Analyse Topologique de Tableaux Multiples

Rafik Abdesselam

Laboratoires ERIC-COACTIS, Université Lumière Lyon 2, France
rafik.abdesselam@univ-lyon2.fr

Résumé. Cet article propose une approche topologique pour explorer et analyser plusieurs tableaux de données simultanément. Il s'agit de tableaux de variables quantitatives et/ou qualitatives mesurées sur différentes thématiques homogènes, collectées sur les mêmes individus. Cette approche, appelée Analyse Topologique de Tableaux Multiples (ATTM), est basée sur la notion de graphes de voisinage dans le cadre d'une analyse conjointe de plusieurs tableaux de données. Elle permet d'étudier les éventuels liens entre plusieurs tables thématiques. La structure des corrélations ou des associations des variables de chaque tableau thématique est analysée selon le type des variables quantitatives, qualitatives ou mixtes considérées. Comme l'Analyse Factorielle Multiple (AFM), l'ATTM permet d'analyser plusieurs tableaux de variables simultanément, d'obtenir des résultats, notamment des représentations graphiques, et d'étudier la relation entre individus, variables et tableaux de données. Il peut s'agir également de tableaux de données temporelles, collectées à différents moments sur les mêmes individus. L'approche ATTM proposée est illustrée à l'aide de données réelles associées à plusieurs thématiques homogènes différentes. Les résultats sont comparés à ceux de la méthode AFM.

Mots-clés. Tableaux multiples, mesure de proximité, graphe de voisinage, matrice d'adjacence, analyse factorielle et classification.

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's 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 the 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 and clustering.

1 Introduction

L'objectif de cet article est de proposer une approche topologique d'analyse des données appliquée à plusieurs tableaux de données croisant les mêmes individus avec différentes variables quantitatives, qualitatives ou encore mixtes.

L'approche TAMT proposée est différente de celles qui existent déjà, notamment l'Analyse Factorielle Multiple (AMF) [Escofier et Pagès (1985), Dazy *et al.* (1996)] avec laquelle elle est comparée, ou de la méthode des Tableaux Structurants à Trois Indices de la Statistique (STATIS) [Lavit (1988)] ou encore de la méthode de l'Analyse en Composantes Doubles Principales (DPCA) [Bouroche (1975)]. Il existe désormais de nombreuses approches topologiques d'analyse factorielle et de clustering [Abdesselam (2021,2022), Aljarah *et al.* (2021), Panagopoulos (2022)] d'un seul tableau de données homogènes, mais à notre connaissance, aucune de ces approches n'a été proposée pour analyser plusieurs tableaux de données simultanément.

Le choix de la mesure de proximité parmi les nombreuses mesures existantes, joue un rôle important dans une analyse de données multidimensionnelles [Batagelj et Bren (1995), Lesot *et al.* (2009), Zighed *et al.* (2012)]. Elle a un fort impact sur les résultats de toute opération de structuration, de regroupement ou de clustering d'objets.

La structure de corrélation ou de dépendance des variables quantitatives ou qualitatives de chaque tableau de données dépend des données considérées. Les résultats peuvent changer selon la mesure de proximité choisie dans chaque tableau de données, qui permet de mesurer la similarité ou la dissemblance entre deux objets individus ou variables.

Ce document est organisé comme suit. Dans la section 2, nous rappelons brièvement la notion de base des graphes de voisinage, nous définissons et montrons comment construire une matrice d'adjacence associée à une mesure de proximité dans le cadre de l'analyse de la structure de corrélation ou de dépendance d'un ensemble de variables d'un tableau de données, et nous présentons les principes de l'approche proposée. Cette dernière est illustrée dans la section 3 à l'aide d'un exemple basé sur des données réelles. Les résultats sont comparés à ceux de la classification appliquée aux résultats de l'AFM. Enfin, la section 4 présente quelques remarques sur ce travail.

2 Topologie et tableaux de données multiples

L'analyse topologique de données est une approche basée sur le concept de graphe de voisinage. L'idée de base est assez simple, pour une mesure de proximité donnée selon le type de variables continues ou binaires et pour une structure topologique choisie, on peut faire correspondre un graphe topologique induit sur l'ensemble des objets.

L'ATTM proposée consiste à analyser simultanément plusieurs tableaux de données collectées sur les mêmes n individus, à partir des différentes variables thématiques de chaque tableau de données : $X_{(n,p_x)}, Y_{(n,p_y)}, Z_{(n,p_z)}, \cdots, T_{(n,p_t)}$.

Par exemple, pour un tableau de données X, on considère $E_x = \{x^1, \cdots, x^j, \cdots, x^{p_x}\}$ un ensemble de p_x variables quantitatives, qualitatives ou encore mixtes [Abdesselam (2021)].

On peut, au moyen d'une mesure de proximité u, définir une relation de voisinage, notée V_u, comme étant une relation binaire sur $E_x \times E_x$. De nombreuses possibilités existent pour construire cette relation de voisinage. Ainsi, pour une mesure de proximité u donnée, nous pouvons construire un graphe de voisinage sur E_x, où les sommets sont les variables et les arêtes sont définies à partir de la propriété de la relation de voisinage.

De nombreuses définitions sont possibles pour construire cette relation binaire selon le graphe de voisinage choisi, l'Arbre de Longueur Minimalle (ALM), le Graphe de Gabriel (GG), ou encore, comme c'est le cas ici, le graphe des voisins relatifs (GVR). Pour une propriété de voisinage donnée (ALM, GG ou GVR) et une mesure de proximité u choisie, parmi les nombreuses mesures existantes, on peut par exemple pour un tableau de données X, générer une structure topologique sur E_x qui est entièrement décrite par la matrice d'adjacence binaire associée V_{u_x}. d'ordre p_x, où toutes les paires de variables voisines dans E_x satisfont la propriété GVR suivante :

$$V_{u_x}(x^k\ ,\ x^l) = \begin{cases} 1 & \text{si } u(x^k\ ,\ x^l) \ \leq \ \max[u(x^k\ ,\ x^t), u(x^t\ ,\ x^l)]\ ; \\ & \qquad\qquad \forall x^k,\ x^l,\ x^t \in E,\ x^t \neq x^k \ \ et \ \ x^t \neq x^l \\ 0 & \text{sinon.} \end{cases}$$

Cela signifie que si deux variables x^k et x^l qui vérifient la propriété GVR sont connectées par une arête, les sommets x^k et x^l sont voisins.

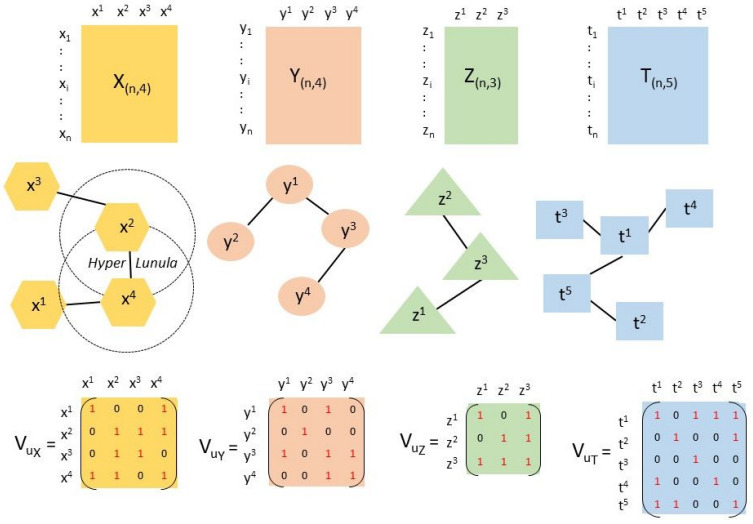

Figure 1: Tablaux multiples - Graphes et matrices d'adjacence associées

La Figure 1 montre un exemple simple dans R^2 de tableaux de variables de quatre thématiques observées sur les mêmes n objets, selon la structure de voisinage GVR et la distance euclidienne pour chaque thématique.

Par exemple, pour le tableau de données de la thématique X, la valeur d'adjacence entre la deuxième et la quatrième variable, $V_{u_x}(x^2, x^4) = 1$, géométriquement, cela signifie que l'hyper-Lunule (intersection entre les deux hypersphères centrées sur les deux variables x^2 et x^4) est vide. Cela génère une structure topologique basée sur les p_x objets dans E_x qui sont complètement décrits par la matrice binaire d'adjacence V_{u_x}.

2.1 Matrices d'adjacence de référence

Nous analysons d'abord de manière topologique les structures de corrélation ou de dépendance des variables de chaque tableau de données, pour réaliser une analyse factorielle globale et conjointe de ces multiples tableaux, puis nous établissons sur cette analyse simultanée, une classification des individus.

Pour chaque tableau de données, X par exemple, on construit la matrice d'adjacence de référence notée $V_{u_x^\star}$, soit à partir de la matrice de corrélation pour des variables quantitatives, soit à partir des profils du tableau de Burt pour des variables qualitatives. Les définitions et expressions des matrices d'adjacence de référence selon le type de variables considérées sont données dans [Abdesselam (2021), (2008)].

Pour analyser la structure de corrélation entre les variables quantitatives du tableau de données X par exemple, nous examinons la signification de leur coefficient de corrélation linéaire. La matrice d'adjacence peut s'écrire comme suit en utilisant le test t de Student du coefficient de corrélation linéaire ρ de Bravais-Pearson :

Pour les variables quantitatives du tableau de données X, la matrice d'adjacence de référence $V_{u_x^\star}$ associée à la mesure de référence $u_x^\star$ est définie comme suit :

$$\mathrm{V}_{u_x^\star}(x^k, x^l) = \begin{cases} 1 & \text{si} \quad \text{p-valeur } = P[\mid T_{n-2} \mid > \text{t-valeur }] \leq \alpha \ ; \ \forall k, l = 1, p \\ 0 & \text{sinon.} \end{cases} \tag{1}$$

Test de Student de signification du coefficient de corrélation linéaire. L'hypothèse nulle H_0 d'absence de corrélation est rejetée avec une p-valeur inférieure ou égale à un risque d'erreur α choisi, par exemple $\alpha = 5\%$. La p-valeur est la probabilité d'accepter ou de rejeter H_0.

Quelle que soit le type des variables du tableau X, la matrice d'adjacence de référence construite $V_{u_x^\star}$ sera associée à une mesure de proximité inconnue notée $u_x^\star$. On obtient ainsi autant de matrices d'adjacence de référence que de tableaux de données considérés.

La robustesse dépend du risque d'erreur α choisi pour l'hypothèse nulle, pas corrélation linéaire dans le cas de variables quantitatives ou indépendance dans le cas de variables qualitatives. On peut fixer un seuil minimum afin d'analyser la sensibilité des résultats. Certes les résultats numériques seront différents, mais probablement pas leur interprétation.

2.2 Analyse factorielle & Classification

Pour définir l'ATTM, nous utiliserons les notations suivantes :

- On note $G_{(n,p)} = [X_{(n,p_x)}|\cdots|Y_{(n,p_y)}|\cdots|T_{(n,p_t)}]$ le tableau de données global, juxtaposition de tous les tableaux de données considérés, à n lignes-individus et $p = p_x + p_y + \cdots + p_t$ colonnes-variables,

- $X_{(n,p_x)}$ est le tableau de données à n individus et p_x variables,

- $V_{u^\star{}_x}$ est la matrice d'adjacence symétrique d'ordre p_x, associée à la mesure de référence $u_x\star$ qui structure au mieux les corrélations des variables du tableau de données X,

- $V_{u^\star(p)} = Diag[V_{u^\star x}, V_{u^\star y}, \cdots, V_{u^\star t}]$ est la matrice d'adjacence globale diagonale d'ordre p, associée à la matrice de données globale G,

- $\widehat{G}_{(n,p)} = GV_{u^\star}$ est la matrice des données projetées à n individus et p variables,

- M_p est la matrice des distances d'ordre p dans l'espace des individus,

- $D_n = \frac{1}{n} I_n$ est la matrice diagonale des poids d'ordre n dans l'espace des variables.

Définition 1 : L'ATTM qui analyse simultanément les structures de corrélation de tous les tableaux de données, consiste à réaliser l'ACP standardisée du triplet $(\widehat{G}, M_p, D_n)$ [Caillez et Pagès (1976), Lebart (1989)] de la matrice de données projetée $\widehat{G} = GV_{u^\star}$.

Définition 2 : La classification ATTM consiste à appliquer une CAH selon le critère de Ward[1] sur les facteurs significatifs de l'analyse factorielle ATTM.

L'analyse factorielle ATTM est comparée à la méthode AFM et la classification ATTM à la méthode CAH-MFA [Fowlkes et Mallows (1983), Hubert et Arabie (1985)].

Enfin, l'approche ATTM et son dendrogramme sont facilement programmables à partir des procédures ACP et CAH des logiciels SAS, SPAD ou R.

3 Exemple illustratif

Panorama des régions métropolitaines de France en 2021 : pour illustrer l'approche ATTM, on a utilisé les données de l'Insee[2][Bilan économique, Rte, Inégalités, Empreinte (2021)] sur l'état des 13 régions métropolitaines de France. On a considéré quatre thématiques régionales : les Energies Renouvelables, le Climat & Environnement, le Dynamisme Economique et la Cohésion Sociale. Les libellés et satistiques sommaires des variables thématiques sont consignés dans la Table 1.

Le tableau 2 présente la matrice d'adjacence globale de référence $V_{u_\star}$ associée à la mesure de proximité $u_\star$, mesure la plus adaptée à chacun des quatre tableaux de données considérés, construite selon l'expression (1).

A noter que dans ce cas de variables quantitatives, on considère que deux variables corrélées positivement sont liées et que deux variables corrélées négativement sont également liées, mais distantes, on prendra donc en compte le signe de la corrélation des variables dans la matrice d'adjacence.

On a d'abord effectué une ATTM pour identifier la structure des corrélations des variables thématiques, puis réalisé une CAH sur les principaux facteurs de cette approche pour établir une typologie des régions selon les différentes thématiques. Les résultats de l'approche ATTM sont comparés à ceux de l'AFM.

A titre de comapaison, la Figure 2 et le Tableau 3 présentent sur le premier plan factoriel, les corrélations entre les facteurs principaux et les variables initiales. Comme on peut le

[1] Agrégation basée sur le critère de perte d'inertie minimale.

[2] Insee - Institut National de la Statistique et des Etudes Economiques

constater, ces corrélations sont légèrement différentes, tout comme les pourcentages d'inerties expliquées sur les premiers plans principaux des méthodes ATTM et AFM.

La Figure 2 montre que les deux premiers facteurs de l'ATTM expliquent respectivement 41.39% et 26.96%, représentant 68.34% de la variation totale, alors que les deux premiers facteurs de l'AFM totalisent que 59.34%.

Table 1: Dictionnaire & Statistiques sommaires des variables thématiques

Identifier	Variable label
	Theme: Renewable Energies (RE)
CCRE	Coverage of Electricity Consumption by RE Production (%)
CCWP	Coverage of Consumption of Wind power Production (%)
CCSP	Coverage of Consumption of Solar Production (%)
CCHP	Coverage of Consumption of Hydraulic Production (%)
CCBP	Coverage of Consumption of Bioenergy Production (%)
	Theme - Climate & Environment (CE)
HSUN	Hours of Sunshine (h)
HRAI	Height of Rainfall (mm)
NPSI	Number of Polluted Sites - Pollution
CARB	Carbon Footprint (tCO2e/hab)
CFOR	Cover Forests (%)
	Theme: Economic Dynamism (ED)
BCRE	Business Creation
BFAI	Business Failure
GDPC	GDP per capita (M.€.)
EMPL	Employment France (%)
	Theme: 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.€.)

Identifier	Mean	Std.Deviation	Min	Max
CCRE	24,28	12,78	1,90	44,70
CCWP	8,01	6,23	0,40	20,60
CCSP	3,65	3,32	0,20	10,80
CCHP	10,98	12,29	0,00	39,30
CCBP	1,65	0,65	0,20	3,00
HSUN	2068,46	292,68	1847,00	2781,00
HRAI	606,92	84,94	442,00	731,00
NPSI	479,15	349,09	8,00	1131,00
CARB	9,69	0,72	9,02	11,27
CFOR	30,56	14,62	12,47	63,36
BCRE	20689,08	19151,84	1452,00	79656,00
BFAI	2319,23	1877,07	142,00	6672,00
GDPC	33,62	8,488	29.12	62,11
EMPL	0,08	0,05	0,01	0,23
UNEM	0,08	0,01	0,07	0,09
POVE	8,21	1,62	5,80	10,70
BASI	0,08	0,05	0,00	0,21
RABO	0,08	0,04	0,00	0,15
SAEL	0,08	0,03	0,01	0,13
SADP	0,08	0,04	0,01	0,14
CSAS	0,08	0,04	0,00	0,15
MSLH	21,52	0,88	20,11	23,86

Table 2: Matrice d'adjacence globale de référence

$$V_{u^*} = \left(\begin{array}{ccccc|ccccc|cccc|cccccccc}
1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 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 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 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 & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & 1 & -1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 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 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{array}\right)$$

Le premier plan factoriel de l'ATTM fournit une synthèse de la structure des corrélations des variables thématiques des régions françaises en 2021. Les corrélations significatives entre les variables initiales et les facteurs principaux des deux approches sont légèrement différentes. A titre de comparaison, la Figure 3 présente les dendrogrammes des classifications ATTM et AFM des régions de France selon les quatre thématiques considérées. A noter que les typologies en 4 classes de régions des deux approches sont identiques. En effet, les compositions des 4 classes sont identiques tandis que les caractérisations de ces classes sont légèrement différentes. De plus, le pourcentage de variance totale expliqué par l'approche

ATTM, $R^2 = 72.82\%$, est bien supérieur à celui de l'approche AFM, $R^2 = 66.47\%$, indiquant ainsi que les classes de l'approche ATTM sont plus homogènes que celles de l'AFM.

Table 3: Corrélations Variables & Facteurs

ATTM	Facteur	
Variable	F1	F2
CCRE	0,025	**-0,736**
CCWP	-0,227	0,417
CCSP	-0,294	**-0,724**
CCHP	0,025	**-0,736**
CCBP	0,012	0,486
HSUN	0,280	**-0,822**
HRAI	0,211	**0,854**
NPSI	0,320	**-0,716**
CARB	-0,101	-0,001
CFOR	-0,250	**-0,879**
BCRE	**0,938**	0,017
BFAI	**0,938**	0,017
GDPC	**0,938**	0,017
EMPL	**0,938**	0,017
UNEM	0,342	**-0,746**
POVE	0,317	**-0,757**
BASI	**0,946**	0,032
RABO	**0,951**	0,051
SAEL	**0,951**	0,051
SADP	**0,951**	0,051
CSAS	**0,951**	0,051
MSLH	**0,649**	0,282

AFM	Facteur	
Variable	F1	F2
CCRE	-0,276	-0,480
CCWP	0,098	**0,652**
CCSP	**-0,619**	**-0,531**
CCHP	-0,185	**-0,713**
CCBP	0,295	**0,503**
HSUN	**-0,676**	**-0,674**
HRAI	0,373	**0,564**
NPSI	**0,615**	-0,019
CARB	-0,080	0,154
CFOR	**-0,545**	**-0,733**
BCRE	**0,792**	**-0,521**
BFAI	**0,683**	**-0,602**
GDPC	**0,677**	-0,411
EMPL	**0,884**	-0,418
UNEM	0,224	-0,361
POVE	-0,069	**-0,725**
BASI	**0,856**	-0,423
RABO	**0,887**	-0,380
SAEL	**0,759**	-0,376
SADP	**0,896**	-0,332
CSAS	**0,927**	-0,161
MSLH	**0,619**	-0,216

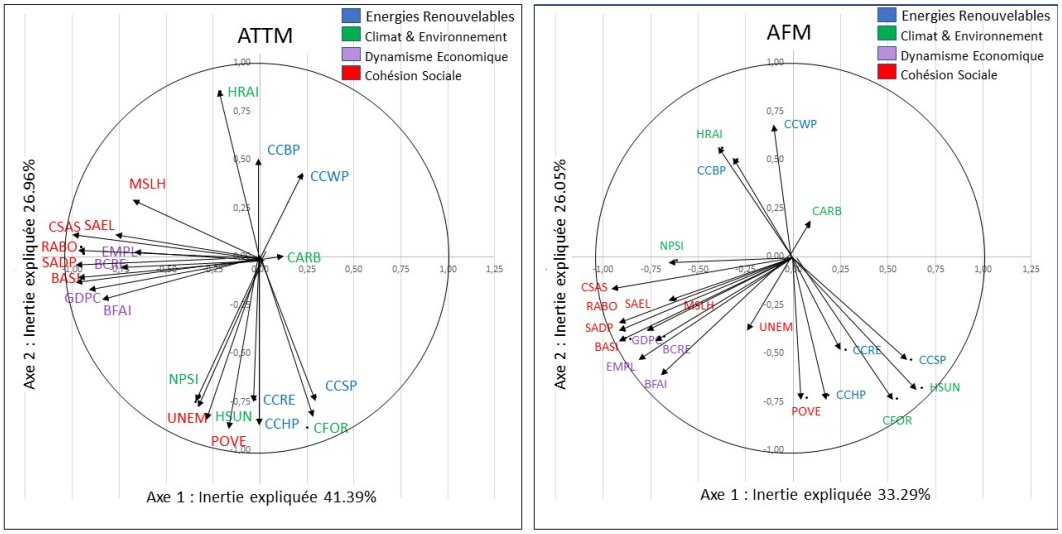

Figure 2: Représentations des variables thématiques

La figure 4 illustre en 4 couleurs, les typologies ATTM et AFM sur la carte des régions métropolitaines de France. A titre de comparaison, la Figure 4 résume les résultats significatifs des profils (+) et anti-profils (-) des deux typologies, avec un risque d'erreur inférieur ou égal à 5%. Les caractérisations diffèrent très peu, les différences sont repérées et précisées en gras et avec un astérisque.

Le premier classe ATTM, composée de six régions (Auvergne-Rhône-Alpes, Grand-Est, Occitanie, Provence-Alpes-Côte-Azur), est caractérisée par une forte couverture de la consommation électrique par les EnR, notamment par la Production Hydraulique, relativement

à la moyenne nationale des variables thématiques. Elle compte un nombre important de sites pollués nuisibles au Climat et à l'Environnement. Ces régions comptent une proportion significativement élevée de bénéficiaires du RSA, de la prime d'activité, des aides sociales aux personnes âgées, aux personnes handicapées et à l'enfance.

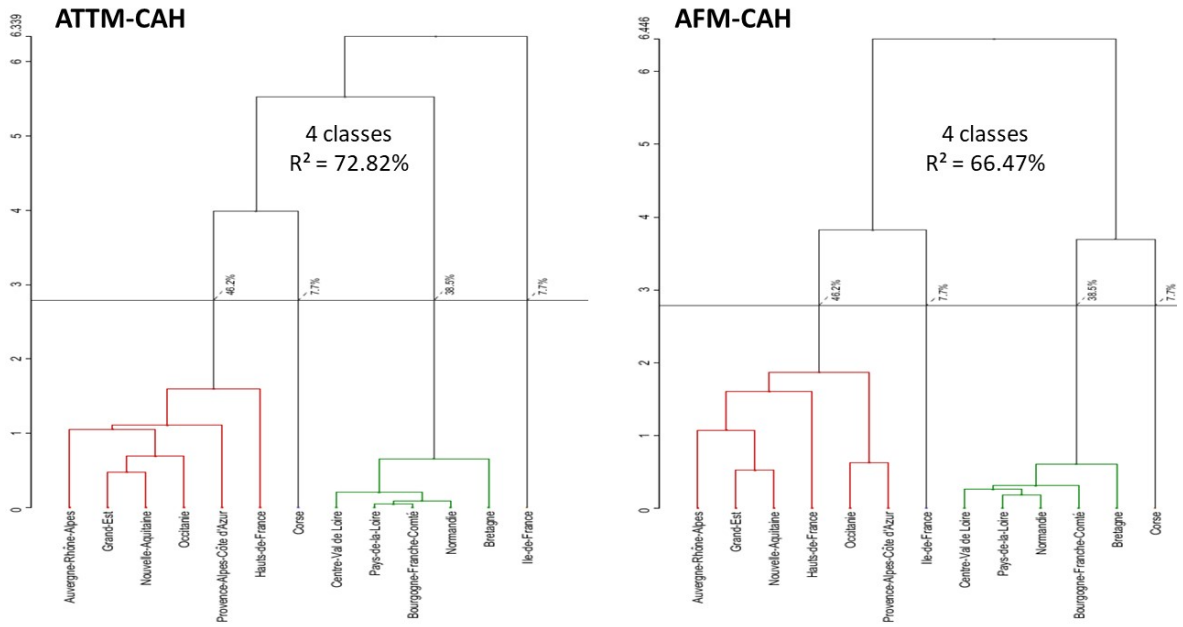

Figure 3: Arbres hiérarchiques des régions métropolitaines de France

La deuxième classe représente la région Corse, qui se caractérise par une couverture importante de la consommation d'électricité solaire et une forte couverture forestière. Elle présente également une faible couverture de la consommation d'électricité bioénergétique et de faibles précipitations d'un point de vue climatique. Cette région compte une faible proportion de bénéficiaires du RSA, de la prime d'activité, et des aides sociales aux personnes âgées, aux personnes handicapées et à l'enfance.

La troisième classe, regroupant 5 régions (Bretagne, Centre Val-de-Loire, Pays de la Loire, Bougogne-Franche-Comté et Normandie), est caractérisée par une faible couverture de la consommation électrique par les EnR et plus particulièrement par la Production Hydraulique, par rapport à la moyenne de la France métropolitaine. Les régions de cette classe présentent un faible nombre de sites pollués et un faible ensoleillement. Ces régions comptent une proportion significativement faible de bénéficiaires du RSA, de la prime d'activité, des aides sociales aux personnes âgées, aux personnes handicapées et l'enfance. Ainsi que des taux de pauvreté et de chômage faibles par rapport au niveau national.

La dernière classe représente la région Ile-de-France caractérisée par un nombre important de créations et de faillites d'entreprises, un PIB par habitant élevé et un pourcentage d'emplois élevé par rapport à la moyenne nationale. Cette région compte une proportion significativement élevée de bénéficiaires du RSA, de la prime d'activité, des aides sociales aux personnes âgées, aux personnes handicapées et à l'enfance. Le niveau de vie médian des ménages y est également très élevé.

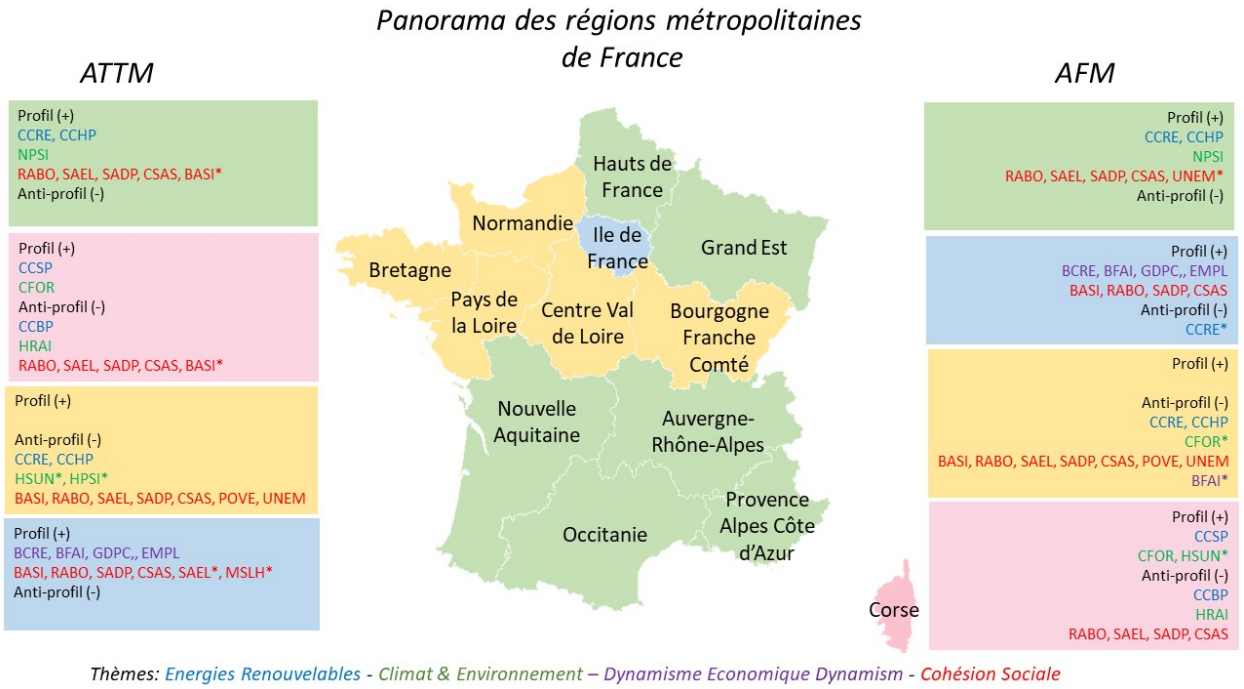

Figure 4: Typologies des classes régionales selon les thématiques

4 Conclusion

Cet article propose une nouvelle approche topologique pour analyser simultanément plusieurs tableaux de données, qui peut enrichir les méthodes classiques d'analyse de données. Les résultats de cette approche factorielle et clustering, basée sur la notion de graphe de voisinage, sont meilleurs que ceux de la méthode classique AFM, selon les pourcentages d'inerties expliquées par les facteurs principaux, et le R^2. Il serait intéressant de réaliser un Benchmark pour évaluer les résultats de cette approche topologique sur des tableaux de données massives (big data). Les travaux futurs consistent à étendre cette approche topologique à d'autres méthodes d'analyse de données, notamment dans le cadre des modèles de prédiction.

Bibliographie

Abdesselam, R. (2022) *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.

Abdesselam, R. (2021) *A Topological Clustering of variables.* Journal of Mathematics and System Science. David Publishing Company,Vol.11, Issue 2, pp.1-17, 2021.

Abdesselam, R. (2008) *Analyse en Composantes Principales Mixte*. Classification : points de vue croisés, RNTI-C-2, Cépaduès Editions, 31-41, 2008.

Aljarah, I., Faris, H. and Mirjalali S. (2021) *Evolutionary data clustering: algorithms and applications*, Springer, 2021.

Panagopoulos, D. (2022) *Topological data analysis and clustering*. Chapter for a book, Algebraic Topology (math.AT) arXiv:2201.09054, Machine Learning, 2022.

Dazy, F., Le Barzic, J.F., Saporta, G., Lavallard F. (1996) *L'analyse des données évolutives – Méthodes et applications*. Editions TECHNIP, 1996.

Escofier, B. et Pagès, J. (1985) *Mise en oeuvre de l'AFM pour des tableaux numériques, qualitatifs, ou mixtes*. Publication interne de l'IRISA, 429, 1985.

Bouroche, J.M. (1975) *Analyse des données ternaires : la double analyse en composantes principales*. Thèse, 1975.

Lavit, C. (1988) *Analyse conjointe de tableaux quantitatifs*. Editions Masson, 1988.

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

Caillez, F. and Pagès, J.P. (1976) *Introduction à l'Analyse des données*. SMASH, Paris, 1976.

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

Lesot, M. J., Rifqi, M. and Benhadda, H. (2009) *Similarity measures for binary and numerical data: a survey*. In IJKESDP, 1, 1, 63-84, 2009.

Zighed, D., Abdesselam, R., and Hadgu, A. (2012) *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.

Bilans économiques 2021 des régions françaises.
https://www.insee.fr/fr/information/6456000

Panorama de l'électricité renouvelable 31/12/2021,
https://assets.rte-france.com/prod/public/2022-02/Pano-2021-T4.pdf.

La pauvreté dans les régions. Observatoire des inégalités, 2021.
https://www.inegalites.fr/La-pauvrete-dans-les-regions.

Carte de France de l'empreinte carbone par région, édition 2021.
https://www.hellocarbo.com/empreinte-carbone-francais-2021-par-region/.

Fowlkes, E.B., Mallows, C.L. (1983) *A Method for Comparing Two Hierarchical Clusterings*. Journal of the American Statistical Association, 78(383), 53-–569, 1983.

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