Ecological Informatics

Generalizing resemblance coefficients to accommodate incomplete data

--Manuscript Draft--

Manuscript Number:	ECOINF-D-21-00299R1
Article Type:	Research Paper
Keywords:	Cluster analysis; Distance; Dissimilarity; Missing data; Ordination; Similarity
Corresponding Author:	Janos Podani L. Eötvös University Budapest, HUNGARY
First Author:	Janos Podani
Order of Authors:	Janos Podani
	Dénes Schmera
Abstract:	Large ecological data matrices may be incomplete for various reasons, preventing the use of standard multidimensional scaling (ordination) and cluster analysis packages. Although there exist a few resemblance functions that allow missing scores, there is no theoretical background and software support for most distance and similarity coefficients potentially applied in multivariate data analysis. We provide a general framework for a precise mathematical redefinition of a large set of resemblance functions originally developed for complete data sets with presence-absence (binary) or ratio-scale variables. Included are coefficients which consider double absences in abundance data. Potential problems with the use of these functions are discussed, with the conclusion that incompleteness of data would rarely if ever influence greatly the interpretability of ordinations and classifications. An R function described in the Appendix represents a link to R. We also provide a stand-alone WINDOWS application for users of other computer programs. The new software will allow users of standard data analysis packages to perform multivariate analysis using a wide variety of resemblance coefficients even if the data are incomplete for whatever reason.
Suggested Reviewers:	Carlo Ricotta carlo.ricotta@uniroma1.it Expertise in distance coefficients
	Pierre Legendre pierre.legendre@umontreal.ca Expertise in ecological data analysis in general
	Enrico Feoli feoli@units.it Expertise in numerical analysis of ecological data
Response to Reviewers:	Comments from the Reviewers: Reviewer #1: Authors propose a framework for defining different kind of dissimilarities when missing data are present. The work can be very useful for practitioners, since it is also accompanied by software. It is very well- explained and clear. I have only minor points: *I miss a very well-known and old reference, Dixon (1979) about missing data that also propose to compute distance between two vectors with missing cells and then normalize to compensate for blanks. Furthermore, as stated in that work, it is preferable because of its consistent performance, ease of implementation, and fast running speed. I think that reference should be commented. Dixon, J. K. (1979). Pattern recognition with partly missing data. IEEE Transactions on Systems, Man, and Cybernetics 9 (10), 617-621. This important reference, which we did not know earlier, is added to the paper at several places. More importantly, his “Normal” method for scaling up unbounded

measures has been included in both packages….
the modified text:
“Alternatively, unbounded measures can be scaled up (the “Normal” method of Dixon, 1979; Datta et al., 2018) according to n/Wjk, the ratio of the number of all variables to those known for a given pair of objects. Dixon (1979) found that this operation greatly improved the results, therefore we also offer this option for the Euclidean, and Manhattan metrics and the Faith’s intermediate coefficient. “

The Partial Distance Strategy (PDS) is implemented in a very well-known R package: in daisy function of cluster package.
Yes, we have checked this. In daisy Euclidean distance and Manhattan distance are available only. The scaling up operation is adapted, but thy use the square root of the ratio when multiplying Euclidean distances, while we use the raw ratio instead for all 3 functions concerned (Euclidean, and Manhattan metrics and the Faith’s intermediate coefficient).

*p.7, l.151: "by" is missing in "given by Tamás et al. (2001).
Done

*p. 7, l.157: More works that also defend the idea that directly estimating distances result in more accurate results than calculating distances from an imputed data set is Eirola et al. (2013).
We now cite this paper.

Eirola, E., G. Doquire, M. Verleysen, and A. Lendasse (2013). Distance estimation in numerical data sets with missing values. Information Sciences 240, 115 - 128.

*p.9, l.183: could you, please, give a certain value about what is considered a high missingness ratio?
Of course, there is no objectively defined threshold, so we rephrased the sentence:

“…missing scores / nm) is considered to be high.”

*p.9, l.185 and 186: a small answer is explored in Epifanio (2020): results obtained in an unsupervised statistical learning methodology by using imputation, PDS with multidimensional scaling or other alternatives are compared. They show that using PDS is a very competitive alternative.

I. Epifanio, M.V. Ibáñez and A. Simó (2020) Archetypal Analysis with Missing Data: See All Samples by Looking at a Few Based on Extreme Profiles, The American Statistician, 74(2), 169–183.

This work is now cited here.

*p.13, l. 282: please remove the end bracket.
done

*p.13 , l. 287: please remove the intro space.
corrected

*Although not essential, but I find recommendable to include the R function in an R package in CRAN for being used easier by analysts.

Yes, we plan to do that if the ms is accepted for publication in Ecological Informatics.

Reviewer #2:
This manuscript addresses an important issue of resemblance indices applied in community ecology, namely how to deal with missing values. While modifications of a few resemblance indices had been previously proposed by others, the authors generalize here the formulation of a panoply of indices to incomplete data, filling an important gap in the area.

I would suggest the authors to contact J. Oksanen (if they haven't already) to replace

	the calculation of resemblance indices with missing data in the R package 'vegan' to include this proposal, for function 'vegdist()'. The current implementation of pairwise deletion may give the same result for some indices, but fails with others (e.g. chord). Yes, we think that we would contact vegan authors only if the ms is accepted for publication in Ecological Informatics. I also have a suggestion for the presentation of indices. I present the cross-product Ajj(k) not with Ajk, but in section 3 below Tj(k). Ajj(k) is the square of the vector norm for object j, and could be named Aj(k). In this way, it is shown the chord distance can be calculated from the Euclidean distance, and the chord and hellinger transformations are analogous. The reviewer is right in stating that it is also possible to present Ajj(k) with Tj(k) and not with Ajk. However, we keep the original way of presentation because we can present 15 resemblance coefficients with Ajj(k) and without Tj(k). The introduction of Tj(k) is required only for indices incorporating standardization (section 3.1). We argue that the distinction between Ajk and Ajj(k) is extremely important. The first is the cross-product for two vectors j and k for all variables that are known for both j and k. The second term corresponds to the cross-product (norm) of j with itself based on data that are also known for observation k. In other words, Ajj(k) differs with k, because it would not be logical to use the same cross-product for all comparisons. The index jj(k) expresses this better than j(k). I liked your discussion about the effect of unbalanced ressemblance on multivariate analyses. Your arguments about agglomerative hierarchical clustering seem fine to me, but what will happen with divisive hierarchical clustering, or non-hierarchical (e.g. k-means) clustering? In the latter, negative distances to the centroids could arise. There is a study (Himmelspach, L., & Conrad, S. (2010). Clustering approaches for data with missing values: Comparison and evaluation. 2010 Fifth International Conference on Digital Information Management (ICDIM)) evaluating the performance of k-means and fuzzy c-means with various levels of missing data. Since these strategies do not use direct resemblances between data points we felt that the discussion of this area would be far beyond the scope of our ms. Miquel De Cáceres *****

Comments from the Reviewers:

Reviewer #1: Authors propose a framework for defining different kind of dissimilarities when missing data are present. The work can be very useful for practitioners, since it is also accompanied by software. It is very well- explained and clear.

I have only minor points:

*I miss a very well-known and old reference, Dixon (1979) about missing data that also propose to compute distance between two vectors with missing cells and then normalize to compensate for blanks. Furthermore, as stated in that work, it is preferable because of its consistent performance, ease of implementation, and fast running speed. I think that reference should be commented.

Dixon, J. K. (1979). Pattern recognition with partly missing data. IEEE Transactions on Systems, Man, and Cybernetics 9 (10), 617-621.

This important reference, which we did not know earlier, is added to the paper at several places. More importantly, his "Normal" method for scaling up unbounded measures has been included in both packages….
the modified text:
"Alternatively, unbounded measures can be scaled up (the "Normal" method of Dixon, 1979; Datta et al., 2018) according to n/W_{jk}, the ratio of the number of all variables to those known for a given pair of objects. Dixon (1979) found that this operation greatly improved the results, therefore we also offer this option for the Euclidean, and Manhattan metrics and the Faith's intermediate coefficient. "

The Partial Distance Strategy (PDS) is implemented in a very well-known R package: in daisy function of cluster package.
Yes, we have checked this. In *daisy Euclidean distance and Manhattan distance are available only.* The scaling up operation is adapted, but thy use the square root of the ratio when multiplying Euclidean distances, while we use the raw ratio instead for all 3 functions concerned (Euclidean, and Manhattan metrics and the Faith's intermediate coefficient).

*p.7, l.151: "by" is missing in "given by Tamás et al. (2001).
Done

*p. 7, l.157: More works that also defend the idea that directly estimating distances result in more accurate results than calculating distances from an imputed data set is Eirola et al. (2013).
We now cite this paper.

Eirola, E., G. Doquire, M. Verleysen, and A. Lendasse (2013). Distance estimation in numerical data sets with missing values. Information Sciences 240, 115 - 128.

*p.9, l.183: could you, please, give a certain value about what is considered a high missingness ratio?
Of course, there is no objectively defined threshold, so we rephrased the sentence:

"…missing scores / *nm*) is considered to be high."

*p.9, l.185 and 186: a small answer is explored in Epifanio (2020): results obtained in an unsupervised statistical learning methodology by using imputation, PDS with multidimensional scaling or other alternatives are compared. They show that using PDS is a very competitive alternative.

I. Epifanio, M.V. Ibáñez and A. Simó (2020) Archetypal Analysis with Missing Data: See All Samples by Looking at a Few Based on Extreme Profiles, The American Statistician, 74(2), 169–183.

This work is now cited here.

*p.13, l. 282: please remove the end bracket.
done

*p.13 , l. 287: please remove the intro space.
corrected

*Although not essential, but I find recommendable to include the R function in an R package in CRAN for being used easier by analysts.

Yes, we plan to do that if the ms is accepted for publication in Ecological Informatics.

Reviewer #2:
This manuscript addresses an important issue of resemblance indices applied in community ecology, namely how to deal with missing values. While modifications of a few resemblance indices had been previously proposed by others, the authors generalize here the formulation of a panoply of indices to incomplete data, filling an important gap in the area.

I would suggest the authors to contact J. Oksanen (if they haven't already) to replace the calculation of resemblance indices with missing data in the R package 'vegan' to include this proposal, for function 'vegdist()'. The current implementation of pairwise deletion may give the same result for some indices, but fails with others (e.g. chord).

Yes, we think that we would contact vegan authors only if the ms is accepted for publication in Ecological Informatics.

I also have a suggestion for the presentation of indices. I present the cross-product Ajj(k) not with Ajk, but in section 3 below Tj(k). Ajj(k) is the square of the vector norm for object j, and could be named Aj(k). In this way, it is shown the chord distance can be calculated from the Euclidean distance, and the chord and hellinger transformations are analogous.

The reviewer is right in stating that it is also possible to present Ajj(k) with Tj(k) and not with Ajk. However, we keep the original way of presentation because we can present 15 resemblance coefficients with Ajj(k) and without Tj(k). The introduction of Tj(k) is required only for indices incorporatiooning standardization (section 3.1). We argue that the distinction between A_{jk} and $A_{jj(k)}$ is extremely important. The first is the cross-product for two vectors j and k for all variables that are known for both j and k. -The second term corresponds to the

Formatted: Font color: Light Blue

cross-product (norm) of j with itself based on data that are also known for observation k. In other words, $A_{jj(k)}$ -differs with k, because it would not be logical to use the same cross-product for all comparisons. The index jj(k) expresses this better than j(k).

I liked your discussion about the effect of unbalanced ressemblance on multivariate analyses. Your arguments about agglomerative hierarchical clustering seem fine to me, but what will happen with divisive hierarchical clustering, or non-hierarchical (e.g. k-means) clustering? In the latter, negative distances to the centroids could arise.

There is a study (Himmelspach, L., & Conrad, S. (2010). *Clustering approaches for data with missing values: Comparison and evaluation. 2010 Fifth International Conference on Digital Information Management (ICDIM))* evaluating the performance of k-means and fuzzy c-means with various levels of missing data. Since these strategies do not use direct resemblances between data points we felt that the discussion of this area would be far beyond the scope of our ms.

Miquel De Cáceres

*****

MethodsX (optional)

We invite you to submit a method article alongside your research article. This is an opportunity to get full credit for the time and money spent on developing research methods, and to increase the visibility and impact of your work. If your research article is accepted, we will contact you with instructions on the submission process for your method article to MethodsX. On receipt at MethodsX it will be editorially reviewed and, upon acceptance, published as a separate method article. Your articles will be linked on ScienceDirect.

Please prepare your paper using the MethodsX Guide for Authors: https://www.elsevier.com/journals/methodsx/2215-0161/guide-for-authors (and template available here: https://www.elsevier.com/MethodsX-template) Open access fees apply.

Have questions or need assistance?
For further assistance, please visit our customer service site: http://help.elsevier.com/app/answers/list/p/9435/. Here you can search for solutions on a range of topics, find answers to frequently asked questions, and learn more about Editorial Manager via interactive tutorials. You can also talk 24/5 to our customer support team by phone and 24/7 by live chat and email.

Highlights

- A general framework is provided to redefine resemblance coefficients for incomplete data
- Coefficients that consider double absences in abundance data are included
- An R Funcion and a stand-alone Windows application are provided for potential users

Generalizing resemblance coefficients to accommodate incomplete data

János Podani[a], Dénes Schmera[b]

[a]Department of Plant Systematics, Ecology and Theoretical Biology, Institute of Biology, Eötvös University, Pázmány P. s. 1/C, H-1117 Budapest, Hungary

Corresponding author. E-mail: podani@ludens.elte.hu; ORCID: 0000-0002-1452-1486

[b]Balaton Limnological Research Institute, Klebelsberg K. u. 3, H-8237 Tihany, Hungary

ABSTRACT

Large ecological data matrices may be incomplete for various reasons, preventing the use of standard multidimensional scaling (ordination) and cluster analysis packages. Although there exist a few resemblance functions that allow missing scores, there is no theoretical background and software support for most distance and similarity coefficients potentially applied in multivariate data analysis. We provide a general framework for a precise mathematical redefinition of a large set of resemblance functions originally developed for complete data sets with presence-absence (binary) or ratio-scale variables. Included are coefficients which consider double absences in abundance data. Potential problems with the use of these functions are discussed, with the conclusion that incompleteness of data would rarely if ever influence greatly the interpretability of ordinations and classifications. An R function described in the Appendix represents a link to R. We also provide a stand-alone WINDOWS application for users of other computer programs. The new software will allow users of standard data analysis packages to perform multivariate analysis using a wide variety of resemblance coefficients even if the data are incomplete for whatever reason.

Keywords: Cluster analysis; Distance; Dissimilarity; Missing data; Ordination; Similarity

1. Introduction

Ecological data matrices are often incomplete because some scores are unknown for a variety of reasons. Data entries may be missing due to measurement errors, unavailability or loss of parts of the observations, malfunctioning of recording devices and human mistakes (Sneath and Sokal, 1973, p. 178; Dixon, 1979; Legendre and Legendre, 2012, p. 54). Furthermore,

incompleteness may also be caused by variables that are undefined, illogical and inapplicable to certain objects or observations, for example, seed mass for ferns in a matrix of plant functional traits. These two situations are distinguished as *unstructured* and *structured missingness*, respectively (Chechik et al., 2008, Zhang et al., 2012). Incomplete data have long been a source of nuisance to users of multivariate methods because the absence of even a single value from the data matrix prevents the use of standard ordination and clustering techniques. The vast majority of multivariate methods operate via calculating resemblance (sensu Orloci, 1972, referring in general to any type of distances, dissimilarities, similarities, correlation, association or proximity measures) among the study objects (in multidimensional scaling and cluster analysis) or variables (principal components analysis) which requires full data arrays in most computer implementations currently in use.

There are several methods to circumvent the problem of missingness. As a brute force solution, removing entire rows and/or columns from the data matrix may come to our mind. In this way large amounts of data may be lost, and therefore removals are not recommended. Instead, empty cells in the data array may be filled by estimation or simulation, called *imputation*, an option most often used in ordination studies (Legendre and Legendre, 2012, Dray and Josse, 2015). Although many relatively simple (e.g., k-nearest neighbour, Dixon, 1979) and more sophisticated algorithms ~~have been suggested for this purpose~~ (Nelson et al., 1996; Grung and Manne, 2005; Oba et al., 2005; Stanimirova et al., 2007; Serneels and Verdonck, 2008)~~,~~ have been suggested for this purpose, imputation remains arbitrary for many, and is illogical to use in case of structured missingness. The third option is to estimate distances between observations with missing data from other distance values, rather than from the data (Eirola et al. 2013). Finally, ~~to calculate~~ each resemblance value may be calculated based on all available information, i.e., using the subset of data that are known for both items being compared (Partial Distance Strategy, PDS, Dixon, 1979). This approach was

first taken by Gower (1971) in developing a general function of dissimilarity which is often used in multidimensional scaling and clustering. This measure corresponds to the mean character difference for range-standardized ratio-scale variables and to the complement of the Jaccard index or the simple matching coefficient for presence-absence data. The idea was also adapted subsequently to ~~range-standardized~~ Euclidean distance (Dixon, 1979) and its range-standardized version (Podani, 1980, Wills, 1998), to the Manhattan metric (Wishart, 2003), as well as to a modified product moment correlation coefficient and covariance (Legendre and Legendre, 2012, Dray and Josse, 2015, Podani et al., 2021) which serve as a basis for computing principal component analysis. However, researchers may want to select other formulae from the large arsenal of resemblance coefficients – but those are not yet modified for this purpose and no computer program is available for their calculation. This paper provides a brief theory of generalizing resemblance coefficients to incomplete data and introduces an R function developed for computations. Emphasis is placed on functions applicable to ratio-scale variables, and to presence-absence coefficients.

2. Basic notations

Most of the resemblance coefficients available in the statistical literature apply to ratio-scale variables (Anderberg, 1973) for which all basic arithmetic operations (summation, subtraction, multiplication, and division) are meaningful. First, we shall be concerned with such variables. Let $\mathbf{X} \equiv \{x_{ij}\}$ denote the data matrix with n rows corresponding to variables, features or descriptors (e.g., species, functional traits or morphological characters) and m columns representing objects (e.g., sample units, individuals or other entities of interest). Another array of the same size, $\mathbf{V} \equiv \{v_{ij}\}$ is an indicator matrix in which $v_{ij} = 1$ if x_{ij} is known, and $v_{ij} = 0$ otherwise. Based on these indicator scores, for each variable i and every pair of objects, j and k, we can calculate the weight $w_{ijk} = v_{ij}\, v_{ik}$ which is zero whenever the two objects are incomparable for that variable due to incomplete information and equals to 1

otherwise. To allow mathematical formalism, assume that lacking scores are represented by a negative dummy value in the data matrix, such as –1, depending on computer program implementation.

2.1. A new set of parameters

Although all resemblance coefficients can be formulated directly using denotations of raw scores, indicator values and weights, we feel that for simplicity, for ease of calculations and for generalization purposes a new set of six parameters will be useful. Each of them expresses some relationship of objects *j* and *k* by considering only those variables that are known for both. These parameters will be abbreviated by capital letters subscripted with *j* and *k*, as follows:

Sum of squared differences: $E_{jk} = \sum_i w_{ijk}(x_{ij} - x_{ik})^2$

Sum of absolute differences: $M_{jk} = \sum_i w_{ijk}\left|x_{ij} - x_{ik}\right|$

Sum of differences: $G_{jk} = \sum_i w_{ijk}(x_{ij} - x_{ik})$

Sum of weights: $W_{jk} = \sum_i w_{ijk}$

Sum of minima: $F_{jk} = \sum_i w_{ijk}\min\{x_{ij}, x_{ik}\}$

Sum of maxima: $H_{jk} = \sum_i w_{ijk}\max\{x_{ij}, x_{ik}\}$

Cross product: $A_{jk} = \sum_i w_{ijk}x_{ij}\,x_{ik}$; $A_{jj(k)} = \sum_i w_{ijk}{x_{ij}}^2$

3. Formulae for ratio-scale variables and their presence-absence variants

Using the six parameters described above, many resemblance coefficients can be rewritten to cope with incomplete data sets. These are summarized in Table 1 in their original and

99 modified form as well. If the data matrix contains presence-absences such that $x_{ij} = 1$ stands
100 for presence and $x_{ij} = 0$ for absence of variable *i* in object *j*, then the results produced by these
101 functions will be identical to those provided by coefficients explicitly developed for
102 presence-absence (binary) data. These latter coefficients are written in terms of the 2 × 2
103 contingency table in which *a* stands for the number of mutual presences, *b* is the number of
104 variables (e.g. species) present in object *j* but absent from object *k*, *c* is the number of
105 variables present in object *k* but absent from object *j*, and *d* is the number of variables absent
106 from both *j* and *k*, but present in at least one object in the data set.

107 *3.1. Indices incorporating standardization*

108 Coefficients that do not fit directly the system outlined in Table 1 use some type of
109 standardization over objects or variables. Fortunately, after modifying the original scores
110 using the following auxiliary parameters, the new values, denoted by x'_{ij} may be substituted
111 into an appropriate equation of the table. The auxiliary parameters are given below.

112 Range of variable *i* for all known data: $R_i = \max_j\{v_{ij}x_{ij}\} - \min_j\{v_{ij}x_{ij}\}$

113 Total of object *j* in relation to *k*: $T_{j(k)} = \sum_i w_{ijk}x_{ij}$

114 Variable total: $V_i = \sum_j v_{ij}x_{ij}$

115 As seen, if normalization is over the objects being compared, then we use the weight w_{ijk} to
116 include variables that are known for both objects. That is, object total $T_{j(k)}$ is applicable to this
117 pair only. If standardization is over variables, then we use the weight v_{ij} so that we consider
118 all available data for calculation. Variable total V_i is therefore valid for all pairs of objects.

119 Then, some notable coefficients will correspond to equations in the table as follows:

120 Gower distance (GOW) for ratio-scale variables is eq. (3) with $x'_{ij} = x_{ij}/R_i$

Hellinger distance (HEL) is eq. (1) with $x'_{ij} = \sqrt{x_{ij}/T_{j(k)}}$

Canberra metric (CAN) is eq. (2) with $x'_{ij} = \frac{x_{ij}}{|x_{ij}+x_{ik}|}$ if $|x_{ij} + x_{ik}| > 0$

otherwise $x'_{ij} = 0$

Kulczynski index (KUL) is eq. (12) with $x'_{ij} = \frac{x_{ij}}{(T_{j(k)}+T_{k(j)})/2}$

Chi-square distance (CHI) is eq. (1) with $x'_{ij} = \frac{x_{ij}}{V_i T_{j(k)}}$

Renkonen index (REN) is eq. (12) with $x'_{ij} = \frac{x_{ij}}{T_{j(k)}}$

Note that x'_{ij} and x'_{ik} refer only to the pair *j*, *k* of objects. Furthermore, there are a couple of functions which may be expressed using the new parameters and/or x'_{ij} but do not correspond to any equation in Table 1:

Coefficient of divergence (DIV) is given by $\sqrt{\frac{E_{jk}}{W_{jk}}}$ with $x'_{ij} = \frac{x_{ij}}{|x_{ij}+x_{ik}|}$ if $|x_{ij} + x_{ik}| > 0$

otherwise $x'_{ij} = 0$

Covariance for objects becomes $COV_{jk} = \frac{A_{jk} - \frac{T_{j(k)}T_{k(j)}}{W_{jk}}}{W_{jk}-1}$

Correlation between objects is $COR_{jk} = \frac{A_{jk} - \frac{T_{j(k)}T_{k(j)}}{n}}{\sqrt{\left(A_{jj(k)} - \frac{T^2_{j(k)}}{W_{jk}}\right)\left(A_{kk(j)} - \frac{T^2_{k(j)}}{W_{jk}}\right)}}$

3.2. Backward extensions to ecological abundances

Table 1 shows presence/absence coefficients to which ratio-scale formulae are reduced when the data set contains only 1-s and 0-s. Generalization may be achieved in the reverse direction

as well, with meaningful applications in ecological data analysis. According to Tamás et al. (2001), coefficients for presence-absence data in which parameter *d* is used can be extended to abundance data (i.e. ratio-scale variables). Between two localities or sites *j* and *k*, *a'* corresponds to overlap in abundances and *b'* + *c'* is the symmetric difference – these are directly comparable to the classical 2 × 2 contingency table parameters *a*, and *b* + *c*. Parameter *d*, the number of "double zeros" or double absences may be understood here as the "potential abundance that could be reached in the study area" (Tamás et al., 2001) and is calculated as the sum of differences between the maxima in *j* and *k* and the maxima reached in the entire sample:

Overlap: $a' = F_{jk}$

Difference in favour of site *j*: $b' = \sum_i w_{ijk}(\max\{x_{ij}, x_{ik}\} - x_{ik})$

Difference in favour of site *k*: $c' = \sum_i w_{ijk}(\max\{x_{ij}, x_{ik}\} - x_{ij})$

"Potential" abundance: $d' = \sum_i w_{ijk}\ \left(\max_j\{v_{ij}x_{ij}\} - \max\{x_{ij}, x_{ik}\}\right)$

Total: $n' = a' + b' + c' + d' = \sum_i \max_j\{v_{ij}x_{ij}\}$

These parameters may be substituted into coefficients formerly applied only to the presence-absence case. There are many indices which incorporate the *d* parameter, Table 2 is a list of ten such functions also given by Tamás et al. (2001). The reader may also consult with Goodall (1973), Ludwig and Reynolds (1988), and Kenkel and Booth (1987) for more.

6. Discussion

The new mathematical formalism developed in this paper allows precise redefinition of many resemblance coefficients to be calculated in multivariate analysis of incomplete data sets. A summary of mathematical details is presented in Electronic Supplement 2. The use of these modified forms frees the data analyst from the drastic operation of data reduction and the

arbitrary solution of imputation: each pair of coefficients is calculated based on the maximum amount of information available in the dataset. That is, our approach corresponds to the Partial Distance Strategy formerly suggested for a small subset of distance metrics, now extended to a wide range of coefficients.

We first defined seven parameters, *E*, *M*, *G*, *W*, *F*, *H* and *A* to express various aspects of mutual relationships between two objects to be used in compact mathematical formulae for 15 resemblance coefficients. Most of these equally apply to presence-absence data with binary coding. Three other parameters, *R*, *T*, and *V* are used for standardization, after which further six coefficients become identical to one of the above-mentioned formulae, and three additional ones are defined. The framework may also be used to generalize presence-absence coefficients that count double zeros (*d*) to abundance data. Although we were concerned with ratio-scale and presence-absence variables, the idea presented here can be readily applied to coefficients for nominal and ordinal scale types as well. Further index families may also be modified, such as the coefficients of similarity between ecological sample plots in which species similarities are also accounted for (Ricotta and Pavoine, 2015; Ricotta et al., 2016; Podani et al., 2018).

The present approach implies that pairwise comparisons are based on different numbers of variables, and we can say that the resemblance matrix is therefore "unbalanced". This is unavoidable in case of structured missingness, when the actual values themselves are the best estimates of between-object resemblance. Unstructured missingness may be handled in different ways. If the resemblance function is bounded, such as the chord distance between 0 and $\sqrt{2}$ and the Bray-Curtis dissimilarity between 0 and 1, then it follows that each "true" resemblance value (i.e., if all variables were known) may be either under- or overestimated and there is a chance that deviations compensate one another. If there is no upper bound (e.g. Euclidean distance) then distances calculated from a reduced number of variables will never

overestimate the true value. One must consider this potential systematic bias when selecting a coefficient and we recommend a bounded measure when the missingness ratio (no. of missing scores / *nm*) is considered to be high. Alternatively, unbounded measures can be scaled up (the "Normal" method of Dixon, 1979; Datta et al., 2018) according to n/W_{jk}, the ratio of the number of all variables to those known for ~~unobserved values for each~~a given pair of objects. Dixon (1979) found that this operation greatly improved the results, therefore we also offer this option for the Euclidean, and Manhattan metrics and the Faith's intermediate coefficient.

Formatted: Font: Italic

Formatted: Font: Italic, Subscript

The question of how clustering and multidimensional scaling are affected by unbalanced resemblances remains to be answered. Earlier studies (e.g., Epifanio et al., 2020) indicated that in this regard the Partial Distance Strategy is a competitive alternative to other methods. We can safely say that hierarchical clustering is less prone to any problem in general because the approach handles the input resemblance matrix without imposing any theoretical, matrix algebraic restrictions. The relative robustness of clustering is demonstrated, for example, by the following arguments: 1) within-dendrogram (ultrametric) distances distort the original distances anyway, 2) several algorithms (e.g., single link, complete link) use only a subset of resemblances during calculations, and 3) the commonly used strategy of constrained clustering simply disregards a large subset of resemblance values. Nonmetric multidimensional scaling is less affected as well because the method relies on the rank order of resemblances and disregards their differences. Metric multidimensional scaling (principal coordinates analysis) is more sensitive, however. Due to unbalanced input resemblance values the object points may not be embedded into a Euclidean space (the matrix is not positive semi-definite), and the analysis may produce negative eigenvalues as well (Gower and Legendre, 1986). Whenever these are negligible in magnitude in comparison with the largest positive eigenvalues, the ordination of points in the first few dimensions remains

perfectly interpretable (Cailliez and Pagès, 1976). Otherwise, the user may consult with Legendre and Legendre (2012) or Li (2015) to select a method for adjusting the resemblance matrix in order to diminish the influence of negative eigenvalues.

Authors' contributions

JP conceived the ideas and wrote the paper, DS contributed to the text, performed literature search and wrote the R script.

Funding

This study was supported financially by the NKFIH-~~872~~ 471/3/2021 and NKFIH K128496 grants.

Declaration of competing interest

The authors declare that they have no competing interest.

Electronic Supplements

Electronic Supplement 1. R function *incomp* for calculating resemblance coefficients based on incomplete data.

The code may be downloaded from the web site of the publisher.

Electronic Supplement 2. A summary of resemblance coefficients for complete and incomplete data matrices

The ElectronicSupplement2.pdf ~~pdf~~ file may be downloaded from the web site of the publisher.

References

Anderberg, M. R., 1973. Cluster Analysis for Applications. Academic Press, London.

Cailliez, F., Pagès, J.-P., 1976. Introduction à l'analyse des données. Société de Mathématiques appliquées et de Sciences humaines, Paris.

Chechik, G., Heitz, G., Elidan, G., Abbeel, P., Koller, D., 2008. Max-margin classification of data with absent features. Journal of Machine Learning Research 9, 1–21.

Datta, S., Bhattacharjee, S., Das, S., 2018. Clustering with missing features: a penalized dissimilarity measure based approach. Machine Learning 107, 1987–2025. https://doi.org/10.1007/s10994-018-5722-4

Dixon, J. K., 1979. Pattern recognition with partly missing data. IEEE Transactions on Systems, Man, and Cybernetics 9(10), 617–621.

Eirola, E., Doquire, Verleysen, M., Lendasse, A., 2013. Distance estimation in numerical data sets with missing values. Information Sciences 240, 115–128.

Epifanio, I., Ibáñez, M.V., Simó, A., 2020. Archetypal analysis with missing data: see all samples by looking at a few based on extreme profiles. The American Statistician 74(2), 169–183.

Dray, S., Josse, J., 2015. Principal component analysis with missing values: a comparative survey of methods. Plant Ecology 216, 657–667.

Faith, D. P., 1983. Asymmetric binary similarity measures. Oecologia (Berl.) 57, 287–290.

Goodall, D. W., 1973. Sample similarity and species correlation. In R. H. Whittaker (Ed.), Ordination and Classification of Vegetation. Junk, The Hague, pp. 107-156.

Gower, J. C., 1971. A general coefficient of similarity and some of its properties. Biometrics 27, 857–871.

Field Code Changed

Formatted: Font:

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

Grung, B., Manne, R., 1998. Missing values in principal component analysis. Chemometrics and Intelligent Laboratory Systems 42, 125–139.

Kenkel, N. C., Booth, T., 1987. A comparison of presence/absence coefficients for use in biogeographical studies. Coenoses 2, 25–30.

Legendre, P., Legendre, L., 2012. Numerical Ecology. 3rd ed. Elsevier, Amsterdam.

Li, W., 2015. Estimating Jaccard index with missing observations: a matrix calibration approach. In: C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, R. Garnett (Eds.), Advances in Neural Information Processing Systems 28, 1–7.

Ludwig, J. A., Reynolds, J. F., 1988. Statistical Ecology. Wiley, New York.

Nelson, P. R. C., Taylor, P. A., MacGregor, J. F., 1996. Missing data methods in PCA and PLS: Score calculations with incomplete observations. Chemometrics and Intelligent Laboratory Systems 35, 45–65. https://doi.org/10.1016/S0169-7439(96)00007-X

Oba, S., Sato, M., Takemasa, I., Monden, M., Matsubara, K., Sin, I., 2005. A Bayesian missing value estimation method for gene expression profile data. Bioinformatics 19, 2088–2096.

Oksanen, J., Blachet, F. G., Friendly, M., Kindt, R., Legendre, P., McGill, D., Minich, P. R., O'Hara. R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E., Wagner, H., 2020. vegan: Community Ecology Package. R package version 2.5-7.

Orloci, L., 1972. On objective functions of phytosociological resemblance. American Midland Naturalist 88, 28–55.

277 Podani, J., 1980. Computer programs for cluster analysis in ecology, phytosociology and
278 systematics. Abstracta Botanica 6, 1-158. (in Hungarian, with English abstract, URL:
279 http://podani.web.elte.hu/PodaniAbsBot6Eng.pdf)

280 Podani, J., 2001. SYN-TAX 2000. Computer Programs for Data Analysis in Ecology and
281 Systematics. Scientia, Budapest. (URL: http://podani.web.elte.hu/syntax2000.pdf)

282 Podani, J., Kalapos, T., Barta, B., Schmera, D., 2021. Principal component analysis of
283 incomplete data – A simple solution to an old problem. Ecological Informatics 61,
284 101235. https://doi.org/10.1016/j.ecoinf.2021.101235

285 Podani, J., Pavoine, S., Ricotta, C., 2018. A generalized framework for analyzing taxonomic,
286 phylogenetic, and functional community structure based on presence-absence data.
287 Mathematics 6(11), 250

288 Ricotta, C., Pavoine, S., 2015. Measuring similarity among plots including similarity among
289 species: an extension of traditional approaches. Journal of Vegetation Science 26,
290 1061–1067.

291 Ricotta, C., Podani, J., Pavoine, S., 2016. A family of functional dissimilarity measures for
292 presence and absence data. Ecology and Evolution 6(15), 5383–5389. doi:
293 10.1002/ece3.2214

294 Serneels, S., Verdonck, T., 2008. Principal component analysis for data containing outliers and
295 missing elements. Computational Statistics & Data Analysis 52, 1712–1727.

296 Sneath, P. H. A., Sokal, R. R., 1973. Numerical Taxonomy. 2nd ed. Freeman, San Francisco.

297 Stanimirova, I., Daszykowski, M., Walczak, B., 2007. Dealing with missing values and outliers
298 in principal component analysis. Talanta 72, 172–178.

Tamás, J., Podani, J., Csontos, P., 2001. An extension of presence/absence coefficients to abundance data: a new look at absence. Journal of Vegetation Science 12, 401–410.

Wills, M. A., 1998. Crustacean disparity through the Phanerozoic: comparing morphological and stratigraphic data. Biological Journal of the Linnean Society 65, 455–500. doi:10.1111/j.1095-8312.1998.tb01149.x)

Wishart, D., 2003. k-means clustering with outlier detection, mixed variables and missing values. In: M. Schwaiger, O. Opitz (Eds.), Exploratory data analysis in empirical research. Studies in classification, data analysis, and knowledge organization. Springer, Berlin, pp. 216-226. https://doi.org/10.1007/978-3-642-55721-7_23

Zhang, W., Yang, Y., Wang, Q.,

2012. A comparative study of absent features and unobserved values in software effort data. International Journal of Software Engineering and Knowledge Engineering 22, 185-202.

Appendix 1

An R function for calculating resemblance coefficients for incomplete data sets

The *incomp* function (Electronic Supplement 1) is based on the *vegan* package. It requires (1) a data matrix as an input file (in agreement with R, rows are the observations or objects, columns are the variables (characters, traits), the cells contain ratio-scale or presence/absence data with missing data (coded as NA), (2) information about the method (see also Electronic Supplement 2 for abbreviations), and whether ratio-scale or binary calculation should be performed (binary= TRUE or FALSE). The output of the function is matrix.

```
#required packages
require(vegan)
```

```
#example data
data(varespec)
data1<-(varespec)

#insert some missing data
data1$Callvulg[1]<-NA
data1$Empenigr[2]<-NA
data1$Rhodtome[3]<-NA

#save as matrix
x<-as.matrix(data1)

#run the function
incomp(x, method="EUC", binary=FALSE)
```

Appendix 2

INDARES: WINDOWS application

For users unfamiliar with R, we provide a stand-alone application, INDARES.EXE, which runs under WINDOWS operation systems. The input for this program is a text file in which the first row is a label, the second row contains two integers, the number of rows and the number of columns. Then follows the data matrix, with each of its rows starting in a new line in the file. Missing scores are coded by –1. The user is prompted for the name of the input filename, for an option to decide whether rows or columns are to be compared and finally for selecting the option for resemblance function, numbers 1-34, which follow the same sequence as in Electronic Supplement 2. Before calculations, the program checks whether all pairs of objects are comparable, and potentially stops with a list of pairs which do not have a single known variable in common. Computations may only be performed if all pairs of objects are comparable. The Euclidean and Manhattan metrics and Faith's intermediate coefficient do not have upper bound, and these can be scaled up upon request by multiplication with the ratio of the number of all and observed values for each pair of objects. The resulting matrix may be saved in full format or as a lower semimatrix with the diagonal values included. If the

formula is a similarity (e.g., Renkonen), correlation or association (e.g. Yule) coefficient, the user is also asked if the values are to be converted into dissimilarities for output. If it contains a semimatrix of distances or dissimilarities, the save file MATRIX.TXT may be directly input to the SYN-TAX 2000 multivariate analysis package (Podani 2001). INDARES.EXE and the SYN-TAX 2000 modules may be downloaded from http://podani.web.elte.hu/SYN2000.html.

Table 1. Resemblance coefficients for complete and incomplete data matrices. Most coefficients for incomplete data are meaningful for binary data as well with $x_{ij} = 1$ for presence and $x_{ij} = 0$ for absence.

No.	Complete data				Incomplete data
	Ratio scale		Presence/absence		Both types
	Name/Author and abbreviation	Formula	Name/Author	Formula	Formula
1.	Euclidean distance – EUC	$\sqrt{\sum_i (x_{ij} - x_{ik})^2}$	Euclidean distance	$\sqrt{b + c}$	$\sqrt{E_{jk}}$
2.	Manhattan (city block) – MAN	$\sum_i \lvert x_{ij} - x_{ik} \rvert$	Symmetric difference, Hamming distance	$b + c$	M_{jk}
3.	Mean character difference – MCD	$\frac{1}{n}\sum_i \lvert x_{ij} - x_{ik} \rvert$	Simple matching dissimilarity	$\frac{b + c}{n}$	$\frac{M_{jk}}{W_{jk}}$
4.	Ruzicka – RUZ	$\frac{\sum_i \min\{x_{ij}, x_{ik}\}}{\sum_i \max\{x_{ij}, x_{ik}\}}$	Jaccard similarity	$\frac{a}{a + b + c}$	$\frac{F_{jk}}{H_{jk}}$
5.	Similarity ratio – SR	$\frac{\sum_i x_{ij}\, x_{ik}}{\sum_i x_{ij}^2 + \sum_i x_{ik}^2 - \sum_i x_{ij}\, x_{ik}}$			$\frac{A_{jk}}{A_{jj(k)} + A_{kk(j)} - A_{jk}}$
6.	Marczewski & Steinhaus – MS	$\frac{\sum_i \lvert x_{ij} - x_{ik} \rvert}{\sum_i \max\{x_{ij}, x_{ik}\}}$	Jaccard dissimilarity	$\frac{b + c}{a + b + c}$	$\frac{M_{jk}}{H_{jk}}$
7.	Bray & Curtis similarity – BCS	$\frac{2\sum_i \min\{x_{ij}, x_{ik}\}}{\sum_i (x_{ij} + x_{ik})}$	Sörensen similarity	$\frac{2a}{2a + b + c}$	$\frac{2F_{jk}}{F_{jk} + H_{jk}}$
8.	Bray & Curtis dissimilarity – BCD	$\frac{\sum_i \lvert x_{ij} - x_{ik} \rvert}{\sum_i (x_{ij} + x_{ik})}$	Sörensen dissimilarity	$\frac{b + c}{2a + b + c}$	$\frac{M_{jk}}{F_{jk} + H_{jk}}$
9.	Chord distance – CHO	$\sqrt{2\left(1 - \frac{\sum_i x_{ij}\, x_{ik}}{\sqrt{\sum_i x_{ij}^2 \sum_i x_{ik}^2}}\right)}$	Chord distance	$\sqrt{2\left(1 - \frac{a}{\sqrt{(a + b)(a + c)}}\right)}$	$\sqrt{2\left(1 - \frac{A_{jk}}{\sqrt{A_{jj(k)} A_{kk(j)}}}\right)}$
10.	Angular separation – ANG	$1 - \frac{\sum_i x_{ij}\, x_{ik}}{\sqrt{\sum_i x_{ij}^2 \sum_i x_{ik}^2}}$	Ochiai	$1 - \frac{a}{\sqrt{(a + b)(a + c)}}$	$1 - \frac{A_{jk}}{\sqrt{A_{jj(k)} A_{kk(j)}}}$
11.	Geodesic metric – GEO	$\arccos \frac{\sum_i x_{ij}\, x_{ik}}{\sqrt{\sum_i x_{ij}^2 \sum_i x_{ik}^2}}$	-	-	$\arccos \frac{A_{jk}}{\sqrt{A_{jj(k)} A_{kk(j)}}}$
12.	Kendall similarity – KEN	$\sum_i \min\{x_{ij}, x_{ik}\}$	Overlap	a	F_{jk}
13.	Faith intermediate coefficient – FA1	$0.5\sum_i (\lvert x_{ij} - x_{ik} \rvert + \max\{x_{ij}, x_{ik}\} - \min\{x_{ij}, x_{ik}\})$	Symmetric difference	$b + c$	$0.5(M_{jk} + H_{jk} - F_{jk})$
14	Penrose shape – P1	$\frac{1}{n-1}\sum_i (x_{ij} - x_{ik})^2 - \frac{1}{n^2 - n}(\sum_i x_{ij} - x_{ik})^2$	-	-	$\frac{1}{W_{jk} - 1} E_{jk} - \frac{1}{W_{jk}^2 - W_{jk}} G_{jk}^2$
15	Penrose size – P2	$\frac{1}{n^2}(\sum_i x_{ij} - x_{ik})^2$	-	-	$\frac{1}{W_{jk}^2} G_{jk}^2$

Table 2. List of presence-absence coefficients extended to incomplete abundance data. See text for the meaning of *a', b', c', d'* and *n'*.

No.	Name/Author and abbreviation	Formula
1.	Simple matching – SM	$\frac{a'+d'}{n'}$
2.	Rogers & Tanimoto – RT	$\frac{a'+d'}{a'+2b'+2c'+d'}$
3.	Sokal & Sneath – SS1	$\frac{2a'+2d'}{2a'+b'+c'+2d'}$
4.	Anderberg 1 – A1	$\sqrt{\frac{a'}{a'+b'}\times\frac{a'}{a'+c'}\times\frac{d'}{b'+d'}\times\frac{d'}{c'+d'}}$
5.	Anderberg 2 – A2	$\frac{1}{4}\left(\frac{a'}{a'+b'}\times\frac{a'}{a'+c'}\times\frac{d'}{b'+d'}\times\frac{d'}{c'+d'}\right)$
6.	Faith 2 – FA2	$\frac{a'+0.5d'}{n'}$
7.	Russell & Rao – RR	$\frac{a'}{n'}$
8.	Raroni-Urbani & Buser – BB2	$\frac{\sqrt{a'd'}+a'}{\sqrt{a'd'}+a'+b'+c'}$
9.	Yule 1 – Y1	$\frac{\sqrt{a'd'}-\sqrt{b'c'}}{\sqrt{a'd'}+\sqrt{b'c'}}$
10.	Yule 2 – Y2	$\frac{a'd'-b'c'}{a'd'+b'c'}$

Click here to access/download

Supporting File

ElectronicSupplement1.R

Click here to access/download

Supporting File

ElectronicSupplement2.docx