PART 2

Classification Data Analysis and Methods

6

Selection of Proximity Measures for a Topological Correspondence Analysis

In this chapter, we propose a new topological approach to analyze the associations between two qualitative variables in the context of correspondence analysis. It compares and classifies proximity measures to select the best one according to the data under consideration. Similarity measures play an important role in many domains of data analysis. The results of any investigation into whether association exists between variables or any operation of clustering or classification of 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 two qualitative variables, through the topological structure induced by each measure. 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 qualitative data. The principle of the proposed approach is illustrated using a real data set with conventional proximity measures for qualitative variables.

6.1. Introduction

In order 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

Chapter written by Rafik ABDESSELAM.

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, 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 remaining measures 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 in order to identify those that are similar, they would no longer need to try out all measures.

This chapter proposes a new framework for comparing proximity measures in order to choose the best one in a context of association between two qualitative variables.

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.

Comparing 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 in the context of correspondence analysis. The eventual links or associations between modalities of two qualitative variables partly depend on the learning database being used. The results of correspondence analysis can change according to the selected proximity measure. Here, we are interested in characterizing a topological equivalence index of independence between two qualitative variables. The greater this topological index is, the more independent the variables are, according to the proximity measure u_i chosen.

Several studies on the topological equivalence of proximity measures have been proposed, [BAT 92, RIF 03, BAT 95, LES 09, ZIG 12], but none of these propositions has an association objective.

Therefore, this chapter focuses on how to construct the best adjacency matrix [ABD 14] induced by a proximity measure, taking into account the independence between two qualitative variables. A criterion for statistically selecting the best correspondence proximity measure is defined in this chapter.

This chapter is organized as follows. In section 2, after recalling the basic notions of structure, graph and topological equivalence, we present the proposed approach. How to build an adjacency matrix for no association between two qualitative variables, the choice of a measure of the degree of topological equivalence between two proximity measures and the selection criterion for picking the best association measure are discussed in this section. Section 3 presents an illustrative example using qualitative economic data. The conclusion and some perspectives of this work are given in section 4.

Table 6.1 shows some classic proximity measures used for binary data [WAR 08]; we give on $\{0,1\}^n$ the definition of 22 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	$s_8 = \frac{a+d}{a+b+c+d}$	$u_8 = 1 - s_8$
Russel 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	$s_{22} = \frac{2(a+d)}{2(a+d)+b+c}$	$u_{SS1} = 1 - s_{22}$

Table 6.1. *Some proximity measures*

$\{x^j; j = 1,..,p\}$ and $\{y^k; k = 1,..,q\}$ are sets of two qualitative variables, partition of $n = \sum_{j=1}^{p} n_j = \sum_{k=1}^{q} n_k$ individuals-objects into p

and q modalities-subgroups. The interest lies in whether there is a topological association between these two variables. Let us denote:

– $X_{(n,p)}$ the data matrix associated with the p dummy variables $\{x^j; j = 1, p\}$, of a qualitative variable x with n rows-objects and p columns-variables;

– $Y_{(n,q)}$ the data matrix associated with the q dummy variables $\{y^k; k = 1, q\}$ of a qualitative variable y with n rows-objects and q columns-variables;

– $Z_{(n,r)} = [\,X\,|\,Y\,] = [\,z^1 = x^1, \cdots, z^j = x^j, \cdots, z^p = x^p \,|\, z^{p+1} = y^1, \cdots, z^k = y^k, \cdots, z^r = y^q\,]$ the full binary table, juxtaposition of X and Y binary tables, with n rows-objects and $r = p + q$ columns-modalities;

– $K_{(p,q)} = {}^tX\ Y$ the contingency table;

– $M_{B_{(r,r)}} = {}^tZ\ \ Z = \left(\begin{array}{c|c} {}^tX\ X & {}^tX\ Y \\ \hline {}^tY\ X & {}^tY\ Y \end{array}\right) = \left(\begin{array}{c|c} {}^tX\ X & K \\ \hline {}^tK & {}^tY\ Y \end{array}\right)$ the symmetric Burt matrix of the two-way cross-tabulations of the two variables. The diagonals are the cross-tabulations of each variable with itself;

– $W_{(r,r)} = Diag[M_B] = \left(\begin{array}{c|c} {}^tX\ X & 0 \\ \hline 0 & {}^tY\ Y \end{array}\right) = \left(\begin{array}{c|c} W_p & 0 \\ \hline 0 & W_q \end{array}\right)$ the diagonal matrix of $r = p + q$ frequencies. The diagonal terms are the frequencies of the modalities of x and y, totals rows and columns of contingency table K.

– $U = \mathbb{1}_r\ {}^t\mathbb{1}_r$ is the $r \times r$ matrix of 1s, I_r the $r \times r$ identity matrix where $\mathbb{1}_r$ denotes the r vector of 1s and $\mathbb{1}_n$ the n vector of 1s.

The dissimilarity matrices associated with proximity measures are computed from data given by the contingency table K. The attributes of any two points' modalities' z^j and z^k in $\{0,1\}^n$ of the proximity measures can be easily written and calculated from the following matrices. Computational complexity is thus considerably reduced.

- $A_{(r,r)} = (a_{jk}) = M_B$

whose element $a_{jk} = |Z^j \cap Z^k| = \sum_{i=1}^n z_i^j z_i^k$ is the number of attributes common to both points z^j and z^k;

- $B_{(r,r)} = (b_{jk}) = {}^tZ\ (\mathbb{1}_n\ {}^t\mathbb{1}_r - Z) = {}^tZ\ \mathbb{1}_n\ {}^t\mathbb{1}_r\ - \ {}^tZ\ Z$

$$= W\ \mathbb{1}_r\ {}^t\mathbb{1}_r - A = W\ U - A$$

whose element $b_{jk} = |Z^j - Z^k| = |Z^j \cap \overline{Z^k}| = \sum_{i=1}^{n} z_i^j(1 - z_i^k)$ is the number of attributes present in z^j but not in z^k;

- $C_{(r,r)} = (c_{jk}) = {}^t(\mathbb{1}_n\, {}^t\mathbb{1}_r - Z)\, Z = {}^t(\mathbb{1}_n\, {}^t\mathbb{1}_r)\, Z - {}^tZ\, Z$

$$= \mathbb{1}_r\, {}^t\mathbb{1}_n\, Z - {}^tZ\, Z = UW - A$$

whose element $c_{jk} = |Z^k - Z^j| = |Z^k \cap \overline{Z^j}| = \sum_{i=1}^{n} z_i^k(1 - z_i^j)$ is the number of attributes present in z^k but not in z^j;

- $D_{(r,r)} = (d_{jk}) = {}^t(\mathbb{1}_n\, {}^t\mathbb{1}_r - Z)\,(\mathbb{1}_n\, {}^t\mathbb{1}_r - Z)$

$$= \mathbb{1}_r\, {}^t\mathbb{1}_n\, \mathbb{1}_n\, {}^t\mathbb{1}_r - \mathbb{1}_r\, {}^t\mathbb{1}_n\, Z - {}^tZ\, \mathbb{1}_n\, {}^t\mathbb{1}_r + {}^tZ\, Z$$

$$= n\mathbb{1}_r\, {}^t\mathbb{1}_r - UW - WU + A = nU - UW - WU + A$$

$$= nU - (A + B + C)$$

whose element $d_{jk} = |\overline{Z^j} \cap \overline{Z^k}| = \sum_{i=1}^{n}(1 - z_i^j)(1 - z_i^k)$ is the number of attributes in neither z^j nor z^k.

$Z^j = \{i/z_i^j = 1\}$ and $Z^k = \{i/z_i^k = 1\}$ are the sets of attributes present in data point-modality z^j and z^k, respectively, and $|.|$ is the cardinality of a set.

The attributes are linked by the relation:

$$\forall j = 1, p\ ;\ \forall k = 1, q \quad a_{jk} + b_{jk} + c_{jk} + d_{jk} = n.$$

Together, the four dependent quantities a_{jk}, b_{jk}, c_{jk} and d_{jk} are presented in Table 6.2, where the information can be summarized by an index of similarity (affinity, resemblance, association, coexistence).

	$z^k = 1$	$z^k = 0$	Total
$z^j = 1$	a_{jk}	b_{jk}	$a_{jk} + b_{jk}$
$z^j = 0$	c_{jk}	d_{jk}	$c_{jk} + d_{jk}$
Total	$a_{jk} + c_{jk}$	$b_{jk} + d_{jk}$	n

Table 6.2. *The 2×2 contingency table between modalities z^j and z^k*

6.2. Topological correspondence

Topological equivalence is based on the concept of the topological graph also referred to as the 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 more precisely define 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 $E = \{z^1 = x^1, \ldots, z^p = x^p, z^{p+1} = y^1, \ldots, z^r = y^q\}$ of $r = |E|$ modalities in $\{0,1\}^n$, associated with the variables x and y. 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. We can choose the minimal spanning tree (MST) [KIM 03], the Gabriel graph (GG) [PAR 06] or, as is the case here, the relative neighborhood graph (RNG) [TOU 80, JAR 92].

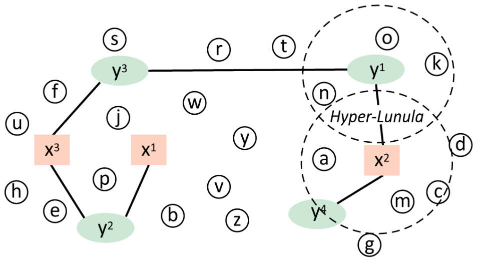
For any given proximity measure u, we construct the associated adjacency binary symmetric matrix V_u of order $r = p + q$, where all pairs of neighboring modalities (z^j, z^k) satisfy the following RNG property.

PROPERTY 6.1.– Relative neighborhood graph (RNG)

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

This means that if two modalities z^j and z^k which verify the RNG property are connected by an edge, the vertices z^j and z^k are neighbors.

Thus, for any proximity measure given, u, we can associate an adjacency matrix V_u, of binary and symmetrical order $r = p + q$. Figure 6.1 illustrates a set of n object-individuals around seven modalities associated with two qualitative variables x and y with three and four modalities, respectively.



$$\left(\begin{array}{c|ccc|cccc} V_u & z^1 = x^1 & z^2 = x^2 & z^3 = x^3 & z^4 = y^1 & z^5 = y^2 & z^6 = y^3 & z^7 = y^4 \\ \hline z^1 = x^1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ z^2 = x^2 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ z^3 = x^3 & 1 & 0 & 1 & 0 & 1 & 1 & 0 \\ \hline z^4 = y^1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ z^5 = y^2 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ z^6 = y^3 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ z^7 = y^4 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \end{array}\right)$$

Figure 6.1. *RNG example with seven groups-modalities – associated adjacency matrix. For a color version of this figure, see www.iste.co.uk/makrides/data3.zip*

For example, if $V_u(z^2 = x^2, z^4 = y^1) = 1$, then on the geometrical plane, the hyper-Lunula (intersection between the two hyperspheres centered on the two modalities x^2 and y^1) 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 chapter, we chose to use the relative Neighborhood graph (RNG) [TOU 80]. AQ1

6.2.1. *Comparison and selection of proximity measures*

First, we compare different proximity measures according to their topological similarity in order 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}, respectively, are different or not. The degree of topological equivalence between two proximity measures is measured by the following property of concordance.

PROPERTY 6.2.– Topological equivalence index between two adjacency matrices

$$S(V_{u_i}, V_{u_j}) = \frac{\sum_{k=1}^{r}\sum_{l=1}^{r}\delta_{kl}(z^k, z^l)}{r^2} \text{ with } \delta_{kl}(z^k, z^l)$$
$$= \begin{cases} 1 \text{ if } V_{u_i}(z^k, z^l) = V_{u_j}(z^k, z^l) \\ 0 \text{ otherwise.} \end{cases}$$

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

The 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 note $V_{u_*} = I_r$, and the $r = p + q$ identity matrix. It is a perfect adjacency matrix, which corresponds to the null hypothesis H_0 of independence: no association between the two variables.

$$V_{u_*} = \left(\begin{array}{c|c} I_p & 0 \\ \hline 0 & I_q \end{array}\right) = I_r$$

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

Thus, with this reference proximity measure we can establish the topological independence index $TII_i = S(V_{u_i}, V_{u_*})$ – the degree of topological equivalence of no association between the two variables – by measuring the percentage of similarity between the adjacency matrix V_{u_i} and the reference adjacency matrix V_{u_*}. The greater this topological index is and tends to 1, the more independent the variables are, according to the proximity measure u_i chosen.

In order to visualize the similarities among all the 22 proximity measures considered, 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 in order to see which measures are close to one another.

We can use any classic visualization technique 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 Laplacian projection, to map the 22 proximity measures into a two-dimensional space.

Finally, in order to evaluate and select the no-association proximity measures, we project the reference measure u_* as a supplementary element into the methodological chain of data analysis methods (PCA and HAC), positioned by the dissimilarity vector with 22 components $[D]_{*i} = 1- S(V_{u_*}, V_{u_i})$.

6.2.2. *Statistical comparisons between two proximity measures*

In a metric framework, there are several ways of testing the null hypothesis, H_0, of no association between two variables, and many of the tests are based on the chi-square statistic.

In this paragraph, we use Cohen's kappa coefficient to test statistically the degree of topological equivalence between two proximity measures. This non-parametric test compares these measures based on their associated adjacency matrices.

A comparison between indices of proximity measures has also been studied by [SCH 07a, SCH 07b] and [DEM 06] from a statistical perspective. These authors proposed an approach that compares similarity matrices obtained by each proximity measure, using Mantel's test [MAN 67], in a pairwise manner.

Cohen's non-parametric Kappa test [COH 60] is the statistical test best suited to our problem, as 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 measures of proximity, unlike the coefficients of Kendall or Spearman, for example, which evaluate the degree of concordance between quantitative values. The Kappa concordance rate between two adjacency matrices is estimated to evaluate the degree of topological equivalence between their 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 non-parametric test of paired data. These binary and symmetric matrices of order r are unfolded in two vector-matched components, consisting of $\frac{r(r+1)}{2}$ values: the r diagonal values and the $\frac{r(r-1)}{2}$ values above or below the diagonal.

The degree of topological equivalence between two proximity measures is estimated from the Kappa coefficient of concordance, computed on a 2×2 contingency table $N = (n_{kl})_{k,l=0,1}$ formed by the two binary vectors, using the following relation:

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

where

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

$P_e = \frac{4}{r^2(r+1)2} \sum_{k=0}^{1} n_{k.} n_{.k}$ 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 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 AQ2 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 each proximity measure u_i with the perfect measure u_* by comparing the adjacency matrices V_{u_i} and V_{u_*} to estimate the degree of topological equivalence of independence of each measure.

6.3. Application to real data and empirical results

The data displayed in Table 6.3 are from an INSEE[1] study concerning the 554,000 enterprise births in France 2016 [INS 16]. The question was whether there was any association between the type of enterprise and the sector of activity of the enterprise's operation.

	Type of enterprise			
Activity sector	Company	Traditional Individual Enterprise	Micro Entrepreneur	Total
Industry	8,6	7,7	8,3	24,6
Construction	26,5	18,6	16,5	61,6
Trade, Transport, Accommodation and Restoration	64	48,7	48,7	161,5
Information and communication	11,1	2,1	14,5	27,6
Financial and insurance activities	12,6	1,3	2	15,8
Real estate activities	11,3	5,1	2,5	18,9
Specialized, scientific and technical activities	27,6	11,9	51	90,6
Education and Health	6,5	26,4	36,4	69,4
Service activities	20,6	20,6	42,9	84
Total	188,8	142,4	222,8	554

Table 6.3. *Contingency table - Enterprise births in France 2016 (in thousands)*

[1] National Institute of Statistics and Economic Studies.

In a metric context, the null hypothesis of the chi-square independence test is clearly rejected with a risk of error $\alpha \leq 5\%$. Therefore, there is a strong association between the type of enterprise and the activity sector. We can also perform a factorial correspondence analysis to locate and visualize any significant links between all the modalities of these two variables.

In a topological context, the main results of the proposed approach are presented in the following tables and graphs, which allow us to visualize proximity measures close to each other in the context of no association between the type of enterprise and the activity sector.

Table 6.4 summarizes the similarities and Kappa statistic values between the reference measure u_* and each of the 22 proximity measures in a topological framework.

HAC Class	Letter	Measure	TII_i	$\widehat{\kappa}(V_{u_i}, V_{u_*})$	$p - value$
4	A	Jaccard	0.625	0.308	$< .0001$
4	A	Dice, Czekanowski	0.625	0.308	$< .0001$
4	A	Kulczynski	0.625	0.308	$< .0001$
4	A	Driver, Kroeber and Ochiai	0.625	0.308	$< .0001$
4	A	Sokal-Sneath-2	0.625	0.308	$< .0001$
4	A	Braun and Blanquet	0.625	0.308	$< .0001$
4	A	Simpson	0.625	0.308	$< .0001$
4	A	Russel and Rao	0.625	0.308	$< .0001$
4	A	Sokal and Sneath 5	0.625	0.308	$< .0001$
4	A	Y-Yule	0.625	0.308	$< .0001$
2	A	Baroni, Urbani and Buser	0.625	0.308	$< .0001$
2	A	Q-Yule	0.625	0.308	$< .0001$
3	B	Sokal and Sneath 4	0.708	0.397	$< .0001$
3	C	Pearson	0.736	0.432	$< .0001$
2	D	Michael	0.736	0.432	$< .0001$
1	E	Simple Matching	0.847	0.606	$< .0001$
1	E	Rogers and Tanimoto	0.847	0.606	$< .0001$
1	E	Hamann	0.847	0.606	$< .0001$
1	E	BC	0.847	0.606	$< .0001$
1	E	Sokal and Sneath 3	0.847	0.606	$< .0001$
1	E	Gower and Legendre	0.847	0.606	$< .0001$
1	E	Sokal and Sneath 1	0.847	0.606	$< .0001$

Table 6.4. *Topological Index of Independence & Kappa test. For a color version of this table, see www.iste.co.uk/makrides/data3.zip*

The proximity measures are given in ascending order of the topological independence index $S(V_{u_i}, V_{u_*})$. Therefore, greater this index is, further we are getting closer the independence position, and more the null hypothesis will be rejected. All the 22 proximity measures considered reject the null hypothesis $H_0 : \kappa = 0$ (no concordance, independence), so they all conclude that there is a link between the type of enterprise and the activity sector.

The results of similarities and statistical Kappa tests between all pairs of proximity measures are given in the appendix, Table 6.7. 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 the $\alpha \leq 5\%$ level of significance. The similarities in pairs between the 22 proximity measures somewhat differ: some are closer than others. In Table 6.4, proximity measures with the same letter are in perfect topological equivalence $S(V_{u_i}, V_{u_j}) = 1$ with a perfect concordance $\widehat{\kappa}(V_{u_i}, V_{u_j}) = 1$ and proximity measures with the same number are in the same HAC class.

An HAC algorithm based on the Ward criterion[2] [WAR 63] was used in order to characterize classes of proximity measures relative to their similarities. The reference measure u_* is projected as a supplementary element.

The dendrogram in Figure 6.2 represents the hierarchical tree of the 22 proximity measures considered.

Table 6.5 summarizes the main results of the chosen partition into four homogeneous classes of proximity measures, obtained from the cut of the hierarchical tree of Figure 6.2. Moreover, in view of the results in Table 6.5, the reference measure u_* is closer to the measures of the first class, measures for which there is a weak association between the two variables among the 22 proximity measures considered. We will have a stronger association between the type of enterprise and the activity sector if we choose one proximity measure among those of class 4.

It was shown in [ABD 14] and [ZIG 12], 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.

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

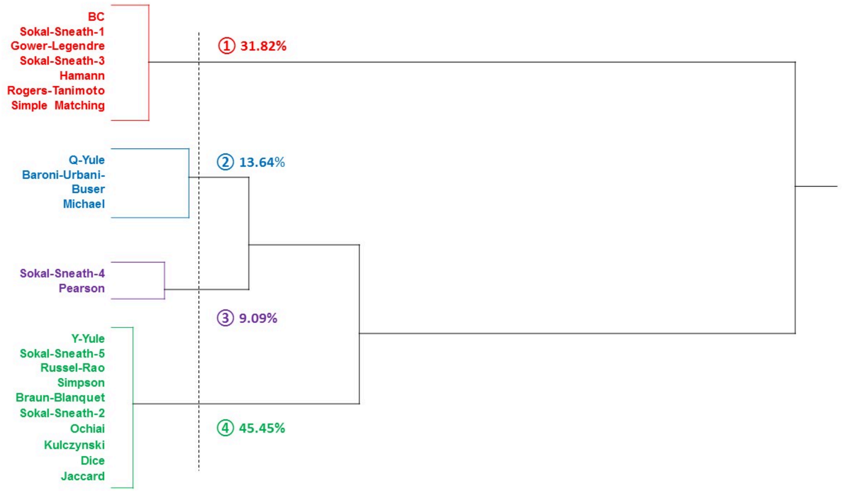


Figure 6.2. *Hierarchical tree of the proximity measures. For a color version of this figure, see www.iste.co.uk/makrides/data3.zip*

Number	Class 1	Class 2	Class 3	Class 4
Frequency	7	3	2	10
Proximity measures	$u_{Simple-Matching}$ $u_{Rogers-Tanimoto}$ u_{Hamann} u_{BC} $u_{Sokal-Sneath-3}$ $u_{Gower-Legendre}$ $u_{Sokal-Sneath-1}$	$u_{Michael}$ $u_{Baroni-Urbani-Buser}$ u_{Q-Yule}	$u_{Pearson}$ $u_{Sokal-Sneath-4}$	$u_{Jaccard}$ u_{Dice} $u_{Kulczynski}$ u_{Ochiai} $u_{Sokal-Sneath-2}$ $u_{Braun-Blanquet}$ $u_{Simpson}$ $u_{Russel-Rao}$ $u_{Sokal-Sneath-5}$ u_{Y-Yule}
Reference	u_*			

Table 6.5. *Assignment of the reference measure. For a color version of this table, see www.iste.co.uk/makrides/data3.zip*

For any proximity measure given in Table 6.1, we will show how to build and apply the Kappa test in order to compare two adjacency matrices to measure and test their topological equivalence $\kappa(V_{u_i}, V_{u_j})$ and their degree of independence $\kappa(V_{u_i}, V_{u_*})$.

Let V_{u*} and $V_{Jaccard}$, the reference and Jaccard adjacency matrices, respectively, be $n \times n$ binary symmetric matrices with lower similarity $S(V_{u*}, V_{Jaccard}) = 62.50\%$. These matrices are unfolded to two vectors comprising the $r(r+1)/2 = 78$ diagonal and upper-diagonal values. These two binary vectors are two dummy variables represented in the same sample size of 78 pairs of objects. We then formulated the null hypothesis, $H_0 : \kappa = 0$ (independence), that there is no association between the two variables.

Table 6.6 shows the contingency table observed between the two binary vectors associated with the reference and Jaccard proximity measures. Thus, for this example, the calculated Kappa value $\widehat{\kappa} = 0.3077$ corresponds to a p-value of less than 0.01%. Since this probability is lower than a pre-specified significance level of 5%, the null hypothesis of independence is rejected. We can therefore conclude that the Jaccard measure and reference measure are not independent.

	$V_{Jaccard} = 0$	$V_{Jaccard} = 1$	Total
$V_{u*} = 0$	39	27	66
$V_{u*} = 1$	0	12	12
Total	39	39	78

Table 6.6. *The 2×2 contingency table – Reference and Jaccard measures*

6.4. Conclusion and perspectives

The choice of a proximity measure is very subjective; it is often based on habits or on criteria such as the interpretation of the *a posteriori* results.

This work proposes a new approach to select the best proximity measure in a context of topological independence between two qualitative variables, for the purpose of performing a topological correspondence analysis (TCA). The proposed approach is based on the concept of neighborhood graphs induced by a proximity measure in the case of qualitative data. Results obtained from a real data set highlight the effectiveness of selecting the best proximity measure(s).

Future research will focus on developing TCAs with the best proximity measure selected and on extending this approach to analyze associations between more than two categorical variables, called topological multiple correspondence analysis (TMCA).

6.5. Appendix

Measure	Sokal-Sneath-1	Gower-Legendre	Sokal-Sneath-3	Sokal-Sneath-4	Y-Yule	Q-Yule	Baroni-Urbani-Buser	Michael	Sokal-Sneath-5	BC	Hamann	Pearson	Rogers-Tanimoto	Russel-Rao	Simple Matching	Simpson	Braun-Blanquet	Sokal-Sneath-2	Ochiai	Kulczynski	Dice	Jaccard	Measure
Jaccard	0.18	0.18	0.18	0.79	1	1	1	0.64	1	0.18	0.18	0.74	0.18	1	0.18	1	1	1	1	1	1	1	
Dice	0.18	0.18	0.18	0.79	1	1	1	0.64	1	0.18	0.18	0.74	0.18	1	0.18	1	1	1	1	1	1	1	
Kulczynski	0.18	0.18	0.18	0.79	1	1	1	0.64	1	0.18	0.18	0.74	0.18	1	0.18	1	1	1	1	1	1	1	
Ochiai	0.18	0.18	0.18	0.79	1	1	1	0.64	1	0.18	0.18	0.74	0.18	1	0.18	1	1	1	1	1	1	1	
Sokal-Sneath-2	0.18	0.18	0.18	0.79	1	1	1	0.64	1	0.18	0.18	0.74	0.18	1	0.18	1	1	1	1	1	1	1	
Braun-Blanquet	0.18	0.18	0.18	0.79	1	1	1	0.64	1	0.18	0.18	0.74	0.18	1	0.18	1	1	1	1	1	1	1	
Simpson	0.18	0.18	0.18	0.79	1	1	1	0.64	1	0.18	0.18	0.74	0.18	1	0.18	1	1	1	1	1	1	1	
Simple Matching	1	1	1	0.34	0.18	0.18	0.18	0.50	0.18	0.94	1	0.38	1	0.18	1	0.56	0.56	0.56	0.56	0.56	0.56	0.56	
Russel-Rao	0.18	0.18	0.18	0.79	1	1	1	0.64	1	0.18	0.18	0.74	0.18	1	0.56	1	1	1	1	1	1	1	
Rogers-Tanimoto	1	1	1	0.34	0.18	0.18	0.18	0.50	0.18	0.94	1	0.38	1	0.56	1	0.56	0.56	0.56	0.56	0.56	0.56	0.56	
Pearson	0.38	0.38	0.38	0.95	0.74	0.74	0.74	0.89	0.74	0.38	0.38	1	0.69	0.86	0.69	0.86	0.86	0.86	0.86	0.86	0.86	0.86	
Hamann	1	1	1	0.34	0.18	0.18	0.18	0.50	0.18	0.94	1	0.69	1	0.56	1	0.56	0.56	0.56	0.56	0.56	0.56	0.56	
BC	1	1	1	0.34	0.18	0.18	0.18	0.50	0.18	1	0.97	0.69	0.97	0.56	0.97	0.56	0.56	0.56	0.56	0.56	0.56	0.56	
Sokal-Sneath-5	0.18	0.18	0.18	0.79	1	1	1	0.64	1	0.56	0.56	0,86	0.56	1	0.56	1	1	1	1	1	1	1	
Michael	0.50	0.50	0.50	0.84	0.64	0.64	0.64	1	0.81	0.75	0.75	0.94	0.75	0.81	0.75	0.81	0.81	0.81	0.81	0.81	0.81	0.81	
Baroni-Urbani-Buser	0.18	0.18	0.18	0.79	1	1	1	0.81	1	1	1	1	1	1	0.56	1	1	1	1	1	1	1	
Q-Yule	0.18	0.18	0.18	0.79	1	1	1	0.81	1	1	1	1	1	1	0.56	1	1	1	1	1	1	1	
Y-Yule	0.18	0.18	0.18	0.79	1	1	1	0.56	1	0.56	0.56	0.86	0.56	1	0.56	1	1	1	1	1	1	1	
Sokal-Sneath-4	0.34	0.34	0.34	1	0.89	0.89	0.89	0,92	0.89	0.67	0.67	0.97	0.67	0.89	0.67	0.89	0.89	0.89	0.89	0.89	0.89	0.89	
Sokal-Sneath-3	1	1	1	0.67	0.56	0.56	0.56	0.75	0.56	1	1	0.69	1	0.56	1	0.56	0.56	0.56	0.56	0.56	0.56	0.56	
Gower-Legendre	1	1	1	0,67	0.56	0.56	0.56	0.75	0.56	1	1	0.69	1	0.56	1	0.56	0.56	0.56	0.56	0.56	0.56	0.56	
Sokal-Sneath-1	1	1	1	0.67	0.56	0.56	0.56	0.75	0.56	1	1	0.69	1	0.56	1	0.56	0.56	0.56	0.56	0.56	0.56	0.56	
Measure	Sokal-Sneath-1	Gower-Legendre	Sokal-Sneath-3	Sokal-Sneath4	Y-Yule	Q-Yule	Baroni-Urbani-Buser	Michael	Sokal-Sneath-5	BC	Hamann	Pearson	Rogers-Tanimoto	Russel-Rao	Simple Matching	Simpson	Braun-Blanquet	Sokal-Sneath-2	Ochiai	Kulczynski	Dice	Jaccard	

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

$$\text{Example}: \; S(u_{Simple\ matching}\ ,\ u_{Jaccard}) = 0.56$$

$$\widehat{\kappa}(u_{Jaccard}\ ,\ u_{Simple\ matching}) = 0.18 \;\; ; \;\; p - value = 0.0411$$

Table 6.7. *Similarities $S(V_{u_i}, V_{u_j})$ & Kappa coefficient $\widehat{\kappa}(V_{u_i}, V_{u_j})$. For a color version of this table, see www.iste.co.uk/makrides/data3.zip*

6.6. References

[ABD 14] ABDESSELAM R., "Proximity measures in topological structure for discrimination", SKIADAS C.H. (ed.), *In a Book Series SMTDA-2014, 3nd Stochastic Modeling Techniques and Data Analysis, International Conference*, Lisbon, Portugal, ISAST, pp. 599–606, 2014.

[BAT 92] BATAGELJ V., BREN M., "Comparing resemblance measures", In *Proc. International Meeting on Distance Analysis (Distancia'92)*, 1992.

[BAT 95] BATAGELJ V., BREN M., "Comparing resemblance measures", *In Journal of Classification*, vol. 12, pp. 73–90, 1995.

[COH 60] COHEN J., "A coefficient of agreement for nominal scales", *Educ Psychol Meas*, vol. 20, pp. 27–46, 1960.

[DEM 06] DEMSAR J., "Statistical comparisons of classifiers over multiple data sets", *The journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[INS 16] INSEE 2016. Available at: https://www.insee.fr/fr/statistiques/2562977.

[JAR 92] JAROMCZYK J.-W., TOUSSAINT G.-T., "Relative neighborhood graphs and their relatives", *Proceedings of IEEE*, vol. 80, no. 9, pp. 1502–1517, 1992.

[KIM 03] KIM J.H., LEE S., "Tail bound for the minimal spanning tree of a complete graph", *In Statistics & Probability Letters*, vol. 4, no. 64, pp. 425–430, 2003.

[LES 09] LESOT M.J., RIFQI M., BENHADDA H., "Similarity measures for binary and numerical data: A survey", *In IJKESDP*, vol. 1, no. 1, pp. 63–84, 2009.

[MAN 67] MANTEL N., "A technique of disease clustering and a generalized regression approach", *In Cancer Research*, vol. 27, pp. 209–220, 1967.

[PAR 06] PARK J.C., SHIN H., CHOI B.K., "Elliptic gabriel graph for finding neighbors in a point set and its application to normal vector estimation", *In Computer-Aided Design Elsevier*, vol. 38, no. 6, pp. 619–626, 2006.

[RIF 03] RIFQI M., DETYNIECKI M., BOUCHON-MEUNIER B., "Discrimination power of measures of resemblance", *IFSA'03 Citeseer*, 2003.

[SCH 07a] SCHNEIDER J.W., BORLUND P., "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*, vol. 58, no. 11, pp. 1586–1595, 2007.

[SCH 07b] SCHNEIDER J.W., BORLUND P., "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*, vol. 11, no. 58, pp. 1596–1609, 2007.

[TOU 80] TOUSSAINT G.T., "The relative neighbourhood graph of a finite planar set", *In Pattern Recognition*, vol. 12, no. 4, pp. 261–268, 1980.

[WAR 63] WARD J.R., “Hierarchical grouping to optimize an objective function”, *In Journal of the American Statistical Association JSTOR*, vol. 58, no. 301, pp. 236–244, 1963.

[WAR 08] WARRENS M.J., “Bounds of resemblance measures for binary (presence/absence) variables”, *Journal of Classification*, vol. 25, no. 2, pp. 195–208, 2008.

[ZIG 12] ZIGHED D., ABDESSELAM R., HADGU A., “Topological comparisons of proximity measures”, In TAN P.-N. *et al.* (eds) , *The 16th PAKDD 2012 Conference*. Part I, LNAI 7301, Springer-Verlag, Berlin Heidelberg, pp. 379–391, 2012.

[AQ1] “Relative neighbors graph (GNR)” has been changed to “relative neighborhood graph (RNG)”. Please check.

[AQ2] This sentence is unclear. Please check.