MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **83** (2020) 239-260

ISSN 0340 - 6253

Similarity/Diversity Indices on Incidence Matrices Containing Missing Values

Cecile Valsecchi[a], Roberto Todeschini[a,*]

[a]Milano Chemometrics and QSAR Research Group, Department of Environmental Sciences, University of Milano-Bicocca. P.zza della Scienza 1, 20126, Milano, Italy.

(Received July 10, 2019)

Abstract

Quantifying the diversity content of an incidence matrix is challenging in several scientific fields. The existing indices capture diverse facets of diversity and thus comparing their behaviour is not a straightforward task. For example, an application of diversity measures involves ensembles of classifiers which usually in real applications contain missing values. Therefore, we analysed 14 statistics and, after making them comparable and able to deal with missing values, we applied them on more than one hundred incidence matrices in order to examine the relationships among the measures themselves. In particular, we highlighted the importance of the interrow agreement of factors, the general agreement of incident factors, as well as the influence on the indices of the proportion of missing values and matrix dimensions, the sensitivity to missing values, the uniform distribution of entries and the invariance to matrix transposition.

* Corresponding author. E-mail: roberto.todeschini@unimib.it

1. Introduction

Incidence matrices are numerical tables where the presence of a relationship between a row object and a column factor is simply denoted by a numerical value equal to one, while the absence of this relationship is denoted by zero. For example, the activity of a molecule (row) with respect to a pharmacological target (columns) is denoted by 1 if the molecule is active and by 0 is the molecule is not active or the existence of a relationship between an individual (row) with another individual (column) is denoted by 1 while the absence of a relationship is denoted by 0. In the latter case, typically the matrices are squared, and the row and column elements are the same. When the existence of a relationship between row and column elements is not known, missing values are used.

Incidence matrices are used in several fields and some examples of their application are summarized in Table 1, which explicates also the meanings for each dimension and entry of the incidence matrix.

Table 1. Structure and meaning of some incidence matrices.

Incidence matrix type	*Rows/objects*	*Columns/factors*	*value 1*	*value 0*	*Missing values*
Generic data set	objects	properties	yes	no	unknown
Classifier ensemble	objects	classifiers	correct	wrong	unclassified
Ecological data	species	sites	presence	absence	unknown
Molecule activities	molecules	activities	active	not active	unknown
Molecule structure	molecules	fingerprint	presence	absence	-
Databases comparison	objects	databases	included	non-included	-
Graph theory	vertices	edges	belonging	not belonging	-
Tuning matrix	targets	actions	effective	not effective	no concern
Molecule connectivity	atoms	atoms	bonded	not bonded	-
Social network	individuals	individuals	connected	not connected	unknown
Road/train map	cities	cities	connection	not connection	-
Tournament matrix	players	players	win	defeat	no game

In order to numerically characterize the structure of the incidence matrices, several indices able to capture similar or different facets of diversity were proposed [1]. These indices, usually divided into pairwise and non-pairwise, have been used especially in pruning of ensembles of classifiers. In this case, objects (*i.e.* rows) are the samples while factors (*i.e.* columns) are the classifiers. The entry in the *i*-th row and *j*-th column is 1 if the *i*-th sample

is correctly classified by the *j*-th classifier and 0 if it is not. Although the most used indices seem to be inadequate to evaluate the overall performance [2,3], they had been successfully applied in combination with accuracy and margin theory to select classifiers in several areas, including remote sensing, social media semantic analysis and online classifications [4–10].

However, the existing diversity indices are challenging in real applications because of the issue of missing values regarded in the incidence matrix as the *i*-th sample not classified by the *j*-th classifier (*i.e.* NaN value in MATLAB code) which so far were not considered in the formulas. In addition, these indices are often calculated in different ranges and meanings, which make their relationships not immediately intelligible.

We have studied statistics which can measure similarity or diversity of an incidence matrix: three averaged pairwise coefficients (the Yule's Q statistic, the Sokal-Michener similarity and the mutual average difference) and eight non-pairwise measures (the Wave-Edge distance, the Soergel distance, the Kohavi-Wolpert variance, inter-factor agreement, the generalized similarity, the coincident failure similarity, the multivariate correlation index and the average agreement) [1,11–17].

These statistics were made comparable and able to deal with the presence of missing values and finally, were applied to several both patterned and random matrices to investigate their behaviours, correlations and sensitivities to missing values through the analysis of principal components.

2. Materials and methods

2.1 Similarity/diversity indices

In the sections from now on we will refer in the general case of incidence matrix where the columns are factors and the rows are the objects. The procedure can be easily generalized or adapted case by case, *e.g.* following Table 1.

The data matrix **T'** is constituted by *n'* objects and *L'* factors; the entry t'_{ij} of the **T'** matrix is equal to 1 if there is a relationship between the *i*-th object and the *j*-th factor and 0 if there is not a relationship. When no information is available for an object, the entries are missing values denoted as m_{ij} (some special code is used, *e.g.* 'NaN' in MATLAB code).

As first step to deal with missing values, the rows entirely filled by missing values are deleted, *i.e.* the objects not related to any factor; analogously, all the columns entirely filled by missing values are deleted, *i.e.* factors not able to relate to any object.
The reduced data matrix **T** is then defined by n objects and L factors. All the quantities below are defined in terms of these new dimensions, *i.e.*

n = number of rows not entirely filled by missing values
L = number of columns not entirely filled by missing values

We divided the similarity/diversity indices into three groups based on their algorithm. The first group includes the indices calculated as an average of column pairwise statistics such as Yule's Q statistic and Sokal-Michener similarity. The indices deriving from the row sums and thus focused on intra-row differences are collected in the second group which includes the Wave-Edge distance, the Soergel distance, the Kohavi-Wolpert variance, the interfactor agreement, the generalized similarity, the average agreement and the coincident failure similarity. Finally, the multivariate correlation index and the mutual difference are collected in the third group, since they are calculated from the whole matrix.
The similarity/diversity indices are defined below. Note that each index has been rescaled between 0 and 1 for ease of interpretation. Moreover, each diversity measure increases when diversity decreases, excepting mutual average difference and Kohavi-Wolpert variance.

2.1.1 Column pairwise measures

In pairwise measures, firstly the diversity between all pairs (j,k) of factors is calculated computing the so-called contingency table, constituted by the parameters a, b, c and d, where a and d are the number of objects both equal to 1 and both equal to 0, respectively, while b and c are the number of samples having 1 in the j-column and 0 in the k-column and vice versa. Thereafter, the overall diversity measure values are computed as the mean of the pairwise values.

Yule's Q statistic (QY)
For each pair of factors (j,k) the Yule correlation similarity (QY) is calculated by the contingency table as:

$$Q_{jk} = \frac{a \cdot d - b \cdot c}{a \cdot d + b \cdot c} \tag{1}$$

The final index *QY* is calculated as the average value of the pairwise Yule correlation similarity as:

$$QY = \frac{1}{C} \cdot \sum_{j=1}^{L-1} \sum_{k=j+1}^{L} Q_{jk} \qquad -1 \le QY \le +1 \tag{2}$$

The term *C* is the count of valid column pairs, *i.e.* all the cases where $a \cdot d + b \cdot c > 0$; if no missing values are present, all the contributions are valid and $C = \frac{L \cdot (L-1)}{2}$.

Finally, *QY* was rescaled in range [0, 1]:

$$QY' = \frac{QY + 1}{2} \tag{3}$$

Sokal-Michener similarity (SM)

For each pair of factors (*j*,*k*) the Sokal-Michener similarity is calculated by the contingency table as:

$$SM_{jk} = \frac{a + d}{a + b + c + d} \tag{4}$$

The global index is then calculated as:

$$SM = \frac{1}{C} \cdot \sum_{j=1}^{L-1} \sum_{k=j+1}^{L} SM_{jk} \qquad 0 \le SM \le 1 \tag{5}$$

The term *C* is the count of valid column pairs, *i.e.* all the cases where $a + b + c + d > 0$; if no missing values are present, all the contributions are valid, and *C* is calculated as in the previous case.

2.1.2 Indices based on the row sum

Most of the similarity/diversity indices are based on the different distribution of 1 and 0 among the entries of each row. For these statistics the following quantities must be previously defined for each *i*-th row of the **T** matrix with *L* columns:

$$u_i = \sum_{j=1}^{L} t_{ij} \qquad 0 \le u_i \le L \qquad z_i = L - u_i - m_i \qquad 0 \le z_i \le L \tag{6}$$

$u_i + z_i = L_i \le L \qquad L_i + m_i = L$

where m_i is the total number of missing values in the i-th row, *i.e.* denoting the number of factors whose relationship with the selected example is unknown; u_i is the number of 1 in the i-th row while z_i is the number of 0.

Row sum probabilities can thus be defined as the probabilities of u_i or z_i to be equal to a number ranging from 0 to L.

The row sum probability is calculated as:

$$pu_m = \frac{\sum_{i=1}^{n} \delta_{im}}{n} \qquad 0 \le m \le L \qquad \delta_{im} = \begin{cases} 1 & if \quad u_i = m \\ 0 & otherwise \end{cases} \qquad \wedge \qquad \sum_{m=0}^{L} pu_m = 1$$

$$pz_m = \frac{\sum_{i=1}^{n} \delta_{im}}{n} \qquad 0 \le m \le L \qquad \delta_{im} = \begin{cases} 1 & if \quad z_i = m \\ 0 & otherwise \end{cases} \qquad \wedge \qquad \sum_{m=0}^{L} pz_m = 1 \qquad (7)$$

Note that, if no missing values are present, $pu_m = pz_{L-m}$.

Soergel concordance (SOC)

Derived from the Soergel distance [17], it is here defined as a complementary quantity, here called concordance:

$$SOC = \frac{\sum_{i=1}^{n} |u_i - z_i|}{\sum_{i=1}^{n} max(u_i, z_i)} \qquad (8)$$

where u_i and z_i are defined above in equation (6).

SOC takes into account the intra-row differences, *i.e.* how the factors outputs differ for the same i-th object.

Wave-edge concordance (WEC)

Derived from the Wave-edge distance [17], it is here defined as a complementary quantity, here called concordance:

$$WEC = 1 - \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{min\{u_i, z_i\}}{max\{u_i, z_i\}} \qquad (9)$$

As *SOC*, *WEC* explicitly considers the different contributions of u_i and z_i in the i-th row.

Kohavi-Wolpert variance (KWV)

We use the general idea expressed in [1,4] and derived from [13] in the following way:

$$KWV = \frac{4}{n \cdot \bar{L}} \cdot \sum_{i=1}^{n} \frac{u_i \cdot z_i}{L_i} \qquad \bar{L} = \frac{\sum_{i=1}^{n} L_i}{n} \tag{10}$$

$\bar{L}$ is equal to L if there are not missing values, otherwise it considers only the factors which provide information for each object. We multiplied the original formula by 4 in order to rescale the *KWV* between 0 and 1.

Higher the variability of each row, higher the value of *KWV*.

Measure of interfactor agreement (IA)

As it is suggested in [18], the formula for the interfactor agreement (*IA*) takes into account the mean of present relationships ($\bar{u}$), the mean square between objects (*BMS*) and the mean square within objects (*WMS*) defined as:

$$\bar{u} = \frac{1}{n} \cdot \sum_{i=1}^{n} u_i$$

$$BMS = \frac{\sum_{i=1}^{n} (u_i - \bar{u})^2}{n \cdot L} \qquad WMS = \frac{\sum_{i=1}^{n} u_i \cdot z_i}{n \cdot L \cdot (L-1)} \tag{11}$$

From the two previous quantities, *IA* is calculated as:

$$IA = \frac{BMS - WMS}{BMS + (L-1) \cdot WMS} \qquad -\frac{1}{L-1} \le IA \le 1 \tag{12}$$

Finally, *IA* is rescaled in the range [0, 1]:

$$IA' = \frac{IA \cdot (L-1) + 1}{L} \qquad 0 \le IA' \le 1 \tag{13}$$

Generalized similarity (GSu and GSz)

The generalized diversity defined in [14] is here replaced by the complementary quantity, here called generalized similarity:

$$GSu = \frac{\sum_{m=1}^{L}\left(\frac{m}{L}\cdot\frac{(m-1)}{(L-1)}\right)\cdot pu_m}{\sum_{m=1}^{L}\frac{m}{L}\cdot pu_m} \tag{14}$$

$$GSz = \frac{\sum_{m=1}^{L}\left(\frac{m}{L}\cdot\frac{(m-1)}{(L-1)}\right)\cdot pz_m}{\sum_{m=1}^{L}\frac{m}{L}\cdot pz_m} \tag{15}$$

Since we defined two row sum probabilities *pu* and *pz* (equations (7)), there are also two generalized similarity indices. *GSu* considers the probability of 1 while *GSz* that of 0. Clearly the difference between the two indices becomes significant in presence of missing values.

Coincidence failure similarity (CFSu and CFSz)

Similarly to the generalized diversity, two complementary quantities are defined as coincident failure similarity indices from [14]:

$$CFDu = \begin{cases} 0 & if\ \ pu_0 = 1 \\ \frac{1}{1-pu_0}\cdot\sum_{m=1}^{L}\left(\frac{L-m}{L-1}\right)\cdot pu_m & if\ \ pu_0 < 1 \end{cases} \tag{16}$$

$$CFDz = \begin{cases} 0 & if\ \ pz_0 = 1 \\ \frac{1}{1-pz_0}\cdot\sum_{m=1}^{L}\left(\frac{L-m}{L-1}\right)\cdot pz_m & if\ \ pz_0 < 1 \end{cases} \tag{17}$$

The two diversity indices defined above are then converted into similarities:

$$CFSu = 1 - CFDu \tag{18}$$

$$CFSz = 1 - CFDz \tag{19}$$

Average agreement (AAu and AAum)

As an index of the degree of incidence, we calculated also the average agreement as the average of the proportion of 1 in the rows, once without considering missing values (*AAu*) and once taking them into account (*AAum*).

$$AAu = \frac{\sum_{i=1}^{n}\left(\frac{u_i}{L_i}\right)}{n} \tag{20}$$

$$AAum = \frac{\sum_{i=1}^{n}\left(\frac{u_i}{L_i + m_i}\right)}{n} \qquad (21)$$

In presence of missing values, by definition, $AAum < AAu$, otherwise $AAum = AAu$.
For example, in case of classifiers ensemble each *i*-th row is the proportion of correct classifications, while, in the case of active molecules each *i*-th row is the proportion in which a molecule is active with respect to the *L* different targets.

2.1.3 Global indices

We considered the two similarity/diversity indices described below which take into account the whole matrix content, not taking into account explicitly difference between rows and columns.

Multivariate correlation index (K)

To evaluate the correlation content into the matrix, we considered the multivariate correlation index which was proposed for the evaluation of the global correlation of a dataset [15,16]. After substituting the missing value codes by a numerical value, the total quantity of correlation is estimated from the eigenvalue distribution obtained from the eigenvalue decomposition of the corresponding matrix **M**, calculated by symmetrisation of **T**:

$$t_{ij} = \begin{cases} t_{ij} & if\ t_{ij} \neq m_{ij} \\ 0.25 & if\ t_{ij} = m_{ij} \end{cases} \qquad \begin{array}{ll} \mathbf{M} = \mathbf{T}^{\mathbf{T}} \cdot \mathbf{T} & if\ min(n,L) = L \\ \mathbf{M} = \mathbf{T} \cdot \mathbf{T}^{\mathbf{T}} & if\ min(n,L) = n \end{array}$$

$$K = \frac{\sum_{k=1}^{p}\left|\frac{\lambda_k}{\sum_{k=1}^{p}\lambda_k} - \frac{1}{p}\right|}{\frac{2\cdot(p-1)}{p}} \qquad (22)$$

where λ are the eigenvalues of the matrix **M** and *p* the minimum rank between *n* and *L*. The assumed empirical value of 0.25 is chosen in such a way to give $K \cong 0.5$ for total random matrices.

Mutual average difference (Δ)

The mutual average difference (Δ) for the *i*-th row takes into account the average of the modules of the differences between each pair of different entries in the row (excluding missing values) as:

$$\Delta_i = \frac{1}{C} \cdot \sum_{j=1}^{L-1} \sum_{k=j+1}^{L} \left| t_{ij} - t_{ik} \right| \qquad t_{ij} = \{0,1\} \wedge t_{ik} = \{0,1\} \tag{23}$$

The term *C* is the count of valid differences; if no missing values are present, all the differences are valid, and *C* is calculated as in the previous cases.

The final index is calculated as the average over all the objects as:

$$\Delta = \frac{\sum_{i=1}^{n} \Delta_i}{n} \tag{24}$$

2.1.4 Numeric example

To better understand the calculation of the indices we provided an example using the incidence matrix **T** reported in Table 2 where the last two columns collect the *i*-th row sum of 1 (u_i) and 0 (z_i).

In this case, $n = 6$ and $L = 4$.

Table 2. Example of incidence matrix.

T	**F1**	**F2**	**F3**	**F4**	***u***	***z***
O1	1	1	1	0	3	1
O2	1	0	1	1	3	1
O3	0	1	0	0	1	3
O4	1	1	1	1	4	0
O5	1	0	0	1	2	2
O6	0	0	1	0	1	3

For the pairwise indices, six comparisons between each pair of the four factors are carried out. For example, Table 3 reports the parameters *a*, *b*, *c* and *d* considering factor F1 and factor F2.

Table 3. Example of parameters needed to calculate pairwise statistics. The case of the pair of factors F1 and F2 is reported.

Q_{12}	**F1 = 1**	**F1 = 0**
F2 = 1	a = 2	b = 1
F2 = 0	c = 2	d = 1

In this case,

$$Q_{12} = \frac{2 \cdot 1 - 1 \cdot 2}{2 \cdot 1 + 1 \cdot 2} = 0$$

which denotes a middle value of similarity, since Q_{jk} ranges from -1 to 1.
The global *QY* index is calculated and rescaled between 0 and 1 as it follows.

$$QY = \frac{2}{L \cdot (L-1)} \cdot (Q_{12} + Q_{13} + Q_{14} + Q_{23} + Q_{24} + Q_{34}) = \frac{1}{6} \cdot (0 + 0.5 + 1 + 0 - 0.6 + 0) = 0.15$$

$$QY' = \frac{QY + 1}{2} = 0.58$$

The parameters *a*, *b*, *c* and *d* calculated for each pair of factors are used also to calculate *SM* as:

$$SM = \frac{2}{L \cdot (L-1)} \cdot (SM_{12} + SM_{13} + SM_{14} + SM_{23} + SM_{24} + SM_{34}) = 0.56$$

The statistics derived from the row sum of 0 and 1 (*i.e.* last two columns of Table 2) are reported below.

$$SOC = \frac{|3-1| + |3-1| + |1-3| + |4-0| + |2-2| + |1-3|}{3+3+3+4+2+3} = 0.67$$

$$WEC = 1 - \frac{1}{6} \cdot \left(\frac{1}{3} + \frac{1}{3} + \frac{1}{3} + 0 + \frac{2}{2} + \frac{1}{3} \right) = 0.61$$

$$KWV = \frac{4}{6 \cdot 4} \cdot \left(\frac{3}{4} + \frac{3}{4} + \frac{3}{4} + 0 + \frac{4}{4} + \frac{3}{4} \right) = 0.67$$

For *IA* it is necessary to calculate first $\bar{u}$, *BMS* and *WMS* as it follows:

$$\bar{u} = \frac{1}{6} \cdot (3+3+1+4+2+1) = 2.33$$

$$BMS = \frac{(3-2.33)^2 + (3-2.33)^2 + (1-2.33)^2 + (4-2.33)^2 + (2-2.33)^2 + (1-2.33)^2}{6 \cdot 4} = 0.31$$

$$WMS = \frac{3 \cdot 1 + 3 \cdot 1 + 1 \cdot 3 + 4 \cdot 0 + 2 \cdot 2 + 1 \cdot 3}{6 \cdot 4 \cdot (4-1)} = 0.22$$

Then, *IA* is computed and rescaled:

$$IA = \frac{0.31 - 0{,}22}{0.31 + (4-1) \cdot 0.22} = 0.09 \qquad IA' = \frac{0.09 \cdot (4-1) + 1}{4} = 0.32$$

To calculate *GSu*, *GSz*, *CFSu* and *CFSz* the row sum probability is needed. Table 4 summarises these values for the matrix **T**.

Table 4. Summary of the row probabilities for the matrix T defined in Table 2.

	m = 0	*m = 1*	*m = 2*	*m = 3*	*m = 4*
pu_m	0/6	2/6	1/6	2/6	1/6
pz_m	1/6	2/6	1/6	2/6	0/6

$$GSu = \frac{\left[\frac{1}{4} \cdot \left(\frac{1-1}{4-1}\right)\right] \cdot pu_1 + \left[\frac{2}{4} \cdot \left(\frac{2-1}{4-1}\right)\right] \cdot pu_2 + \left[\frac{3}{4} \cdot \left(\frac{3-1}{4-1}\right)\right] \cdot pu_3 + \left[\frac{4}{4} \cdot \left(\frac{4-1}{4-1}\right)\right] \cdot pu_4}{\frac{1}{4} \cdot pu_1 + \frac{2}{4} \cdot pu_2 + \frac{3}{4} \cdot pu_3 + \frac{4}{4} \cdot pu_4} = 0.62$$

$$GSz = \frac{\left[\frac{1}{4} \cdot \left(\frac{1-1}{4-1}\right)\right] \cdot pz_1 + \left[\frac{2}{4} \cdot \left(\frac{2-1}{4-1}\right)\right] \cdot pz_2 + \left[\frac{3}{4} \cdot \left(\frac{3-1}{4-1}\right)\right] \cdot pz_3 + \left[\frac{4}{4} \cdot \left(\frac{4-1}{4-1}\right)\right] \cdot pz_4}{\frac{1}{4} \cdot pz_1 + \frac{2}{4} \cdot pz_2 + \frac{3}{4} \cdot pz_3 + \frac{4}{4} \cdot pz_4} = 0.47$$

$$CFDu = \frac{1}{1 - pu_0} \cdot \left[\left(\frac{4-1}{4-1}\right) \cdot pu_1 + \left(\frac{4-2}{4-1}\right) \cdot pu_2 + \left(\frac{4-3}{4-1}\right) \cdot pu_3 + \left(\frac{4-4}{4-1}\right) \cdot pu_4\right] = 0.56$$

$$CFSu = 1 - 0.56 = 0.44$$

$$CFDz = \frac{1}{1 - pz_0} \cdot \left[\left(\frac{4-1}{4-1}\right) \cdot pz_1 + \left(\frac{4-2}{4-1}\right) \cdot pz_2 + \left(\frac{4-3}{4-1}\right) \cdot pz_3 + \left(\frac{4-4}{4-1}\right) \cdot pz_4\right] = 0.67$$

$$CFSz = 1 - 0.67 = 0.33$$

The multivariate correlation index *K* takes into account the eigenvalues λ_j calculated on the matrix $\mathbf{M} = \mathbf{T}^{T} \cdot \mathbf{T}$, since the minimum rank is *L* = 4. The four eigenvalues are: 10.274, 2.189, 1.253, and 0.284.

If we define the sum of the eigenvalues as $\Lambda = \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 14$, *K* can be calculated as:

$$K = \frac{\left|\frac{10.274}{14} - \frac{1}{4}\right| + \left|\frac{2.189}{14} - \frac{1}{4}\right| + \left|\frac{1.253}{14} - \frac{1}{4}\right| + \left|\frac{0.284}{14} - \frac{1}{4}\right|}{\frac{2 \cdot (4-1)}{4}} = 0.65$$

Finally, since **T** does not contain missing values, the average agreement is equal to:

$$AAu = AAum = \frac{\frac{3}{4} + \frac{3}{4} + \frac{1}{4} + \frac{4}{4} + \frac{2}{4} + \frac{1}{4}}{6} = 0.58$$

We can conclude that **T** has a diversity content not far from a random matrix.

2.2 Data sets

We generated 129 matrices in order to test the similarity/diversity indices explained above. To study the relationships between the indices we used 19 patterned matrices, *i.e.* matrices with a peculiar conformation not containing any missing value, which can be divided as it follows.

1. Matrices with rows entirely composed by either 1 or 0 with dimensions 1000 × 50 (1000 rows and 50 columns). The proportions of 1 considered are 10% (labelled as Row10 in the figures), 30% (Row30), 50% (Row50), 70% (Row70) and 100% (Row100). The latter is a matrix entirely composed by 1.
2. Diagonal matrices. Two squared 100 × 100 matrices, one composed by 0 with the main diagonal entries equal to 1 (Diag1) and the other composed by 1 with the main diagonal entries equal to 0 (Diag0).
3. Matrices derived from bi- and tri-diagonal. Two 50 ×10 matrices composed by the vertical concatenation of five 10 × 10 banded matrices with one-entries along the diagonal above or the diagonal below the main diagonal, in one case the main diagonal is composed by 0 (Bidiag1) and in the other by 1 (Tridiag1). Analogously,

two 50 × 10 composed by the vertical concatenation of five 10 × 10 banded matrices with 0 entries along the diagonal above or the diagonal below the main diagonal and 1 elsewhere, in one case the main diagonal is composed by 1 (Bidiag0) and in the other by 0 (Tridiag0).

4. Matrices derived from orthogonal matrices. Five matrices with dimensions 100 × 10 were created. One is made by a vertical concatenation of ten sub-matrices 10 × 10, the first of which is entirely composed by 0 excepting for the first column entirely filled with 1, the second sub-matrix indeed is composed by 0 excepting the second column entirely filled with 1, and so on to form Orth1_1. Orth1_2 follows the same scheme but considering each time two columns filled with 1 instead of one. This procedure was repeated reversing 0 and 1 to obtain Orth0_1 and Orth0_2. Finally, Orth5 is composed by the vertical concatenation of two 50 × 10 sub-matrices, the first with five columns of 1 and the other five of 0, while the second with 5 columns of 0 followed by 5 columns of 1.
5. A 1000 × 50 (Col50) and a 50 × 2 (TwoCol) matrix which alternate columns of 1 and columns of 0.
6. A 100 × 100 chequered matrix where 1 and 0 are distributed like white and black on a chessboard (Chess).

In order to study the behaviour of the different indices in relation both to an increase percentage of missing values and to an increase dimensions, we created 10 matrices with random distributions of 1 and 0 (100 × 10, 1000 × 10, 10000 × 10, 100 × 50, 1000 × 50, 10000 × 50, 1000 × 100, 50 × 50, 100 × 100, 50 × 100). For each of these matrices we formed other ten matrices with growing percentage of missing values from 5% to 50% in random positions, for a total of 110 matrices.

3. Results and discussion

The defined data sets were analysed by means of PCA with the aim to understand the different behaviours of the studied indices. Similar results were gained with a correspondence analysis. The numerical results are collected in Tables S1, S2 and S3 in the supplementary material. First, the 19 patterned data sets were examined; in Fig. 1 the loading plots of the first 3 PCs are shown (the indices), while in Fig. 2 the corresponding

score plots (the datasets) are shown. In the latter plots, representative images of the patterned matrices are reported where black and grey stand for 0 and 1, respectively.

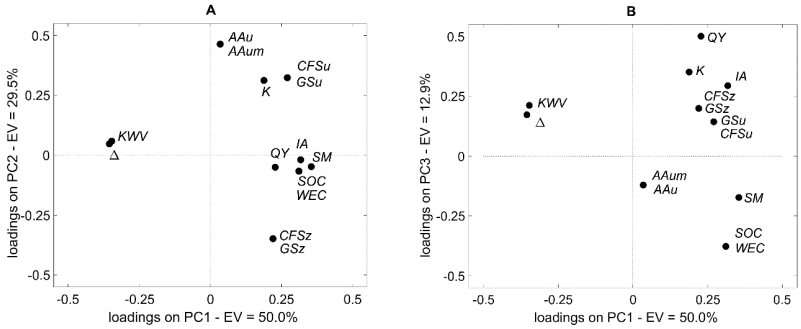

Figure 1. Loading plots PC1 *vs* PC2 (A) and PC1 *vs* PC3 (B) for the 19 patterned matrices. The total explained variance is 92.5%. See Table S1 for numerical results.

As expected, analysing PC1, all the indices indicate the maximum similarity (agreement) when an object is related to all factors, except for *AAum* and *AAu*. In our exemplificative matrices, *QY*, *SM*, *IA*, *WEC*, SOC, *GSz*, *GSu*, *CFSz*, *CFSu* and *K* are equal to 1 and Δ and *KWV* are equal to 0 for Row10, Row30, Row50 and Row70. On the other hand, they designate maximum diversity for TwoCol, namely when half factors are related to the object. By definition, *AAu* and *AAum* are dependent on the row percentages of 1.

The first principal component (PC1) interprets well the interrows agreement, *i.e.* the concordance between factors. Matrices with rows entirely composed by either 1 or 0 indeed have high values on this axis (*e.g.* Row70, Row50, Row30 …), while TwoCol has the lowest PC1 value because in this matrix the two columns (*i.e.* factors) always disagree. High values of *QY*, *SM*, *IA*, *WEC*, *SOC*, *GSz*, *GSu*, *CFSz*, *CFSu* and *K* and low values of Δ and *KWV* corresponds to high concordance between factors.

The second component (PC2) reflects the agreement of incident factors, *i.e.* the percentage of 1, which influence positively *K*, *GSu*, *CFSu* and *AAu* (these last three indices are clearly correlated to the presence of 1) and negatively *GSz* and *CFSz* (which obviously are dependent to the percentage of 0). Along the second PC, the indices *KWV*, Δ, *IA*, *QY*, *SM*, *SOC* and *WEC* do not influence the patterned matrices. The third PC highlights the contrast (13% of variance) among *SOC*, *WEC* and, partially, *SM* (negative scores) with *QY*, *K*, *IA* and, partially, *KWV*, Δ, *GSz*, *GSu*, *CFSz*, *CFSu* (positive scores). The third component (PC3) takes into account the structure of the patterned matrices, independently from the

percentage of ones and zeros, distinguishing checkboard structures from matrices with diagonal square structures.

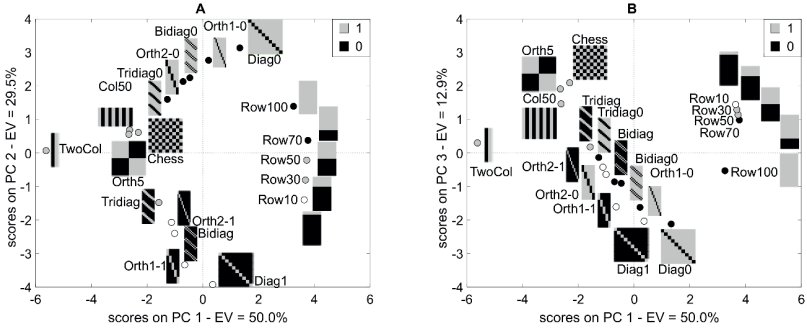

Figure 2. Score plots PC1 *vs* PC2 (A) and PC1 *vs* PC3 (B) for the 19 patterned matrices. The points are filled according to the percentage of 1 in the matrices: black indicates high percentage of 1 (above 70%), grey intermediate (between 25% and 75%) and white low (less than 25%). Near each point there are the label and a pictorial representation of the matrices where black refers to 0 and grey to 1.

The application of the 14 indices to the 110 random matrices lead to the loading and score plots of PC1 *vs* PC2 shown in Fig. 3. In this case, analysing the PC1 loadings, the pairwise indices (*QY*, *SM*), the global indices (*K*, Δ) and *AAu* are not influenced by the percentage of missing values; indeed, their values are always more or less equal to 0.50. High loadings values of PC1 are related to low percentage of missing values, while high values of PC2 are linked to the number of factors; in particular, *QY*, *SM* and *KWV* decrease and *SOC*, *WEC*, *IA*, Δ, *AAu*, *AAum*, *GSu* and *GSz* increase as factors decrease (blue data sets in Figure 3-B).

To better investigate the role of missing values, we took a random binary matrix 100 × 10 and we created other 50 matrices by adding every time 1% of missing values in random positions, always preserving the previous missing values. The PCA loading and score plots are reported in Fig. 4. PC1 reflects PC1 of Figure 3-B (*i.e.* the percentage of missing values), while PC2 seems to reflect the uniform distribution between 1, 0 and missing values. Lower PC2 values indeed were reached for matrices with a proportion of missing values around 33%. In this case, since the distribution of values is random, it is reasonable to assume that more or less there are equal number of 1, 0 and missing values. On the other hand, when the proportion of missing values is low (below 10%) or high (above 40%), the

classes are unbalanced. Δ decreases and *QY*, *SM* and *K* increase when the number of 1, 0 and missing values is unbalanced.

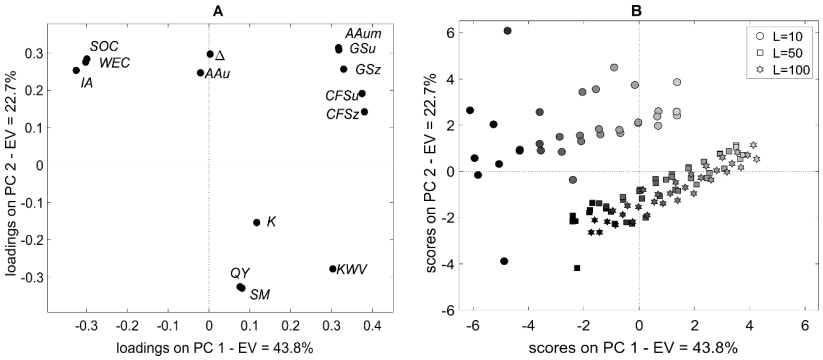

Figure 3. Loading (A) and score (B) plots of PC1 *vs* PC2 for the 110 random matrices. The explained variance is 66.5%. In the score plot different shapes indicate different number of factors in the matrices: circle indicates 10 columns, square 50 and star 100. The points darkness is proportional to the percentage of missing values (black filled = 50% of missing values). See Table S2 for numerical results.

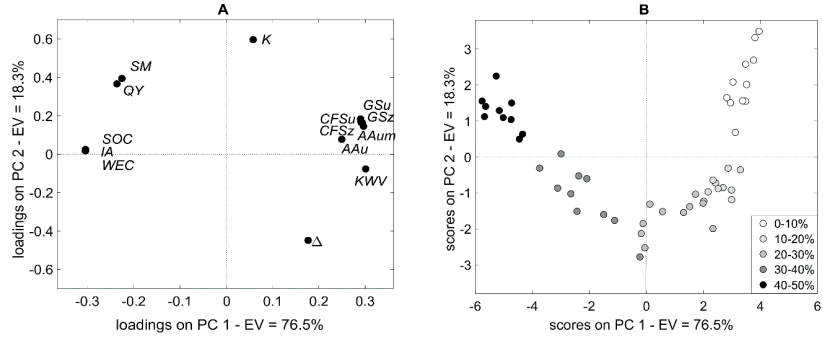

Figure 4. Loading (A) and score plot (B) of PC1 *vs* PC2 for the 51 matrices 100 × 10 derived from the same random matrix with increasing percentage of missing values from 1% to 50% with step of 1. In the score plot the darkness of the points is proportional to the percentage of missing values. The explained variance is 94.8%. See Table S3 for numerical results.

Table 5 summarizes the different behaviours of the 14 diversity indices derived from PCA. First, it can be noted that *WEC* and *SOC* as well as *QY* and *SM* measures express the same information. In column A the trends of the indices for an increasing inter-row agreement is reported.

AAu and *AAum* differ only when we are considering matrices containing missing values. *GSu* and *GSz* as well as *CFSu* and *CFSz*, as it could be expected, have a different meaning only in describing the agreement of incidence factors (column *B*). In column C, it is shown the behaviour of the indices in presence of missing values, while in column D their behaviour for an increasing number of factors.

Columns *E*, *G* and *H* of Table 5 refer to the following sections in which the sensitivity to missing values and the transposition invariance of the indices were quantitatively analysed.

Table 5. *A*: inter-row agreement of factors; *B*: agreement of incident factors; *C*: proportion of missing values; *D*: total number of factors; *E*: sensitivity to missing values; *F*: uniform distribution of {0, 1, missing}; *G*: invariance to matrix transposition (n > L); *H*: invariance to matrix transposition in presence of missing values (n > L).
VH: very high; H: high; L: low; I: invariant; N: no; Y: yes.

ID	*Index*	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*
1	*QY*	H	I	I	H	L	L	Y	N
2	*SM*	H	I	I	H	L	L	Y	Y
3	*WEC*	H	I	H	L	H	I	N	N
4	*SOC*	H	I	H	L	H	I	N	N
5	*KWV*	L	I	L	H	L	I	Y	Y
6	*IA*	H	I	H	L	VH	I	N	N
7	*GSu*	H	H	L	L	H	I/L	Y	Y
8	*GSz*	H	L	L	L	H	I/L	Y	Y
9	*CFSu*	H	H	L	I	H	I/L	Y	Y
10	*CFSz*	H	L	L	I	H	I/L	Y	Y
11	*Δ*	L	I	I	L	L	H	Y	Y
12	*K*	H	H	I	I	L	L	Y	Y
13	*AAu*	H	H	I	L	L	I	Y	Y
14	*AAum*	H	H	L	L	H	I/L	Y	Y

Looking the analyses carried out, the meaning for each index can be roughly expressed as reported in Table 6.

Table 6. General meaning of the studied incidence indices.

ID	*Index*	*Index meaning*
1	*QY*	average factor pairwise agreement including disagreement
2	*SM*	average factor pairwise agreement
3	*WEC*	Wave-edge factor concordance
4	*SOC*	Soergel factor concordance
5	*KWV*	average standardized variance between factors
6	*IA*	inter-factor agreement
7	*GSu*	generalized similarity of incident factors
8	*GSz*	generalized similarity of non-incident factors
9	*CFSu*	generalized similarity of incident factors
10	*CFSz*	generalized similarity of non-incident factors
11	Δ	average mutual difference between factors
12	*K*	global correlation on incident factors
13	*AAu*	global incidence
14	*AAum*	global incidence including missing values

3.1 Sensitivity analysis

The sensitivity analysis of the studied indices was performed by using the matrix with a fixed dimension (100 × 10) and by evaluating the relative difference of the value of an index *x* including a percentage *p* of missing values compared with the value of the same index where no missing values are present. Then, the formula used is:

$$Sn(x,p)=\frac{x(p)-x(0)}{x(0)} \qquad p=1,2,3,,...,50\%$$

In Fig. 5 the sensitivities of all the indices are graphically shown in three different graphs for sake of an easier visualisation. *IA* has the greatest variation in sensitivity and, in particular, its sensitivity increases with the percentage of missing values (Figure 5-C). The sensitivities of *GSz*, *GSu*, *CFSz*, *CFSu* and *AAum* slightly decrease and *WEC* and *SOC* slightly increase, while the other indices seem to be insensitive to the percentage of missing values. A possible explanation is that *IA*, *SOC* and *WEC* by definition take into account the number of 1 and 0 for the *i*-th row without a normalization on L_i (*i.e.* the difference between L and m_i) while for instance *KWV* or *AAu* normalize u_i and z_i by L_i. This consideration is also supported by the difference between *AAu* and *AAum*. In real applications, when dealing

with a large number of missing values it could be better to apply one of the indices which are insensitive to the percentage of missing values.

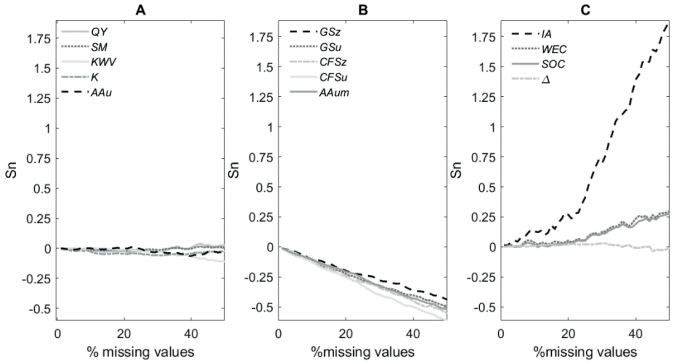

Figure 5. Sensitivity versus percentage of missing values for the 14 diversity indices.

3.2 Matrix transposition

The invariance to matrix transposition was also evaluated with and without the presence of missing values.

It was assumed that the rows of the starting matrix are larger than the columns ($n > L$) and then the transpose matrix contains a number of columns larger than the number of rows ($n < L$).

The invariance to matrix transposition was quantitatively evaluated as:

$$I(x) = \frac{min(x, x^T)}{max(x, x^T)} \qquad 0 \le I(x) \le 1$$

where x is the index and values near/equal to 1 indicate invariance.

For matrices with a different number of columns but without missing values, the invariance was estimated directly by using the quantity $I(x)$,

For matrices with a different number of columns but having also different percentages of missing values comprised between 0% - 50% (step 5%), the invariance was evaluated, for each index, by the coefficient of variation, defined as:

$$CV(x) = \frac{s_x}{\overline{I}(x)}$$

where s_x and $\overline{I}(x)$ are the standard deviation and the average, respectively, calculated from the $I(x)$ values obtained by matrices having the same dimension but eleven different percentages of missing values. Values of CV near to zero indicate invariance.

In a matrix without missing values, only the indices *K*, *AAu* and *AAum* are invariant (Table 5, column *G*).
In matrices containing missing values, all the indices are not constant under transposition, although *K*, *AAu*, *AAum*, *SM*, and Δ show only small variations.
Graphs showing the dependence of *CV* and *I* from *L* are reported in supplementary material (Fig. S1).
As expected, for square random matrices, all the indices show very small variations.

4. Conclusions

The behaviour of the considered 14 measures of diversity with both patterned and random matrices was studied together with their sensitivity to the presence of missing values and invariance to transposition.
The pairs *CFSz* and *GSz*, *CFSu* and *GSu*, *SOC* and *WEC* give, in all the cases both with and without missing values, almost the same information. *CFSz*, *GSz*, *CFSu* and *GSu* are almost invariant to transposition and decrease with the increase of missing values. *CFSz* and *GSz* as *CFSu* and *GSu* differ only because of the opposite dependence of factors agreement. The former pair expresses the generalized similarity of incidence in factors while the latter pair of non-incidence in factors. *SOC* and *WEC* are weakly variant to transposition especially for matrix with lower dimensions (namely $L < 20$) and they weakly increase with the percentage of missing values. *IA* has a similar behaviour of *WEC* and *SOC*, but it is more sensitive both to missing values and to transposition. The column pairwise statistics *QY* and *SM* carry more or less the same information, being defined in a similar way, and are influenced by the dimensions and by the inter-row agreement of factors, but only the second is invariant to matrix transposition. By definition, *AAum* is equal to *AAu*, in the case without missing values, and they are sensible to global incidence of the factors. Since *KWV* and Δ increase with diversity, they are related to low values of the other indices. Furthermore, the column pairwise statistics such as the global indices are sensible to the unbalanced distribution of 0, 1 and missing values in the incidence matrix.
It can be concluded that the choice of the similarity/diversity indices is dependent on the problem under study since they carry different useful information, but, in presence of a significant number of missing values, the indices *QY*, *SM*, *KWV*, *K*, *AAu* and Δ should be preferred being invariant to missing values.

References

[1] L. I. Kuncheva, C. J. Whitaker, Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy, *Mach. Learn.* **51** (2003) 181–207.

[2] M. N. Kapp, R. Sabourin, P. Maupin, An empirical study on diversity measures and margin theory for ensembles of classifiers, *IEEE Xplore* (2007) #9902469.

[3] Y. Bi, The impact of diversity on the accuracy of evidential classifier ensembles, *Int. J. Approx. Reason.* **53** (2012) 584–607.

[4] A. Mellor, S. Boukir, Exploring diversity in ensemble classification: Applications in large area land cover mapping, ISPRS *J. Photogramm. Remote Sens.* **129** (2017) 151–161.

[5] S. Bhatt, B. Minnery, S. Nadella, B. Bullemer, V. Shalin, A. Sheth, *Enhancing Crowd Wisdom Using Measures of Diversity Computed From Social Media Data*, ACM Press, Providence, 2017.

[6] M. A. O. Ahmed, L. Didaci, B. Lavi, G. Fumera, Using diversity for classifier ensemble pruning: an empirical investigation, *Theor. Appl. Informatics* **29** (2018) 25–39.

[7] Q. Dai, R. Ye, Z. Liu, Considering diversity and accuracy simultaneously for ensemble pruning, *Appl. Soft Comput.* **58** (2017) 75–91.

[8] H. Guo, H. Liu, R. Li, C. Wu, Y. Guo, M. Xu, Margin & diversity based ordering ensemble pruning, *Neurocomputing* **275** (2018) 237–246.

[9] I. Visentini, L. Snidaro, G. L. Foresti, Diversity-aware classifier ensemble selection *via* f-score, *Inf. Fusion* **28** (2016) 24–43.

[10] G. D. C. Cavalcanti, L. S. Oliveira, T. J. M. Moura, G. V. Carvalho, Combining diversity measures for ensemble pruning, *Pattern Recogn. Lett.* **74** (2016) 38–45.

[11] G. Giacinto, F. Roli, Design of effective neural network ensembles for image classification purposes, *Image Vis. Comput.* **19** (2001) 699–707.

[12] L. K. Hansen, P. Salamon, Neural network ensembles, *IEEE Trans. Pattern Anal. Mach. Learn.* **12** (1990) 993–1001.

[13] R. Kohavi, D. H. Wolpert, Bias plus variance decomposition for zero-one loss functions, in: L. Saitta (Ed.), *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning*, Morgan Kaufmann Pub., San Francisco, 1996, pp. 275–283.

[14] D. Partridge, W. Krzanowskib, Software diversity: practical statistics for its measurement and exploitation, *Inf. Softw. Technol.* **39** (1996) 707–717.

[15] R. Todeschini, Data correlation, number of significant principal components and shape of molecules. The K correlation index, *Anal. Chim. Acta* **348** (1997) 419–430.

[16] R. Todeschini, V. Consonni, A. Maiocchi, The K correlation index: theory development and its application in chemometrics, *Chemom. Intell. Lab. Syst.* **46** (1999) 13–29.

[17] R. Todeschini, D. Ballabio, V. Consonni, Distances and other dissimilarity measures in chemometrics, in: R A. Meyers (Ed.), *Encyclopedia of Analytical Chemistry: Applications, Theory and Instrumentation*, Wiley, New York, 2015, pp. 1–34.

[18] J. L. Fleiss, B. Levin, M. C. Paik, *Statistical Methods for Rates and Proportions*, Wiley, New York, 2003.