Metadata of the chapter that will be visualized online

Chapter Title	A Topological Approach of Clustering	
Copyright Year	2023	
Copyright Holder	The Author(s), under exclusive license to Springer Nature Switzerland AG	
Corresponding Author	Family Name	**Abdesselam**
	Particle	
	Given Name	**Rafik**
	Suffix	
	Division	University of Lyon, Lumière Lyon 2, ERIC - COACTIS Laboratories
	Organization	Department of Economics and Management
	Address	Lyon, France
	Email	rafik.abdesselam@univ-lyon2.fr
	URL	http://perso.univ-lyon2.fr/~rabdesse/fr/
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 approach of clustering, called Topological Clustering of Individuals (TCI), studies a homogeneous set of individuals-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 the topological context of principal component analysis (PCA) or multiple correspondence analysis (MCA), depending on the type of variable, then classifies individuals into homogeneous groups relative to the structure of the variables considered. The proposed TCI method is presented and illustrated here using a simple real dataset with quantitative variables; however, it can also be applied with qualitative or mixed variables.	
Keywords (separated by "-")	Hierarchical clustering - Proximity measure - Neighborhood graph - Adjacency matrix - Multivariate data analysis	

Chapter 22
A Topological Approach of Clustering

Rafik Abdesselam

22.1 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 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 (Zighed et al., 2012; Batagelj and Bren, 1995; Lesot et al., 2009). 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 (Abdesselam, 2019), simple and multiple correspondence analyses (Abdesselam, 2020, 2019), principal component analysis (Abdesselam, 2021)) and in the context of clustering of variables (Abdes-

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

C. H. Skiadas, C. Skiadas (eds.), *Quantitative Demography and Health Estimates*,
The Springer Series on Demographic Methods and Population Analysis 55,
https://doi.org/10.1007/978-3-031-28697-1_22

selam, 2021), clustering of individuals (Panagopoulos, 2022) and this proposed TCI approach.

This paper is organized as follows. In Sect. 22.2, 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 Sect. 22.3 using an example based on real data. The TCI results are compared with those of the well-known classical clustering of individuals. Finally, Sect. 22.4 presents the concluding remarks on this work.

22.2 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 Abdesselam (2021) 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.

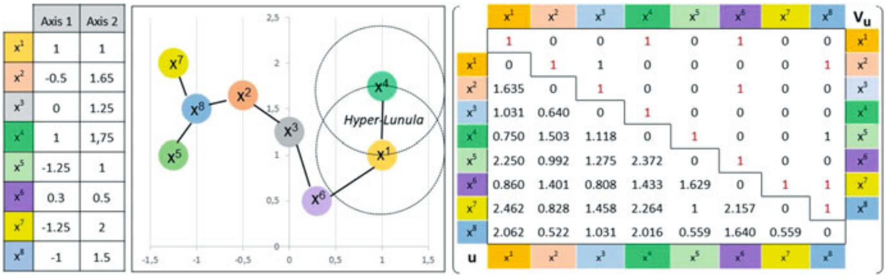
Thus, for a given proximity measure u, we can build a neighborhood graph on E, where the vertices are the variables and the edges are defined by a property of the neighborhood relationship.

Many definitions are possible to build this binary neighborhood relationship. One can choose the Minimal Spanning Tree (MST) (Kim and Lee, 2003), the Gabriel Graph (GG) (Park et al., 2006) or, as is the case here, the Relative Neighborhood Graph (RNG) (Toussaint, 1980).

For any given proximity measure u, for continuous or binary data listed in Table 22.5 given in the Appendix, 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:

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

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.



	Axis 1	Axis 2
x^1	1	1
x^2	-0.5	1.65
x^3	0	1.25
x^4	1	1,75
x^5	-1.25	1
x^6	0.3	0.5
x^7	-1.25	2
x^8	-1	1.5

	x^1	x^2	x^3	x^4	x^5	x^6	x^7	x^8	V_u
	1	0	0	1	0	1	0	0	x^1
x^1	0	1	1	0	0	0	0	1	x^2
x^2	1.635	0	1	0	0	1	0	0	x^3
x^3	1.031	0.640	0	1	0	0	0	0	x^4
x^4	0.750	1.503	1.118	0	1	0	0	1	x^5
x^5	2.250	0.992	1.275	2.372	0	1	0	0	x^6
x^6	0.860	1.401	0.808	1.433	1.629	0	1	1	x^7
x^7	2.462	0.828	1.458	2.264	1	2.157	0	1	x^8
x^8	2.062	0.522	1.031	2.016	0.559	1.640	0.559	0	
u	x^1	x^2	x^3	x^4	x^5	x^6	x^7	x^8	

Fig. 22.1 Data—RNG structure—Euclidean distance—Associated adjacency matrix

Figure 22.1 shows a simple illustrative example in $\mathbb{R}^2$ of a set of eight quantitative variables $\{x^1, \cdots, x^2, \cdots, x^8\}$, that verify the structure of the RNG graph with Euclidean distance as proximity measure: $u(x^k\ ,\ x^l) = \sqrt{\sum_{j=1}^{2}(x_j^k - x_j^l)^2}$.

For example, for the first and the fourth variables, $V_u(x^1, x^4) = 1$, 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 which are completely described by the adjacency binary matrix V_u.

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.

22.2.1 Reference Adjacency Matrices

Three topological factorial approaches are described in Abdesselam (2021) according to the type of variables considered, quantitative, qualitative or a mixture of both. We treat 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 (Abdesselam, 2021), from which the classification of individuals will then be established.

We construct the reference adjacency matrix noted $V_{u_\star}$, in the case of quantitative variables, from the correlation matrix. The expressions of the suitable adjacency reference matrices in the case of qualitative variables or mixed variables are given in Abdesselam (2021).

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 22.1 For quantitative variables, the reference adjacency matrix $V_{u_\star}$ associated to reference measure $u_\star$ is defined as:

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

Where p-value is the significance test of the 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 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.

22.2.2 Topological Equivalence

The different proximity measures given in Table 22.5 in appendix, can be compared according to their topological equivalence in order to better visualize their similarities and their proximity with the reference measure $u_\star$.

The topological equivalence between two proximity measures u_i and u_j is measured using the associated adjacency matrices V_{u_i} and V_{u_j}. It is based on the following concordance index:

$$S(V_{u_i}, V_{u_j}) = \frac{\sum_{k=1}^{r} \sum_{l=1}^{r} \delta_{kl}(z^k, z^l)}{r^2}$$

$$\text{with } \delta_{kl}(z^k, z^l) = \begin{cases} 1 \text{ if } \mathrm{V}_{u_i}(z^k, z^l) = \mathrm{V}_{u_j}(z^k, z^l) \\ 0 \text{ otherwise.} \end{cases}$$

The greater this topological index is and tends to 1, the more the proximity measures are equivalent. $S(V_{u_i}, V_{u_\star})$ measures the similarity and resemblance between any proximity measure u_i and the reference measure $u_\star$.

22.2.3 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.

We assume that we have at our disposal $\{x^k; k = 1, .., p\}$ 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,
- $\widehat{X}_{(n,p)} = X V_{u_\star}$ 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 (Caillez and Pagès, 1976; Lebart, 1989) triplet $(\,\widehat{X}\,,\ M_p\,,\ D_n\,)$ of the projected data matrix $\widehat{X} = X V_{u_\star}$ 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.

Figure 22.2 shows the duality diagram corresponding to the Topological PCA according to the standardized PCA triplet $(\,\widehat{X}\,,\ M_p\,,\ D_n\,)$ of the projected data matrix $\widehat{X} = X V_{u_\star}$, and for comparison, the duality diagram of the Classical standardized PCA triplet $(\,X\,,\ M_p\,,\ D_n\,)$ of the initial data matrix X.

$$
\begin{array}{ccc}
E = R^p & \overset{{}^t\widehat{X}}{\longleftarrow} & F^\star = R^n \\
M_p \downarrow\uparrow V & \textit{Topological PCA} & W \downarrow\uparrow D_n \\
E^\star = R^p & \underset{\widehat{X}_{(n,p)} = X V_{u_\star}}{\longrightarrow} & F = R^n
\end{array}
\qquad
\begin{array}{ccc}
E = R^p & \overset{{}^tX_{(n,p)}}{\longleftarrow} & F^\star = R^n \\
M_p \downarrow\uparrow V & \textit{Classical PCA} & W \downarrow\uparrow D_n \\
E^\star = R^p & \underset{X_{(n,p)}}{\longrightarrow} & F = R^n
\end{array}
$$

Fig. 22.2 Duality diagrams

Table 22.1 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

Definition 22.2 TCI consist to perform a HAC based on to the Ward[1] (Ward, 1963), criterion on the significant factors of the standardized PCA of the triplet $(\widehat{X}, M_p, D_n)$.

We compare the proposed TCI to the most used method of individuals clustering, the Cluster procedure (SAS Institute Inc., 2016) of the SAS software.

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

22.3 Illustrative Example

The data used (Selectra, 2020) 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. Simple statistics relating to the variables are displayed in Table 22.1.

AQ1

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 22.2 according to Definition 22.1.

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

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

Table 22.2 Correlation matrix (p-value)—Reference adjacency matrix $V_{u_\star}$

Production	1						
Consumption	0.575 (0.040)	1					
Coverage	0.798 (0.001)	0.090 (0.771)	1				
Hydroelectricity	0.720 (0.006)	0.138 (0.653)	0.872 (0.000)	1			
Solar	−0.272 (0.369)	−0.477 (0.099)	0.105 (0.734)	0.168 (0.582)	1		
Wind	−0.408 (0.167)	−0.305 (0.311)	−0.524 (0.066)	−0.772 (0.002)	−0.395 (0.181)	1	
Biomass	−0.365 (0.220)	0.489 (0.090)	−0.609 (0.027)	−0.459 (0.114)	−0.149 (0.627)	−0.135 (0.660)	1

$$V_{u^*} = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & -1 \\ 1 & 0 & 1 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

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

Table 22.3 Topological equivalences

Rank	u_i	$S(V_{u_i}; V_{u_\star})$
1	*Size distance*	*83.67%*
2	Euclidean	75.51%
2	Minkowski	75.51%
2	Cosine dissimilarity	75.51%
2	Squared chord	75.51%
2	Doverlap measure	75.51%
2	Shape distance	75.51%
2	Lpower	75.51%
3	Tchebychev	71.43%
3	Pearson correlation	71.43%
4	Manhattan	67.35%
4	Normalized Euclidean	67.35%
4	Canberra	67.35%
4	Weighted Euclidean	67.35%
4	Gower's dissimilarity	67.35%

Table 22.3 summarizes the topological equivalence between the reference measure $u_\star$ with the usual proximity measures for continuous data. Size Distance is the closest measure to the reference measure $u_\star$ with a topological equivalence of 83.67%.

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 and classical PCA.

Figure 22.3 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 and Classic PCA.

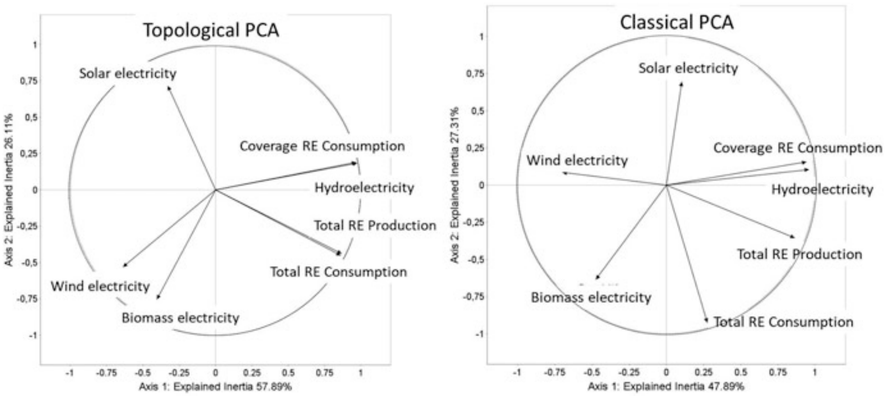


Fig. 22.3 Topological and Classical PCA of the RE of the French regions

Table 22.4 Topological and classical PCA—correlations variables and factors

Topological PCA		
Eigenvalue	Proportion (%)	Cumulative (%)
4.052	*57.89*	57.89
1.827	*26.11*	*83.99*
0.858	12.25	96.24
0.246	3.52	99.76
0.017	0.24	100.00
0.000	0.00	100.00
0.000	0.00	100.00
7.000	100.00	100.00

Correlation	Factor	
Variable	F1	F2
Total RE production	*0.867*	−0.439
Total RE consumption	*0.860*	−0.452
Coverage RE consumption	*0.966*	0.189
Hydroelectricity	*0.974*	0.184
Solar electricity	−0.329	*0.715*
Wind electricity	*−0.637*	*−0.531*
Biomass electricity	−0.405	*−0.754*

Classical PCA		
Eigenvalue	Proportion (%)	Cumulative (%)
3.352	*47.89*	47.89
1.912	*27.31*	*75.20*
1.345	*19.22*	94.42
0.275	3.93	98.35
0.098	1.40	99.75
0.017	0.25	100.00
0.000	0.00	100.00
7.000	100.00	100.00

Correlation	Factor	
Variable	F1	F2
Total RE production	*0.863*	−0.355
Total RE consumption	0.274	*−0.925*
Coverage RE consumption	*0.942*	0.155
Hydroelectricity	*0.959*	0.105
Solar electricity	0.103	*0.694*
Wind electricity	*−0.700*	0.084
Biomass electricity	−0.475	*−0.636*

Table 22.4 shows that 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.

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

For comparison, Fig. 22.4 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-Aquitaine region is found in the second cluster with the Grand-Est, Occitanie and Provence-Alpes-Côte-d'Azur (PACA) regions; however, in the

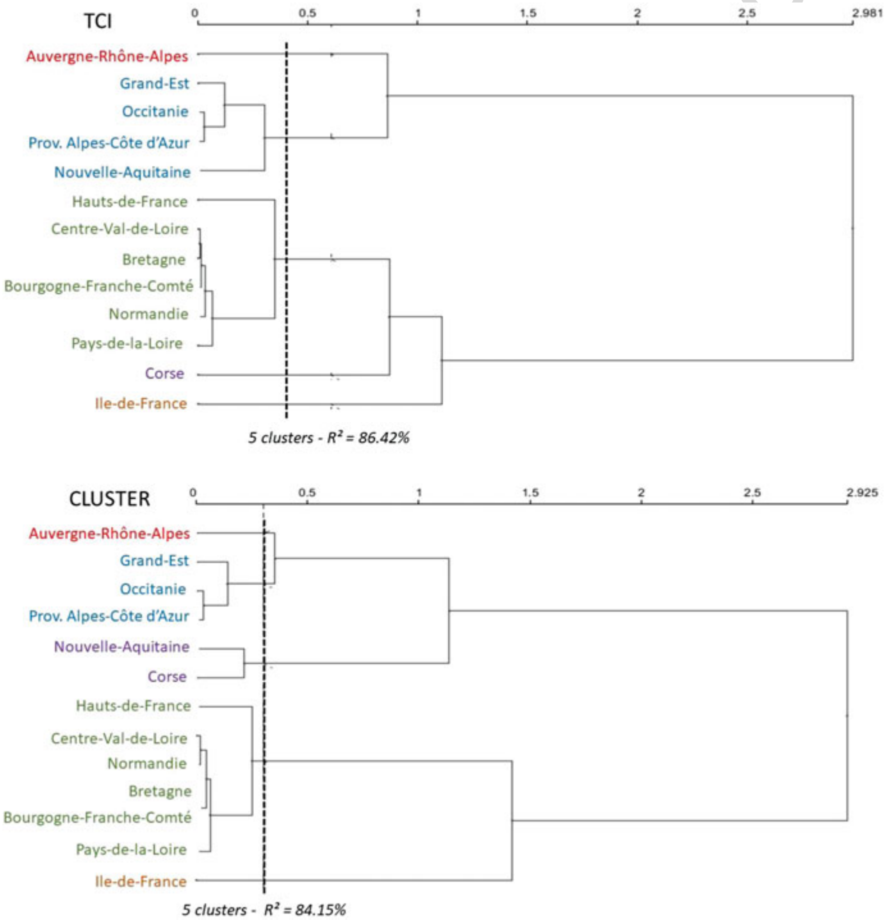


Fig. 22.4 Topological and classical dendrograms of the French regions

Classical clustering, these two regions—Corse and Nouvelle-Aquitaine—together constitute the third cluster.

Figure 22.5 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.

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.

22.4 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.

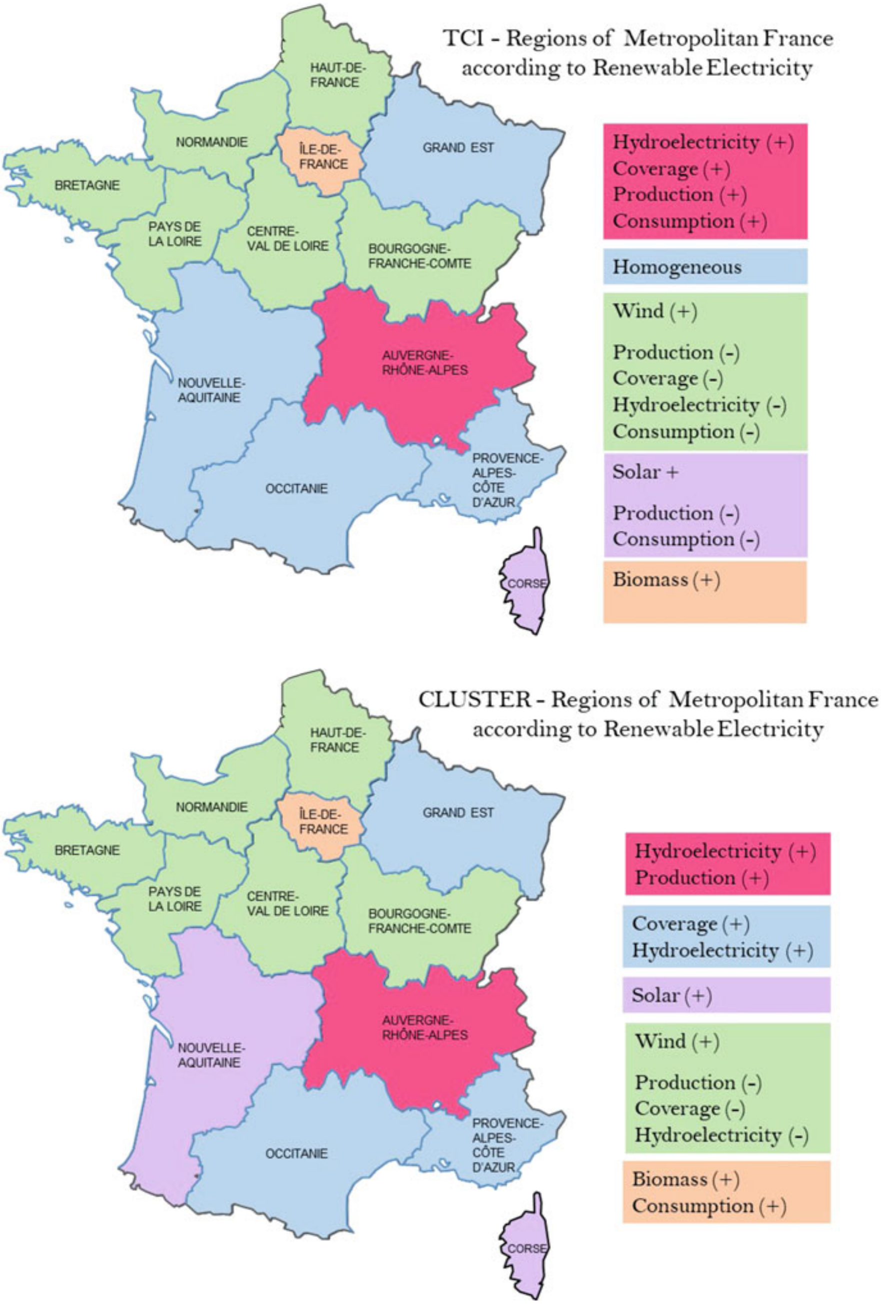


Fig. 22.5 Characterization of TCI and classical clusters

Appendix

 See Table 22.5.

Table 22.5 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	$u_{Cos}(x, y) = 1 - \frac{\sum_{j=1}^{p} x_j y_j}{\sqrt{\sum_{j=1}^{p} x_j^2}\sqrt{\sum_{j=1}^{p} y_j^2}} = 1 - \frac{<x, y>}{\lVert x \rVert \lVert y \rVert}$
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	$u_{Cor}(x, y) = 1 - \frac{(\sum_{j=1}^{p}(x_j - \overline{x})(y_j - \overline{y}))^2}{\sum_{j=1}^{p}(x_j - \overline{x})^2 \sum_{j=1}^{p}(y_j - \overline{y})^2} = 1 - \frac{(<x - \overline{x}, y - \overline{y}>)^2}{\lVert x - \overline{x} \rVert^2 \lVert y - \overline{y} \rVert^2}$
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	$u_{Siz}(x, y) = \mid \sum_{j=1}^{p}(x_j - y_j) \mid$

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	$s_{12} = \frac{a+d-b-c}{a+b+c+d}$	$u_{12} = \frac{1-s_{12}}{2}$
Michael	$s_{13} = \frac{4(ad-bc)}{(a+d)^2+(b+c)^2}$	$u_{13} = \frac{1-s_{13}}{2}$
Baroni, Urbani and Buser	$s_{14} = \frac{a+\sqrt{ad}}{a+b+c+\sqrt{ad}}$	$u_{14} = 1 - s_{14}$
Yule Q	$s_{15} = \frac{ad-bc}{ad+bc}$	$u_{15} = \frac{1-s_{15}}{2}$
Yule Y	$s_{16} = \frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}$	$u_{16} = \frac{1-s_{16}}{2}$
Sokal and Sneath 4	$s_{17} = \frac{1}{4}(\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{d+b} + \frac{d}{d+c})$	$u_{17} = 1 - s_{17}$
Gower and Legendre	$s_{18} = \frac{a+d}{a+\frac{(b+c)}{2}+d}$	$u_{18} = 1 - s_{18}$
Sokal and Sneath 1	$s_{19} = \frac{2(a+d)}{2(a+d)+b+c}$	$u_{19} = 1 - s_{19}$

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

Abdesselam, R. (2019). A topological multiple correspondence analysis. *Journal of Mathematics and Statistical Science, 5*(8), 175–192. , ISSN 2411-2518, Science Signpost Publishing Inc., USA.

Abdesselam, R. (2019). A topological discriminant analysis. In: *Data analysis and applications 2, utilization of results in Europe and other topics* (Vol. 3, Part 4, pp. 167–178). Wiley.

Abdesselam, R. (2020). Selection of proximity measures for a topological correspondence analysis. In: *Data analysis and applications 3, Computational, classification, financial, statistical and stochastic methods, Vol. 5, Part 2. Classification data analysis and methods* (pp. 103–120). Wiley.

Abdesselam, R. (2021). A topological clustering of variables. *Journal of Mathematics and System Science, 11*(2), 1–17. David Publishing Company, USA.

Abdesselam, R. (2021). A topological principal component analysis. *International Journal of Data Science and Analysis, 7*(2), 20–31. Science Publishing Group, USA.

AQ3

Batagelj, V., & Bren, M. (1995). Comparing resemblance measures. *Journal of Classification, 12*, 73–90.

Benzécri, J. P. (1976). *L'Analyse des Données. Tome 1: La Taxinomie. Tome 2: L'analyse des correspondances* (2ème édition). Paris: Dunod.

Caillez, F., & Pagès, J. P. (1976). Introduction à l'Analyse des données. *S.M.A.S.H., Paris*.

Demsar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research, 7*, 1–30.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2*, 193–218.

Kim, J. H., & Lee, S.: Tail bound for the minimal spanning tree of a complete graph. *Statistics & Probability Letters, 64*(4), 425–430.

Lebart, L. (1989). Stratégies du traitement des données d'enquêtes. *La Revue de MODULAD, 3*, 21–29.

Lesot, M. J., Rifqi, M., & Benhadda, H. (2009). Similarity measures for binary and numerical data: a survey. *International Journal of Knowledge Engineering and Soft Data Paradigms, 1*(1), 63–84.

Mantel, N. (1967). A technique of disease clustering and a generalized regression approach. *Cancer Research, 27*, 209–220.

Panagopoulos, D. (2022). Topological data analysis and clustering. In *Algebraic topology* (math.AT) arXiv:2201.09054, Machine Learning.

Park, J. C., Shin, H., & Choi, B. K. (2006). Elliptic Gabriel graph for finding neighbors in a point set and its application to normal vector estimation. *Computer-Aided Design, 38*(6), 619–626. Elsevier.

Rifqi, M., Detyniecki, M., & Bouchon-Meunier, B. (2003). Discrimination power of measures of resemblance. In *IFSA'03 Citeseer*.

SAS Institute Inc. *SAS/STAT software, the cluster procedure*. https://support.sas.com/documentation/onlinedoc/stat/142/cluster.pdf

Schneider, J. W., & Borlund, P. (2007). Matrix comparison, Part 1: Motivation and important issues for measuring the resemblance between proximity measures or ordination results. *Journal of the American Society for Information Science and Technology, 58*(11), 1586–1595.

Schneider, J. W., & Borlund, P. (2007). Matrix comparison, Part 2: Measuring the resemblance between proximity measures or ordination results by use of the Mantel and Procrustes statistics. *Journal of the American Society for Information Science and Technology, 58*(11), 1596–1609.

Selectra. (2020). Electricité renouvelable: quelles sont les régions les plus vertes de France? https://selectra.info/energie/actualites/expert/electricite-renouvelable-regions-plus-vertes-france

Toussaint, G. T. (1980). The relative neighbourhood graph of a finite planar set. *Pattern Recognition, 12*(4), 261–268.

Ward, J. R. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association JSTOR, 58*(301), 236–244.

Warrens, M. J. (2008). Bounds of resemblance measures for binary (presence/absence) variables. *Journal of Classification, 25*(2), 195–208. Springer.

Zighed, D., Abdesselam, R., & Hadgu, A. (2012). Topological comparisons of proximity measures. In P.-N. Tan et al. (Eds.), *In the 16th PAKDD 2012 Conference* (pp. 379–391). Part I, LNAI 7301. Berlin Heidelberg: Springer-Verlag.

AUTHOR QUERIES

AQ1. Please provide the significance for given italic values in Tables "22.2, 22.3, and 22.4".

AQ2. As per style, a citation for Table 8.5 is inserted in the appendix section. Kindly check and confirm.

AQ3. Refs. Benzécri (1976); Demsar (2006); Hubert and Arabie (1985); Mantel (1967); Rifqi et al. (2003); Schneider and Borlund (2007,?); Warrens (2008) are not cited in the text. Please provide the citations or delete them from the list.