Journal of AppliedMath 2024, 2(1), 424.
https://doi.org/10.59400/jam.v2i1.424

Article

Topological analysis of multiple tables

Rafik Abdesselam

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

CITATION

Abdesselam R. Topological analysis of multiple tables. Journal of AppliedMath. 2024; 2(1): 424. https://doi.org/10.59400/jam.v2i1.424

ARTICLE INFO

Received: 21 December 2023
Accepted: 17 January 2024
Available online: 7 February 2024

Abstract: The paper proposes a topological approach in order to explore several data tables simultaneously. These data tables of quantitative and/or qualitative variables measured on different homogeneous themes, collected from the same individuals. This approach, called topological analysis of multiple tables (TAMT), is based on the notion of neighborhood graphs in the context of a joint analysis of several data tables. It allows the simultaneous study of possible links between several thematic tables. The structure of the correlations or associations of the variables in each thematic table is analyzed according to quantitative, qualitative or mixed variables considered. Like the multiple factorial analysis (MFA), the TAMT allows several tables of variables to be analyzed simultaneously, and to obtain results, in particular graphical representations, which make it possible to study the relationship between individuals, variables and tables of data. These can also be tables of temporal data, collected at different times on the same individuals. The proposed TAMT approach is illustrated using real data associated with several and different homogeneous themes. Its results are compared to those from the MFA method.

Keywords: multiple data tables; proximity measure; neighborhood graph; adjacency matrix; factorial analysis; clustering

1. Introduction

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

The proposed TAMT approach is different from those that already exist, in particular, the multiple factorial analysis (MFA) [1,2] with which it is compared, or also the structuring tables with three indices of the statistic (STATIS) [3,4] method or the double principal component analysis (DPCA) [5] method.

There are now many topological approaches to factor analysis and clustering [6–9] of a single table of homogeneous data, but as far as we know, none of these approaches has been proposed to analyze multiple data tables simultaneously.

The choice of proximity measure among the many existing measures plays an important role in multidimensional data analysis [10–12]. It has a strong impact on the results of any operation of structuring, grouping, or clustering of objects.

The structure of correlation or dependence of the quantitative or qualitative variables of each data table depends on the considered data.

Results may change depending on the proximity measure chosen for each data table, which allows for a measure of 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 an adjacency

matrix associated with a proximity measure, within the framework of the analysis of the correlation structure or dependence of a set of variables of a data table, and we present the principle of the proposed approach. This is illustrated in section 3 using an example based on real data. The results 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 multiple data tables 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.

The proposed TAMT consists of simultaneously analyzing several data tables collected on the same n individuals, from different thematic variables of each data table: $X(n, p_x)$, $Y(n, p_y)$, $Z(n, p_z)$, …, $T(n, p_t)$.

For example, for the data table X, we consider $E_x = \{x^1, \ldots, x^j, \ldots, x^{px}\}$ a set of p_x quantitative, qualitative, or even mixed variables [7].

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

Thus, for a given proximity measure u, we can build a neighborhood graph on E_x, 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) [13], the Gabriel Graph (GG) [14], or, as is the case here, the Relative Neighborhood Graph (RNG) [15].

Given a set E_x of p_x variables of the data table X and a proximity measure u, for continuous or binary data, we can construct the associated adjacency binary symmetric matrix Vu_x of order p_x, where, all pairs of neighboring variables in E_x satisfy the following RNG property:

$$Vu_x(x^k, x^l) = \begin{cases} 1, \text{if } u(x^k, x^l) \leq \max[u(x^k, x^t), u(x^t, x^l)] \\ \forall x^k, x^l, x^t \in E_x, x^t \neq x^k, x^t \neq x^l \\ 0, \text{otherwise} \end{cases}$$

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.

Figure 1 shows a simple example in IR^2 of four sets of thematic variables observed on the same n objects, which check the structure of the RNG graph with the Euclidean distance for each table, to establish the adjacency matrix of each thematic.

For example, for the data table X, we see that the adjacency value between the second and fourth variables, $Vu_x(x^2, x^4) = 1$, this means that geometrically, the hyper-Lunula (intersection between the two hyperspheres centered on the two variables x^2 and x^4 is empty.

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

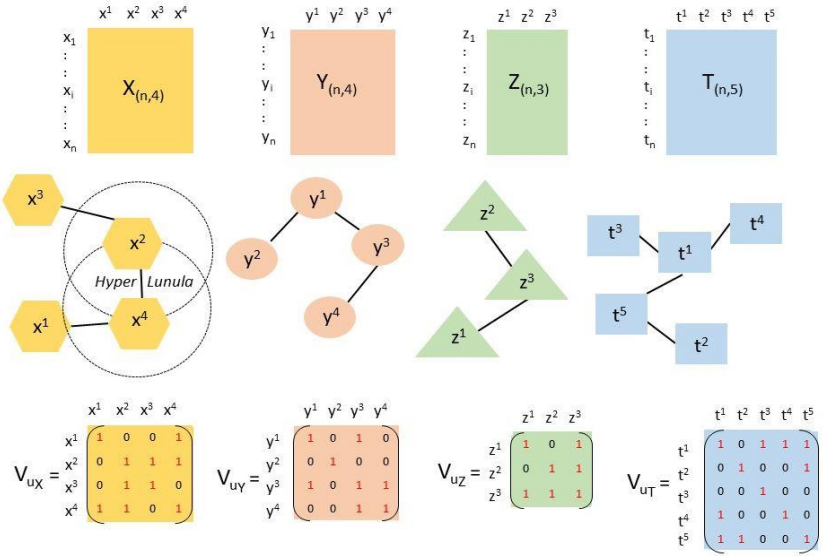

Figure 1. Multiple data tables—associated graphs and adjacency matrices.

For a given neighborhood property (MST, GG, or RNG), a proximity measure u chosen, among the numerous measures given in the Appendix in **Tables A1** and **A2**, we can generate a topological structure on the objects of E_x for the data table X, which is completely described by the associated binary adjacency matrix Vu_x.

2.1. Reference adjacency matrices

We first analyze in a topological way the correlation structures of the variables of each data table, to carry out a global and joint factorial analysis of these multiple data tables, then we establish on this simultaneous analysis, a clustering of individuals.

For each data table, X for example, we construct the reference adjacency matrix noted Vu_x^*, either from the correlation matrix or from the Burt's table profiles, depending on the type of variables in the data table X.

The definitions and expressions of adjacency reference matrices in the case of quantitative, qualitative, or mixed variables are given in Abdesselam [7,16].

To examine the correlation structure between the variables in data table X, 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 of data table X, the reference adjacency matrix Vu_{x^*} associated to reference measure u_x^* is defined as:

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

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: $\rho(x^k, x^l) \neq 0$.

The null hypothesis H_0 of no correlation is rejected with a p-value less than or equal to a chosen α significance level, for example α = 5%. The p-value is the probability of accepting or rejecting H_0.

Whatever the type of variables in table X, the constructed reference adjacency matrix Vu_x^* will be associated with an unknown reference proximity measure denoted

u_x^* We thus obtain as many reference adjacency matrices as multiple data tables considered.

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. TAMT factorial analysis and clustering & notations

We will use the following notations:

- We denote $G_{(n,p)} = [X_{(n,px)}|\ldots|Y_{(n,py)}|\ldots|T_{(n,pt)}]$ the global table, the juxtaposition of all the data tables considered, with n rows-individuals and $p = p_x + p_y + \ldots + p_t$ columns-variables,
- $X_{(n,px)}$ is the data table with n individuals and p_x variables,
- Vu_x^* is the symmetric adjacency matrix of order p_x, associated with the reference measure u_x^* which best structures the correlations of the variables of the data table X,
- $Vu^*(p) = \text{Diag}[Vu_x^*, Vu_y^*, \ldots, Vu_t^*]$ is the global diagonal adjacency matrix of order p, associated with the global data matrix G,
- $\hat{G}_{(n,p)} = GVu^*$ 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,
- $D_n = (1/n)\, I_n$ is the diagonal matrix of weights of order n in the space of variables.

Definition 2. The TAMT which analyze simultaneously the correlation structures of all the data tables, consists of carrying out the standardized PCA of the triplet ($\hat{G}$, M_p, D_n) PCA [17,18] of the projected data matrix $\hat{G} = GVu^*$.

Definition 3. The TAMT clustering consists of performing a HAC based on the Ward [19] criterion (aggregation based on the criterion of the loss of minimal inertia), on the significant factors of the TAMT factorial.

The TAMT factorial analysis is compared with the MFA method and the TAMT clustering with the HAC-MFA [20,21].

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

3. Illustrative example: Panorama of French metropolitan regions in 2021

To illustrate the TAMT approach, we use Insee (National Institute of Statistics and Studies) data [22–25] on the state of the 13 metropolitan regions of France in 2021. We consider four regional themes: The energy transition centered on Renewable Energies, Climate & Environment, Economic Dynamism, and Social Cohesion.

The description of the thematic variables is given in **Table 1** and summary statistics of these variables are presented in **Table 2**.

Table 1. Dictionary of thematic variables of metropolitan regions of France.

Theme	Identifier	Variable label
Renewable energies (RE)	CCRE	Coverage of electricity consumption by RE production (%)
	CCWP	Coverage of consumption of wind production (%)
	CCSP	Coverage of consumption of solar production (%)
	CCHP	Coverage of consumption of hydraulic production (%)
	CCBP	Coverage of consumption of bioenergy production (%)
Climate & environment (CE)	HSUN	Hours of sunshine (h)
	HRAI	Height of rainfall (mm)
	NPSI	Number of polluted sites—Pollution
	CARB	Carbon footprint (tCO2e per capita)
	CFOR	Cover forests (%)
Economic dynamism (ED)	BCRE	Business creation
	BFAI	Business failure
	GDPC	GDP per capita (M.€.)
	EMPL	Employment France (%)
Social cohesion (SC)	UNEM	Unemployment rate (%)
	POVE	Poverty rate (%)
	BASI	Beneficiaries of active solidarity income—RSA (%)
	RABO	Recipients of the activity bonus (%)
	SAEL	Social assistance for the elderly (%)
	SADP	Social assistance for disabled people (%)
	CSAS	Child social assistance (%)
	MSLH	Median standard of living of households (M.€.)

Table 2. Summary statistics of thematic variables of metropolitan regions.

Identifier	Mean	Standard deviation	Coefficient of variation (%)	Min	Max
CCRE	24.28	12.78	52.62	1.90	44.70
CCWP	8.01	6.23	77.79	0.40	20.60
CCSP	3.65	3.32	90.73	0.20	10.80
CCHP	10.98	12.29	111.89	0.00	39.30
CCBP	1.65	0.65	39.09	0.20	3.00
HSUN	2068.46	292.68	14.15	1847.00	2781.00
HRAI	606.92	84.94	13.99	442.00	731.00
NPSI	479.15	349.09	72.86	8.00	1131.00
CARB	9.69	0.72	7.44	9.02	11.27
CFOR	30.56	14.62	47.84	12.47	63.36
BCRE	20,689.08	19,151.84	92.57	1452.00	79,656.00
BFAI	2319.23	1877.07	80.94	142.00	6672.00
GDPC	33.62	8.49	25.25	29.12	62.11
EMPL	0.08	0.05	68.97	0.01	0.23
UNEM	0.08	0.01	12.47	0.07	0.09
POVE	8.21	1.62	19.78	5.80	10.70

Table 2. (*Continued*).

Identifier	Mean	Standard deviation	Coefficient of variation (%)	Min	Max
BASI	0.08	0.05	66.15	0.00	0.21
RABO	0.08	0.04	49.59	0.00	0.15
SAEL	0.08	0.03	45.13	0.01	0.13
SADP	0.08	0.04	46.77	0.01	0.14
CSAS	0.08	0.04	50.87	0.00	0.15
MSLH	21.52	0.88	4.07	20.11	23.86

Figure 2 presents the global reference adjacency matrix Vu^*associated with the proximity measure u^*, the most adapted to the four data tables considered, is constructed from the individual adjacency matrices associated with the multiple tables according to the Definition 1.

$$Vu^* = \begin{bmatrix} V_u{}^*RE & 0 & 0 & 0 \\ 0 & V_u{}^*CE & 0 & 0 \\ 0 & 0 & V_u{}^*ED & 0 \\ 0 & 0 & & V_u{}^*SC \end{bmatrix}$$

$$Vu^* = \left[\begin{array}{cccc}
10010 & 00000 & 0000 & 00000000 \\
01000 & 00000 & 0000 & 00000000 \\
00100 & 00000 & 0000 & 00000000 \\
10010 & 00000 & 0000 & 00000000 \\
00001 & 00000 & 0000 & 00000000 \\
\hline
00000 & 1\text{-}1101 & 0000 & 00000000 \\
00000 & \text{-}11000 & 0000 & 00000000 \\
00000 & 10100 & 0000 & 00000000 \\
00000 & 00010 & 0000 & 00000000 \\
00000 & 10001 & 0000 & 00000000 \\
\hline
00000 & 00000 & 1111 & 00000000 \\
00000 & 00000 & 1111 & 00000000 \\
00000 & 00000 & 1111 & 00000000 \\
00000 & 00000 & 1111 & 00000000 \\
\hline
00000 & 00000 & 0000 & 11100000 \\
00000 & 00000 & 0000 & 11000000 \\
00000 & 00000 & 0000 & 10111110 \\
00000 & 00000 & 0000 & 00111110 \\
00000 & 00000 & 0000 & 00111110 \\
00000 & 00000 & 0000 & 00111110 \\
00000 & 00000 & 0000 & 00111110 \\
00000 & 00000 & 0000 & 00000001
\end{array}\right]$$

Figure 2. Global reference adjacency matrix.

Note that in the case of quantitative variables, we consider that two positively correlated variables are linked and that two negatively correlated variables are linked, but distant, we will therefore take into account the sign of the correlation between variables in the adjacency matrix.

We established a TAMT to identify the correlation structure of the thematic variables, then carried out a CAH on the main factors of the TAMT, to give a typology of the regions according to the different themes.

The results of the TAMT approach and the MFA method were compared.

Figure 3 and **Table 3** present, for comparison on the first factorial plane, the correlations between principal components-factors and original variables. As can be seen, these correlations are slightly different, as are the percentages of inertia explained on the first principal planes of the TAMT and MFA method.

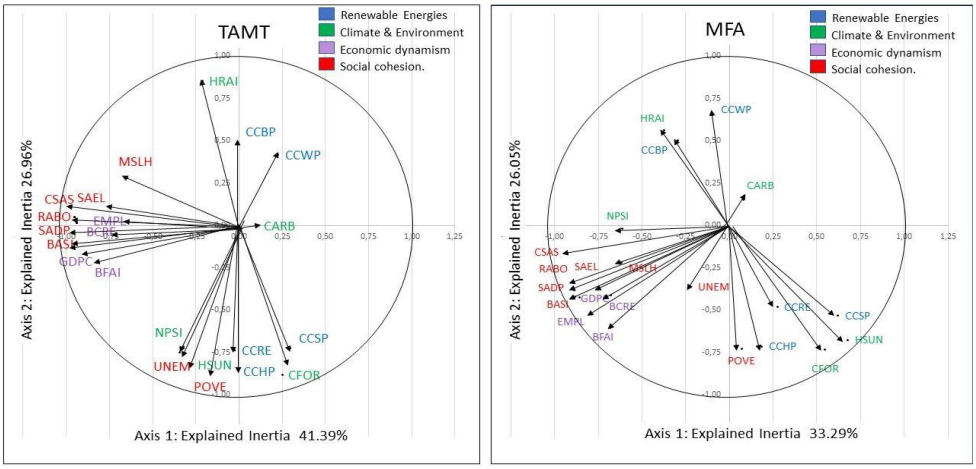

Figure 3. Representations of thematic variables.

Table 3. Eigenvalues of TAMT and MFA analyses.

TAMT N°	Eigen-value	Proportion (%)	Cumulative (%)	MFA N°	Eigen-value	Proportion (%)	Cumulative (%)
1	9.105	41.39	41.39	1	2.173	33.29	33.29
2	5.931	26.96	68.34	2	1.701	26.05	59.33
3	2.725	12.39	80.73	3	1.009	15.45	74.79
4	1.716	7.80	88.53	4	0.571	8.74	83.53
5	0.897	4.08	92.61	5	0.407	6.20	89.76
6	0.635	2.89	95.50	6	0.242	3.71	93.47
7	0.573	2.60	98.10	7	0.172	2.63	96.10
8	0.185	0.84	98.94	8	0.092	1.41	97.51
9	0.135	0.62	99.56	9	0.077	1.18	98.69
10	0.092	0.42	99.97	10	0.068	1.04	99.73
11	0.006	0.03	100.00	11	0.016	0.25	99.98
12	0.000	0.00	100.00	12	0.001	0.02	100.00
:	:	:	:	Total	6.529	100.00	
22	0.000	0.00	100.00				
Total	22.000	100.00					

Table 3 shows that the two first factors of the TAMT explain 41.39% and 26.96%, respectively, accounting for 68.34% of the total variation in the data set; however, the two first factors of the MFA add up to 59.34%.

Thus, the first two factors provide an adequate synthesis of the data, that is, of the four themes of the metropolitan regions of France in 2021. We restrict the comparison to the first significant factorial plan. In **Table 4**, the significant correlations between the initial variables and the principal factors in the two analyses are quite different.are quite different.

Table 4. Correlations initial variables & factors.

TAMT			MFA		
Variable	**F1**	**F2**	**Variable**	**F1**	**F2**
CCRE	−0.025	**−0.736**	CCRE	0.276	−0.480
CCWP	0.227	0.417	CCWP	−0.098	**0.652**
CCSP	0.294	**−0.724**	CCSP	0.619	**−0.531**
CCHP	−0.025	**−0.736**	CCHP	0.185	**−0.713**
CCBP	−0.012	0.486	CCBP	−0.295	**0.503**
HSUN	−0.280	**−0.822**	HSUN	**0.676**	**−0.674**
HRAI	−0.211	**0.854**	HRAI	−0.373	**0.564**
NPSI	−0.320	**−0.716**	NPSI	**−0.615**	−0.019
CARB	0.101	−0.001	CARB	0.080	0.154
CFOR	0.250	**−0.879**	CFOR	**0.545**	**−0.733**
BCRE	**−0.938**	0.017	BCRE	**−0.792**	**−0.521**
BFAI	**−0.938**	0.017	BFAI	**−0.683**	**−0.602**
GDPC	**−0.938**	0.017	GDPC	**−0.677**	−0.411
EMPL	**−0.938**	0.017	EMPL	**−0.884**	−0.418
UNEM	−0.342	−0.746	UNEM	−0.224	−0.361
POVE	−0.317	−0.757	POVE	0.069	**−0.725**
BASI	**−0.946**	0.032	BASI	**−0.856**	−0.423
RABO	**−0.951**	0.051	RABO	**−0.887**	−0.380
SAEL	**−0.951**	0.051	SAEL	**−0.759**	−0.376
SADP	**−0.951**	0.051	SADP	**−0.896**	−0.332
CSAS	**−0.951**	0.051	CSAS	**−0.927**	−0.161
MSLH	**−0.649**	0.282	MSLH	**−0.619**	−0.216

For comparison, **Figure 4** shows dendrograms of the TAMT and MFA clustering of the metropolitan regions of France according to the four themes considered. Note that the partitions TAMT and MFA chosen into 4 clusters of regions are identical.

Indeed, the compositions of the clusters are identical while the characterizations of these clusters are slightly different. Furthermore, the percentage of total variance explained by the TAMT approach, R^2 = 72.82%, is much higher than that of the MFA approach, R^2 = 66.47%, thus indicating that the clusters of the TAMT approach are more homogeneous than those of the MFA.

Figure 5 illustrates the typology in 4 colors on the map of the metropolitan regions of France. For comparison, **Figure 5** also summarizes the results of the tests of significant profiles (+) and anti-profiles (−) of the two typologies, with a risk of error less than or equal to 5%. The characterizations are very little different, differences are located and specified in bold and with an asterisk.

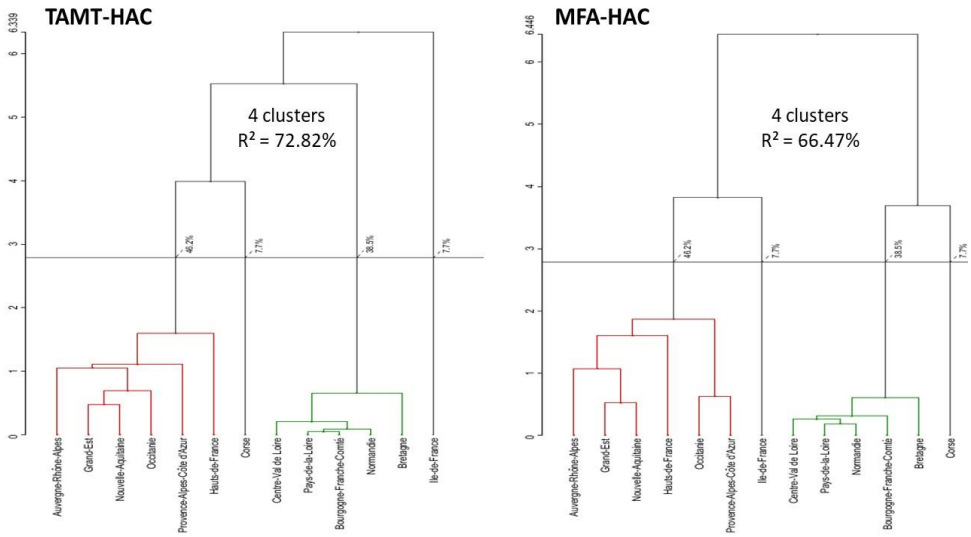

Figure 4. Hierarchical trees of metropolitan regions of France.

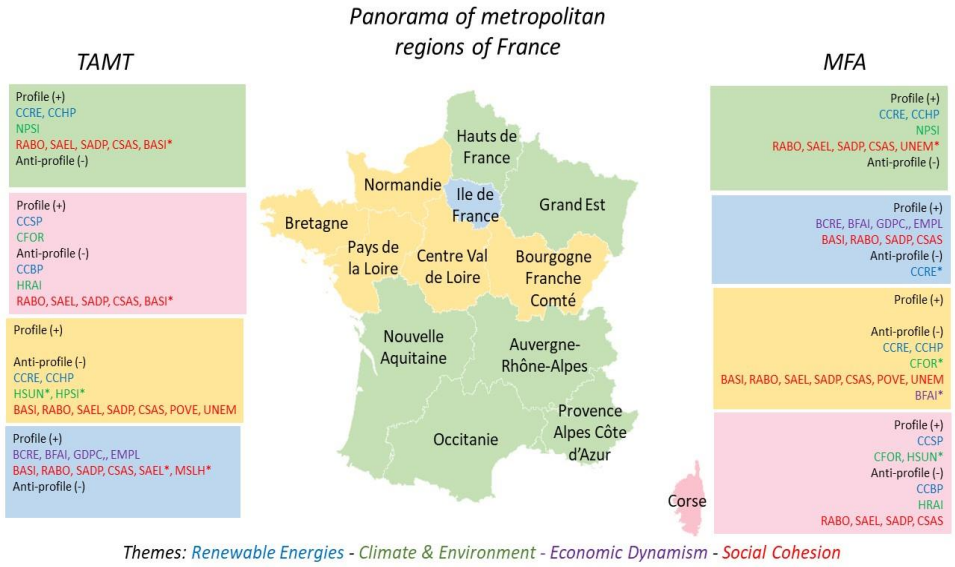

Figure 5. Typologies of regional clusters according to themes.

The first TAMT cluster, composed of six regions (Auvergne-Rhône-Alpes, Grand-Est, Occitanie, Provence-Alpes-Côte-Azur), is characterized by high coverage of electricity consumption by RE and in particular by Hydraulic production, relative to the national average of the ER theme variables. It has a significant number of polluted sites that are harmful to the Climate & Environment. These regions have a significantly high proportion of recipients of the RSA, the activity bonus, social assistance for the elderly, and disabled people, and social assistance for children.

The second cluster represents the Corsica region only, which is characterized by significant coverage of solar electricity consumption and high forest coverage. It also has low coverage of bioenergy electricity consumption and low precipitation from a climatic point of view. This region has a low proportion of beneficiaries of RSA, activity bonuses, and social assistance for the elderly, disabled people, and children.

The third cluster, bringing together the regions of Bretagne, Center Val-de-Loire,

Pays de la Loire, Bourgogne-Franche-Comté, and Normandy, is characterized by a low coverage of electricity consumption by RE and more particularly by Hydraulic Production, compared to the average for Metropolitan France. It has a low number of polluted sites and hours of sunshine. These regions have a significantly low proportion of beneficiaries of the RSA, the activity bonus, social assistance for the elderly, and disabled people and social assistance for children. As well as low poverty and unemployment rates compared to the national level.

The last fourth cluster represents the Ile-de-France region characterized by a significant number of business creations and failures, a high GDP per capita and a high percentage of jobs in France. This region has a significantly high proportion of beneficiaries of the RSA, the activity bonus, social assistance for the elderly, disabled people and children. It also has a significantly high median household standard of living.

4. Conclusion

This paper proposes a new topological approach to analyze simultaneous multiple data tables, which can enrich classical data analysis methods. The results of this factorial and clustering approach, based on the notion of a neighborhood graph, are better than those of the classic MFA method, according to the results of the percentages of inertia explained by the principal factors, and according to the R-squared. It would be interesting to make a Benchmark to evaluate the results of this topological approach on massive data tables (big data). Future work consists of extending this topological approach to other methods of data analysis, in particular in the context of prediction models.

Conflict of interest: The author declares no conflict of interest.

References

1. Dazy F, Le Barzic JF, Saporta G, et al. Evolutionary Data Analysis—Methods and Applications (French). Editions TECHNIP; 1996.
2. Escofier B, Pagès J. Implementation of MFA for Numerical, Qualitative, or Mixed Tables (French). Publication Interne de l'IRISA; 1985. p. 429.
3. Lavit C. Joint Analysis of Quantitative Tables (French). Editions Masson; 1988.
4. des Planttes HLH. Structuring Tables with Three Statistical Indices (French). Université des Sciences et Techniques du Languedoc; 1976.
5. Bouroche JM. Analysis of Ternary Data: Double Principal Component Analysis (French) [PhD thesis]. Université de Paris VI; 1975.
6. Abdesselam R. A topological clustering of individuals. In: Brito P, Dias JG, Lausen B, et al. Classification and Data Science in the Digital Age. Springer; 2022. pp. 1–8.
7. Abdesselam R. A Topological Clustering of Variables. Journal of Mathematics and System Science. 2021, 11(2): 1–17. doi: 10.17265/2159-5291/2021.02.001
8. Aljarah I, Faris H, Mirjalili S, et al. Evolutionary Data Clustering: Algorithms and Applications. Springer Singapore, 2021. doi: 10.1007/978-981-33-4191-3
9. Panagopoulos D. Topological data analysis and clustering. In: Daras NJ, Pardalos PM, Rassias M (editors). Analysis, Cryptography and Information Science. World Scientific; 2023.
10. Batagelj V, Bren M. Comparing resemblance measures. Journal of Classification. 1995, 12(1): 73-90. doi: 10.1007/bf01202268

11. Lesot MJ, Rifqi M, Benhadda H. Similarity measures for binary and numerical data: a survey. International Journal of Knowledge Engineering and Soft Data Paradigms. 2009, 1(1): 63. doi: 10.1504/ijkesdp.2009.021985
12. Zighed DA, Abdesselam R, Hadgu A. Topological Comparisons of Proximity Measures. Lecture Notes in Computer Science. Published online 2012: 379-391. doi: 10.1007/978-3-642-30217-6_32
13. Kim JH, Lee S. Tail bound for the minimal spanning tree of a complete graph. Statistics & Probability Letters. 2003, 64(4): 425-430. doi: 10.1016/s0167-7152(03)00208-6
14. Park JC, Shin H, Choi BK. Elliptic Gabriel graph for finding neighbors in a point set and its application to normal vector estimation. Computer-Aided Design. 2006, 38(6): 619-626. doi: 10.1016/j.cad.2006.02.008
15. Toussaint GT. The relative neighbourhood graph of a finite planar set. Pattern Recognition. 1980, 12(4): 261-268. doi: 10.1016/0031-3203(80)90066-7
16. Abdesselam R. Mixed principal component analysis (French). In: Classification: Crossed Points of View (French). Cépaduès; 2008. pp. 31–41.
17. Caillez F, Pagès JP. Introduction to Data Analysis (French). SMASH; 1976.
18. Lebart L. Survey data processing strategy (French). La Revue de MODULAD. 1989; 3: 21–29.
19. Ward JH. Hierarchical Grouping to Optimize an Objective Function. Journal of the American Statistical Association. 1963, 58(301): 236-244. doi: 10.1080/01621459.1963.10500845
20. Fowlkes EB, Mallows CL. A Method for Comparing Two Hierarchical Clusterings. Journal of the American Statistical Association. 1983, 78(383): 553-569. doi: 10.1080/01621459.1983.10478008
21. Hubert L, Arabie P. Comparing partitions. Journal of Classification. 1985, 2(1): 193-218. doi: 10.1007/bf01908075
22. 2021 economic reports of French regions (French). Available online: https://www.insee.fr/fr/information/6456000 (accessed on 31 December 2021).
23. Overview of renewable electricity (French). Available online: https://assets.rte-france.com/prod/public/2022-02/Pano-2021-T4.pdf (accessed on 31 December 2021).
24. Poverty in the regions (French). Available online: https://www.inegalites.fr/La-pauvrete-dans-les-regions (accessed on 31 December 2021).
25. France map of carbon footprint by region (French). Available online: https://www.hellocarbo.com/empreinte-carbone-francais-2021-par-region/ (accessed on 31 December 2021).

Appendix

Table A1. Some proximity measures for continuous data.

Measure	Formula: Distance—Dissimilarity—Continuous data
Euclidean	$u_{\text{Euclidean}}(x,y) = \sqrt{\sum_{j=1}^{p}(x_j - y_j)^2}$
Manhattan	$u_{\text{Manhattan}}(x,y) = \sum_{j=1}^{p}\lvert x_j - y_j\rvert$
Minkowski-ν	$u_{\text{Minkowski}}(x,y) = \left(\sum_{j=1}^{p}\lvert x_j - y_j\rvert^{\nu}\right)^{1/\nu}$
Cosine dissimilarity	$u_{\text{Cosine}}(x,y) = 1 - \frac{\sum_{j=1}^{p} x_j y_j}{\sqrt{\sum_{j=1}^{p} {x_j}^2}\sqrt{\sum_{j=1}^{p} {y_j}^2}}$
Pearson correlation	$u_{\text{Correlation}}(x,y) = 1 - \frac{(\sum_{j=1}^{p}(x_j - \bar{x})(y_j - \bar{y}))^2}{\sum_{j=1}^{p}(x_j - \bar{x})^2 \sum_{j=1}^{p}(y_j - \bar{y})^2}$
Squared chord	$u_{\text{Chord}}(x,y) = \sum_{j=1}^{p}(\sqrt{x_j} - \sqrt{y_j})^2$
Doverlap measure	$u_{\text{Doverlap}}(x,y) = \max\left(\sum_{j=1}^{p} x_j, \sum_{j=1}^{p} y_j\right) - \sum_{j=1}^{p}\min(x_j, y_j)$
Gower	$u_{\text{Gower}}(x,y) = \frac{1}{p}\sum_{j=1}^{p}\lvert x_j - y_j\rvert$
Shape distance	$u_{\text{Shape}}(x,y) = \sqrt{\sum_{j=1}^{p}\left[(x_j - \bar{x}_j) - (y_j - \bar{y}_j)\right]^2}$
Size distance	$u_{\text{Size}}(x,y) = \left\lvert\sum_{j=1}^{p}(x_j - y_j)\right\rvert$
Lpower	$u_{\text{Lpower}}(x,y) = \sum_{j=1}^{p}\lvert x_j - y_j\rvert^{\nu}$
Tchebytchev	$u_{\text{Tchebychev}}(x,y) = \max_{1\le j\le p}\lvert x_j - y_j\rvert$
Normalized Euclidean	$u_{\text{NEuclidean}}(x,y) = \sqrt{\sum_{j=1}^{p}\frac{1}{\sigma^2}\left[(x_j - \bar{x}_j) - (y_j - \bar{y}_j)\right]^2}$
Canberra	$u_{\text{Canberra}}(x,y) = \sum_{j=1}^{p}\frac{\lvert x_j - y_j\rvert}{\lvert x_j\rvert + \lvert y_j\rvert}$
Weighted Euclidean	$u_{\text{WEuclidean}}(x,y) = \sqrt{\sum_{j=1}^{p}\alpha_j(x_j - y_j)^2}$

where, p is the dimension of space, $x = (xj)_j = 1, \ldots, p$ and $y = (yj)_j = 1, \ldots, p$ two points in Rp, xj the mean, σj the standard deviation, $\alpha j = 1\backslash\sigma j 2$ and $\nu > 0$.

Table A2. Some proximity measures for binary data.

Measure	Similarity—Binary data	Dissimilarity
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	$s_{14} = \frac{4(ad-bc)}{(a+b)^2+(b+c)^2}$	$u_{14} = \frac{1-s_{14}}{2}$
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	$s_{17} = \frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}$	$u_{17} = \frac{1-s_{17}}{2}$
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, $a = |X \cap Y|$ is the number of attributes common to both points x and y, $b = |X - Y|$ is the number of attributes present in x but not in y, $c = |Y - X|$ is the number of attributes present in y but not in x and $d = |\overline{X} \cap \overline{Y}|$ is the number of attributes in neither x or y and $|.|$ the cardinality of a set.