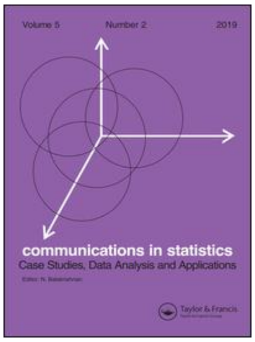


Communications in Statistics: Case Studies, Data Analysis and Applications

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ucas20

A topological approach of multiple correspondence analysis

Rafik Abdesselam

To cite this article: Rafik Abdesselam (2020): A topological approach of multiple correspondence analysis, Communications in Statistics: Case Studies, Data Analysis and Applications, DOI: 10.1080/23737484.2020.1830733

To link to this article: https://doi.org/10.1080/23737484.2020.1830733

Published online: 16 Oct 2020.

COMMUNICATIONS IN STATISTICS: CASE STUDIES, DATA ANALYSIS AND APPLICATIONS
https://doi.org/10.1080/23737484.2020.1830733

Taylor & Francis
Taylor & Francis Group

A topological approach of multiple correspondence analysis

Rafik Abdesselam

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

ABSTRACT
Topological multiple correspondence analysis (TMCA) studies a group of categorical variables defined on the same set of individuals. It is a topological method of data analysis that consists of exploring, analyzing, and representing the associations between several qualitative variables in the context of multiple correspondence analysis (MCA). It compares and classifies proximity measures to select the best one according to the data under consideration, then analyzes, interprets, and visualizes with graphic representations, the possible associations between several categorical variables relating to the known problem of MCA. Based on the notion of neighborhood graphs, some of these proximity measures are more-or-less equivalent. A topological equivalence index between two measures is defined and statistically tested according to the degree of description of the associations between the modalities of these qualitative variables.
We compare proximity measures and propose a topological criterion for choosing the best association measure, adapted to the data considered, from among some of the most widely used proximity measures for categorical data. The principle of the proposed approach is illustrated using a real dataset with conventional proximity measures for binary variables from the literature. The first step is to find the proximity measure that can best be adapted to the data; the second step is to use this measure to perform the TMCA.

ARTICLE HISTORY
Received 19 November 2019
Accepted 28 September 2020

KEYWORDS
Burt table; proximity measure; neighborhood graph; adjacency matrix; topological equivalence; associations; graphical representations

1. Introduction

Similarity measures play an important role in many areas of data analysis. The results of any operation involving structuring, clustering or classifying objects are strongly dependent on the proximity measure chosen. The user has to select one measure among many existing ones. Yet, according to the notion of topological equivalence chosen, some measures are more-or-less equivalent. The concept of topological equivalence uses the basic notion of local neighborhood. We define the topological equivalence between two proximity measures, in the context of association between several categorical variables, through the topological structure induced by each measure.

CONTACT Rafik Abdesselam rafik.abdesselam@univ-lyon2.fr Department of Economics and Management, ERIC – COACTIS Laboratories, Lumière Lyon 2, University of Lyon, 69365 Lyon, France.

Multiple correspondence analysis (MCA) is an important methodology among factorial techniques due to the extent of its field of application (Benzécri 1976; Lebart, Morineau, and Warwick 1984). It allows us, among others things, to describe large binary tables, such as socio-economic surveys, and usually answers questions on modalities.

This method is a generalization of correspondence analysis (CA); it concerns the relations between or within a set of p ($p > 2$) qualitative variables simultaneously observed in n individuals. Generally the variables are homogeneous in the sense that they revolve around a particular theme.

To understand and act on situations that are represented by a set of objects, very often we are required to compare them. Humans perform this comparison subconsciously, using the brain. In the context of artificial intelligence, however, we should be able to describe how the machine might perform this comparison. In this context, one of the basic elements that must be specified is the proximity measure between objects.

Certainly, the application context, prior knowledge, data type and many other factors can help in identifying the appropriate measure. For instance, if the objects to be compared are described by Boolean vectors, we can restrict our comparisons to a class of measures specifically devoted to this type of data. However, the number of candidate measures may still remain quite large. Can we consider that all those measures remaining are equivalent and just pick one of them at random? Or are there some that are equivalent and, if so, to what extent? This information might interest a user when seeking a specific measure. For instance, in information retrieval, choosing a given proximity measure is an important issue. We effectively know that the result of a query depends on the measure used. For this reason, users may wonder which one is more useful. Very often, users try many of them, randomly or sequentially, seeking a "suitable" measure. If we could provide a framework that allows the user to compare proximity measures to identify those that are similar, they would no longer need to try out all measures.

The present study proposes a new framework for comparing proximity measures to choose the best one in the context of association between a set of qualitative variables. The aim is to establish a TMCA.

We deliberately ignore the issue of the appropriateness of the proximity measure, as it is still an open and challenging question currently being studied. The comparison of proximity measures can be analyzed from various angles.

The comparison of objects, situations or ideas is an essential task to assess a situation, to rank preferences, to structure a set of tangible or abstract elements, and so on. In a word, to understand and act, we have to compare. These comparisons that the brain naturally performs, however, must be clarified if we want them to be done by a machine. For this purpose, we use proximity measures. A proximity measure is a function which measures the similarity or dissimilarity between two objects within a set. These proximity measures have mathematical properties and specific axioms. But are such measures equivalent? Can they be used in practice in an undifferentiated way? Do they produce the same learning database that will serve to find the membership class of a new object? If we know that the answer is negative, then how do we decide which one to use? Of course, the context of the study and the type of data being considered can help in selecting a few possible proximity measures, but which one should we choose from this selection as the best measure for summarizing the association?

We find this problematic also in the context of TMCA. The eventual links or associations between all the qualitative variables partly depends on the learning database being used. The results of MCA can change according to the selected proximity measure.

Several studies have been proposed on the topological equivalence of proximity measures (Batagelj and Bren 1992; Rifqi, Detyniecki, and Bouchon-Meunier 2003; Batagelj and Bren 1995; Lesot, Rifqi, and Benhadda 2009; Zighed, Abdesselam, and Hadgu 2012), and on the discrimination context (Abdesselam 2018b), but none of these propositions has the objective of identifying an association between several categorical variables. An approach in the case of association between two qualitative variables has been proposed in Abdesselam (2018a).

Therefore, this article focuses on how to construct the best adjacency matrix induced by a proximity measure, taking into account the association between all the modalities of the qualitative variables.

This article is organized as follows. In Sec. 2, after recalling the basic notions of structure, graph and topological equivalence, we present the proposed method, how to build an adjacency matrix associated with a proximity measure in the context of association between several qualitative variables, how to compare and statistically test the degree of topological equivalence between proximity measures and how to select the best measure to describe multiple associations. Sec. 3 presents an illustrative example using real data. The conclusion of this work is given in Sec. 4.

Table A1, shown in Appendix A, summarizes some classic proximity measures used for binary data (Warrens 2008); we give on $\{0, 1\}^n$ the definition of 22 of them.

We assume that we have at our disposal $\{x^k; k = 1, \ldots, p\}$, a set of $p > 2$ qualitative variables and partitions of $n = \sum_{k=1}^{p} n_k$ individuals-objects into m_k modalities-subgroups. The interest lies in whether there is a topological association between all these variables. Let us denote:

- $X_k = X_{(n,m_k)}$ the disjunctive table, data matrix associated to the m_k dummy variables of the qualitative variable x^k with n rows-objects and m_k columns-modalities. We check that $\Sigma_{k=1}^{m_k} x_i^k = 1,\ \forall_i$ and $\Sigma_{i=1}^{n} x_i^k = n_k$.

- $X_{(n,m)} = [\,X_1|X_2|\cdots|X_p\,]$ the indicator matrix, juxtaposition of the p binary tables X_k, with n rows-objects and $m = \sum_{k=1}^{p} m_k$ columns-modalities. We check that $\Sigma_{k=1}^{m_k} x_i^k = p,\ \forall_i$ and $\Sigma_{i=1}^{n}\Sigma_{k=1}^{m_k} x_i^k = np$.

An alternative coding of such data is as a Burt matrix, a square symmetric modality-by-modality matrix formed from all two-way contingency tables of pairs of variables, including on the block diagonal the cross-tabulations of each variable with itself.

- $\mathcal{B}_{(m,m)} = {}^tX\ X$ is the symmetric Burt matrix of the two-way cross-tabulations of the p variables.
- $W_{(m,m)} = \text{diag}[\mathcal{B}]$, is the diagonal marginal frequency matrices.
- $U = \mathbb{1}_m\,{}^t\mathbb{1}_m$ is the $m \times m$ matrix of 1s.
- I_m, is the $m \times m$ identity matrix where $\mathbb{1}_m$ denotes the m indicator vector of 1s and $\mathbb{1}_n$ the n indicator vector of 1s.

The dissimilarity matrices associated with proximity measures are computed from data given by the Burt table $\mathcal{B}$.

The attributes of the modalities of any two points x^k and x^l in $\{0,1\}^n$ of the proximity measures can be easily written and calculated from the following matrices. Computational complexity is thus considerably reduced.

- $A_{(m,m)} = \mathcal{B}$, the Burt matrix whose element, $a_{kl} = |x^k \cap x^l| = \sum_{i=1}^{n} x_i^k x_i^l$ is the number of attributes common to both points x^k and x^l,
- $$\begin{aligned} B_{(m,m)} &= {}^tX\,(\mathbb{1}_n\,{}^t\mathbb{1}_m - X) = {}^tX\,\mathbb{1}_n\,{}^t\mathbb{1}_m - {}^tX\,X \\ &= W\,\mathbb{1}_m\,{}^t\mathbb{1}_m - A = W\,U - A \end{aligned}$$
  whose element, $b_{kl} = |X^k - X^l| = |X^k \cap \overline{X^l}| = \sum_{i=1}^{n} x_i^k(1 - x_i^l)$ is the number of attributes present in x^k but not in x^l,
- $$\begin{aligned} C_{(m,m)} &= {}^t(\mathbb{1}_n\,{}^t\mathbb{1}_m - X)\,X = {}^t(\mathbb{1}_n\,{}^t\mathbb{1}_m)\,X - {}^tX\,X \\ &= \mathbb{1}_m\,{}^t\mathbb{1}_n\,X - {}^tX\,X = UW - A \end{aligned}$$
  whose element, $c_{kl} = |X^l - X^k| = |X^l \cap \overline{X^k}| = \sum_{i=1}^{n} x_i^l(1 - x_i^k)$ is the number of attributes present in x^l but not in x^k.
- $$\begin{aligned} D_{(m,m)} &= {}^t(\mathbb{1}_n\,{}^t\mathbb{1}_m - X)\,(\mathbb{1}_n\,{}^t\mathbb{1}_m - X) \\ &= \mathbb{1}_m\,{}^t\mathbb{1}_n\,\mathbb{1}_n\,{}^t\mathbb{1}_m - \mathbb{1}_m\,{}^t\mathbb{1}_n\,X - {}^tX\,\mathbb{1}_n\,{}^t\mathbb{1}_m + {}^tX\,X \\ &= n\mathbb{1}_m\,{}^t\mathbb{1}_m - UW - WU + A = nU - UW - WU + A \\ &= nU - (A + B + C) \end{aligned}$$
  whose element, $d_{kl} = |\overline{X^k} \cap \overline{X^l}| = \sum_{i=1}^{n}(1 - x_i^k)(1 - x_i^l)$ is the number of attributes in neither x^k or x^l.

$X^k = \{i/x_i^k = 1\}$ and $X^l = \{i/x_i^l = 1\}$ are the sets of attributes present in data point-modality x^k and x^l respectively, and $|.|$ the cardinality of a set.

The attributes are linked by the relation:

$$\forall k = 1, p\ ;\ \forall l = 1, p \quad a_{kl} + b_{kl} + c_{kl} + d_{kl} = n.$$

Together, the four dependent quantities a_{kl}, b_{kl}, c_{kl}, and d_{kl} can be used to construct the 2×2 contingency table, where the information can be summarized by an index of similarity (affinity, resemblance, association, coexistence). As a general symbol for a similarity coefficient the capital letter S will be used. A list of 22 similarity coefficients is given in Table A1 in Appendix A.

2. Topological correspondence

Topological equivalence is based on the concept of the topological graph, also referred to as a neighborhood graph. The basic idea is actually quite simple: two proximity measures are equivalent if the corresponding topological graphs induced on the set of objects remain identical. Measuring the similarity between proximity measures involves comparing the neighborhood graphs and measuring their similarity. We will first define more precisely what a topological graph is and how to build it. Then, we propose a measure of proximity between topological graphs that will subsequently be used to compare the proximity measures.

Consider a set $E = \{x^{11}, \ldots, x^{1m_1}, \ldots, x^{p1}, \ldots, x^{pm_p}\}$ of $m = \sum_{j=1}^{p} m_j$ modalities in $\{0,1\}^n$, associated with the p qualitative variables x^j with m_j modalities. We can, by means of a proximity measure u, define a neighborhood relationship V_u to be a binary relationship on $E \times E$. There are many possibilities for building this neighborhood binary relationship.

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

Many definitions are possible to build this binary neighborhood relationship. One can choose the minimal spanning tree (MST) (Kim and Lee 2003), the Gabriel graph (GG) (Park, Shin, and Choi 2006) or, as is the case here, the relative neighborhood graph (RNG) (Toussaint 1980).

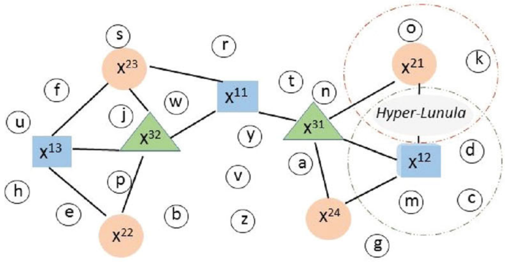
For any given proximity measure u, we construct the associated adjacency binary symmetric matrix V_u of order $m = \sum_{j=1}^{p} m_j$, where all pairs of neighboring modalities (x^{kr}, x^{ls}) and where $k, l = 1, p$; $r = 1, m_k$ and $s = 1, m_l$ satisfy the following RNG definition.

Definition 1. Relative neighborhood graph (RNG)

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

This means that if two modalities, x^{kr} and x^{ls} which verify the RNG property are connected by an edge, the vertices x^{kr} and x^{ls} are neighbors.

Thus, for any given proximity measure u, we can associate an adjacency matrix V_u, of binary and symmetrical order m. Figure 1 illustrates an example of RNG in $\mathbb{R}^2$ of a set of n objects-individuals around nine modalities associated



V_u	x^{11}	x^{12}	x^{13}	x^{21}	x^{22}	x^{23}	x^{24}	x^{31}	x^{32}
x^{11}	1	0	0	0	0	1	0	1	1
x^{12}	0	1	0	1	0	0	1	1	0
x^{13}	0	0	1	0	1	1	0	0	1
x^{21}	0	1	0	1	0	0	0	1	0
x^{22}	0	0	1	0	1	0	0	0	1
x^{23}	1	0	1	0	0	1	0	0	1
x^{24}	0	1	0	0	0	0	1	1	0
x^{31}	1	1	0	1	0	0	1	1	0
x^{32}	1	0	1	0	1	1	0	0	1

Figure 1. RNG example with nine groups-modalities and associated adjacency matrix.

with three qualitative variables x^1, x^2, and x^3 with three, four, and two modalities, respectively.

For example, for the second modality of the first variable and the first modality of the second variable, $V_u(x^{12}, x^{21}) = 1$, it means that on the geometrical plane, the hyper-Lunula (intersection between the two hyperspheres centered on the two modalities x^{12} and x^{21}) is empty.

For a given neighborhood property (MST, GG, or RNG), each measure u generates a topological structure on the objects in E which are totally described by the adjacency binary matrix V_u. In this article, we chose to use the relative neighbors graph (GNR).

2.1. Comparison and selection of proximity measures

First we compare different proximity measures according to their topological similarity to regroup them and to better visualize their resemblances.

To measure the topological equivalence between two proximity measures, u_i and u_j, we propose to test if the associated adjacency matrices, V_{u_i} and V_{u_j}, are different or not. The degree of topological equivalence between two proximity measures is measured by the following definition of concordance.

Definition 2. Topological equivalence index between two adjacency matrices

$$S(V_{u_i}, V_{u_j}) = \frac{1}{m^2} \sum_{k=1}^{p} \sum_{r=1}^{m_k} \sum_{l=1}^{p} \sum_{s=1}^{m_l} \delta_{kr\,ls}(x^{kr}, x^{ls})$$

with $\delta_{kr\,ls}(x^{kr}, x^{ls}) = \begin{cases} 1 \text{ if } V_{u_i}(x^{kr}, x^{ls}) = V_{u_j}(x^{kr}, x^{ls}) \\ 0 \text{ otherwise.} \end{cases}$

Then, in our case, we want to compare these different proximity measures according to their topological equivalence in a context of association. So we define a criterion for measuring the spacing from the independence or no association position.

A contingency table is one of the most common ways to summarize categorical data. Generally, interest lies in whether there is an association between the row variable and the column variable that produce the table; sometimes there is further interest in describing the strength of that association. The data can arise from several different sampling frameworks, and the interpretation of the hypothesis of no association depends on the framework. The question of interest is whether there is an association between the two variables.

We construct the adjacency matrix denoted by V_{u_*}, which corresponds best to the Burt table. Thus, to examine similarities between the modalities we examine the gap between each profile-modality and its average profile, that is, the gap to independence. This best adjacency matrix can be written as follows:

Definition 3. Reference adjacency matrix

$$\begin{cases} V_{u_*}(x^{kr}, x^{ls}) = 1 & \text{if } \frac{\mathcal{B}_{kr\,ls}}{\mathcal{B}_{kr\,..}} \geq \frac{\mathcal{B}_{kr\,..}}{np^2};\ \ \forall k, l = 1, p\ ;\ r = 1, m_k \text{ and } s = 1, m_l \\ V_{u_*}(x^{kr}, x^{ls}) = 0 & \text{otherwise.} \end{cases}$$

$\mathcal{B}_{kr\,ls} = \Sigma_{i=1}^{n} x_i^{kr} x_i^{ls}$ is the element of the Burt matrix that corresponds to the number of individuals who have the modality r of the variable k and the modality s of the variable l,

$\mathcal{B}_{kr\,..} = \Sigma_{l=1}^{p} \Sigma_{s=1}^{m_s} b_{kr\,ls}$ is the row margin of the modality r of the variable k,

$\frac{\mathcal{B}_{kr\,ls}}{\mathcal{B}_{kr\,..}}$ is the row profile of the modality r of the variable k,

$\frac{\mathcal{B}_{kr\,..}}{np^2}$ is the average profile of the modality r of the variable k, np^2 being the total number.

The binary and symmetric adjacency matrix V_{u_*} is associated with an unknown proximity measure denoted u_* and called a reference measure.

The robustness of this positive deviation from independence can be studied by setting a minimum threshold to analyze the sensitivity of the results. Certainly the numerical results will change, but probably not their interpretation.

Thus, with this reference proximity measure we can establish $S(V_{u_i}, V_{u_*})$, the topological equivalence of association between the modalities of the p variables, by measuring the percentage of similarity between the adjacency matrix V_{u_i} and the reference adjacency matrix V_{u_*}.

To graphically describe the similarities between proximity measures, we can, for example, apply the notion of themascope (Lebart 1989), which is a methodological sequence of a clustering method on the results of a factorial method. In this case of this article, a principal component analysis (PCA) followed by a hierarchical ascendant classification (HAC) were performed upon the 22 component dissimilarity matrix defined by $[D]_{ij} = D(V_{u_i}, V_{u_j}) = 1 - S(V_{u_i}, V_{u_j})$ to partition them into homogeneous groups and to view their similarities to see which measures are close to one another.

We can use any classic visualization techniques to achieve this. For example, we can build a dendrogram of hierarchical clustering of the proximity measures. We can also use multidimensional scaling or any other technique, such as a

Laplacian projection, to map the 22 proximity measures into a two-dimensional space.

Finally, to evaluate and determine the closest class of proximity measures to the reference measure u_*, we project the latter as a supplementary element into the two data analysis methods, positioned by the dissimilarity vector with 22 components $[D]_{*i} = 1 - S(V_{u_*}, V_{u_i})$.

2.2. Statistical comparisons between two proximity measures

In this section, we use Cohen's kappa coefficient (Cohen 1960) to test statistically the degree of topological equivalence between two proximity measures. This nonparametric test compares these measures based on their associated adjacency matrices.

The comparison between indices of proximity measures has also been studied by Schneider and Borlund (2007a, 2007b) and Demsar (2006) from a statistical perspective. The authors proposed an approach that compares similarity matrices obtained by each proximity measure, using Mantel's test (Mantel 1967), in a pairwise manner.

Cohen's nonparametric Kappa test is the statistical test best suited to compare matched binary data, while another good option is the Fisher's exact test (Fisher 1922), which is an alternative to the chi-square test when the size m of the sample is small. The Kendall or Spearman coefficient compares matched continuous data. It makes it possible in this context to measure the agreement or the concordance of the binary values of two adjacency matrices associated with two proximity measures.

Let V_{u_i} and V_{u_j} be adjacency matrices associated with two proximity measures, u_i and u_j. To compare the degree of topological equivalence between these two measures, we propose to test if the associated adjacency matrices are statistically different or not, using a nonparametric test of paired data.

These binary and symmetric matrices of order m are unfolded in two vector-matched components, consisting of $\frac{m(m+1)}{2}$ values: the m diagonal values and the $\frac{m(m-1)}{2}$ values are above or below the diagonal.

The degree of topological equivalence between two proximity measures is estimated from the kappa coefficient, computed on the 2×2 contingency table formed by the two binary vectors, using the following definition:

Definition 4. Kappa coefficient

$$\widehat{\kappa} = \widehat{\kappa}(V_{u_i}, V_{u_j}) = \frac{P_o - P_e}{1 - P_e},$$

where,

$P_o = \frac{2}{m(m+1)} \sum_{k=0}^{1} n_{kk}$ is the observed proportion of concordance, and

$P_e = \frac{4}{m^2(m+1)^2} \sum_{k=0}^{1} n_{k.} n_{.k}$ represents the expected proportion of concordance under the assumption of independence.

The kappa coefficient is a real number, without dimension, between -1 and $+1$. The concordance is higher the closer the value of Kappa is to 1 and the maximum concordance is reached ($\widehat{\kappa} = 1$) when $P_o = 1$ and $P_e = 0.5$. When there is perfect independence, $\widehat{\kappa} = 0$ with $P_o = P_e$, and in the case of total mismatch, $\widehat{\kappa} = -1$ with $P_o = 0$ and $P_e = 0.5$.

The true value of the kappa coefficient in the population is a random variable that approximately follows a Gaussian law of mean $E(\kappa)$ and variance $Var(\kappa)$. The null hypothesis H_0 is $\kappa = 0$ against the alternative hypothesis $H_1 : \kappa > 0$. We formulate the null hypothesis $H_0 : \kappa = 0$, independence of agreement or concordance. The concordance becomes higher as κ tends towards 1, and is a perfect maximum if $\kappa = 1$. It is equal to -1 in the case of a perfect discordance.

We also test the topological equivalence between each proximity measure u_i and the reference measure u_* by comparing the adjacency matrices V_{u_i} and V_{u_*}.

2.3. Graphical representation of the topological associations

To represent graphically the possible topological links between the m modalities of the p qualitative variables, we use multidimensional scaling (MDS). This allows us to visualize a proximity matrix (similarity or dissimilarity) and makes it possible to pass from a proximity matrix between a set of n objects to the coordinates of these same objects in a p-dimensional space. We propose to carry out the classical MDS, namely factorial analysis on the similarity V_{u*} or dissimilarity $D_{u*} = U - V_{u*}$ table Cailliez and Pagès (1976). The topological multiple correspondence analysis (TMCA) returns to perform the following PCA:

Definition 5. TMCA consists to perform the PCA of the triple $\{V_{u*} \,;\, M \,;\, D_m\}$, where, V_{u*} is the adjacency matrix associated with the proximity measure $u*$, the most appropriate measure for the considered data, $M = I_m$ is the identity matrix of order m and $D_m = \frac{W}{np}$ is the weighted diagonal matrix of the m modality weights.

One can also opt for a normalized PCA if one wishes to give the same weight to all the variables in the calculation of the distance between two modalities.

This topological analysis leads to the spectral decomposition of the M-symmetric and positive matrix ${}^tV_{u*}\, D_m\, V_{u*}\, M$, whose explained inertia is equal to $\frac{1}{np} trace({}^tV_{u*}\, W\, V_{u*})$, with the first $m - p - 1$ nonzero eigenvalues.

We can thus establish the topological correspondence analysis of each of the 22 proximity measures u_i considered, by carrying out a PCA of the V_{u_i} adjacency data table.

The PCA aids in the interpretation of TMCA results. Graphical representations of factorial plans allow the visualization and identification of the topological links between the modalities of the variables. As in weighted PCA, we consider the most significant modalities on the axes, that is, the modalities which

Table 1. Burt table—Female entrepreneurship in Dakar, Senegal.

	Variables												
Modalities	Age			Marital status				Number of children			Level of study		
Under 25	22	0	0	18	2	1	1	13	3	6	3	1	18
25–50 years	0	80	0	16	9	21	34	14	11	55	58	5	17
Over 50	0	0	51	3	8	24	16	8	35	8	30	10	11
Single	18	16	3	37	0	0	0	20	3	14	9	1	27
Divorcee	2	9	8	0	19	0	0	3	10	6	13	5	1
Monogamous bride	1	21	24	0	0	46	0	7	21	18	26	5	15
Polygamous bride	1	34	16	0	0	0	51	5	15	31	43	5	3
No children	13	14	8	20	3	7	5	35	0	0	11	5	19
From 1 to 3 children	3	11	35	3	10	21	15	0	49	0	27	9	13
More than 3 children	6	55	8	14	6	18	31	0	0	69	53	2	14
Illiterate-Primary	3	58	30	9	13	26	43	11	27	53	91	0	0
Secondary	1	5	10	1	5	5	5	5	9	2	0	16	0
Higher	18	17	11	27	1	15	3	19	13	14	0	0	46

have both a strong relative contribution and a good quality of representation, measured by the square cosine of the angle formed by the point-modality and its projection on the factorial plane being considered.

3. Application to real data

To illustrate the TMCA, we considered the data displayed in Table 1 of a study on female entrepreneurship conducted in Dakar, Senegal in 2014. These data were collected from 153 female entrepreneurs of the Dakar region, and their objective here is to give a topological description of the sample's demographic features: age, marital status, number of children and level of study.

In a metric and classical context, we simply have to apply an MCA on the homogeneous set of the four characteristics of the female entrepreneurs. The main numerical and graphical results of this MCA, given in Table A3 in Appendix A and in Fig. 4, will be compared to those of the proposed TMCA.

In a topological context, the main results of the proposed method are presented in the following tables and graphs, which allow us to visualize proximity measures close to each other and to select the one that best describes the associations between the modalities of the four characteristics of the sample population.

An HAC algorithm based on the Ward (1963) criterion[1] was used to characterize classes of proximity measure relative to their similarities.

[1] Aggregation based on the criterion of the loss of minimal inertia. Ward's method is a criterion applied in hierarchical cluster analysis; it is a general agglomerative hierarchical clustering procedure. With the square of the Euclidean distance, this criterion allows one to minimize the total within-cluster variance or, equivalently, maximize the between-cluster variance.

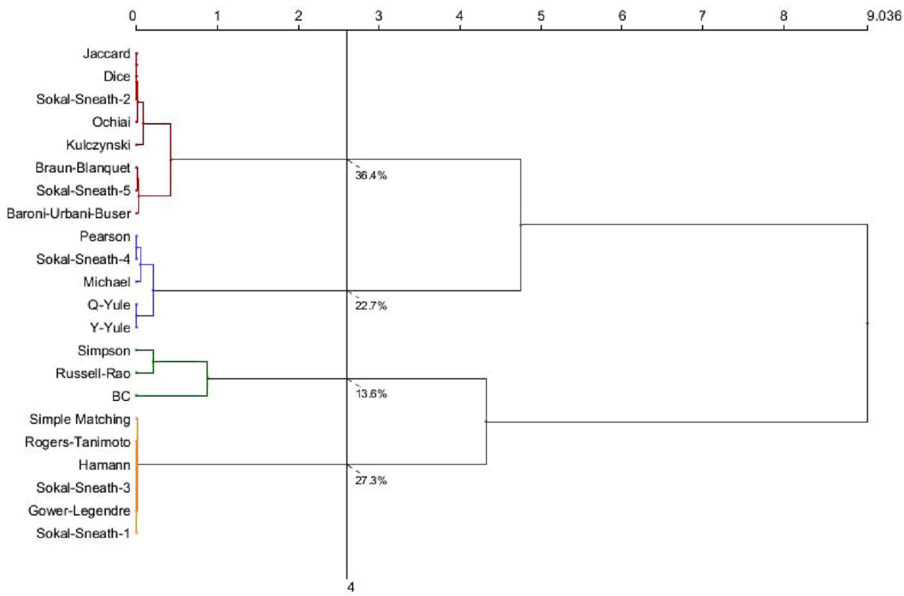


Figure 2. Hierarchical tree of the proximity measures.

Table 2. PCA & HAC results—Assignment of the reference measure.

Class number Frequency	Class 1 8	Class 2 5	Class 3 3	Class 4 6
Proximity measure	$u_{Jaccard}$ u_{Dice} $u_{Sokal\text{-}Sneath\text{-}2}$ u_{Ochiai} $u_{Kulczynski}$ $u_{Baroni\text{-}Urbani\text{-}Buser}$ $u_{Sokal\text{-}Sneath\text{-}5}$ $u_{Braun\text{-}Blanquet}$	$u_{Pearson}$ $u_{Sokal\text{-}Sneath\text{-}4}$ $u_{Q\text{-}Yule}$ $u_{Y\text{-}Yule}$ $u_{Michael}$	$u_{Russell\text{-}Rao}$ $u_{Simpson}$ u_{BC}	$u_{Simple\text{-}Matching}$ $u_{Rogers\text{-}Tanimoto}$ u_{Hamann} $u_{Sokal\text{-}Sneath\text{-}3}$ $u_{Gower\text{-}Legendre}$ $u_{Sokal\text{-}Sneath\text{-}1}$
Reference measure			u_*	

The reference measure u_* is projected as a supplementary element. The dendrogram in Fig. 2 represents the hierarchical tree of the 22 proximity measures considered.

Table 2 summarizes the main results of the chosen partition into four homogeneous classes of proximity measure, obtained from the cut of the hierarchical tree in Fig. 2.

Moreover, in view of the results in Table 2, the reference measure u_* is closer to the third class consisting of Russell-Rao, Simpson, and BC measures for which there is a strong topological association between the modalities of the variables among the 22 proximity measures considered. We will have a stronger association between the variables of the typical profile of the entrepreneur in Dakar, Senegal.

Table 3. Measures with perfect topological equivalence.

Group 1	Group 2	Group 3
$u_{Jaccard}$	$u_{Pearson}$	$u_{Simple\text{-}Matching}$
u_{Dice}	$u_{Sokal\text{-}Sneath\text{-}4}$	$u_{Rogers\text{-}Tanimoto}$
$u_{Sokal\text{-}Sneath\text{-}2}$		u_{Hamann}
		$u_{Sokal\text{-}Sneath\text{-}3}$
		$u_{Gower\text{-}Legendre}$
		$u_{Sokal\text{-}Sneath\text{-}1}$

It was shown in Zighed, Abdesselam, and Hadgu (2012), by means of a series of experiments, that the choice of proximity measure has an impact on the results of a supervised or unsupervised classification.

For any pair of proximity measures given in Table A1 in Appendix A, we built and applied the Kappa test to compare the two adjacency matricies and to measure and test their topological equivalence. Thus, for example, for the pair $(u_* \ ; \ u_{RR})$, reference and Russell-Rao proximity measures, the calculated Kappa value $\widehat{\kappa} = 0.5939$ corresponds to a p-value less than 0.01%. Since this probability is lower than a prespecified significance level of 5%, the null hypothesis that $\kappa = 0$ for these data (no agreement) is rejected. We can therefore conclude that the topological equivalence between the two proximity measures, measured by $S(V_{u_*} \ ; \ V_{RR}) = 79.88\%$, is significant.

Table A2, given in Appendix A, summarizes the similarities and Kappa statistic values between all pairs of proximity measures formed with the 22 measures considered and the unknown reference measure u_*, in a topological framework. The values below the diagonal correspond to the similarities $S(V_{u_i}, V_{u_j})$ and the values above the diagonal are the kappa coefficients $\widehat{\kappa}(V_{u_i}, V_{u_j})$. All Kappa statistical tests are significant with $\alpha \leq 5\%$ level of significance.

The similarities in pairs between the 22 proximity measures differ somewhat: some are closer than others. Some measures are in perfect topological equivalence $S(V_{u_i}, V_{u_j}) = 1$ with a perfect concordance $\widehat{\kappa}(V_{u_i}, V_{u_j}) = 1$; these are therefore identical for the data considered, as is the case with the measure groups presented in Table 3.

The adjacency matrix V_{u*} associated with the proximity measure best adapted to the considered data, $u*$, is established from the profile of Table 4.

Figure 3 shows on the main first TMCA plan, the significant links between the modalities of the signage of female entrepreneurship. The links are materialized by geometric shapes.

Figure 4 presents, for comparison, on the first factorial plan, a graphical representation of the classical multiple correspondence analyse (MCA) (Escofier 1979; Greenacre and Jörg 2006).

Unlike the MCA method, which describes only three strong links, the TMCA highlights four: two opposing on the first factorial axis (56.35%) and the other two on the second factorial axis (32.01%).

Table 4. Row and average profiles.

Row-profiles	Age			Marital status				Number of child			Level of study		
Under 25 years	0.25	0	0	0.205	0.023	0.011	0.011	0.148	0.034	0.068	0.034	0.011	0.205
25 to 50 years	0	0.25	0	0.050	0.028	0.066	0.106	0.044	0.034	0.172	0.181	0.016	0.053
Over 50 years	0	0	0.25	0.015	0.039	0.118	0.078	0.039	0.172	0.039	0.147	0.049	0.054
Single	0.122	0.108	0.020	0.25	0	0	0	0.135	0.020	0.095	0.061	0.007	0.182
Divorcee	0.026	0.118	0.105	0	0.25	0	0	0.040	0.132	0.079	0.171	0.066	0.013
Monogamous	0.005	0.114	0.130	0	0	0.25	0	0.038	0.114	0.098	0.141	0.027	0.082
Polygamous	0.005	0.167	0.078	0	0	0	0.25	0.025	0.074	0.152	0.211	0.025	0.015
No children	0.093	0.100	0.057	0.143	0.021	0.050	0.036	0.25	0	0	0.079	0.036	0.136
From 1 to 3 child	0.015	0.056	0.179	0.015	0.051	0.107	0.077	0	0.25	0	0.138	0.046	0.066
More than 3 child	0.022	0.199	0.029	0.051	0.022	0.065	0.112	0	0	0.25	0.192	0.007	0.051
Illiterate-Primary	0.008	0.159	0.082	0.025	0.036	0.071	0.118	0.030	0.074	0.146	0.25	0	0
Secondary	0.016	0.078	0.156	0.016	0.078	0.078	0.078	0.078	0.141	0.031	0	0.25	0
Higher	0.098	0.092	0.060	0.147	0.005	0.082	0.016	0.103	0.071	0.076	0	0	0.25
Average profile	0.036	0.131	0.083	0.061	0.031	0.075	0.083	0.057	0.080	0.113	0.149	0.026	0.075

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

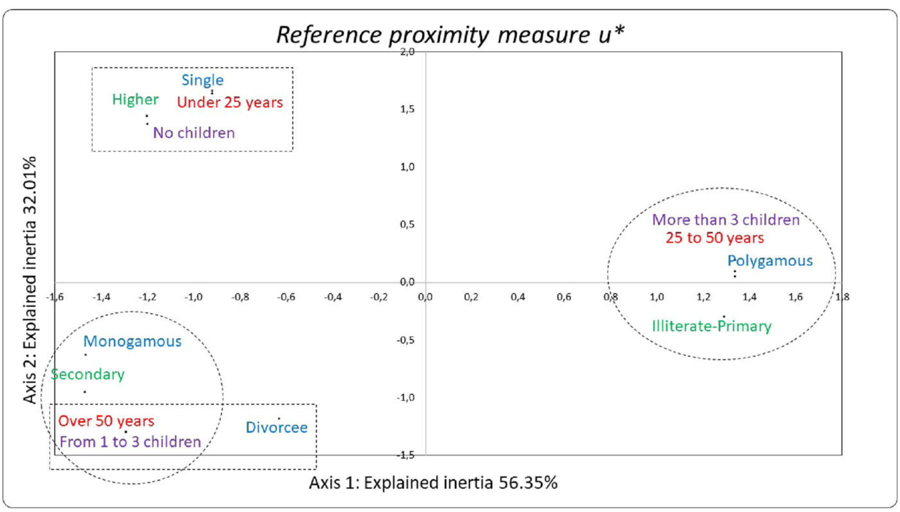


Figure 3. TMCA—Adjacency matrix and graphical representation.

Considering percentages of inertia presented in Table A3 in Appendix A, which represent the associations between all modalities, we restrict the comparison of the graphical representations to the two first factorial axes.

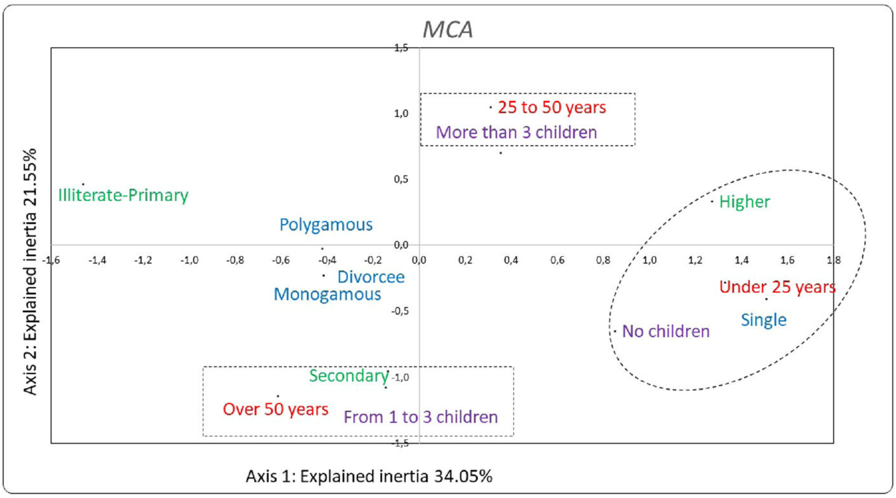


Figure 4. MCA—Graphical representation.

One can also represent the graphical representation associated with a perfect topological independence built from the adjacency identity matrix.

We give, in Fig. A1 in Appendix A, the different TMCA graphical representations associated with 4 of the 22 proximity measures considered.

4. Conclusion

This work proposes a new topological method of MCA, TMCA, that enriches the classical methods of qualitative data analysis. This work compares existing proximity measures to perform a TMCA based on the notion of neighborhood graphs according to the considered data. Future work involves extending this topological approach to other factorial methods of data analysis, especially to analyze the correlation structure of a set of continuous variables, topological principal component analysis, or to synthesize the relations existing between two groups of continuous variables, topological canonical analysis.

References

Abdesselam, R. 2018a. "Selection of Proximity Measures for a Topological Correspondence Analysis." In *5th Stochastic Modeling Techniques and Data Analysis*, edited by C. H. Skiadas, 11–24, International Conference, Chania, Greece.

Abdesselam, R. 2018b. "A Topological Discriminant Analysis." In *Data Analysis and Applications 2: Utilization of Results in Europe and Other Topics*, edited by J. Bozeman and C. Skiadas, vol. 3, 167–78. London: ISTE Science Publishing.

Batagelj, V., and M. Bren. 1992. "Comparing Resemblance Measures." In *Proc. International Meeting on Distance Analysis (Distancia'92)*.

Batagelj, V., and M. Bren. 1995. "Comparing Resemblance Measures." *Journal of Classification* 12:73–90.

Benzécri, J. P. 1976. *L'Analyse des Données. Tome 1: La Taxinomie. Tome 2: L'analyse des correspondances*. 2nd ed. Paris: Dunod.

Cailliez, F. and J. P. Pagès. 1976. "Introduction à l'Analyse des données." In *S.M.A.S.H., Paris*.

Cohen, J. 1960. "A Coefficient of Agreement for Nominal Scales." *Educational and Psychological Measurement* 20:27–46.

Demsar, J. 2006. "Statistical Comparisons of Classifiers over Multiple Data Sets." *The Journal of Machine Learning Research* 7:1–30.

Escofier, B. 1979. "Une représentation des variables dans l'analyse des correspondances multiples." *Revue de statistique Appliquées* 27:37–47.

Fisher, R.-A. 1922. "On the Interpretation of χ^2 from Contingency Tables, and the Calculation of P." *Journal of the Royal Statistical Society* 85 (1):87–94.

Greenacre, M. and B. Jörg. 2006. *Multiple Correspondence Analysis and Related Methods*. London: Chapman & Hall/CRC.

Kim, J. H., and S. Lee. 2003. "Tail Bound for the Minimal Spanning Tree of a Complete Graph." *Statistics & Probability Letters* 4 (64):425–30.

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

Lebart, L., A. Morineau, and K. Warwick. 1984. *Multivariate Descriptive Statistical Analysis*. New York: Wiley.

Lesot, M. J., M. Rifqi, and H. Benhadda. 2009. "Similarity Measures for Binary and Numerical Data: A Survey." *International Journal of Knowledge Engineering and Soft Data Paradigms* 1 (1):63–84.

Mantel, N. 1967. "A Technique of Disease Clustering and a Generalized Regression Approach." *Cancer Research* 27:209–20.

Park, J. C., H. Shin, and B. K. Choi. 2006. "Elliptic Gabriel Graph for Finding Neighbors in a Point Set and Its Application to Normal Vector Estimation." *Computer-Aided Design* 38 (6):619–26.

Rifqi, M., M. Detyniecki, and B. Bouchon-Meunier. 2003. "Discrimination Power of Measures of Resemblance." In *IFSA'03*.

Schneider, J. W., and P. Borlund. 2007a. "Matrix Comparison, Part 1: Motivation and Important Issues for Measuring the Resemblance between Proximity Measures or Ordination Results." *Journal of the American Society for Information Science and Technology* 58 (11):1586–95.

Schneider, J. W., and P. Borlund. 2007b. "Matrix Comparison, Part 2: Measuring the Resemblance between Proximity Measures or Ordination Results by Use of the Mantel and Procrustes Statistics. *Journal of the American Society for Information Science and Technology* 11 (58):1596–609.

Toussaint, G. T. 1980. "The Relative Neighbourhood Graph of a Finite Planar Set." *Pattern Recognition* 12 (4):261–68.

Ward, J. R. 1963. "Hierarchical Grouping to Optimize an Objective Function." *Journal of the American Statistical Association* 58 (301):236–44.

Warrens, M. J. 2008. "Bounds of Resemblance Measures for Binary (Presence/Absence) Variables." *Journal of Classification* 25 (2):195–208.

Zighed, D., Abdesselam, R., and Hadgu, A. 2012. "Topological Comparisons of Proximity Measures." In *The 16th PAKDD 2012 Conference*, edited by P.-N. Tan et al., Part I, LNAI, vol. 7301, 379–91. Berlin, Heidelberg: Springer-Verlag.

Appendix A

Table A1. Some proximity measures.

Measures	Similarity	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	$s_{12} = \frac{a+d-b-c}{a+b+c+d}$	$u_{12} = \frac{1-s_{12}}{2}$
bc		$u_{13} = \frac{4bc}{(a+b+c+d)^2}$
Sokal and Sneath 5	$s_{14} = \frac{ad}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$	$u_{14} = 1 - s_{14}$
Michael	$s_{15} = \frac{4(ad-bc)}{(a+d)^2+(b+c)^2}$	$u_{15} = \frac{1-s_{15}}{2}$
Baroni, Urbani and Buser	$s_{16} = \frac{a+\sqrt{ad}}{a+b+c+\sqrt{ad}}$	$u_{16} = 1 - s_{16}$
Yule Q	$s_{17} = \frac{ad-bc}{ad+bc}$	$u_{17} = \frac{1-s_{17}}{2}$
Yule Y	$s_{18} = \frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}$	$u_{18} = \frac{1-s_{18}}{2}$
Sokal and Sneath 4	$s_{19} = \frac{1}{4}(\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{d+b} + \frac{d}{d+c})$	$u_{19} = 1 - s_{19}$
Sokal and Sneath 3		$u_{20} = \frac{b+c}{a+d}$
Gower and Legendre	$s_{21} = \frac{a+d}{a+\frac{(b+c)}{2}+d}$	$u_{21} = 1 - s_{21}$
Sokal and Sneath 1	$s_{22} = \frac{2(a+d)}{2(a+d)+b+c}$	$u_{22} = 1 - s_{22}$

Table A2. Similarities $S(V_{u_i}, V_{u_j})$ and kappa coefficient $\widehat{\kappa}(V_{u_i}, V_{u_j})$.

Measure	Jaccard	Dice	Kulczynski	Ochiai	Sokal-Sneath-2	Braun-Blanquet	Simpson	Simple Matching	Russell-Rao	Rogers-Tanimoto	Pearson	Hamann	BC	Sokal-Sneath-5	Michael	Baroni-Urbani-Buser	Q-Yule	Y-Yule	Sokal-Sneath4	Sokal-Sneath-3	Gower-Legendre	Sokal-Sneath-1
Reference measure u_*	0.69	0.69	0.67	0.67	0.69	0.64	0.57	0.67	0.59	0.67	0.67	0.67	0.62	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67
Jaccard	1	1	0.92	0.97	1	0.90	0.66	0.72	0.74	0.71	0.77	0.71	0.56	0.87	0.77	0.92	0.72	0.72	0.77	0.71	0.71	0.71
Dice	1	1	0.92	0.97	1	0.90	0.66	0.72	0.74	0.71	0.77	0.71	0.56	0.87	0.77	0.92	0.72	0.72	0.77	0.71	0.71	0.71
Kulczynski	0.96	0.96	1	0.95	0.92	0.82	0.74	0.68	0.81	0.68	0.79	0.68	0.58	0.84	0.79	0.85	0.74	0.74	0.79	0.68	0.68	0.68
Ochiai	0.99	0.99	0.98	1	0.97	0.87	0.68	0.68	0.76	0.68	0.74	0.68	0.53	0.89	0.79	0.90	0.69	0.69	0.74	0.68	0.68	0.68
Sokal-Sneath-2	1	1	0.96	0.99	1	0.90	0.66	0.71	0.74	0.71	0.77	0.71	0.56	0.87	0.77	0.92	0.72	0.72	0.77	0.71	0.71	0.71
Braun-Blanquet	0.95	0.95	0.92	0.94	0.95	1	0.56	0.71	0.74	0.71	0.82	0.71	0.56	0.87	0.82	0.97	0.77	0.77	0.82	0.71	0.71	0.71
Simpson	0.85	0.85	0.,88	0.86	0.85	0.80	1	0.63	0.76	0.63	0.64	0.63	0.74	0.68	0.64	0.59	0.69	0.69	0.64	0.63	0.63	0.63
Simple Matching	0.87	0.87	0.86	0.86	0.87	0.87	0.83	1	0.60	1	0.79	1	0.79	0.74	0.74	0.69	0.74	0.74	0.79	1	1	1
Russell-Rao	0.88	0.88	0.92	0.89	0.88	0.88	0.89	0.82	1	0.60	0.66	0.60	0.65	0.70	0.66	0.71	0.66	0.66	0.66	0.60	0.60	0.60
Rogers-Tanimoto	0.87	0.87	0.86	0.86	0.87	0.87	0.83	1	0.82	1	0.79	1	0.79	0.74	0.74	0.69	0.74	0.74	0.79	1	1	1
Pearson	0.89	0.89	0.91	0.88	0.89	0.92	0.83	0.91	0.85	0.91	1	0.79	0.69	0.85	0.95	0.85	0.95	0.95	1	0.79	0.79	0.79
Hamann	0.87	0.87	0.86	0.86	0.87	0.87	0.83	1	0.82	1	0.91	1	0.79	0.74	0.74	0.69	0.74	0.74	0.79	1	1	1
BC	0.80	0.80	0.81	0.79	0.80	0.80	0.88	0.91	0.85	0.91	0.86	0.91	1	0.58	0.64	0.54	0.74	0.74	0.69	0.79	0.79	0.79
Sokal-Sneath-5	0.94	0.94	0.93	0.95	0.94	0.94	0.86	0.88	0.87	0.88	0.93	0.88	0.81	1	0.90	0.90	0.79	0.79	0.85	0.74	0.79	0.79
Michael	0.89	0.89	0.91	0.91	0.89	0.92	0.83	0.88	0.85	0.88	0.98	0.88	0.83	0.95	1	0.85	0.90	0.90	0.95	0.74	0.74	0.74
Baroni-Urbani-Buser	0.96	0.96	0.93	0.95	0.96	0.99	0.81	0.86	0.87	0.86	0.93	0.86	0.79	0.95	0.93	1	0.80	0.80	0.85	0.69	0.69	0.69
Q-Yule	0.87	0.87	0.88	0.86	0.87	0.89	0.86	0.88	0.85	0.88	0.98	0.88	0.88	0.91	0.95	0.91	1	1	0.95	0.74	0.74	0.74
Y-Yule	0.87	0.87	0.88	0.86	0.87	0.89	0.86	0.88	0.85	0.88	0.98	0.88	0.88	0.91	0.95	0.91	1	1	0.95	0.74	0.74	0.74
Sokal-Sneath-4	0,89	0,89	0.91	0.88	0.89	0.92	0.83	0.91	0.85	0.91	1	0.91	0.86	0.93	0.98	0.93	0.98	0.98	1	0.79	0.79	0.79
Sokal-Sneath-3	0.87	0.87	0.86	0.86	0.87	0.87	0.83	1	0.82	1	0.91	1	0,91	0.88	0.88	0.86	0.88	0.88	0.91	1	1	1
Gower-Legendre	0.87	0.87	0.86	0.86	0.87	0.87	0.83	1	0.82	1	0,91	1	0.91	0.88	0.88	0.86	0.88	0.88	0.91	1	1	1
Sokal-Sneath-1	0.87	0.87	0.86	0.86	0.87	0.87	0.83	1	0.82	1	0.91	1	0,91	0.88	0.88	0.86	0.88	0.88	0.91	1	1	1
Reference measure u_*	0.85	0.85	0.83	0.83	0.85	0.82	0.79	0.83	0.80	0.83	0.83	0.83	081	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83
Measure	Jaccard	Dice	Kulczynski	Ochiai	Sokal-Sneath-2	Braun-Blanquet	Simpson	Simple Matching	Russell-Rao	Rogers-Tanimoto	Pearson	Hamann	BC	Sokal-Sneath-5	Michael	Baroni-Urbani-Buser	Q-Yule	Y-Yule	Sokal-Sneath-4	Sokal-Sneath-3	Gower-Legendre	Sokal-Sneath-1

Examples:

$S(u_{Kulczynski}\ ,\ u_{Jaccard}) = 0.96$

$\widehat{\kappa}(u_{Jaccard}\ ,\ u_{Kulczynski}) = 0.92\ ;\ \text{p}-value < 0.01\%$

All kappa statistical tests are significant with $\alpha \leq 5\%$ level of Significance.

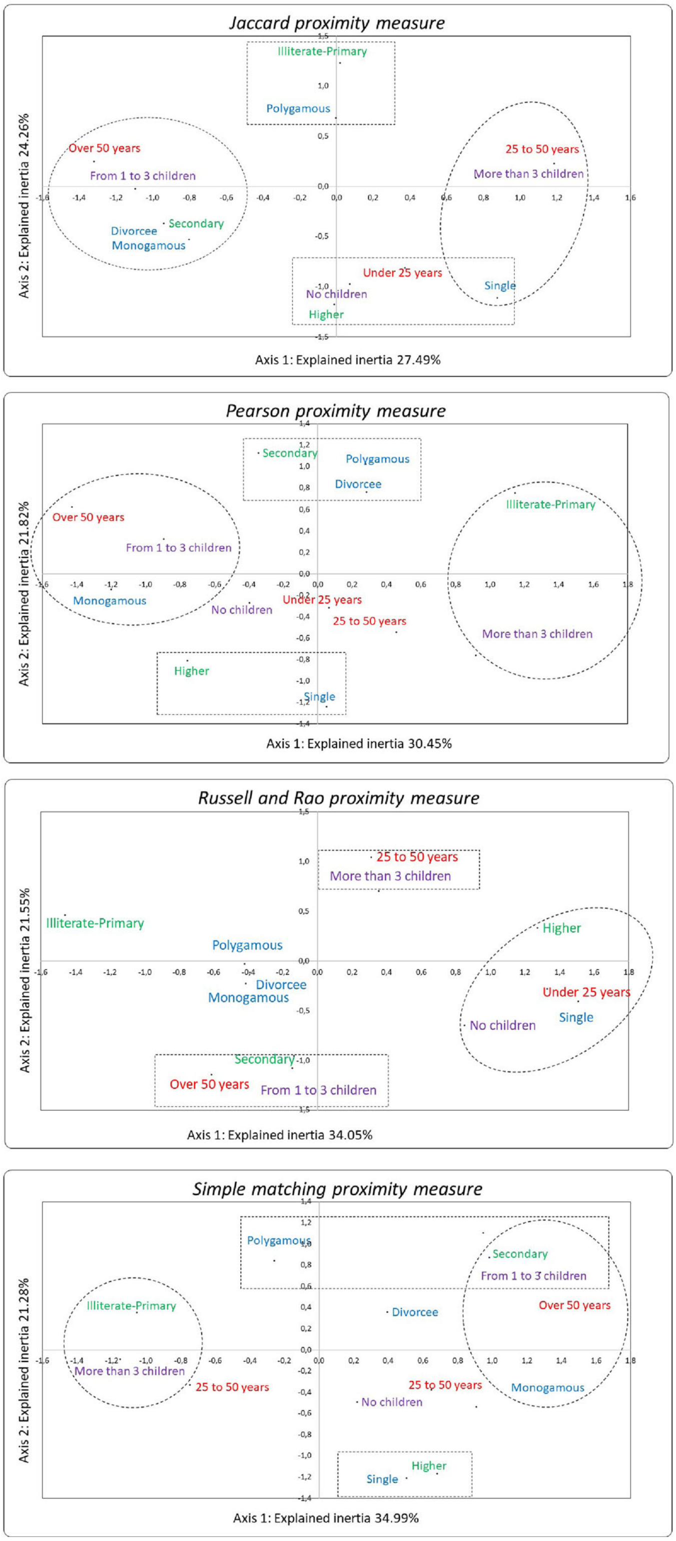


Figure A1. Jaccard (Class 1), Pearson (Class 2), Russel & Rao (Class 3), and Simple Matching (Class 4) proximity measures.

Table A3. Eigenvalues associated with the topological and classical multiple correspondence analyses.

TMCA	Axis	Eigenvalue	Proportion	Cumulative
	1	1.609	56.35%	56.35%
	2	0.914	32.01%	88.36%
	3	0.159	5.56%	93.91%
	4	0.087	3.06%	96.97%
	5	0.032	1.12%	98.10%
	6	0.027	0.95%	99.05%
	7	0.015	0.53%	99.59%
$m - p - 1 \rightarrow$	8	0.012	0.41%	100.00%
	Total	2.855	100.00%	100.00%
MCA	**Axis**	**Eigenvalue**	**Proportion**	**Cumulative**
	1	0.585	26.01%	26.01%
	2	0.462	20.52%	46.53%
	3	0.285	12.67%	59.20%
	4	0.222	9.85%	69.05%
	5	0.212	9.40%	78.45%
	6	0.166	7.39%	85.84%
	7	0.126	5.60%	91.44%
	8	0.101	4.48%	95.92%
$m - p \rightarrow$	9	0.092	4.08%	100.00%
	Total	2.250	100.00%	100.00%