A Topological Clustering of Individuals

Rafik Abdesselam

Abstract The clustering of objects-individuals 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 Topological Clustering of Individuals, or TCI, studies a homogeneous set of individual rows of a data table, based on the notion of neighborhood graphs; the columns-variables are more-or-less correlated or linked according to whether the variable is of a quantitative or qualitative type. It enables topological analysis of the clustering of individual variables which can be quantitative, qualitative or a mixture of the two. It first analyzes the correlations or associations observed between the variables in a topological context of principal component analysis (PCA) or multiple correspondence analysis (MCA), depending on the type of variable, then classifies individuals into homogeneous group, relative to the structure of the variables considered. The proposed TCI method is presented and illustrated here using a real dataset with quantitative variables, but it can also be applied with qualitative or mixed variables.

Keywords: hierarchical clustering, proximity measure, neighborhood graph, adjacency matrix, multivariate data analysis

1 Introduction

The objective of this article is to propose a topological method of data analysis in the context of clustering. The proposed approach, Topological Clustering of Individuals (TCI) is different from those that already exist and with which it is compared. There

University of Lyon, Lyon 2, ERIC - COACTIS Laboratories
Department of Economics and Management, 69365 Lyon, France
e-mail: rafik.abdesselam@univ-lyon2.fr

are approaches specifically devoted to the clustering of individuals, for example, the Cluster procedure implemented in SAS software, but as far as we know, none of these approaches has been proposed in a topological context.

Proximity measures play an important role in many areas of data analysis [16, 5, 9]. The results of any operation involving structuring, clustering or classifying objects are strongly dependent on the proximity measure chosen.

This study proposes a method for the topological clustering of individuals whatever type of variable is being considered: 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 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.

Several topological data analysis studies have been proposed both in the context of factorial analyses (discriminant analysis [4], simple and multiple correspondence analyses [3], principal component analysis [2]) and in the context of clustering of variables [1], clustering of individuals [10] and this proposed TCI approach.

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 structure of a set of quantitative variables, and we present the principles of TCI according to continuous data. This is illustrated in Section 3 using an example based on real data. The TCI results are compared with those of the well-known classical clustering of individuals. Finally, Section 4 presents the concluding remarks on 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.

In the case of continuous data, we consider $E = \{x^1, \cdots, x^j, \cdots, x^p\}$, a set of p quantitative variables. We can see in [1] cases of qualitative or even mixed variables.

We can, by means of a proximity measure u, define a neighborhood relationship, V_u, to be a binary relationship based 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) [7], the Gabriel Graph (GG) [11] or, as is the case here, the Relative Neighborhood Graph (RNG) [14].

For any given proximity measure u, we can construct the associated adjacency binary symmetric matrix V_u of order p, where, all pairs of neighboring variables in E satisfy the following RNG property:

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

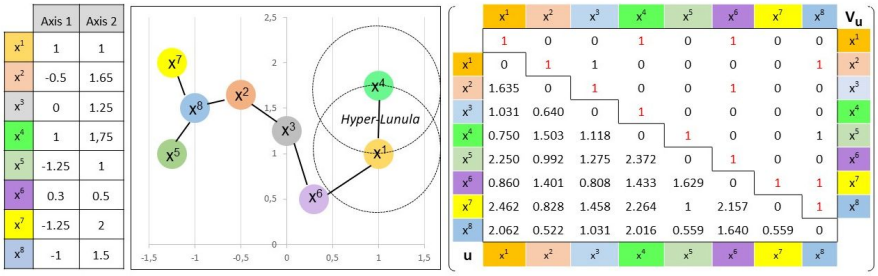

Fig. 1 Data - RNG structure - Euclidean distance - Associated adjacency matrix

Figure 1 shows a simple illustrative example in $\mathbb{R}^2$ of a set of quantitative variables that verify the structure of the RNG graph with Euclidean distance as proximity measure: $u(x^k\ ,\ x^l) = \sqrt{\sum_{j=1}^{2}(x_j^k - x_j^l)^2}$.

This generates a topological structure based on the objects in E which are completely described by the adjacency binary matrix V_u.

2.1 Reference adjacency matrices

Three topological factorial approaches are described in [1] according to the type of variables considered: quantitative, qualitative or a mixture of both. We consider here the case of a set of quantitative variables.

We assume that we have at our disposal a set $E = \{x^j; j = 1, \cdots, p\}$ of p quantitative variables and n individuals-objects. The objective here is to analyze in a topological way, the structure of the correlations of the variables considered [2], from which the clustering of individuals will then be established.

We construct the reference adjacency matrix named $V_{u\star}$ from the correlation matrix. Expressions of suitable adjacency reference matrices for cases involving qualitative variables or mixed variables are given in [1].

To examine the correlation structure between the variables, we look at the significance of their linear correlation. The reference adjacency matrix $V_{u\star}$ associated

with reference measure $u_\star$, can be written using the Student's t-test of the linear correlation coefficient ρ of Bravais-Pearson:

Definition 1 For quantitative variables, $V_{u_\star}$ is defined as:

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

where the p-value is the significance test of the linear correlation coefficient for the two-sided test of the null and alternative hypotheses, $H_0 : \rho(x^k , x^l) = 0$ 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 if the p-value is less than or equal to a chosen α significance level, for example, $\alpha = 5\%$. Using a linear correlation test, if the p-value is very small, it means that there is a very low likelihood that the null hypothesis is correct, and consequently we can reject it.

2.2 Topological analysis - Selective review

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

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.

$\sqrt[4]{We assume that we have at our disposal}\{\ \mathrm{x}^k; k = 1, .., p\}$ a set of p homogeneous quantitative variables measured on n individuals. We will use the following notations:

- $X_{(n,p)}$ is the data matrix with n rows-individuals and p columns-variables,
- $V_{u_\star}$ is the symmetric adjacency matrix of order p, associated with the reference measure $u_\star$ which best structures the correlations of the variables,
- $\widehat{X}_{(n,p)} = XV_{u_\star}$ 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 [6, 8] triplet $(\ \widehat{X}\ ,\ M_p\ ,\ D_n\ )$ of the projected data matrix $\widehat{X} = XV_{u_\star}$ and, for comparison, the duality diagram of the Classical standardized PCA triplet $(\ X\ ,\ M_p\ ,\ D_n\ )$ of the initial data matrix X. We then proceed with a clustering of individuals based on the significant principal components of the previous topological PCA.

Definition 2 TCI consist of performing a HAC, based on the Ward criterion[1] [15], on the significant factors of the standardized PCA of the triplet $(\widehat{X}, M_p, D_n)$.

3 Illustrative example

The data used [13] to illustrate the TCI approach describe the renewable electricity (RE) of the 13 French regions in 2017, described by 7 quantitative variables relating to RE. The growth of renewable energy in France is significant. Some French regions have expertise in this area; however, the regions' profiles appear to differ.

The objective is to specify regional disparities in terms of RE by applying topological clustering to the French regions in order to identify which were the country's greenest regions in 2017. Statistics relating to the variables are displayed in Table 1.

Table 1 Summary statistics of renewable energy variables

Variable	Frequency	Mean	Standard Deviation (N)	Coefficient of variation (%)	Min	Max
Total RE production (TWH)	13	6.84	6.58	96.19	0.59	2.34
Total RE consumption (TWH)	13	3.70	1.87	50.67	2.18	7.06
Coverage RE consumption (%)	13	0.18	0.11	59.01	0.02	0.36
Hydroelectricity (%)	13	0.34	0.30	87.47	0.01	0.89
Solar electricity (%)	13	0.13	0.09	72.57	0.02	0.31
Wind electricity (%)	13	0.39	0.29	76.12	0.01	0.86
Biomass electricity (%)	13	0.15	0.19	130.54	0.01	0.79

Table 2 Correlation matrix (p-value) - Reference adjacency matrix $V_{u_\star}$

Production	1.000						
Consumption	0.575	1.000					
	(0.040)						
Coverage	0.798	0.090	1.000				
	(0.001)	(0.771)					
Hydroelectricity	0.720	0.138	0.872	1.000			
	(0.006)	(0.653)	(0.000)				
Solar	-0.272	-0.477	0.105	0.168	1.000		
	(0.369)	(0.099)	(0.734)	(0.582)			
Wind	-0.408	-0.305	-0.524	-0.772	-0.395	1.000	
	(0.167)	(0.311)	(0.066)	(0.002)	(0.181)		
Biomass	-0.365	0.489	-0.609	-0.459	-0.149	-0.135	1.000
	(0.220)	(0.090)	(0.027)	(0.114)	(0.627)	(0.660)	

Significance level: p−value $\leq \alpha = 5\%$

$$V_{u_\star} = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & -1 \\ 1 & 0 & 1 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

The adjacency matrix $V_{u_\star}$, associated with the proximity measure $u_\star$, adapted to the data considered, is built from the correlations matrix Table 2 according to Definition 1. Note that in this case, which uses 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 any sign of correlation between variables in the adjacency matrix.

[1] Aggregation based on the criterion of the loss of minimal inertia.

We first carry out a Topological PCA to identify the correlation structure of the variables. A HAC, according to Ward's criterion, is then applied to the significant principal components of the PCA of the projected data. We then compare the results of a topological and a classical PCA.

Figure 2 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 and Classic PCA.

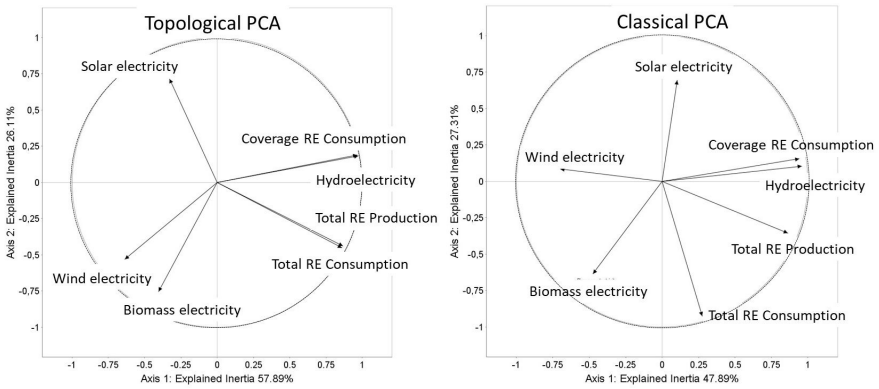

Fig. 2 Topological & Classical PCA of RE of the French regions

The two first factors of the Topological PCA explain 57.89% and 26.11%, respectively, accounting for 83.99% of the total variation in the data set; however, the two first factors of the Classical PCA add up to 75.20%. Thus, the first two factors provide an adequate synthesis of the data, that is, of RE in the French regions. We restrict the comparison to the first significant factorial axes.

For comparison, Figure 3 shows dendrograms of the Topological and Classical clustering of the French regions according to their RE. Note that the partitions chosen in 5 clusters are appreciably different, as much by composition as by characterization. The percentage variance produced by the TCI approach, $R^2 = 86.42\%$, is higher than that of the classic approach, $R^2 = 84.15\%$, indicating that the clusters produced via the TCI approach are more homogeneous than those generated by the Classical one.

Based on the TCI analysis, the Corse region alone constitutes the fourth cluster, and the Nouvelle-Acquitaine region is found in the second cluster with the Grand-Est, Occitanie and Provence-Alpes-Côte-d'Azur (PACA) regions; however, in the Classical clustering, these two regions - Corse and Nouvelle-Aquitaine - together constitute the third cluster.

Figure 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 cluster produced via the TCI approach, consisting of a single region, Auvergne-Rhônes-Alpes (AURA), is characterized by high share of hydroelectricity,

a high level of coverage of regional consumption, and high RE production and consumption. The second cluster - which groups together the four regions of Grand-Est, Occitanie, Provence-Alpes-Côte-d'Azur (PACA) and Nouvelle-Aquitaine - is considered a homogeneous cluster, which means that none of the seven RE characteristics differ significantly from the average of these characteristics across all regions. This cluster can therefore be considered to reflect the typical picture of RE in France.

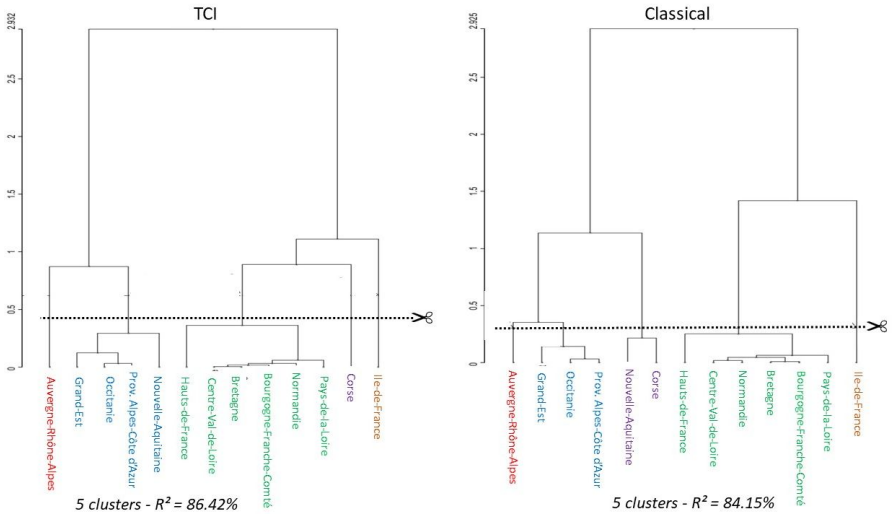

Fig. 3 Topological and Classical dendrograms of the French regions

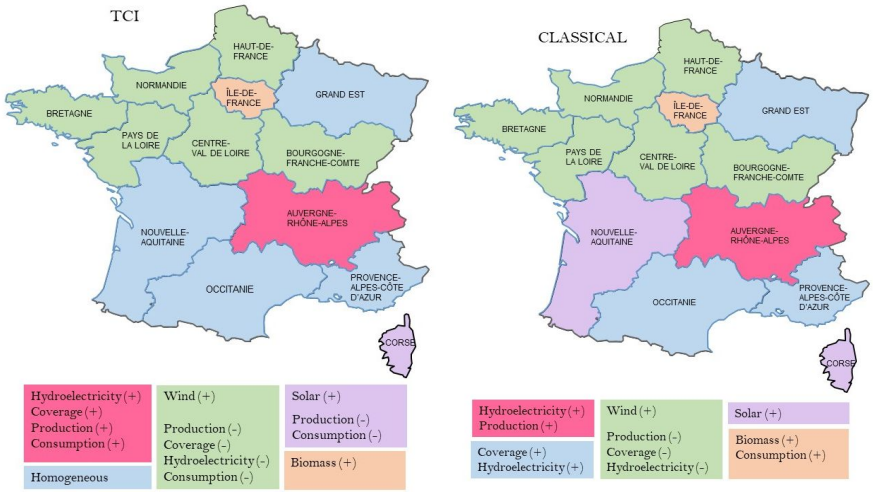

Fig. 4 Typologies - Characterization of TCI & Classical clusters

Cluster 3, which consists of six regions, is characterized by a high degree of wind energy, a low degree of hydroelectricity, low coverage of regional consumption, and low production and consumption of RE compared to the national average. Cluster 4, represented by the Corse region, is characterized by a high share of solar energy and low production and consumption of RE. The last class, represented by the Ile-de-France region, is characterized by a high share of biomass energy. Regarding the other types of RE, their share is close to the national average.

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 TCI approach is be easily programmable from the PCA and HAC procedures of SAS, SPAD or R software. Future work will involve extending this topological approach to other methods of data analysis, in particular in the context of evolutionary data analysis.

References

1. Abdesselam, R.: A Topological Clustering of variables. Journal of Mathematics and System Science. David Publishing Company, USA, Volume 11, Issue 2, pp.1-17, 2021.
2. Abdesselam, R.: A Topological Principal Component Analysis. International Journal of Data Science and Analysis. Vol.7, Issue 2, 20–31 (2021)
3. Abdesselam, R.: A Topological Multiple Correspondence Analysis. Journal of Mathematics and Statistical Science, ISSN 2411-2518, 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. pp. 167–178 Wiley, (2019)
5. Batagelj, V., Bren, M.: Comparing resemblance measures. Journal of classification, 12, 73–90 (1995)
6. Caillez, F. and Pagès, J.P.: Introduction à l'Analyse des données. S.M.A.S.H., Paris (1976)
7. 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)
8. Lebart, L.: Stratégies du traitement des données d'enquêtes. La Revue de MODULAD, 3, 21–29 (1989)
9. Lesot, M. J., Rifqi, M. and Benhadda, H.: Similarity measures for binary and numerical data: a survey. In: IJKESDP, 1, 1, 63-84 (2009)
10. Panagopoulos, D.: Topological data analysis and clustering. Chapter for a book, Algebraic Topology (math.AT) arXiv:2201.09054, Machine Learning, (2022)
11. 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. Computer-Aided Design Elsevier, 38, 6, 619–626 (2006)
12. SAS Institute Inc. SAS/STAT Software, the Cluster Procedure, Available via DIALOG. https://support.sas.com/documentation/onlinedoc/stat/142/cluster.pdf

13. Selectra : Electricité renouvelable : quelles sont les régions les plus vertes de France ? http://selectra.info/energie/actualites/expert/electricite-renouvelable-regions-plus-vertes-france,(2020).
14. Toussaint, G. T.: The relative neighbourhood graph of a finite planar set. Pattern recognition, 12, 4, 261–268 (1980)
15. Ward, J. R.: Hierarchical grouping to optimize an objective function. Journal of the American statistical association JSTOR, 58, 301, 236–244 (1963)
16. Zighed, D., Abdesselam, R., and Hadgu, A.: Topological comparisons of proximity measures. In: Tan et al. (Eds). 16th PAKDD 2012 Conference, pp. 379–391. Springer, (2012)