New Weighed Similarity Indexes for Market Segmentation Using Categorical Variables

Isabella Morlini and Sergio Zani

Abstract In this paper we introduce new similarity indexes for binary and polytomous variables, employing the concept of "information content". In contrast to traditionally used similarity measures, we suggest to consider the frequency of the categories of each attribute in the sample. This feature is useful when dealing with rare categories, since it makes sense to differently evaluate the pairwise presence of a rare category from the pairwise presence of a widespread one. We also propose a weighted index for dependent categorical variables. The suitability of the proposed measures from a marketing research perspective is shown using two real data sets.

1 Introduction

Consider a general setup in which k categorical variables X_s $(s = 1, \ldots, k)$ with nominal scale are of interest and a categorical data set $\mathbf{X} = (\mathbf{x}_1', \ldots, \mathbf{x}_n')'$ is collected from n subjects $u_1, u_2, \ldots, u_n$. Let $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})'$ be the profile of the k attributes for the ith subject. The resemblance between two subjects u_i and u_j is typically measured by pairwise similarity indexes (see, e.g., Sneath and Sokal 1973 and, recently, Warrens 2008). Most of the available similarity indexes in the literature have been developed to deal with binary variables and few measures have been proposed specifically for polytomous attributes. For these variables, distance functions like the Euclidean or the Manhattan are sometimes used, especially for classification purposes. However, the principal difficulty in dealing with nominal categorical data is typically the lack of a metric space in which data points are positioned and the measured distances can be different when a different coding scheme is used for the variables (Zhang et al. 2006). In this paper we extend the original work of Zani (1982) and we first allow the classical similarity measures to deal with polytomous variables. Then we consider the problem of weighting variables in computing similarities between subjects. We propose a criterion for weighting the pairwise presence of a category, on the basis of the Shannon's "information content" of the relative frequency in the sample. Both in marketing research and in other fields, it appears relevant to attach a higher weight to the pairwise presence of a rare

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_61,

category in the sample rather than to the pairwise presence of a widespread one. A similar criterion is used in correspondence analysis, where the effect of increasing the values corresponding to low-frequencies categories relatively more than those corresponding to high-frequencies categories is accomplished with the χ^2 distance. Finally, we provide some numerical examples to illustrate the use of the indexes and we show the suitability of the proposed measures for market segmentation.

2 A Class of Similarity Indexes for Polytomous Variables

Consider k categorical variables X_s $(s = 1, \ldots, k)$ with $h_s \geq 2$ categories. An easy way to numerically code the attributes is through the so called "dummy variables". A binary variable is introduced for each category: the number of dummy variables is $h = \sum_{s=1}^{k} h_s$. With this coding scheme, we obtain a $(n \times h)$ data matrix of the form:

X_1	$\ldots$	X_s	$\ldots$	X_k
$X_{11} \ldots X_{1h_1}$	$\ldots$	$X_{s1} \ldots X_{sv} \ldots X_{sh_s}$	$\ldots$	$X_{k1} \ldots X_{kh_k}$
$\ldots$		$\vdots$ x_{isv} $\vdots$		$\ldots$
$\ldots$		n_{sv}		$\ldots$

where X_{sv} is the dummy variable for the vth category of the sth attribute ($s = 1, \ldots, k$ $v = 1, \ldots, h_s$). The observed value for the ith observation is $x_{isv} = 1$ if $x_{is} = v$ and $x_{isv} = 0$ if $x_{is} \neq v$. The frequency, in the sample, of the vth category of the sth attribute is $n_{sv} = \sum_{i=1}^{n} x_{isv}$ and the relative frequency is $f_{sv} = n_{sv}/n$. When the categorical variable is dichotomous, this coding scheme implies two dummy variables. It is obvious that $x_{is1} = 1 \iff x_{is2} = 0$ and $x_{is1} = 0 \iff x_{is2} = 1$ and the second dummy is superfluous. However, when dealing with mixed polytomous and dichotomous variables, the same coding is needed for both. To evaluate the similarity between subjects u_i and u_j, we introduce the following contingency table:

	1	0	tot
1	α	β	$\alpha + \beta$
0	γ	δ	$\gamma + \delta$
tot	$\alpha + \gamma$	$\beta + \delta$	h

We will call *positive matches* or *agreements* in u_i and u_j the α pairs $1 - 1$ and *disagreements* the $\beta + \gamma$ pairs $1 - 0$ and $0 - 1$. The δ pairs $0 - 0$ (*negative matches*) simply indicate that both u_i and u_j do not share the category corresponding to the dummy variable and are useless in evaluating the similarity between two subjects

since this number only depends on the number of the categories of the original categorical variables. The index:

$$S1_{ij} = \frac{\alpha}{\alpha + \beta + \gamma} \tag{1}$$

is bounded in [0,1] and has the following properties:

- $1 - S1_{ij} = (\beta + \gamma)/(\alpha + \beta + \gamma)$ is a distance. $\beta + \gamma$ is the Manhattan and the square Euclidean distance between u_i and u_j in the dummy variable coding.
- for binary variables, i.e., for $h_s = 2$, $s = 1, \ldots, k$, index $S1_{ij}$ becomes equivalent to the Rogers-Tanimoto index.

We may obtain a more general index by introducing a weight for the *disagreements* (Gower and Legendre 1986):

$$_wS1_{ij} = \frac{\alpha}{\alpha + w(\beta + \gamma)} \tag{2}$$

with $w > 0$. When $w = 0.5$ and $h_s = 2$, $s = 1, \ldots, k$, expression (2) is equivalent to the Sokal-Michener index. Given two subjects u_i and u_j, the probability of an *agreement* in X_{sv}, in a Bernoulli trial, is f_{sv}^2. The weight given to an *agreement* in X_{sv} should be a decreasing function of f_{sv}^2. Assuming ($f_{sv}^2 > 0$), here we propose the weight $w_{sv} = log(1/f_{sv}^2)$ which is a measure of the information content of an *agreement* in X_{sv}. For independent variables, this measure is additive: if subjects u_i and u_j present two *positive matches*, in X_{sv} and X_{kl}, then the joint weight is $w_{sv,kl} = w_{sv} + w_{kl}$. The choice of w_{sv} is for interpretability purposes rather than for numerical ones. This weight is conceived in the light of the information theory. This criterion was first introduced by Burnaby (1970). We do not use the information entropy (see, e.g., MacKay 2002) $e_{sv} = f_{sv}^2 log(1/f_{sv}^2)$, because it is not a decreasing function of f_{sv}^2. By weighting the pairwise *positive matches* with w_{sv}, we obtain the index:

$$E_{ij} = \sum_{s=1}^{k} \sum_{v=1}^{h_s} \tau(i,j)_{sv} log\left(\frac{1}{f_{sv}^2}\right) \tag{3}$$

where $\tau(i,j)_{sv} = 1$ if $x_{isv} = 1$ and $x_{jsv} = 1$, $\tau(i,j)_{sv} = 0$ otherwise. $\sum_{s=1}^{k} \sum_{v=1}^{h_s} \tau(i,j)_{sv} = \alpha$. Expression (3) is equal to zero iff u_i and u_j do not share any *positive match*. However, it is not a similarity index since the condition $E_{ii} = E_{jj} = 1$ for $i, j = 1, \ldots, n$ is not satisfied. A general expression for a similarity index based on (3) is:

$$S2_{ij} = \frac{E_{ij}}{E_{ij} + F_{ij}} \tag{4}$$

with $F_{ij} \geq 0$ depending on the number of *disagreements* in u_i and u_j. We may define F_{ij} in different ways:

1. $F_{ij} = \sum_{s=1}^{k} \sum_{v=1}^{h_s} \sum_{t=1}^{h_s} \phi(i,j)_{svt} log\left(\frac{1}{f_{sv} f_{st}}\right)$ with $v \neq t$ and
   - $\phi(i,j)_{svt} = 1$ if $\begin{cases} x_{isv} = 1, x_{jsv} = 0 \ \& \ x_{ist} = 0, x_{jst} = 1 \\ x_{isv} = 0, x_{jsv} = 1 \ \& \ x_{ist} = 1, x_{jst} = 0 \end{cases}$
   - $\phi(i,j)_{svt} = 0$ otherwise.

   $\sum_{s=1}^{k} \sum_{v=1}^{h_s} \sum_{t=1}^{h_s} \phi(i,j)_{svt} = \beta + \gamma$
2. $F_{ij} = \sum_{s=1}^{k} \phi(i,j)_s log\left(\frac{1}{1-\sum_{v=1}^{h_s} f_{sv}^2}\right)$ and
3. $F_{ij} = \sum_{s=1}^{k} \phi(i,j)_s \sum_{v=1}^{h_s} f_{sv} log\left(\frac{1}{f_{sv}^2}\right)$ with
   - $\phi(i,j)_s = 1$ if $x_{isv} = 1, x_{jst} = 1$ and $x_{ist} = 0, x_{jsv} = 0$ $v \neq t$
   - $\phi(i,j)_s = 0$ otherwise.

   $\sum_{s=1}^{k} \phi(i,j)_s = 0.5(\beta + \gamma)$

In the first case, F_{ij} is equal to the information content of the specific pairwise *disagreements* in u_i and u_j and expression (4) becomes:

$$\frac{\sum_{s=1}^{k} \sum_{v=1}^{h_s} \tau(i,j)_{sv} log\left(\frac{1}{f_{sv}^2}\right)}{\sum_{s=1}^{k} \sum_{v=1}^{h_s} \tau(i,j)_{sv} log\left(\frac{1}{f_{sv}^2}\right) + \sum_{s=1}^{k} \sum_{v=1}^{h_s} \sum_{t=1}^{h_s} \phi(i,j)_{svt} log\left(\frac{1}{f_{sv} f_{st}}\right)} \tag{5}$$

Coefficient (5) is equal to (1) when $h_s = q$ and $f_{sv} = 1/q$, for $s = 1, \ldots, k$ and $v = 1, \ldots, q$. However, for two couples of subjects having the same pairwise *positive matches* but different pairwise *disagreements* it may assume a different value. Given the categorical nature of the data, the evaluation of the similarity should depend on the number of *disagreements* but not on the dummy variables in which they are present. In the second case, F_{ij} is the information content of any *dissimilarity*, without considering the specific dummy variable in which the *dissimilarity* is present. For attribute X_s, the probability of a *positive match* is $\sum_{v=1}^{h_s} f_{sv}^2$ and thus the probability of a *dissimilarity* is its complement $1 - \sum_{v=1}^{h_s} f_{sv}^2$. Considering this second expression, index (4) becomes:

$$\frac{\sum_{s=1}^{k} \sum_{v=1}^{h_s} \tau(i,j)_{sv} log\left(\frac{1}{f_{sv}^2}\right)}{\sum_{s=1}^{k} \sum_{v=1}^{h_s} \tau(i,j)_{sv} log\left(\frac{1}{f_{sv}^2}\right) + \sum_{s=1}^{k} \phi(i,j)_s log\left(\frac{1}{1-\sum_{v=1}^{h_s} f_{sv}^2}\right)} \tag{6}$$

Index (6) assumes the same numerical value for every pair of subjects having the same *positive matches*, regardless of the specific dummy variables in which the *disagreements* are present. In the trivial case in which $h_s = q$ and $f_{sv} = 1/q$ for $s = 1, \ldots, k$ and $v = 1, \ldots, q$, (6) becomes equal to (2) when $w = 0.5 log\frac{q}{q-1} / log(q^2)$. In the third case, F_{ij} is equal to the average of the information content of the pairwise *positive matches* in variables which have *disagreements* in u_i and u_j. Thus, F_{ij} may be perceived as the average loss in the information content due to the lack of

positive matches in u_i and u_j. For each attribute, the information content of a pairwise *positive match* in modality X_{sv} is $log(1/f_{sv}^2)$. The average of the information content is then $\sum_{v=1}^{h_s} f_{sv} log(1/f_{sv}^2)$. This quantity is also the average loss of the information content due to the lack of a *positive match* in X_s. With this expression:

$$S2_{ij} = \frac{\sum_{s=1}^{k} \sum_{v=1}^{h_s} \tau(i,j)_{sv} log\left(\frac{1}{f_{sv}^2}\right)}{\sum_{s=1}^{k} \sum_{v=1}^{h_s} \tau(i,j)_{sv} log\left(\frac{1}{f_{sv}^2}\right) + \sum_{s=1}^{k} \phi(i,j)_s \sum_{v=1}^{h_s} f_{sv} log\left(\frac{1}{f_{sv}^2}\right)} \tag{7}$$

Index (7) satisfies the following properties: (1) $S2_{ij} = 0$ iff u_i and u_j do not share any *positive match* and $S2_{ij} = 1$ iff u_i and u_j share a *positive match* in each attribute. (2) It is invariant to any permutation of the *disagreements*, provided that the *disagreements* are on the same attributes. (3) In the trivial case in which $h_s = q$ and $f_{sv} = 1/q$, for $s = 1, \ldots, k$, $v = 1, \ldots, q$, it becomes equal to (2) with $w = 0.5$. This last index, when $h_s = 2$, $s = 1, \ldots, k$, is equivalent to the Sokal-Michener measure. In order to take into account the possible association between variables, the information content E_{ij} of the pairwise *agreements* between two subjects should be defined in term of frequencies of specific 'sequences' of *agreements*. Let c_{ij} be the sequence of ones corresponding to pairwise *agreements* in u_i and u_j and $fr(c_{ij})$ be the relative frequency, in the sample, of observations holding the sequence c_{ij}. Consider, for example, three attributes having three, three and two categories, respectively. If the profile vectors of the dummy variables in u_i and u_j are $\mathbf{x}_i' = [001\ 100\ 01]$ and $\mathbf{x}_j' = [001\ 100\ 10]$, $fr(c_{ij})$ is the relative frequency, in the sample, of observations having the third category in the first attribute and the first category in the second attribute. Assume $fr(c_{ij})^2$ to be the probability, in a Bernoulli trial, of sampling two subjects with sequence c_{ij}. The information content of the sequence of *agreements* in u_i and u_j is: $L_{ij} = -2log\left(fr(c_{ij})\right)$ with the convention $L_{ij} = 0$ if u_i and u_j do not have any *positive match*. We normalize L_{ij} and we introduce the new similarity index:

$$S3_{ij} = L_{ij}/(L_{ij} + M(L_{ij})) \tag{8}$$

where $M(L_{ij})$ is the average information content of possible *agreements* in categories which have a *dissimilarity* in u_i and u_j. $M(L_{ij})$ may be thought of as the average loss in the information content due to the lack of pairwise *agreements*. Let us consider a number of *disagreements* equal to g, with $1 \leq g < k$. For $g = 0$, we introduce the convention $S3_{ij} = 0$. For $g = k$, $S3_{ij} = 1$. The count of observations having the same sequence of categories is $m_{ij} = n \times fr(c_{ij})$. In the sub sample of these m_{ij} observations, we determine the relative frequencies of each particular sequence of *agreements* for the remaining $(k - g)$ attributes having a *disagreements* in u_i and u_j. Since the number of categories in the sth attribute is h_s, the count of all possible sequences of *agreements* is $p = \prod_{s=1}^{k-g} h_s$ where the product is extended to the attributes having a *disagreement* in u_i and u_j. Let

– $c(ij)_t$ be the tth sequence of *agreements* among the $(k-g)$ attributes having a *disagreement* in u_i and u_j, $t = 1, \ldots, p$.
– $fr(c(ij)_t)$ be the relative frequency of the sequence $c(ij)_t$ in the sub sample of the m_{ij} subjects having the sequence c_{ij}.

The information content of $c(ij)_t$ is $-2log\,(fr(c(ij)_t))$. The average information content of the sequences $c(ij)_t$ is $M(L_{ij}) = \sum_{t=1}^{p} fr(c(ij)_t) - 2log\,(fr(c(ij)_t))$. With this expression, after algebraic simplification, index $S3_{ij}$ can be written as follows:

$$S3_{ij} = \frac{log\left(fr(c_{ij})\right)}{log\left(fr(c_{ij})\right) + \sum_{t=1}^{p} fr(c(ij)_t) log\,(fr(c(ij)_t))} \tag{9}$$

Using the conventions previously introduced, $S3_{ij} = 0$ iff u_i and u_j do not have any *agreement*, $S3_{ij} = 1$ iff u_i and u_j have a *positive match* in all k attributes.

3 Applications in Marketing Research and Discussion

In this section we try to gain insights into the characteristics of $S1$, $S2$ and $S3$ through applications in marketing research. All the analyses are performed in the Matlab environment (programs are available upon request). We also compare the proposed indexes with two popular similarity measures for polytomous variables: the Jaccard index (Sd) and the Hamming similarity index (Sh) (the function $I\{\cdot\} \in \{0, 1\}$ indicates the truth of its argument):

$$Sd_{ij} = \frac{I\{x_{is} = x_{js}, x_{is} \neq 0, x_{js} \neq 0\}}{I\{x_{is} \neq 0, x_{js} \neq 0\}} \qquad s = 1, \ldots, k \tag{10}$$

$$Sh_{ij} = \frac{I\{x_{is} = x_{js}\}}{k} \qquad s = 1, \ldots, k \tag{11}$$

While Sh is independent on the coding scheme, Sd depends on the code. For binary variables indicating the presence or the absence of a feature, we use the code `1` for the presence and `0` for the absence (so that the number of pairwise absences is not counted in Sd). For dichotomous variables in which the categories do not reflect the presence or the absence of a feature and for polytomous variables we use the code `1, 2, ..., s`. It is worth to highlight that Sd differs from $S1$ both in case of all polytomous variables and in case of mixed dichotomous and polytomous variables. We refer to Boriah et al. (2008) for a comparative study of the performances of a variety of similarity measures. The first data set consists of $k = 37$ observed features (technical specifications) of $n = 100$ satellite navigators. Seven variables have three categories and the other 30 variables are binary attributes (presence or absence). Some features, like a CD player, are very rare. Some other features, like a touch screen and a Gps system, are very common (see Table 1). Among the 666

Table 1 Relative frequencies of technical features in the satellite navigators

Attribute	fr	Attribute	fr	Attribute	fr
External slot	0.94	Slot expansion	0.92	Hard Disk	0.07
Mp3	0.58	Audio book	0.17	Picture viewer	0.61
In-built speaker	0.82	Camera	0.24	Automatic router	0.90
Autovelox	0.72	Bluetooth	0.44	Multimedia card	0.43
Touch screen	0.95	Gps	0.93	Internet connection	0.11
Cd player	0.02	Dvd	0.02	Tmb	0.60
Vocal control	0.05	Phone	0.18	Memory stick	0.11
Vocal warnings	0.91	In-built hard disk	0.27	Usb	0.76
Earpiece socket	0.91	Fm tuner	0.11	Tv tuner	0.16
IPod device	0.05	In-built antenna	0.97	High Memory	0.05

pairwise associations measured by the χ^2 statistics, there are 178 values leading to the rejection of the null hypothesis of independency between variables for $\alpha = 0.05$ and 110 for $\alpha = 0.01$. To analyze the behavior of the indexes, we consider the objects `Kenwood Dnx 7,200` (u_1) and `LG Lan 9,600 R` (u_2). They share the same category in 18 attributes (14 dichotomous and 4 polytomous). They both have a DVD and a CD player and the most rare category in two of the polytomous variables. The relative frequencies in the sample of the 18 categories shared by the objects are: 0.02, 0.24, 0.22, 0.03, 0.92 0.83, 0.61, 0.97, 0.91, 0.44, 0.95, 0.07, 0.02, 0.02, 0.6, 0.82 0.57, 0.99. The values of the similarity indexes are:

$$Sd_{12} = 0.38 \quad Sh_{12} = 0.48 \quad S1_{12} = 0.35 \quad S2_{12} = 0.69 \quad S3_{12} = 0.85$$

The Objects `Alpine Pmdb 100 P Blackbird` (u_3) and `Nokia E 61` (u_4) also have the same categories in 18 features (14 dichotomous and 4 polytomous). The relative frequencies, in the sample, of the 18 shared categories are: 0.92, 0.71, 0.75, 0.92, 0.91, 0.61, 0.91, 0.24, 0.9, 0.44, 0.93, 0.98, 0.98, 0.6, 0.95, 0.95, 0.82, 0.99. Since the number of *positive matches* and *negative matches* in the binary variables are also equal, the degree of similarity evaluated by Sd, Sh and $S1$ is identical:

$$Sd_{34} = 0.38 \quad Sh_{34} = 0.48 \quad S1_{34} = 0.35 \quad S2_{34} = 0.47 \quad S3_{34} = 0.70$$

but $S2_{34} < S2_{12}$ due to the pairwise presence of five very rare categories. For `Kenwood Dnx 7,200` (u_1) and `Qteck 9,090` (u_5) the values of the indexes are:

$$Sd_{15} = 0.55 \quad Sh_{15} = 0.59 \quad S1_{15} = 0.42 \quad S2_{15} = 0.85 \quad S3_{15} = 0.85$$

These satellite navigators share 22 categories (20 in binary variables and 2 in polytomous variables). Once again, the difference between $S1_{15}$ and $S2_{15}$ is due to the pairwise presence of the category `Ipod Interface`, which has a relative frequency in the sample equal to 0.05. While Sh and $S1$ are increasing functions of

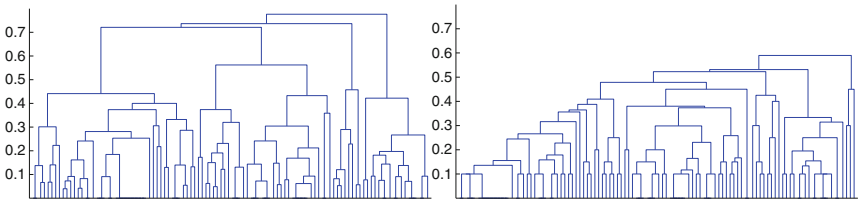

Fig. 1 Dendrograms obtained with the average linkage: in the *left* is used $S2$, in the *right* Sh

the number of shared categories, Sd and $S2$ may also decrease as the number of shared categories increase. Due to the weights, $S2$ is the most variable index. Sd may differ between couples of objects when the number of shared categories is equal but the number of pairwise absences differs. $S2$ may also differ when the number of shared categories and the number of pairwise absences are identical. Index $S3$ is not an increasing function of the number of *positive matches* since agreements in variables which are strongly associated are "penalized".

The second data set consists of $k = 10$ features of $n = 106$ sparkling wines. Five features are dichotomous variables and the others are polytomous. For this data set we obtain partitions with the most common hierarchical methods applied to the complements to one of the indexes. In marketing research, a specific criterion for assessing the performance of similarity measures is the 'segment addressability' suggested by Helsen and Green (1991), related to the degree to which a clustering solution can be explained by variables controlled by marketing managers and helping 'targeting' competitors. In Fig. 1, only two dendrograms are reported, for lack of space. In general, all classifications based on $S2$ and $S3$ readily distinguish four main groups while the other indexes show less ability to provide separation. Moreover, the classification based on $S2$ remains more stable, with respect to the different linkages, than those reached by the other measures. The four groups detected by $S2$ and $S3$ delineate specific segments of products and are easily interpretable also for the size (the smaller group comprehends eight wines and the biggest one 44 wines). These segments are homogeneous with respect to the alcohol content and the sugar level: the R^2 statistics for these variables is always higher in partitions reached with $S2$ and $S3$. Among the features used for classification, the 'taste' and the 'origin' have the rarest categories. In the 4-groups partition with $S3$, wines having the same modality in these two attributes are classified into the same group. With Sh, there is no evidence of a clustering structure in three or four groups: a very small cluster remains isolated until the last aggregation steps.

In conclusion, the major advantage of these indexes is that they are able to handle mixed dichotomous and polytomous variables and the weighted versions are able to give more importance to *agreements* in rare categories. Index $S3$ is designed to take into account the possible associations between variables and there do not appear to be other similarity measures that are directly focused on this goal.

References

Boriah, S., Chandola, V., & Kumar, V. (2008). Similarity measures for categorical data: a comparative evaluation. In *Proceedings of the 8th SIAM international conference on data mining*, Atlanta, pp. 243–254.

Burnaby, T. P. (1970). On a method for caracter weighting a similarity coefficient, employing the concept of information. *Mathematical Geology*, *2*, 25–38.

Gower, J. C., & Legendre, P. (1986). Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification*, *3*, 5–48.

Helsen, K., & Green, P. E. (1991). A computational study of replicated clustering with an application to marketing research. *Decision Science*, *22*, 1124–1141.

MacKay, D. J. C. (2002). *Information theory, inference and learning algorithms*, Cambridge, UK: Cambridge University Press.

Sneath, P. H., & Sokal, R. R. (1973). *Numerical taxonomy*. San Francisco, CA: Freeman.

Warrens, M. J. (2008). Bounds of resemblance measures for binary (presence/absence) variables. *Journal of Classification*, *25*, 195–208.

Zani, S. (1982). Sui criteri di ponderazione negli indici di similarità. In R. Leoni (a cura di) (Ed.), *Alcuni lavori di analisi statistica multivariata (*pp. 187–208). Firenze, Italia: SIS.

Zhang, P., Wang, X., & Song, P. X. (2006). Clustering categorical data based on distance vectors, *Journal of the American Statistical Association*, *101*, 355–367.