International Journal of Mathematics, Statistics and Operations Research
Vol. 6; Number 1; 2026, pp. 143-161
https://doi.org/10.47509/IJMSOR.2026.v06i01.08

ARF INDIA
Academic Open Access Publishing
www.arfjournals.com

Topological Data Analysis Methods Application to European Union Public Finance

ABDESSELAM, Rafik

University of Lyon, Lumière Lyon 2, ERIC - COACTIS Laboratories
Department of Economics and Management, 69365 Lyon, France

ABSTRACT

The objective of this work is to propose Topological Principal Component Analysis (TPCA) and Topological Discriminant Analysis (TDA) according to the type of data: quantitative, qualitative, or mixed variables.

Large datasets are increasingly common across many disciplines. The exponential growth of data requires the development of more advanced data analysis methods in order to process information efficiently. To better visualize data, many methods, such as PCA, allow the extraction of a low-dimensional structure from high-dimensional datasets. The proposed TPCA approach is a multidimensional descriptive method that studies a homogeneous set of continuous variables defined on the same set of individuals.

Topological Discriminant Analysis (TDA) is a supervised classification method that aims to uncover the intrinsic structures and discriminative information embedded in the data. Many predictive techniques exist and are widely applied to a variety of problems across numerous fields. Classification is the process that assigns each individual in a population to one of several predefined classes based on their characteristics, which are considered explanatory variables.

The two proposed topological approaches, TPCA and TDA, are based on the notion of neighborhood graphs: one in an exploratory and descriptive context, and the other in a decision-making and predictive context.

The variables may be more or less correlated or related depending on their type—quantitative, qualitative, or mixed. These topological methods analyze the structure of correlations or dependencies observed among the variables under consideration.

A real-data example illustrates the two topological methods. The results of TPCA are compared with those of classical PCA, while the results of TDA are compared with those of various machine learning predictive models.

KEYWORDS

proximity measure, neighborhood graph, adjacency matrix, principal components analysis, discriminant analysis, predictive modeling, machine learning models.

CONTACT Author. Email:rafik.abdesselam@univ-lyon2.fr

Article History

Received: 27 April 2026; Revised: 23 May 2026; Accepted: 31 May 2026; Published: 30 June 2026

To cite this paper

ABDESSELAM, Rafik (2026). Topological Data Anlysis Methods Application to European Union Public Finance. *International Journal of Mathematics, Statistics and Operations Research.* 6(1), 143-161.

1. Introduction

Large datasets are increasingly common and are often difficult to interpret. Principal Component Analysis (PCA) [15, 21, 25, 10, 20, 13, 26, 17, 8, 5, 3] is a technique for reducing the dimensionality of such datasets, improving interpretability while minimizing information loss. It is an exploratory tool primarily designed for continuous data.

PCA is an adaptive technique for continuous data, and several variants of this method have been developed and tailored to different data types and structures. To properly interpret large datasets, methods are required to significantly reduce their dimensionality in an interpretable way. Many techniques have been developed for this purpose, but PCA remains one of the oldest and most widely used. From a statistical perspective, PCA is a widely used multivariate method for dimensionality reduction and data representation. Its aim is to identify common factors, known as principal components, in the form of linear combinations of the variables under study. It provides insight into the correlation structure among variables, as well as potential similarities in behavior between individuals.

In the context of artificial intelligence, we often compare situations represented by a set of objects. To do so, it is necessary to choose and define a proximity measure between these objects. The study context, the data type, and other factors can help determine an appropriate proximity measure. However, the number of possible measures can still be quite large. Moreover, are all these measures equivalent? Is one measure more appropriate or better suited than another for a given study? In information retrieval, the choice of proximity measure is a crucial issue, as the results strongly depend on it.

The present study proposes a new framework for comparing proximity measures in order to identify those that are similar. This approach avoids the need to test all possible measures. These comparisons are based on proximity measures that evaluate the similarity or dissimilarity between two objects within a dataset. Proximity measures are defined by specific mathematical properties and well-established axioms. The most appropriate measure is selected according to the correlation structure of the set of quantitative variables to be synthesized. The aim is to establish a topological PCA. The results of TPCA may vary depending on the selected proximity measure.

Topological Discriminant Analysis (TDA), proposed for quantitative, qualitative, or mixed explanatory variables, is a predictive model comparable to many existing machine learning techniques [28]. It can be viewed as a form of artificial intelligence (AI) used to build predictive models. TDA is notably compared to classical Discriminant Analysis (DA) [9, 4], a classification method that has long been used in many contexts, as well as to Logistic Regression (LR) [2]. It is also compared with other widely used machine learning models, such as K-nearest neighbors (KNN) [38], Support Vector Machines (SVM) [22], Random Forests (RF) [34], Bayesian Networks (BN) [37], Decision Trees (DT) [35], and Neural Networks (NN) [36].

Machine learning techniques are both explanatory and predictive. They can be used to assess whether predefined groups of individuals are distinct, to identify their characteristics based on explanatory variables, and to predict the group to which a new individual belongs.

Predictive modeling is widely used across various sectors and fields. In marketing, it can be used to predict consumer preferences, personalize promotional offers, and target online advertising. In finance and insurance (e.g., credit scoring), predictive modeling plays a key role in risk analysis, market trend forecasting, and investment management.

In healthcare and medicine, such models are used to support medical diagnosis, identify health trends, personalize patient treatment, and improve the management of medical records. Predictive modeling is also increasingly applied in the technology and internet sectors, human resources, and environmental sciences to support and enhance decision-making processes.

Discriminant Analysis (DA) assumes that the explanatory variables are normally distributed and that the within-group covariance matrices are equal. However, it is known to be relatively robust to violations of these assumptions and generally provides a reliable framework for supervised classification and decision-making.

Although several approaches specific to discriminant analysis have been developed, to the best of our knowledge, none have been proposed within a topological framework.

The objective of this paper is to propose a topological approach to discriminant analysis applicable to quantitative, qualitative, or mixed explanatory variables.

The choice of a proximity measure among the many existing measures plays a crucial role in multidimensional data analysis [1, 18, 30]. It has a strong impact on the results of any operation involving the structuring, grouping, or classification of objects.

The correlation or dependence structure of the quantitative or qualitative variables in each dataset depends on the data under consideration. Consequently, the results may vary depending on the proximity measure selected for each dataset. A proximity measure is a function that quantifies the similarity or dissimilarity between two objects or variables within a set.

This paper is organized as follows. Section 2 briefly reviews the basic concept of neighborhood graphs. It then defines and describes how to construct adjacency matrices associated with proximity measures within the framework of analyzing the correlation structure of a set of explanatory variables. The principles of the TDA approach are also presented. Section 3 illustrates the method using a real-data example. The results of the proposed topological analyses are compared with those obtained from classical Discriminant Analysis (DA), Logistic Regression (LR), and other machine learning predictive models. Finally, Section 4 provides concluding remarks.

2. Topological context of PCA

Topological equivalence is based on the concept of a topological graph, also referred to as a neighborhood graph. The basic idea is straightforward: two proximity measures are considered equivalent if the corresponding topological graphs induced on the set of objects are identical. Measuring the similarity between proximity measures therefore involves comparing their associated neighborhood graphs and quantifying their similarity.

We first define more precisely what a topological graph is and how it is constructed. Then, we propose a measure of similarity between topological graphs, which will subsequently be used to compare proximity measures.

Consider a set $E = \{x^1, x^2, \ldots, x^k, \ldots, x^p\}$ of p objects in $\mathbb{R}^n$,associated with p variables. Using a proximity measure u, we can define a neighborhood relation V_u as a binary relation on $E \times E$. There are many possibilities for building this neighborhood binary relationship.

Thus, for a given proximity measure u, a neighborhood graph can be constructed on the set of object-variables, where the vertices represent the variables and the edges are defined according to a specified neighborhood property.

Many definitions can be used to construct this binary neighborhood relation. One

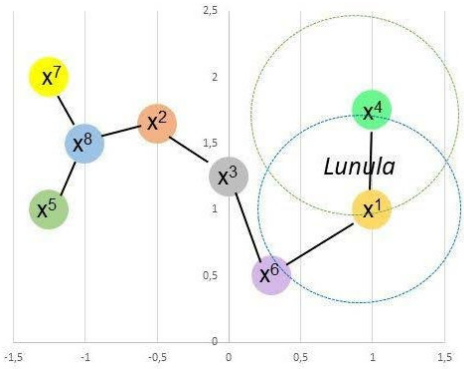
may consider the Minimum Spanning Tree (MST) [16], the Gabriel Graph (GG) [24], or, as in this work, the Relative Neighborhood Graph (RNG) [27].

For a given neighborhood property (MST, GG, or RNG), each proximity measure u_i induces a topological structure on the objects in E, which is fully described by the corresponding binary adjacency matrix V_{u_i}.

For any given proximity measure u, we construct the associated adjacency binary symmetric matrix V_u of order p, where, all pairs of neighboring variables (x^k, x^l), where $k, l = 1, p$, satisfy the following RNG property.

Property 1.

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



V_u	x^1	x^2	x^3	x^4	x^5	x^6	x^7	x^8
x^1	1	0	0	1	0	1	0	0
x^2	0	1	1	0	0	0	0	1
x^3	0	1	1	0	0	1	0	0
x^4	1	0	0	1	0	0	0	0
x^5	0	0	0	0	1	0	0	1
x^6	1	0	1	0	0	1	0	0
x^7	0	0	0	0	0	0	1	1
x^8	0	1	0	0	1	0	1	1

Figure 1. RNG example with eight variables - Adjacency matrix.

This means that if two variables x^k and x^l satisfy the RNG property and are connected by an edge, the vertices x^k and x^l are considered neighbors.

Thus, for any given proximity measure u, we can associate a binary symmetric adjacency matrix V_u of order p. Figure 2 illustrates an example of an RNG in $\mathbb{R}^2$ for a set of $p = 8$ object-variables.

For instance, if $V_u(x^1, x^4) = 1$, this indicates that on the geometric plane, the hyperlune (the intersection of the two hyperspheres centered at the variables x^1 and x^4)is empty.

For a given neighborhood property (MST, GG, or RNG), each proximity measure u induces a topological structure on the objects in E, which is fully described by the binary adjacency matrix V_u.

2.1. *Topological equivalence between two proximity measures*

We compare different proximity measures based on their topological similarity, aiming to group them and better visualize their relationships.

To assess the topological equivalence between two proximity measures, u_i and u_j, we propose testing whether their associated adjacency matrices, V_{u_i} and V_{u_j}, differ. The degree of topological equivalence between two proximity measures is quantified using the following concordance property.

Property 2.

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

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

We can thus compare two proximity measures based on their topological equivalence. The similarity $S(V_{u_i}, V_{u_j})$ between two proximity measures ranges from 0 to 1. It is considered perfect when $S(V_{u_i}, V_{u_j}) = 1$, indicating that the results obtained with the two measures are identical.

Fisher's exact test [12, 23] can be applied with a p-value threshold of $\alpha = 5\%$ to assess the statistical significance of the similarity between two measures. This test is an alternative to the Chi-square test when sample sizes are small. Its principle is to determine whether the observed configuration in the contingency table is extreme relative to all possible configurations, taking into account the marginal distributions.

Fisher's exact test is classified as an exact test because the probabilities are calculated precisely, rather than using an approximation, as is the case with the Chi-square test, which is only asymptotically valid for large samples. Unlike tests based on test statistics with known asymptotic distributions, Fisher's test computes the exact p-value directly.

To statistically evaluate the topological equivalence between two proximity measures, this nonparametric test compares the measures via their associated adjacency matrices. Two proximity measures are considered statistically topologically equivalent if the null hypothesis H_0 of independence is rejected.

The comparison of proximity measure indices has also been studied from a statistical perspective by [1, 31, 32, 11]. These authors proposed an approach that compares the similarity matrices obtained from each proximity measure using Mantel's test [19], in a pairwise manner.

Fisher's exact test is the statistical test most appropriate for comparing matched binary data. Cohen's Kappa test [33] can also be used, but it is generally an asymptotic test. For matched continuous data, the Kendall or Spearman correlation coefficients are appropriate.

In this context, Fisher's exact test allows us to measure the agreement or concordance between the binary values of two adjacency matrices associated with two proximity measures. Specifically, the test evaluates the topological equivalence between the corresponding proximity measures.

We evaluate the topological equivalence of each proximity measure $u_{i=1,15}$, presented in the Appendix in Table 11, with the reference measure u^* by comparing their adjacency matrices V_{u_i} and V_{u^*}.

For quantitative variables, we construct the adjacency matrix $V_{u_\star}$ associated with the reference measure u^*, which corresponds to the correlation matrix. To analyze the correlation structure among the variables, we assess the significance of their linear correlation coefficients. This adjacency matrix can be defined using the t-test for the Pearson correlation coefficient ρ :

Property 3.

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

The p-value represents the significance of the correlation coefficient in a two-sided test of the null and alternative hypotheses: $H_0 : \rho(x^k, x^l) = 0$ vs. $H_1 : \rho(x^k, x^l) \neq 0$.

It quantifies the evidence against the null hypothesis: the smaller the p-value, the stronger the evidence that the null hypothesis should be rejected, indicating that there is a statistically significant correlation between the variables x^k and x^l in the population.

The Student t-test for the significance of the correlation coefficient is given by: $t = r\sqrt{\frac{n-2}{1-r^2}}$ where $r = r(x^k, x^l)$ is the observed linear correlation coefficient between the variables, and $\nu = n2$ is the number of degrees of freedom.

Let T_ν denote a Student t-distributed random variable with $\nu = n2$ degrees of freedom. The null hypothesis is rejected if the p-value is less than or equal to a chosen significance level, e.g., $\alpha = 5\%$. A very small p-value indicates a low probability that the null hypothesis is true, justifying its rejection. In general, statistical significance is achieved when the p-value is smaller than the predetermined α level. The p-value thus represents the probability of observing data at least as extreme as that obtained, assuming that the null hypothesis is true.

The robustness with respect to the chosen significance level α for testing the null hypothesis of no linear correlation can be assessed by introducing a minimum threshold, allowing the sensitivity of the results to be analyzed. Although the numerical results may vary depending on this choice, their overall interpretation is expected to remain stable.

Two additional fundamental properties for constructing the adjacency matrix V_{u^*} in the case of qualitative or mixed variables are presented in [7, 8].

The binary and symmetric adjacency matrix V_{u^*} is associated with an unknown proximity measure, denoted $u*$, referred to as the reference measure. Using this reference proximity measure, we define $S(V_{u_i}, V_{u^*})$, which quantifies the topological equivalence between the proximity measures u_i and $u_\star$. This equivalence is evaluated by measuring the percentage of similarity between the adjacency matrix V_{u_i} and the reference adjacency matrix V_{u^*}.

Definition 1.

TPCA consist to perform the standardized PCA of the triple $\{\widehat{X}\ ;\ M\ ;\ D_p\}$, of the projected data matrix $\widehat{X} = XV_{u^*}$.

Interpretation of TPCA results is supported by concepts analogous to those used in PCA. Graphical representations on factorial planes enable the visualization and identification of the topological structure of the variables.

As in PCA, the representation of variables focuses on the most significant ones along the axes, namely those that are highly correlated with the factors, have strong contributions, and exhibit a high quality of representation. This quality is measured by the squared cosine of the angle between the principal axes and the original axes.

For the representation of active individuals, these are projected onto the factorial planes as illustrative elements. The quality of their representation on the factorial axes is also assessed using their squared cosine.

3. Topological context of TDA

TDA consists of simultaneously analyzing each data table associated with the modality classes of the target variable to be discriminated.

We consider the data table X_k, associated with the set of p explanatory quantitative variables $E_k = \{x^1, \cdots, x^j, \cdots, x^p\}$, observed for n_k individuals out of a total of

$n = \sum_{k=1}^{q} n_k$. These individuals share the modality k of the target variable $y = \{y^k; k = 1, q\}$, which comprises q modalities (groups).

Using a proximity measure u_k, we define a neighborhood relationship V_{u_k}, represented as a binary relation on $E_k \times E_k$. There are several possible ways to construct this binary neighborhood relationship.

Given the set E_k of p variables from the data table X_k and a proximity measure u_k, whether for continuous or binary data, we can construct the associated binary symmetric adjacency matrix V_{u_k} of order p, in which all pairs of neighboring variables in E_k satisfy property 1.

3.1. *Reference adjacency matrices*

The objective is to analyze, from both a topological and a discriminative perspective, the correlation or dependency structures of the explanatory variables in the data under consideration [7, 6].

The construction of the appropriate reference adjacency matrices depends on the type of explanatory variables considered, whether quantitative, qualitative, or a mixture of both.

We assume that we have a set $\{x^j; j = 1, \cdots, p\}$ of p quantitative variables observed on n individuals (objects), as well as a target qualitative variable $y = \{y^k; k = 1, q\}$ to be explained, with q modalities (groups).

The aim is to investigate whether a topological correlation structure exists among the variables under consideration [7].

We construct the adjacency matrix $V_{u^\star}$, which is derived from the correlation matrix. To analyze the correlation structure between variables, we assess the significance of their linear correlation coefficients. This adjacency matrix can be defined using the Student t-test for the Bravais–Pearson correlation coefficient ρ:

For each data table $X_k = (X/Y = k)$, we construct the reference adjacency matrix, denoted $V_{u_k^\star}$, in the case of quantitative variables, from the correlation matrix associated with X_k.

To examine the correlation structure among the variables in X_k, we assess the significance of their linear correlation coefficients. This adjacency matrix can be defined using the Student t-test for the Bravais–Pearson correlation coefficient ρ. For each quantitative data table X_k, the reference adjacency matrix $V_{u_k^\star}$ is associated with a reference measure $u_k^\star$, as defined in Property 3.

Regardless of the type of variables considered, the constructed reference adjacency matrix $V_{u_k^\star}$ is associated with an (unknown) reference proximity measure $u_k^\star$.

3.2. *Topological Discriminant Analysis - Notations*

Figure 2 presents an illustrative example involving five quantitative explanatory variables $\{x^j; j = 1, 5\}$ and a target qualitative variable $y = \{y^k; k = 1, 3\}$ with three modality classes. From the data tables corresponding to each class, $X_1 = (X/Y = 1), X_2 = (X/Y = 2)$ and $X_3 = (X/Y = 3)$, we construct the associated binary adjacency matrices V_{u_1}, V_{u_2} and V_{u_3}, based on the chosen neighborhood structures and proximity measures u_1, u_2 and u_3.

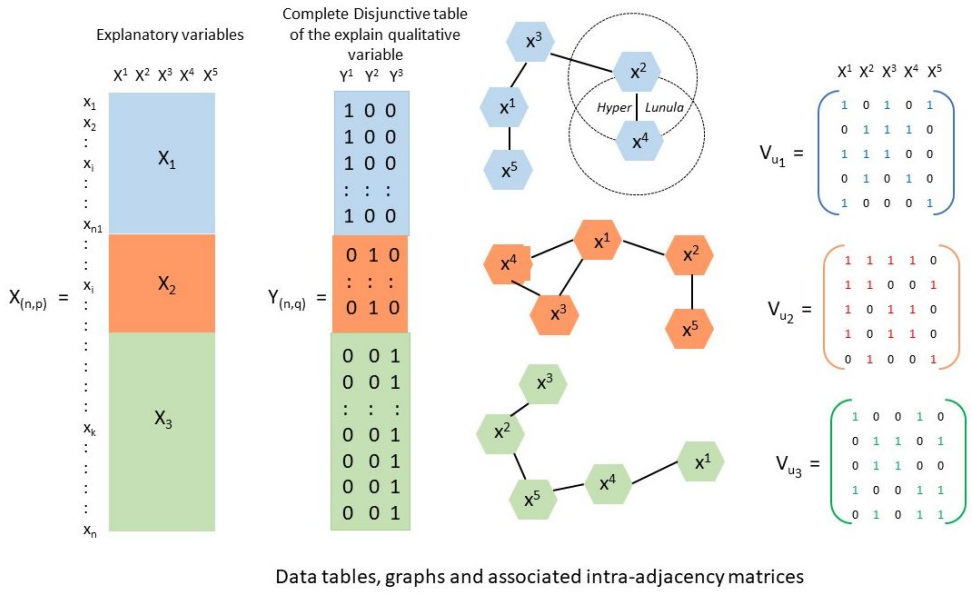


Figure 2. RNG - Within-Adjacency matrices for topological discrimination

For instance, in the data table X_1, we observe that for variables x^1 and x^4, $V_{u_1}(x^1, x^4) = 1$. This indicates that, in the geometric space, the hyper-lunula (i.e., the intersection of the two hyperspheres centered at x^1 and x^4 is empty.

We will use the following matrix notation:

$$X_{(n,p)} = \begin{pmatrix} X_{1_{(n_1,p)}} \\ \cdots\cdots \\ \vdots \\ \cdots\cdots \\ X_{k_{(n_k,p)}} \\ \cdots\cdots \\ \vdots \\ \cdots\cdots \\ X_{q_{(n_q,p)}} \end{pmatrix} \qquad \mathrm{Y}_{(n,q)} = \begin{pmatrix} Y_{1_{(n_1,q)}} \\ \cdots\cdots \\ \vdots \\ \cdots\cdots \\ Y_{k_{(n_k,q)}} \\ \cdots\cdots \\ \vdots \\ \cdots\cdots \\ Y_{q_{(n_q,q)}} \end{pmatrix}$$

$$R_{(qp,p)} = \begin{pmatrix} R_{1_{(p)}} \\ \cdots\cdots \\ \vdots \\ \cdots\cdots \\ R_{k_{(p)}} \\ \cdots\cdots \\ \vdots \\ \cdots\cdots \\ R_{q_{(p)}} \end{pmatrix} \qquad \mathrm{V}_{u^\star{}_{(qp,p)}} = \begin{pmatrix} V_{u^\star_{1(p)}} \\ \cdots\cdots \\ \vdots \\ \cdots\cdots \\ V_{u^\star_{k(p)}} \\ \cdots\cdots \\ \vdots \\ \cdots\cdots \\ V_{u^\star_{q(p)}} \end{pmatrix} \qquad \widehat{X}_{(n,p)} = \begin{pmatrix} X_1 V_{u_1^\star} \\ \cdots\cdots \\ \vdots \\ \cdots\cdots \\ X_k V_{u_k^\star} \\ \cdots\cdots \\ \vdots \\ \cdots\cdots \\ X_q V_{u_q^\star} \end{pmatrix}$$

- $X_(n, p)$ denotes the data matrix of the quantitative explanatory variables,
- $Y_(n, q)$ denotes the data matrix associated with the target qualitative variable to be explained, with q modalities,
- $X_{k_{(n_k,p)}} = (X/Y = k)$ denotes the explanatory data matrix corresponding to the n_k individuals having the $k - th$ modality of the target variable y,
- R_k denotes the correlation matrix of the data table X_k,

- $V_{u_k^\star}$ denotes the symmetric within-class adjacency matrix of order p, associated with the reference proximity measure $u_k^\star$, which best summarizes the correlation structure R_k of the data table X_k,
- $\widehat{X}_{(n,p)} = Diag[X]\, V_{u^\star}$ denotes the projected data matrix with n individuals and p variables,
- M_p denotes the distance matrix of order p in the space of individuals,
- $D_n = \frac{1}{n} I_n$ denotes the diagonal matrix of weights of order n in the space of individuals.

Definition 2.

TDA consists of performing a discriminant analysis on the significant factors of the Non-standardized topological PCA of the triplet $(\widehat{X}, M_p, D_n)$ of the projected data matrix $\widehat{X}$.

4. Illustrative example

To illustrate the two approaches, TPCA and TDA, we use Eurostat data [14] on government finance for the 27 European Union (EU) countries in 2024. We examine how key government finance indicators have evolved in the EU-27, focusing specifically on general government gross debt, deficit/surplus, total revenue, and total expenditure. Summary statistics for the variables considered are presented in Table 1.

Table 1. Summary statistics of public finances variables

Continuous variables	Frequency	Mean	Standard Deviation	Coefficient of variation	Min	Max
Deficit	27	-2.15	3.26	-1.52	-9.30	4.50
Expenditure	27	46.10	6.74	0.15	23.50	57.60
Revenue	27	43.94	5.87	0.13	27.80	53.20
Debt	27	65.22	33.20	0.51	23.60	153.60

4.1. *TPCA - Illustrative Example & Empirical results*

Table 2 shows the adjacency matrix V_{u^*}, associated with the adapted proximity measure u^* for the data under consideration. This matrix is constructed from the correlation matrix R according to Property 3., with a significance level $\alpha \leq 5\%$.

Table 2. Pearson correlation matrix (p-value) - Adjacency matrix

$$R = \begin{pmatrix} 1 & & & \\ (0.000) & & & \\ -0.493 & 1 & & \\ (0.009) & (0.000) & & \\ -0.009 & 0.874 & 1 & \\ (-0.965) & (0.000) & (0.000) & \\ -0.175 & 0.451 & 0.420 & 1 \\ (-0.382) & (-0.018) & (-0.029) & (0.000) \end{pmatrix} \quad \mathrm{V}_{u^*} = \begin{pmatrix} 1 & -1 & 0 & 0 \\ -1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{pmatrix}$$

In a topological framework, Table 3 summarizes the similarities $S(V_{u_i}, V_{u^*})$ and Fisher's Exact p-values between the 15 continuous proximity measures listed in Ap-

pendix Table 11 and the unknown reference measure u^*. The pairwise similarities vary, with some measures being closer to the reference than others.

The table also illustrates the 2×2 contingency table for the Canberra measure compared with the reference measure u^*, which is used to calculate Fisher's exact test. This test is significant, with a p-value of $0.0082(< \alpha = 5\%)$, leading to the rejection of the null hypothesis H_0 of independence.

None of the 15 measures are perfectly similar to the reference measure u^*; however, the Canberra measure is the one most significantly aligned with it.

The main numerical and graphical results of the proposed TPCA are presented in the following Tables and Figures and are compared with those of classical PCA.

Figure 3 shows, on the first factorial planes of both Topological and Classical PCA, the correlations between the principal components (factors) and the original variables. These graphical representations of the variables exhibit slight differences. Specifically, the percentage of inertia explained on the first principal plane of the Topological PCA is greater than that of the Classical PCA, and the variables showing significant correlations with the factors also differ.

Table 4 shows that the first two factors of TPCA explain 86.29% and 13.67% of the variance, respectively, accounting for 99.96% of the total inertia. In comparison, the first two factors of classical PCA explain a total of 82.94%. Therefore, the first two factors provide an adequate summary of the data—namely, the government finance of EU-27 countries—and the comparison of graphical representations is thus restricted to the first factorial plane.

Table 3. Similarities - 2×2 Contingency Table - Fisher's Exact Test

u_i	$S(V_{u_i}, V_{u^*})$	$p - value$
Euclidean	62.50%	0.6044
Manhattan	62.50%	0.6044
Minkovski	62.50%	0.6044
Cosine dissimilarity	62.50%	0.6044
Pearson Correlation	62.50%	0.6044
Squared Chord	62.50%	0.6044
Doverlap measure	62.50%	0.6044
Gower's Dissimilarity	62.50%	0.6044
Shape Distance	62.50%	0.6044
Size Distance	62.50%	0.6044
Lpower	62.50%	0.6044
Tchebytchev	62.50%	0.6044
Normalized Euclidean	62.50%	0.6044
Canberra	87.50%	0.0082
Weighted Euclidean	62.50%	0.6044

Reference measure measure	Canberra measure $V_{u_{canb}} = 0$	$V_{u_{canb}} = 1$
$V_{u^*} = 0$	2	0
$V_{u^*} = 1$	1	7

Fisher's Exact Test: p−Value $= 0.0082 < \alpha = 5\%$

For both PCA methods, the first factor contrasts countries with high expenditure, revenue, and debt against those with a high deficit. The second factor in TPCA represents primarily the deficit, whereas in classical PCA, the second factor is significantly correlated with both the deficit and revenue. This difference is reflected in the correlation tables between the original variables and the principal components.

Table 4. TPCA - Eigenvalues and Correlations Variables & Factors

TPCA - Eigenvalue	Proportion	Cumulative
3.452	86.29%	86.29%
0.547	13.67%	99.96%
0.002	0.04%	100.00%
0.000	0.00%	100.00%
4	100.00%	100.00%

Correlations Variables	Factors F1	F2
Deficit	0.736	0.677
Expenditure	-0.991	0.129
Revenue	-0.982	0.189
Debt	-0.982	0.189

PCA - Eigenvalue	Proportion	Cumulative
2.317	57.93%	57.93%
1.000	25.01%	82.94%
0.683	17.06%	100.00%
0.000	0.00%	100.00%
4	100.00%	100.00%

Correlations Variables	Factors F1	F2
Deficit	0.456	0.890
Expenditure	-0.967	0.056
Revenue	-0.857	0.432
Debt	-0.665	0.136

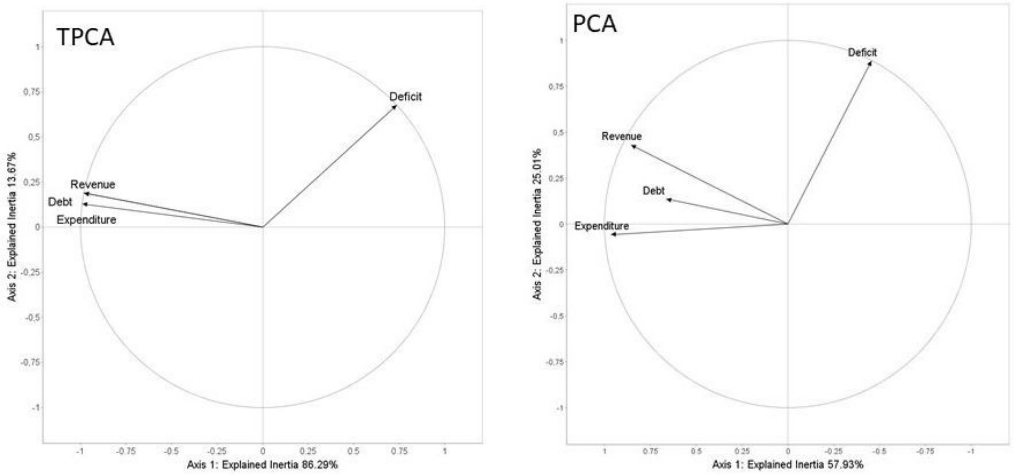


Figure 3. TPCA & PCA - The public finance variables on the first principal plane

For comparison, Figure 4 shows the representations of the 27 EU countries on the first principal planes.

A Hierarchical Ascending Clustering (HAC) algorithm based on Ward's criterion[1] [29] can be applied to the significant factors of TPCA to identify homogeneous groups of EU countries based on their public finances. For comparison, Figure 5 presents dendrograms of both the topological and classical clustering of the EU countries according to their public finances.

It should be noted that the partitions selected into four clusters differ considerably, both in composition and in characterization. The percentage of total variance explained by the HAC-TPCA approach, $R^2 = 83.06\%$, is higher than that of the HAC-PCA approach, $R^2 = 65.46\%$, indicating that the HAC-TPCA clusters are more homogeneous than those obtained from HAC-PCA.

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

The first HAC-TPCA cluster, composed of six countries (Belgium, Finland, France, Greece, Austria, and Italy), is characterized by high shares of expenditures, debt, and revenue.

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

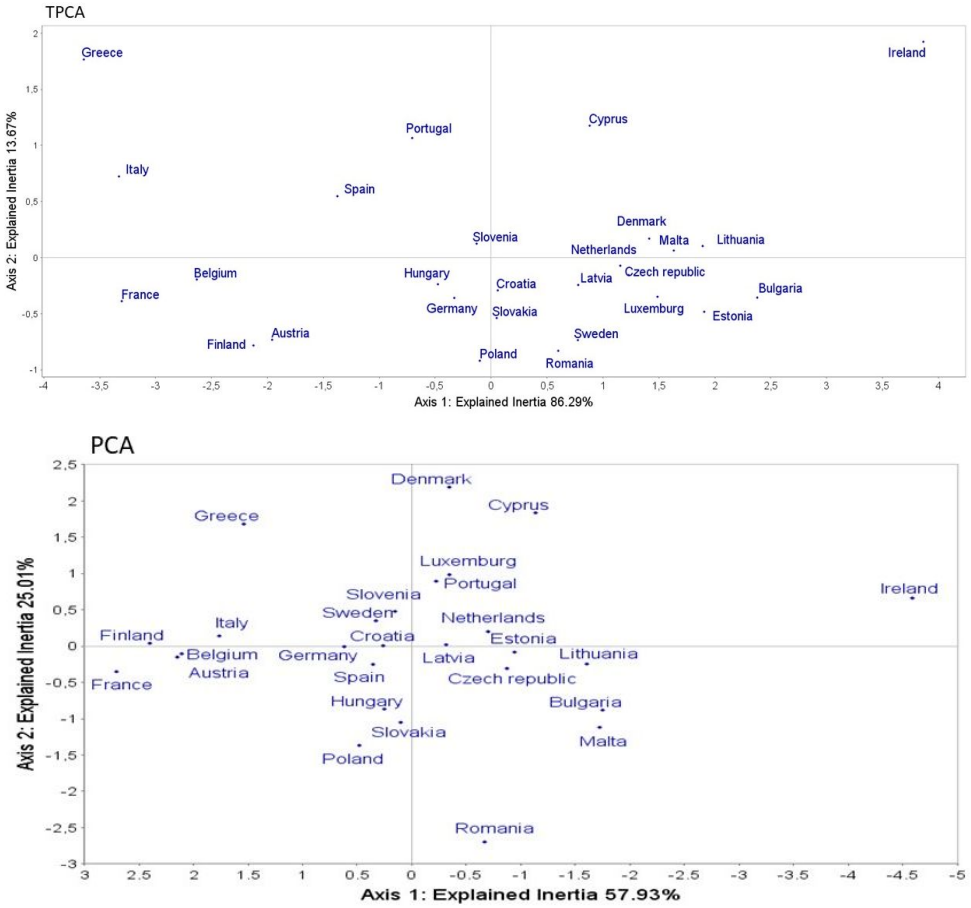


Figure 4. TPCA & PCA - The EU-27 countries on the first principal plane

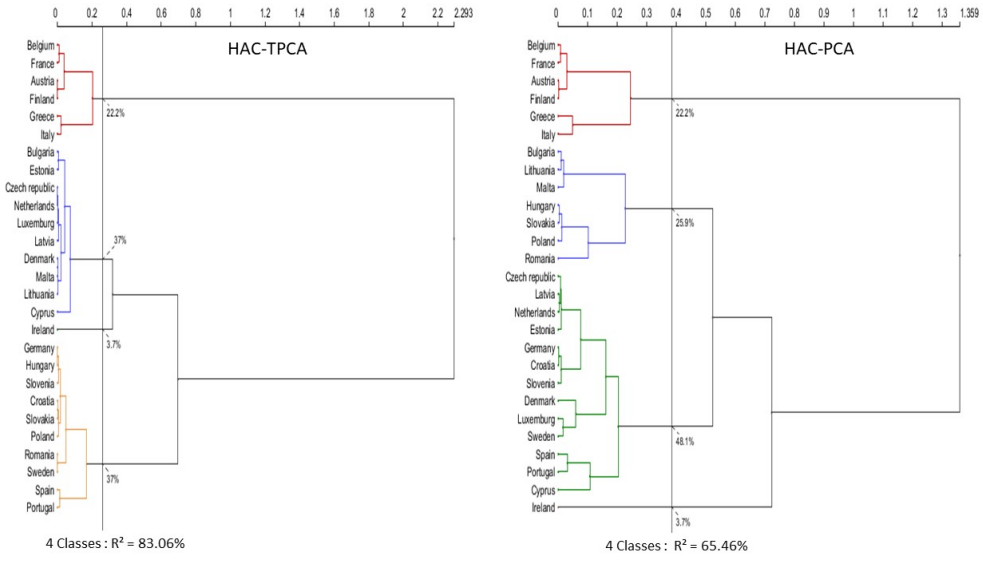


Figure 5. Topological and Classical Dendrograms

Table 5. Characterization of the four clusters of topological and classical clustering

HAC-TPCA				
Frequency (%)	6 (22.22%)	11 (40.74%)	1 (3.70%)	9 (33.33%)
Composition	Belgium,France, Greece, Italy, Austria, Finland	Bulgaria,Danmark, Malta, Luxembourg, Czech republic, Estonia Lithuania,Cyprus Latvia,Sweden Netherlands	Ireland	Germany, Spain, Croatia, Hungary, Poland, Portugal, Romania, Slovenia, Slovakia
Profile (+)	Debt Expenditure Revenue	Deficit	Deficit	
Anti-profile(-)	Deficit	Revenu Debt Expenditure	Expenditure	

HAC-PCA				
Frequency (%)	6 (22.22%)	8 (29.63%)	12 (44.44%)	1 (3.70%)
Composition	Belgium,France, Greece, Italy, Austria, Finland	Bulgaria,Slovakia, Poland, Hungary, Czech republic, Lithuania,Romania Malta	Germany, Spain, Croatia, Luxembourg, Sweden, Portugal, Cyprus, Slovenia, Danmark, Estonia, Latvia, Netherlands	Ireland
Profile (+)	Debt Expenditure Revenue		Deficit	Deficit
Anti-profile(-)		Deficit Revenu	Expenditure Revenu	

The second cluster, which groups eleven EU countries, is characterized by a high share of deficit and low shares of expenditures, debt, and revenue.

The third cluster, consisting solely of Ireland, is characterized by a high share of deficit and a low share of expenditures.

The fourth cluster, composed of nine countries, exhibits public finances that do not differ significantly from the average public finances of the 27 EU countries.

In a classical metric context, we can simply apply a standardized PCA to the homogeneous set of four government finance characteristics of the EU-27 in 2024.

The objective here is to provide a topological synthesis of the public finances of the EU countries in 2024. It was shown in [30], through a series of experiments, that the choice of proximity measure has a significant impact on the results of both supervised and unsupervised classification.

4.2. *TDA - Illustrative Example & Empirical results*

EU-27 public finance data are used to illustrate the proposed TDA method, which is then compared with several widely used machine learning approaches. The binary target variable for group discrimination is presented in Table 6. The goal is to identify the components of public finances that most effectively distinguish between the two groups of EU countries: those in the €urozone and those outside it.

Table 6. Statistics of the target variable

Area modality	Frequency	Percentage	Cummuled
Non-€uro	8	29.63	29.63
€uro	19	70.37	100.00
Total	27	100.00	100.00

The within-class reference adjacency matrices, Vu^{NEuro} and Vu^{Euro}, are presented in Table 7. The global reference adjacency matrix, $Vu^{\star}$, associated with the proximity measure $u^{\star}$ adapted to the data, is constructed from the correlation matrices of the X_{NEuro} and X_{Euro} datasets according to Property 3.

Table 7. Within-class correlation and Adjacency matrices

$$R = \left(\begin{array}{cccc|cccc} 1 & & & & & & & \\ 0.102 & 1 & & & & & & \\ 0.783 & 0.698 & 1 & & & & & \\ -0.365 & 0.354 & -0.184 & 1 & & & & \\ \hline 0 & 0 & 0 & 0 & 1 & & & \\ 0 & 0 & 0 & 0 & -0.715 & 1 & & \\ 0 & 0 & 0 & 0 & -0.439 & 0.942 & 1 & \\ 0 & 0 & 0 & 0 & -0.237 & 0.485 & 0.510 & 1 \end{array}\right)$$

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

According to Definition 2, we first perform a non-standardized Topological PCA to identify the correlation structure of all quantitative variables. A discriminant analysis is then applied to the significant principal components obtained from this Topological PCA of the projected data.

Figure 6 shows, on the first factorial plane of the TPCA, the correlations between the principal component factors and the original variables.

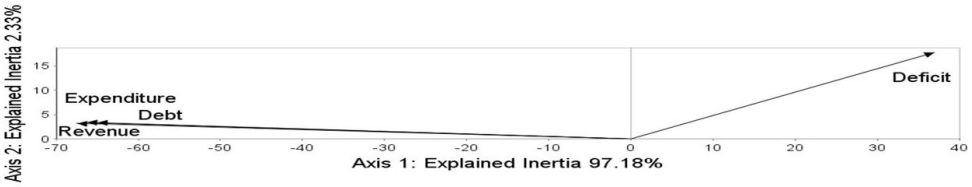


Figure 6. Representation of the explanatory variables

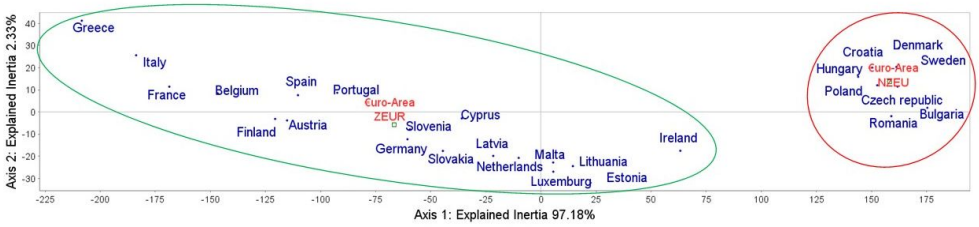


Figure 7. Active countries & Modalities of supplementary target variable

We observe that these correlations differ slightly, as do the percentages of inertia explained on the first principal planes of the Topological PCA. The first two factors explain 76.12% and 19.13% of the variance, respectively, accounting for a total of 95.25% of the variation in the dataset. Therefore, the first two factors provide an adequate summary of the mixed data, i.e., the banking characteristics of the agency's clients.

Figure 7 illustrates the representation of countries on the main factorial plane of the TPCA. The target qualitative variable is projected as a supplementary element. We can see that the first principal component of the TPCA perfectly discriminates between the two groups: the 8 Non-€uro-Area countries and the 19 €uro-Area countries.

TDA is then performed by applying Fisher discriminant analysis to the significant factors obtained from the TPCA of public finances.

Table 8. Confusion matrix

TDA Actual classification	Predicted classification Non-€uro Area	€uro Area	Total
Non-€uro Area	8	0	8
€uro Area	0	19	19
Total	8	19	27

% well classified: 100.00% Area Under Curve-ROC: 1.000

Table 9. TDA - Topological Discriminant Analysis

Fisher's linear function Variable	Coefficient function Discriminant	Standard Deviation	Ratio t-Student
Non-€uro Area			
Deficit	1.037	0.002	11.12**
Expenditure	0.029	0.001	1.46
Revenue	0.013	0.001	0.68
Debt	0.029	0.001	1.51
€uro Area			
Constant	-3.061	0.346	
D2 = 82.232	T2 = 462.933	PROBA = 0.001	

Significance level p-value $\leq \alpha$: $^{**}\alpha \leq 1\%$; $^{*}\alpha \in]1\%; 5\%]$

Table 8 presents the confusion matrix, which allows the performance of the classification model to be evaluated by comparing the predicted values with the actual values of the dataset. The percentage of correctly classified cases is 100%, indicating that the TDA model perfectly discriminates between and separates the two groups of countries.

Table 9 summarizes the significant profiles of the two groups of countries based on their public finances.

The coefficients of Fisher's discriminant function, which significantly distinguish between 'Non-€uro Area' and '€uro Area' countries, were ranked according to the p-values of the Student's t-test.

Among the four variables that comprise public finances, only the deficit significantly discriminates between and best separates the two groups of countries.

Table 10. Comparisons

Machine learning model	Abbreviation	Ranking	% Well classified	AUC - ROC
Topological Discriminant Analysis	TDA	1	100.00	1.000
Random Forest	RF	1	100.00	1.000
K-nearest neighbors	k-NN	3	81.48	0.796
Decision Tree	DT	4	77.78	
Logistic Regression	LR	5	77.77	0.796
Support Vector Machine	SVM	6	70.37	0.053
Neural Networks	NN	6	70.37	
Discriminant Analysis	DA	8	66.67	0.816

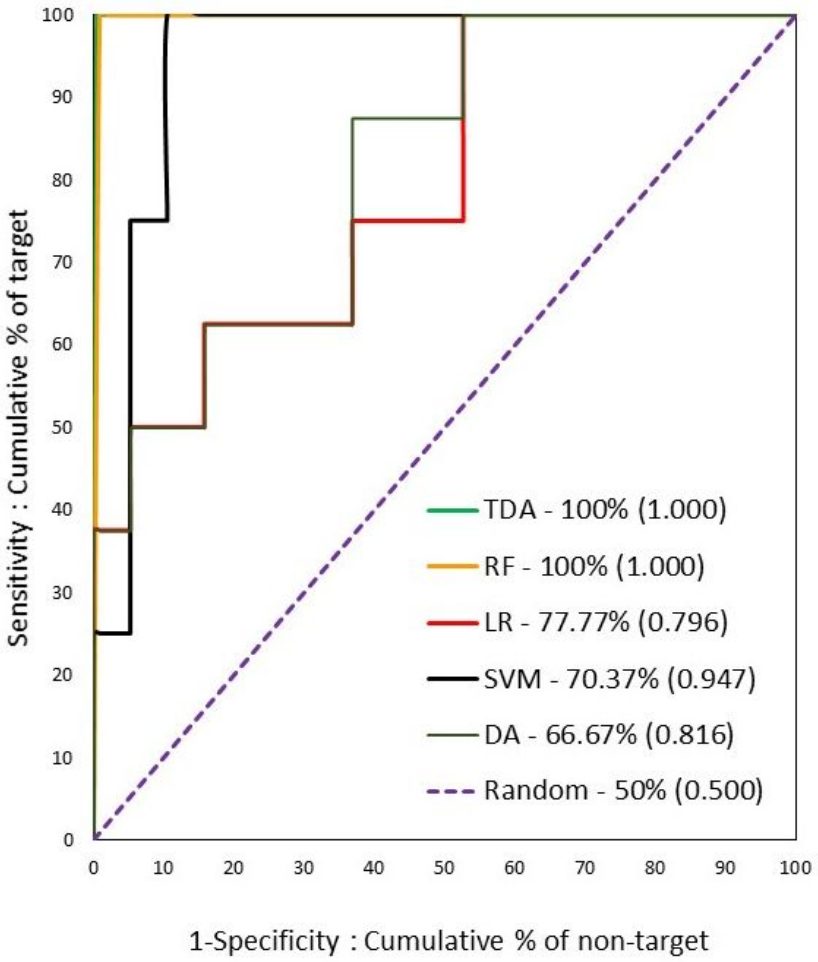


Figure 8. ROC Curves

Accordingly, the first group, consisting of the 8 Non-€uro Area countries—Romania, Bulgaria, Denmark, Sweden, Croatia, Hungary, the Czech Republic, and Poland—is characterized by a high deficit relative to the average deficit of the EU-27. In contrast, the second group, consisting of the 19 €uro Area countries, is characterized by a low deficit.

For comparison, Table 10 and Figure 8 summarize the performance of the TDA model and seven of the most popular supervised learning techniques applied to the public finance data. They present the ROC (Receiver Operating Characteristic) curves and AUC (Area Under the Curve) results for the proposed TDA model as well as for the main machine learning models considered. The topological predictive TDA model and the Random Forest (RF) model, followed by the K-Nearest Neighbors (K-NN) model, achieve excellent discrimination results.

5. Conclusion

This article proposes a novel Topological PCA (TPCA) and a new Topological Discriminant Analysis (TDA), which can complement conventional quantitative data analysis methods in the context of multidimensional descriptive dimension-reduction and predictive modeling.

The performance of the proposed topological models, TPCA and TDA, which are based on the concept of a neighborhood graph, is significantly better than that of classical PCA and existing machine learning models.

These two topological models can be implemented using PCA and DA procedures in SAS, SPAD, or R software. It would be interesting to extend this topological approach to other data analysis models.

6. Appendix

Table 11. Some proximity measures for continuous data

Measures	Formula : Distance - Dissimilarity
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	$u_{Cos}(x,y) = 1 - \frac{\sum_{j=1}^{p} x_j y_j}{\sqrt{\sum_{j=1}^{p} x_j^2}\sqrt{\sum_{j=1}^{p} y_j^2}} = 1 - \frac{<x,y>}{\|x\|\|y\|}$
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	$u_{Cor}(x,y) = 1 - \frac{(\sum_{j=1}^{p}(x_j - \overline{x})(y_j - \overline{y}))^2}{\sum_{j=1}^{p}(x_j - \overline{x})^2 \sum_{j=1}^{p}(y_j - \overline{y})^2} = 1 - \frac{(<x-\overline{x}, y-\overline{y}>)^2}{\|x-\overline{x}\|^2 \|y-\overline{y}\|^2}$
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	$u_{Siz}(x,y) = \mid \sum_{j=1}^{p}(x_j - y_j) \mid$
Lpower	$u_{Lpo_\gamma}(x,y) = \sum_{j=1}^{p} \lvert x_j - y_j \rvert^\gamma$

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 $\gamma > 0$.

References

[1] V. Batagelj, M. Bren, Comparing resemblance measures. *In Journal of classification*, 12, 73–90, 1995.

[2] D.W. Hosmer, and S. Lemeshow, Applied Logistic Regression, *Wiley*, 1989, 2d edition 2000.

[3] L. Lebart, A. Morineau, M. Piron, Statistique exploratoire multidimensionnelle, *3ème édition Dunod*, 2000.

[4] S. Tufféry, Data Mining et statistique décisionnelle – L'intelligence des données. *Editions Technip*, 2007.

[5] G. Govaert, Analyse des données. Hermes Science, Lavoisier, 2003.

[6] R. Abdesselam, 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.

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

[8] R. Abdesselam, Analyse en Composantes Principales Mixte. Classification : points de vue croisés, RNTI-C-2, *Revue des Nouvelles Technologies de l'Information* RNTI, Cépaduès Editions, 31-41, 2008.

[9] R. Abdesselam, Discriminant Analysis on Mixed Predictors. In Book Series "Studies in Classification, Data Analysis, and Knowledge Organization", *Data Analysis and Classification: from the exploratory to the confirmatory approach*, C. Lauro, F. Palumbo, M. Greenacre (eds.), Springer-Verlag Berlin Heidelberg, 113-120, 2010.

[10] F. Caillez and J.P. Pagès, Introduction à l'Analyse des données. *S.M.A.S.H., Paris*, 1976.

[11] J. Demsar, Statistical comparisons of classifiers over multiple data sets. *The journal of Machine Learning Research*, Vol. 7, 1–30, 2006.

[12] R.A. Fisher, The Interpretation of χ^2 from Contingency Tables, and the Calculation of P. *Journal of the Royal Statistical Society*, Published by Wiley, 85, 1, 87–94, 1922.

[13] B. Escofier and J. Pagès, Analyses factorielles simples et multiples : objectifs, méthodes et interprétation, Dunod, 1988.

[14] Eurostat, Data, Datasets, Euro-indicators April 22, 2025, https://ec.europa.eu/eurostat/fr/web/main/home

[15] H. Hotelling, Analysis of a Complex of Statistical Variables into Principal Components. *Journal of Educational Psychology*, Vol. 24, 6, 417–441, 1933.

[16] J.H. Kim, and S. Lee, Tail bound for the minimal spanning tree of a complete graph. *In Statistics & Probability Letters*, 4, 64, 425–430, 2003.

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

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

[19] N. Mantel, A technique of disease clustering and a generalized regression approach. *In Cancer Research*, 27, 209–220, 1967.

[20] J. Pagès, Analyse factorielle de données mixtes. *Revue de Statistique Appliquée* 52(4), 93–111, 2004.

[21] K. Pearson, On lines and Planes of Closest Fit to Systems of Points in Space. *In Philosophical Magazine*, vol. 2, 11, 559–572, 1901.

[22] M. Marjanović, M. Kovačević, B. Bajat, V. Voženílek, Landslide susceptibility assessment using SVM machine learning algorithm. *Engineering Geology, Elsevier,* Volume 123, Issue 3, 225–234, 2011.
[23] B. Mirkin, Eleven Ways to Look at the Chi-Squared Coefficient for Contingency Tables. *The American Statistician*, 55, (2), 111–120, 2001.
[24] J.C. Park, H. Shin, B.K. Choi, Elliptic Gabriel graph for finding neighbors in a point set and its application to normal vector estimation. *In Computer-Aided Design Elsevier*, 38, 6, 619–626, 2006.
[25] G. Saporta, Probabilités, analyse des données et Statistique. *IEditions TECHNIP*, 2011.
[26] M. Tenenhaus, Analyse en composantes principales d'un ensemble de variables nominales ou numériques. *Revue de statistique appliquée*, tome 25, no 2, 39–56, 1977.
[27] G.T. Toussaint, The relative neighbourhood graph of a finite planar set. *In Pattern recognition*, 12, 4, 261–268, 1980.
[28] T. Vafeiadisa, K.I. Diamantarasb, G. Sarigiannidisa, K.Ch. Chatzisavvasa, A comparison of machine learning techniques for customer churn prediction *Simulation Modelling Practice and Theory*, 2015.
[29] J.R. Ward, Hierarchical grouping to optimize an objective function. *In Journal of the American statistical association JSTOR*, 58, 301, 236–244, 1963.
[30] D. Zighed, R. Abdesselam, A. Hadgu, 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.
[31] J.W. Schneider, and P. Borlund, Matrix comparison, Part 1: Motivation and important issues for measuring the resemblance between proximity measures or ordination results. *In Journal of the American Society for Information Science and Technology*, 58, 11, 1586–1595, 2007.
[32] J.W. Schneider and P. Borlund, Matrix comparison, Part 2: Measuring the resemblance between proximity measures or ordination results by use of the Mantel and Procrustes statistics. *In Journal of the American Society for Information Science and Technology*, 11, 58, 1596–1609, 2007.
[33] J. Cohen, A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, Vol 20, 27–46, 1960.
[34] L. Breiman, Rondom Forests. Machine Learning, *Kluwer Academic Publishers. Manufactured in The Netherlands*. Vol. 45, 5-–32, 2001
[35] L. Rokach and 0. Maimon, Decision Trees. *In Maimon, O., Rokach, L. (eds) Data Mining and Knowledge Discovery Handbook*. Springer, Boston, 165–192, 2005.
[36] E. Grossi, M. Busc, Introduction to artificial neural networks, *European Journal of Gastroenterology Hepatology* 19(12), 1046–1054, 2008.
[37] R.E. Neapolitan, Learning Bayesian Networks. KDD'07 *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, Book 2003.
[38] G. Guo, H. Wang., D.A. Bell, Y. Bi, KNN Model-Based Approach in Classification. 1–12, 2004.