greater the value of R^2, the more homogeneous the classes of the partition.

(II) KAPPA TEST (κ)

The kappa coefficient applied to the pairs of objects, provides a new way to measure the agreement between two partitions $\mathcal{C}_1$ and $\mathcal{C}_2$ having the same number r of multi-clusters, from the same sample of size n. The permutation of the maximum kappa value is used to identify the classes of a partition. We study the resemblance between the clusterings using the Cohen's kappa coefficient. The non-parametric statistical test of Kappa [8] allows, in this context of comparison, to measure the agreement or concordance between two classifications. The Kappa agreement rate can be estimated from the contingency table $N = (n_{ij})_{i,j=1,r}$ using the following relation:

$$\widehat{\kappa}(\mathcal{C}_1, \mathcal{C}_2) = \frac{P_o - P_e}{1 - P_e} = \frac{n \sum_{i=1}^{r} n_{ii} - \sum_{i=1}^{r} n_{i.} n_{.i}}{n^2 - \sum_{i=1}^{r} n_{i.} n_{.i}}$$

where, $P_o = \frac{1}{n}\sum_{i=1}^{r} n_{ii}$ is the observed proportion of concordance,
and $P_e = \frac{1}{n^2}\sum_{i=1}^{r} n_{i.}n_{.i}$ represents the expected proportion of concordance under the assumption of independence.

n be the number of individuals, $n_{i.} = \sum_{j=1}^{r} n_{ij}$ and $n_{.j} = \sum_{i=1}^{r} n_{ij}$ designate the number of objects in the ith cluster $\mathcal{C}_1$ and the jth cluster $\mathcal{C}_2$ respectively.

The Kappa coefficient is a real number, without dimension, between -1 and 1. The concordance is higher the value of Kappa is to 1 and the maximum concordance is reached ($\widehat{\kappa} = 1$) when $P_o = 1$ and $P_e = 0.5$. When there is perfect independence, $\widehat{\kappa} = 0$ with $P_o = P_e$, and in the case of total mismatch, $\widehat{\kappa} = -1$ with $P_o = 0$ and $P_e = 0.5$.

(III) RAND INDEX ($\mathcal{R}$)

In order to compare two partitions $\mathcal{C}_1$ and $\mathcal{C}_2$, the most used agreement index is the Rand index (RI) [23]. This index is the overall percentage of pairs in agreement. The Rand index in its contingent form, where all pairs are considered, including identical ones, is written:

$$\mathcal{R}(\mathcal{C}_1, \mathcal{C}_2) = \frac{2\Sigma_{i,j}{n_{ij}}^2 - \Sigma_i {n_{i.}}^2 - \Sigma_j {n_{.j}}^2 + n^2}{n^2}$$

This index takes its values between 0 and 1, it is equal to 1 when the two partitions are identical. Many variants of this index have been proposed.

(IV) ADJUSTED RAND INDEX ($\mathcal{AR}$)

The Adjusted Rand index ($\mathcal{AR}$) proposed by [13] is frequently used in cluster validation since it is a measure of agreement between two partitions. It is a way to compare the similarity of results between two different clustering methods.

The $\mathcal{AR}$ index between these two partitions can be computed from the contingency table $N = (n_{ij})_{i,j=1,r}$ formed by the partitions $\mathcal{C}_1$ and $\mathcal{C}_2$, can be written as:

$$\mathcal{AR}(\mathcal{C}_1, \mathcal{C}_2) = \frac{\Sigma_{i,j} C^2_{n_{ij}} - \frac{[\Sigma_i C^2_{n_{i.}} \, \Sigma_j C^2_{n_{.j}}]}{C^2_n}}{\frac{1}{2}[\Sigma_i C^2_{n_{i.}} + \Sigma_j C^2_{n_{.j}}] - \frac{[\Sigma_i C^2_{n_{i.}} \, \Sigma_j C^2_{n_{.j}}]}{C^2_n}}$$

where, n_{ij} be the number of objects that are in both cluster X_i and Y_j. $n_i = \Sigma_j n_{ij}$ and $n_j = \Sigma_i n_{ij}$ designate the number of objects in cluster X_i and cluster Y_j respectively. The term $C^2_n = \frac{n(n-1)}{2}$ is the binomial coefficients.

The $\mathcal{AR}$ index should be interpreted as follows: $\mathcal{AR} \geq 0.90$ excellent recovery; $0.80 \leq \mathcal{AR} < 0.90$ good recovery; $0.65 \leq \mathcal{AR} < 0.80$ moderate recovery; $\mathcal{AR} < 0.65$ poor recovery.

(V) CHI-SQUARED COEFFICIENT (χ^2)

The measures for comparing clusterings are generally those originally developed for statistical purposes. The chi-square coefficient is one of them, it is one of the most well-known measurements of this type. It is defined as:

$$\chi^2(\mathcal{C}_1, \mathcal{C}_2) = \Sigma_{i=1}^{k}\Sigma_{j=1}^{k}\frac{(n_{ij} - \widehat{n}_{ij})^2}{\widehat{n}_{ij}} \quad \text{where } \widehat{n}_{ij} = \frac{n_{i.}n_{.j}}{n}$$

This measure was proposed by Pearson [20] to test the independence in a bivariate distribution, and not to evaluate in this context of clustering their similarity. The application of this measure for the purpose of comparing clusterings lies in the fact that we must assume the independence of the two clusterings. In general, this is not true and the result of a comparison with such a measure must therefore be put into perspective.

(VI) JACCARD INDEX ($\mathcal{J}$)

The Jaccard index (JI) [15] is an association coefficient known to study the similarity between objects for binary data. It is very similar to the Rand Index, however it disregards the pairs of elements that are in different clusters for both clusterings. It is defined as follows:

$$\mathcal{J}(\mathcal{C}_1, \mathcal{C}_2) = \frac{\Sigma_{i,j} {n_{ij}}^2 - n}{\Sigma_i {n_{i.}}^2 + \Sigma_j {n_{.j}}^2 - \Sigma_{i,j} {n_{ij}}^2 - n}$$

(VII) FOWLKES-MALLOWS INDEX ($\mathcal{FM}$)

The Fowlkes and Mallows index ($\mathcal{FMI}$) [12] is presented as a measure to compare hierarchical clusterings [4]. The generalized Fowlkes–Mallows index is defined by:

$$\mathcal{FM}(\mathcal{C}_1, \mathcal{C}_2) = \frac{\Sigma_{i,j} {n_{ij}}^2 - n}{\sqrt{(\Sigma_i {n_{i.}}^2 - n)(\Sigma_j {n_{.j}}^2 - n)}}$$

Like for the adjusted Rand Index, the "amount" of similarity of two clusterings corresponds to the deviation from the expected value under the null hypothesis of independant clusterings with fixed cluster sizes. Again, the strong assumptions on the distribution make the result hard to interpret.

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

3 Illustrative example

To illustrate the TCED approach, we use data extracted from Eurostat databases [24] on the state of public finances of the 28 countries of the European Union (EU) over the homogeneous period of four years, from 2016 to 2019.

We examine here the evolution of the main characteristics of public finances of the EU28 during the period $2016 - 2019$, which are more precisely, the gross public debt, the deficit, the public expenditure and the public revenue. The same 4 characteristics of public finances were measured on the same 28 EU countries on 4 different years. Simple statistics of the variables considered are displayed in the Table 1.

The global reference adjacency matrix $V_{u_\star}$ associated to the proximity measure $u_\star$ adapted to the evolutionary data considered, is built from the correlation matrices of the evolving tables according to Definition 1. Note that in this case of quantitative variables, it is considered that two positively correlated variables are related and that two negatively correlated variables are related,

Table 1. Summary statistics of EU28 public finances - Period 2016-2019

		2016				2017			
			Standard				Standard		
Variable	Label	Mean	Deviation (N)	Min	Max	Mean	Deviation (N)	Min	Max
Expenditures	EXPE	43.54	6.88	28.10	56.70	42.72	6.87	26.20	56.50
Deficit	DEFI	-0.98	1.59	-4.30	1.90	-0.29	1.76	-3.10	3.30
Revenues	REVE	42.55	6.58	27.30	53.90	42.43	6.67	25.90	53.50
Debt	DEBT	70.90	37.52	10.00	180.50	68.13	37.21	9.10	179.50
		2018				2019			
			Standard				Standard		
Variable	Label	Mean	Deviation (N)	Min	Max	Mean	Deviation (N)	Min	Max
Expenditures	EXPE	43.10	6.48	25.30	55.60	42.94	6.45	24.30	55.40
Deficit	DEFI	-0.27	1.63	-3.60	3.00	-0.11	1.81	-4.30	4.10
Revenues	REVE	42.84	6.46	25.50	53.40	42.81	6.54	24.70	53.80
Debt	DEBT	66.29	38.51	8.20	186.40	64.05	37.53	8.50	180.60

Table 2. Global reference adjacency matrix

$$V_{u^*} = \left(\begin{array}{c|c|c|c} \mathrm{V}_{u*2016} & 0 & 0 & 0 \\ \hline 0 & \mathrm{V}_{u*2017} & 0 & 0 \\ \hline 0 & 0 & \mathrm{V}_{u*2018} & 0 \\ \hline 0 & 0 & 0 & \mathrm{V}_{u*2019} \end{array}\right) = \left(\begin{array}{cccc|cccc|cccc|cccc} 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \end{array}\right)$$

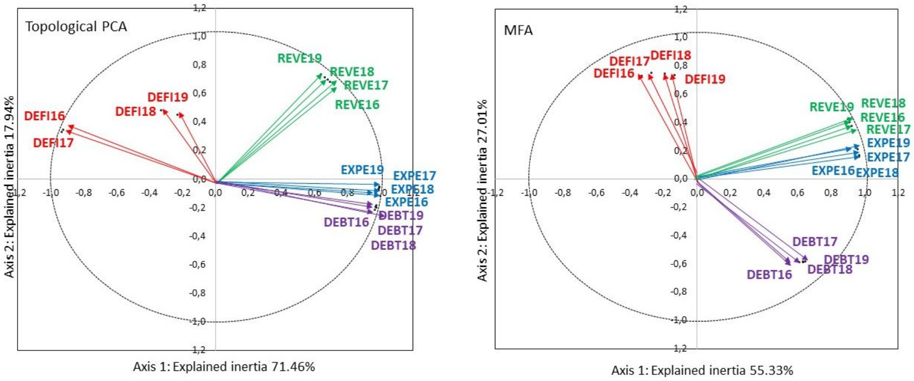


Fig. 3. Topological PCA and MFA of EU-28 public finances

but remote, we will therefore take into account the sign of the correlation between variables in the adjacency matrix.

We first carry out a Topological PCA to identify the correlation structure of the variables, an HAC according to Ward's criterion is then applied on the significant principal components of this PCA of the projected data. We will gradually compare the results of the topological PCA and the MFA method.

Table 3. Topological PCA and MFA - Correlations Variables & Factors

Topological PCA

	Eigenvalue	Proportion (%)	Cumulative (%)
1	11,434	71,46	71,46
2	2.870	17,94	89,40
3	1.373	8,58	97,98
4	0.258	1,61	99,60
5	0.040	0,25	99,85
6	0.012	0,07	99,92
⋮	⋮	⋮	⋮
16	0,000	0,00	100,00
Total	16.000	100.00	

Correlation	Factor	
Variable	F1	F2
EXPE16	0.990	-0.079
DEFI16	-0.924	0.347
REVE16	0.695	0.684
DEBT16	0.969	-0.215
EXPE17	0.,995	-0.055
DEFI17	-0.934	0.331
REVE17	0.684	0.703
DEBT17	0.976	-0.198
EXPE18	0.993	-0.067
DEFI18	-0.330	0.486
REVE18	0.666	0.719
DEBT18	0.975	-0.193
EXPE19	0.994	-0.048
DEFI19	-0.231	0.455
REVE19	0.641	0.739
DEBT19	0.978	-0.180

MFA

	Eigenvalue	Proportion (%)	Cumulative (%)
1	3.939	55,33	55,33
2	1.923	27,01	82,35
3	0.920	12.93	95.28
4	0.156	2.20	97.47
5	0.097	1.36	98.83
6	0.056	0.79	99.62
⋮	⋮	⋮	⋮
15	0.000	0.00	100.00
Total	7.119	100.00	

Correlation	Factor	
Variable	F1	F2
EXPE16	0.965	0.161
DEFI16	-0.330	0.734
REVE16	0.929	0.347
DEBT16	0.610	-0.586
EXPE17	0.966	0.170
DEFI17	-0.270	0.755
REVE17	0.924	0.374
DEBT17	0.632	-0.585
EXPE18	0.956	0.214
DEFI18	-0.189	0.748
REVE18	0.911	0.402
DEBT18	0.634	-0.581
EXPE19	0.947	0.220
DEFI19	-0.133	0.734
REVE19	0.900	0.422
DEBT19	0.646	-0.584

Figure 3 presents, for comparison on the first factorial plane, the correlations between principal components-factors and the original variables.

We can see that these correlations are slightly different, as are the percentages of the inertias explained on the first principal planes of Topological PCA and MFA method.

Table 3 shows that the two first factors of the Topological PCA explain 71.46% and 17.94%, respectively, accounting for 89.40% of the total variation in the data set; however, the two first factors of the MFA add up to 82.35%. Thus, the first two factors provide an adequate synthesis of the data, that is, of the public finance of EU-28 over period 2016-2019. We restrict the comparison to the first significant factorial plan.

The significant correlations between the initial variables and the principal factors in the two analyses are quite different.

For comparison, Figure 4 shows dendrograms of the Topological and MFA clustering of the EU countries according to their public finances.

Note that the partitions chosen in 4 clusters are appreciably different, as much by composition as by characterization. The percentage of the total variance explained by the TCED approach, $R^2 = 70.65\%$, is higher than that of the

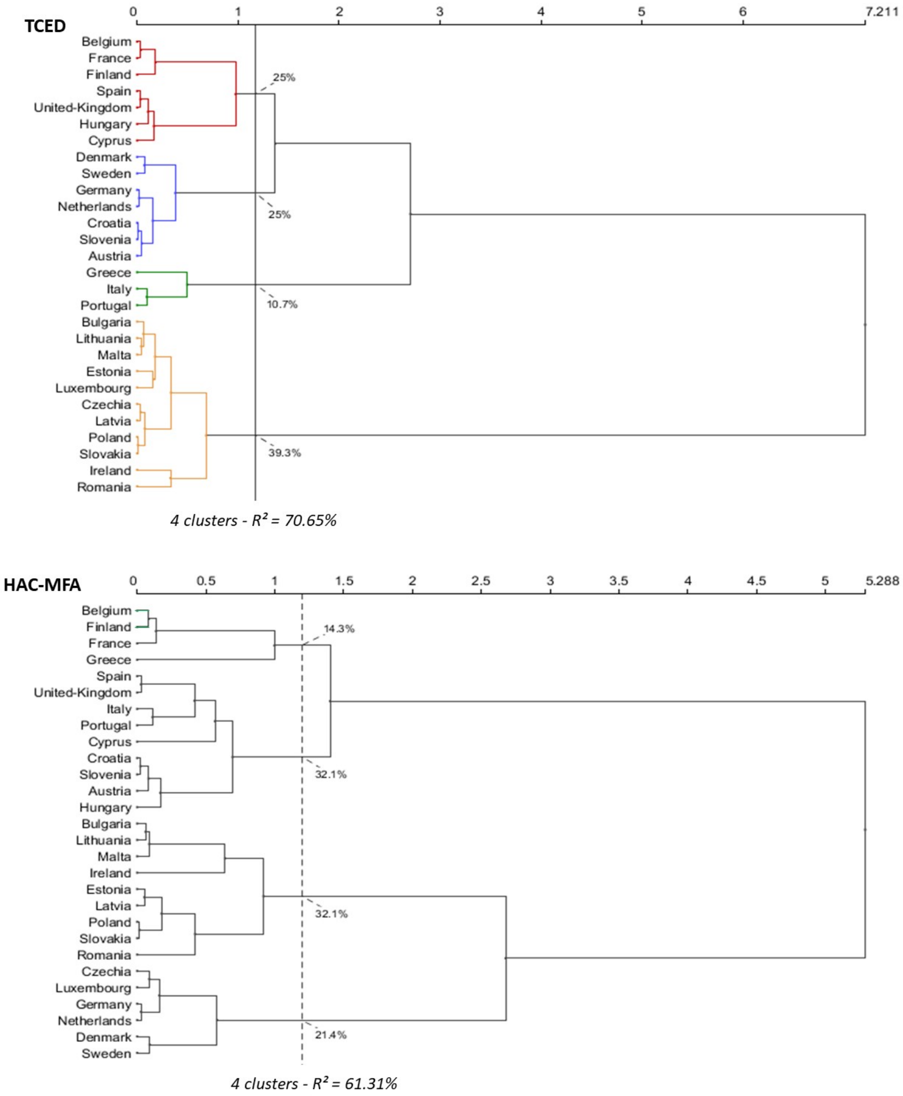


Fig. 4. Topological and HAC-MFA cluster dendrograms of the EU-28 countries

HAC-MFA approach, $R^2 = 61.31\%$, thus indicating that the TCED clusters are more homogeneous than those of HAC-MFA.

Table 4 summarizes the significant profiles (+) and anti-profiles (-) of the two typologies; with a risk of error less than or equal to 5%, they are quite different.

The first TCED cluster, composed of seven countries (Belgium, Spain, Hungary, Finland, France, Cyprus and United Kingdom), is characterized by a high share of expenditure throughout the period $2016-2019$, by a high share of the debt over the years 2018 and 2019 and by a high share of income in 2018 and a low share of the deficit over the entire period 2016 − 2019.

The second cluster which groups together seven EU counties is characterized by a high share of revenues over the entire period 2016 − 2019 and by a high share of deficit in 2018 and 2019.

The third cluster composed of Greece, Italy and Portugal, is characterized by a high share of debts and expenditures over the entire period 2016 − 2019 and a low share of deficit in 2016 and 2017 compared to the EU average.

The last cluster 4, represented by eleven countries, is characterized by a high share of the deficit at the start of the 2016 and 2017 period and a low share of the Debt, revenues and expenditures throughout the 2016 − 2019 period.

TCED				
Cluster Frequency (%)	Cluster 1 7 (25.00%)	Cluster 2 7 (25.00%)	Cluster 3 3 (10.71%)	Cluster 4 11 (39.29%)
Composition	Belgium, Spain, Hungary, Finland, France, Cyprus, United-Kingdom	Denmark,Sweden, Germany, Slovenia Croatia, Austria Netherlands	Greece, Italy, Portugal	Bulgaria, Malta, Latvia, Czechia, Poland, Estonia, Romania, Lithuania, Luxembourg, Ireland,Slovakia
Profile (+)	EXPE16 to 19 DEBT18 to 19 REVE18	REVE16 to 19 DEFI18,19	DEBT16 to 19 EXPE16 to 19	DEFI16,17
Anti-profile(-)	DEFI16 to 19		DEFI16,17	DEBT16 to 19 REVE16 to 19 EXPE16 to 19
HAC-MFA				
Cluster Frequency (%)	Cluster 1 6 (21.43%)	Cluster 2 9 (32.14%)	Cluster 3 7 (25.00%)	Cluster 4 6 (21.43%)
Composition	Italy, Greece, Belgium, France, Finland Austria	Spain, United-Kingdom, Hungary, Cyprus, Slovenia, Croatia, Poland, Slovakia Portugal	Bulgaria, Estonia, Ireland, Latvia, Lithuania, Romania Malta	Netherlands, Sweden, Germany, Denmark, Czechia, Luxembourg
Profile (+)	EXPE16 to 19 REVE16 to 19 DEBT16 to 19			DEFI16 to 19 REVE19
Anti-profile (-)		DEFI16 to 18	DEBT16 to 19 REVE16 to 19 EXPE16 to 19	DEBT16 to 19

Table 4. Characterization of clusters

	HAC-MFA				
TCED	C2	C4	C1	C3	Total
C1	**4**	0	3	0	7
C2	2	**4**	1	0	7
C3	1	0	**2**	0	3
C4	2	2	0	**7**	11
Total	9	6	6	7	28

Table 5. Contingency table - TCED & HAC-MFA

Table 5 shows the contingency table crossing the clusters of the two partitions TCED and HAC-MFA. The classes of HAC-MFA were swapped and renumbered in columns to optimize the concordance indices.

	$\widehat{\kappa}_{max}$	χ^2	$\mathcal{R}$	$\mathcal{AR}$	$\mathcal{J}$	$\mathcal{FM}$
Value	0.473	26.128	0.727	0.241	0.272	0.064
p-value	0.0001	0.0019				

Table 6. Clustering comparison index results

Remember that the goal is not to study whether the two partitions are independent, but whether they are concordant.

Thus, for this example, the calculated Kappa maximal is equal to 0.473 corresponds to a p-value less than 0.01%. Since this probability is lower than a pre-specified significance level of 5%, the null hypothesis, $H_0 : \kappa = 0$ of concordance is rejected. This value indicates a moderate agreement between the two clusterings. The null hypothesis of independence of the chi-square is also rejected.

The $\mathcal{AR}$ value equal to 0.241 is less than 0.65, it is interpreted as especially discordant clusterings.

4 Conclusion

This paper proposes a new topological approach to the clustering of individuals which can enrich classical data analysis methods within the framework of the clustering of objects. The results of the topological clustering approach, based on the notion of a neighborhood graph, are as good - or even better, according to the R-squared results - than the existing classical method. The TCED approach is be easily programmable from the PCA and HAC procedures of SAS, SPAD or R software. It would be interesting to make a Benchmark to evaluate the results of this topological approach on massive evolutionary data (big data). And to compare this topological classification with the time series clustering. Future work consists in extending this topological approach to other methods of data analysis, in particular in the context of prediction models.

References

1. Abdesselam, R.: A Topological Clustering of Individuals. *Classification and Data Science in the Digital Age.* In the Springer book series "Studies in Classification, Data Analysis, and Knowledge Organization". Edts P. Brito, J-G. Dias, B. Lausen, A. Montanari and R. Nugent, 2022.
2. Abdesselam, R.: A Topological Clustering of variables. Journal of Mathematics and System Science. David Publishing Company,Vol.11, Issue 2, pp.1-17, 2021.
3. 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.
4. Aljarah, I., Faris, H. and Mirjalali S. : Evolutionary data clustering: algorithms and applications, Springer, 2021.
5. Batagelj, V., Bren, M.: Comparing resemblance measures. *In Journal of classification*, 12, 73–90, 1995.

6. . Bouroche, J.M.: Analyse des données ternaires : la double analyse en composantes principales. Thèse, 1975.
7. Caillez, F. and Pagès, J.P.: Introduction à l'Analyse des données. *S.M.A.S.H., Paris*, 1976.
8. Cohen, J.: A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, Vol 20, 27–46, 1960.
9. Dazy, F., Le Barzic, J.F., Saporta, G., Lavallard F. : L'analyse des données évolutives – Méthodes et applications. Editions TECHNIP, 1996.
10. Escofier, B. et Pagès, J. : Analyses factorielles simples et multiples : objectifs, méthodes et interprétation, Dunod, 1988.
11. Escofier, B. et Pagès, J. : Mise en oeuvre de l'AFM pour des tableaux numériques, qualitatifs, ou mixtes. Publication interne de l'IRISA, 429, 1985.
12. Fowlkes, E. B., Mallows, C.L.: A Method for Comparing Two Hierarchical Clusterings. *Journal of the American Statistical Association*, 78(383), 53-–569, 1983.
13. Hubert, L. and Arabie, P.: Comparing partitions. *Journal of Classification*, 193-–218, 1985.
14. 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.
15. Jaccard, P., The Distribution of the flora in the alpine zone. *New Phytologist.* 11 (2), 37–50, 1912.
16. Lebart, L.: Stratégies du traitement des données d'enquêtes. *La Revue de MODULAD*, 3, 21–29, 1989.
17. Lavit, C. : Analyse conjointe de tableaux quantitatifs. Editions Masson, 1988.
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. L'Hermier des planttes, H. : Structuration des tableaux à trois indices de la statistique. Thèse de 3ème cycle, Université de Montpellier, 1976.
20. Mirkin, B.: Eleven Ways to Look at the Chi-Squared Coefficient for Contingency Tables. *The American Statistician*, 55, (2), 111–120, 2001.
21. Panagopoulos, D.: Topological data analysis and clustering. Chapter for a book, Algebraic Topology (math.AT) arXiv:2201.09054, Machine Learning, 2022.
22. 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.
23. Rand, W.M., Objective criteria for the evaluation of clustering methods, *Journal of the American Statistical Association, American Statistical Association*, vol. 66, 336, 846–850, 1971.
24. Eurostat, newsrelease, euroindicators, 65/2020 22 April 2020, https://ec.europa.eu/eurostat, 2020.
25. Toussaint, G. T.: The relative neighbourhood graph of a finite planar set. *In Pattern recognition*, 12, 4, 261–268, 1980.
26. Ward, J. R.: Hierarchical grouping to optimize an objective function. *In Journal of the American statistical association JSTOR*, 58, 301, 236–244, 1963.
27. Youness, G., Saporta, G.: Une méthodologie pour la comparaison de partitions. *Revue de statistique appliquée*, tome 52, no 1, 97–120, 2004.
28. 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.