QÜESTIIÓ, Vol 18, 1 pp. 39-49, 1994

NOTES ON THE BIAS OF DISSIMILARITY INDICES FOR INCOMPLETE DATA SETS: THE CASE OF ARCHAEOLOGICAL CLASSIFICATION

ANGELA MONTANARI and STEFANIA MIGNANI

Department of Statistics

University of Bologna, Italy

The problem of missing values is particularly present in archaeological research where, because of the fragmentariness of the finds, only a part of the characteristics of the whole object may be observed. The performance of various dissimilarity indices differently weighting missing values is studied on archaeological data via a simulation. An alternative solution consisting in randomly substituting missing values with character states is also examined. Gower's dissimilarity coefficient seems to be the least biased one either with 25% missing values and 49%; it has not however a constant behaviour as to the sign of the bias. The simulation experiment has also shown that when average linkage cluster analysis is performed on an incomplete data set either using Gower's index or randomly substituting missing values gives satisfactory results while the modified indices fail to detect the cluster structure.

Key words: Dissimilarity indices, missing values, classification, cluster analysis, archaeological research.

–Article presentat al Seventh International Conference on Multivariate Analysis, setembre 1992.

–Acceptat el febrer de 1993.

1. INTRODUCTION

The problem of missing values is particularly present in archaeological research where time and possible geological, climatic and historical changes do not always allow findings of complete remains but only fragments on which just a portion of the characteristics of the whole object may be observed, often different from one object to another. The gain obtained observing, for each unit, the largest number of variables is in fact partially dimmed by the possible increase in missing values, with an ensuing decrease of reliability in comparisons. When one's aim is to detect clusters of observations, each measure of similarity between two units is in fact the more reliable the more it is capable to control for the effect of missing data and to use all the available information.

This paper deals with the definition of dissimilarity indices between archaeological reports of raw pottery, whose age lies between IV century B.C. and XIII century A.D., found out in an excavation nearby Castel Raimondo (Udine, Italy)[1], in order to detect groups homogeneous for characteristics that could allow us to date the finds. The fragmentariness of the manufacts is probably due to the presence in the clayey cob of sometimes big mineral inclusions as well as to the rudimentary production and baking; the work of baking, in fact, took place at very low temperatures (500 – 600°C) and for too short a time to allow a complete clay transformation (Guermandi, 1990).

When one has to deal with incomplete data sets, statistical literature sometimes suggests to restrict the analysis only to the units without missing values. This didn't seem to be a viable solution in our context as we had no complete object on which all the 28 qualitative variables considered could be observed (from kind of pottery to decoration, from brim descriptive parameters to ceramic cob). In order to reduce the proportion of missing values in the data set to be used for statistical analaysis, we selected at first only the 436 units which had no missing values in the four variables that describe brim characteristics. This aspect seems to be the one most closely connected to time variation. (In what follows this data set will be called data1). In this new data set the overall percentage of missings was about 49%; therefore two seemed the possible solutions to the problem of comparison between two units: using a dissimilarity index apt to suitably weight missing values thus sizing the differences between units with a good reliability or substituting missing values with possible states of

[1] The research has been realized by the "Istituto Beni Culturali" of "Regione Emilia Romagna" and by the "Istituto di Archeologia" of the University of Bologna under the direction of Sara Santoro Bianchi. The results have been published in two volumes: *Castel Raimondo. Scavi 1988-90.* I *Lo scavo* (S. Santoro Bianchi ed.) L'Erma di Bretschneider, Roma, 1992;*Castel Raimondo. Scavi 1988-90.* II *I materiali*, L'Erma di Bretschneider, Roma, in press.

the character, in analogy to what is commonly done with quantitative variables (Brothwell and Krzanowski, 1974; Beale and Little, 1975).

An often suggested approach in the statistical literature (Gordon, 1981) is to handle missing observations on qualitative variables by assigning the missing value to the state which occurs most frequently in the set of objects most closely resembling the incompletely recorded unit. In our context however such a solution is not appropriate for two reasons: the high number of missing makes it difficult to determine suitable frequencies for the various states of the characters and, on the other hand, substituting the most frequent state for missing values would lead to underestimate dissimilarities. Infrequent characteristics in observed data may in fact have been very frequent in those reports in which they are now missing because of the manufacts fragmentariness (for example objects with thin brim may poorly be represented in the observed set of objects as such a characteristic favours object brittleness). We have therefore decided to substitute incompletely recorded data with a character state randomly selected among possible ones with probability equal to the reciprocal of the number of states each character can assume.

2. SIMILARITY AND DISSIMILARITY INDICES: A COMPARISON BETWEEN DIFFERENT SOLUTIONS

When, in multivariate statistical analysis, one has to deal with qualitative variables, a similarity index between two generic units i and j may be determined as the ratio between the number of attributes on which the two units have the same category, divided by the total number of attributes considered. Such an index may be easily modified, allowing for missing values, by dividing the number of matches for the number of variables for which comparison is possible and then multiplying by the ratio between the total number of variables and the number of possible comparisons (Seber, 1984; Krzanowski, 1988). If no comparison is possible the similarity between two units is set equal to 0. Yet this index is not normalized and so it is not suitable to make comparisons between different empirical situations: that's why it will not be examined in what follows.

An alternative and more general solution which allows for comparisons between units on which both qualitative and quantitative characters have been examined, has been put forward by Gower in 1971; one version of that coefficient is obtained, for two units i and j, by assigning a score $0 \leq s_{ijk} \leq 1$ and a weight w_{ijk} for character k. The coefficient is described as:

$$S_G = \frac{\sum_{k=1}^{n} s_{ijk} s_{ijk}}{\sum_{k=1}^{n} w_{ijk}}$$

and the weight w_{ijk} is set to 1 when a comparison is considered valid for character k and to 0 when the value of the state for character k is unknown for one or both observational units; in such a case also s_{ijk} is conventionally set equal to 0. For quantitative characters one has:

$$s_{ijk} = 1 - \frac{|x_{ik} - x_{jk}|}{x_{k\,\mathrm{max}} - x_{k\,\mathrm{min}}}$$

where x_{ik} and x_{jk} are the values of unit i and unit j on variable k, while for qualitative characters s_{ijk} is 1 for matches between states for that character and 0 for a mismatch. If no comparison is possible the similarity between two units is set equal to 0. This coefficient, denoted as D_1 in what follows, has recently found fairly extensive archaeological use (Doran and Hodson, 1975; Shennan, 1988).

However, as Sneath and Sokal (1973) underline, the number of states in a multistate character is thus not taken into consideration and Gower's coefficient resembles, in this aspect, a simple matching coefficient applied to a data matrix involving multistate characters. Besides this coefficient does not behave uniformly in all situations: the similarity coefficient between two units with only one match and incomparable for all the other characters is in fact set equal to one, that is maximum similarity, thus showing, in that case, a tendency to overestimate similarity between units; on the contrary, when two units differ in one character state and are incomparable for the remaining ones the coefficient considers them as completely different, tending, in such a case, to underestimate similarity.

An alternative solution Gower's index allows for, is to substitute the weight $w_{ijk} = 0.5$ for $w_{ijk} = 0$ when comparisons are impossible, assuming that, for missing observations, similarity or dissimilarity between two units are equally likely for variable k. (This index will be indicated as D_2 in what follows).

This approach neglects again the different number of states each character can assume and so does not take into account that the probability for a match to occur on character k decreases as the number of such values increases; a further modification may thus be obtained by setting the weight w_{ijk} equal to the reciprocal of the number of states. (From now on this index will be indicated as D_3).

In what follows reference will be made to dissimilarity coefficients, obtained as the complement to 1 of the corresponding similarity index, as the statistical software available for further analysis requires dissimilarity or distance matrices as input.

3. THE SIMULATION

In order to evaluate the performance of the coefficients described in the previous section, a subset of data1 was drawn. The aim was to obtain a set of units with complete records, of suitable size for statistical analysis; so we dropped one variable at a time[2] and counted the number of units with no missing value; we stopped when we got at least 100 units. A set of 122 complete units on 9 variables was thus obtained and the matrix containing the dissimilarities between such units calculated. As no missing values appeared in the data matrix, all the considered indices gave the same dissimilarity values. Under the hypothesis of random distribution of missings in the data set —as random is the finding of one fragment instead of another— a simulation experiment was then run as follows. Some data observations were randomly selected and substituted with missing values (first according to the proportion of missings in the data set of the 436 initial manufacts, *i.e.* 49%, and then in about half that proportion) and the various dissimilarity coefficients were calculated. This process was repeated 400 times, and the bias of each index was calculated as the difference between the mean of the dissimilarity matrices over the 400 replications and the dissimilarity matrix of the 122 real objects. All calculations were implemented in Gauss on a PC.

At last, to determine whether missing data might be better handled by modifying dissimilarity coefficients or by imputing possible values, each missing, generated as previously described, was substituted by a character state[3], chosen at random among all the possible ones, with probability equal to the reciprocal of the number of states and once again, in 400 replicates, dissimilarity matrices and corresponding bias were computed. (From now on this situation will be

[2] One must remember that no entire object was found or assembled and so in data1 the observational units are only fragments.

[3] The same matrix has been used either to replace missings with imputed values or to compute the different dissimilarity coefficients.

denoted as D_4). The results thus obtained have then been compared with those derived using the various dissimilarity indices.

The capability the considered methods have in handling missing values has been further checked studying their performance in cluster analysis (Wishart, 1978). From the set of 9 variables previously used in the simulation experiment, 4 have been chosen, namely "kind of clayey cob" and "brim characteristics" (three), which could well separate between "open" and "closed" shapes[4] of pottery and which had no missings in the considered data set. Thus two populations have been defined and samples of 50 objects have been randomly selected from each population. The couples of samples have then been put together, thus forming a sample of 100 units on which average linkage cluster analysis has been performed (Sneath and Sokal, 1973; Gordon, 1981; Rizzi, 1985). This kind of clustering method was chosen because of its good performances in previous analyses of data1 (see Capitanio, Mignani, Montanari, 1993). Besides, as the dissimilarity indices used are weighted averages, the recourse to average linkage coherently extends the properties of arithmetic mean to the clustering criterion.

Missing values have then been randomly introduced in the sample in the way and in the proportions previously described. Average linkage cluster analysis has been performed again, for each of the different ways of handling missing values, *i.e.* for each of the methods denoted as D_1, D_2, D_3, D_4. The process has been replicated many times and on different samples. In the end the results obtained for each method have been compared with the one derived from the original sample in which no missings were present with the aim of detecting which, among D_1, D_2, D_3, D_4, could best recover the original cluster structure.

4. RESULTS AND CONCLUSIONS

For each coefficient and for the case in which values have been randomly assigned to replace missing data, table 1 and 3 report the overall mean bias, obtained on the values below the principal diagonal in the matrix containing the mean dissimilarity bias for each pair of units (in the 400 replicates), and its range. In symbols, denoted by $\bar{b}_{ij}$ the generic mean dissimilarity bias for units i and j, the overall mean bias is given by:

[4] "Open" shapes are "baking pans", "large bowls" and "bowls"; "closed" shapes are "glasses", "jars" and "little jars".

$$\bar{b} = \frac{\sum_{i,j} \bar{b}_{ij}}{n(n-1)/2}$$

for $i = 1, \ldots, n-1$, $j > i$ and $n = 122$.

Table 2 and 4 report the percentage of cases each index overestimates or underestimates the distance between two units for both 49% and 25% missing values.

As clearly appears from the tables, D_1 has the least mean bias for 25% missing values but not for 49% missings; in that case it is slightly more biased than D_4; its bias has however the minimum range in both situations and so it seems to be the most reliable index among those considered. It half the times overestimates and half times underestimates the "true" distance consistently with what aforesaid in section 2. D_2 and D_3 more often give positively biased estimates but D_2 is more biased than D_3.

Table 1.

Results of the simulation with 49% missing values.

Dissimilarity index	Mean bias	Minimum value of bias	Maximum value of bias
D_1	0.0044	−0.0857	0.1093
D_2	0.2040	0	0.5149
D_3	0.1387	0	0.3613
D_4	0.0034	−0.30	0.4022

Table 2.

Proportion of positive and negative values of bias for each index with 49% missing values.

	D_1	D_2	D_3	D_4
Positive	58	100	100	51
Negative	42	0	0	49

Table 3.

Results of the simulation with 25% missing values.

Dissimilarity index	Mean bias	Minimum value of bias	Maximum value of bias
D_1	−0.0008	−0.0859	0.0943
D_2	0.1108	−0.0109	0.2816
D_3	0.0649	0.0147	0.1716
D_4	0.0371	−0.15	0.2922

Table 4.

Proportion of positive and negative values of bias for each index with 25% missing values.

	D_1	D_2	D_3	D_4
Positive	49	100	99	63
Negative	51	0	1	37

The presence of 100% positive values in table 2 for the two indices can be explained considering that the quoted values are expected values over 400 replicates and that positive biases are larger than negative ones. D_4 gives a small mean bias but has a large bias range; so replacing missings by randomly selected character states does not seem to be a correct way to handle missing values.

In conclusion Gower's dissimilarity coefficient in its original version is the one giving the least biased results even if it has not a constant behaviour as to the sign of the bias.

Cluster analysis[5] for D_1, D_2, D_3, D_4 seems to confirm the results previously stated; Gower's index succeeds in recovering the original cluster structure better than all the other indices although also D_4 works fairly well particularly with 49% missing values. A previous study on the relationships between the variables showed that the characters considered are significantly connected; in this context the good performance of D_4 strengthens the assumption of random distribution of missing values. Further checks are however going on.

[5] Cluster analysis has been performed using SAS/STAT CLUSTER PROCEDURE.

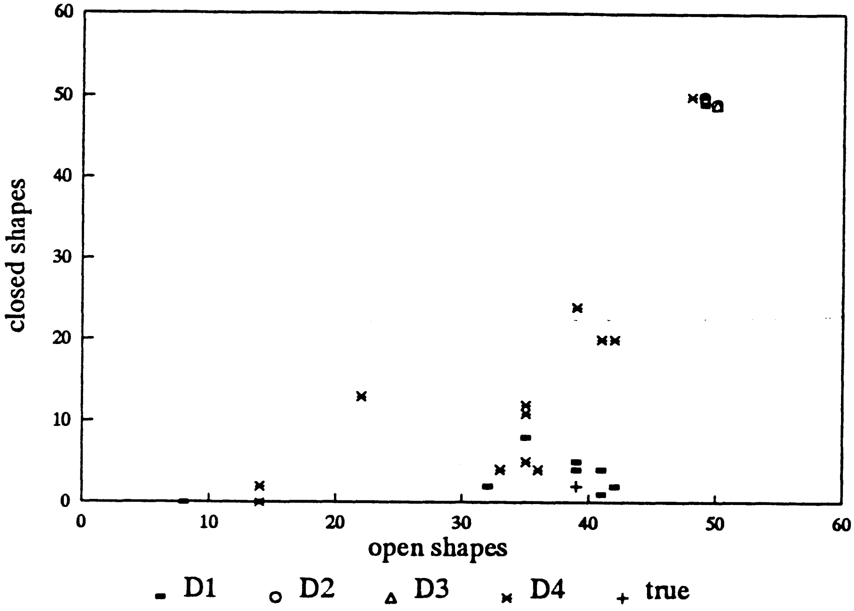

Figure 1.

Cluster results for each index with 49% missing values.

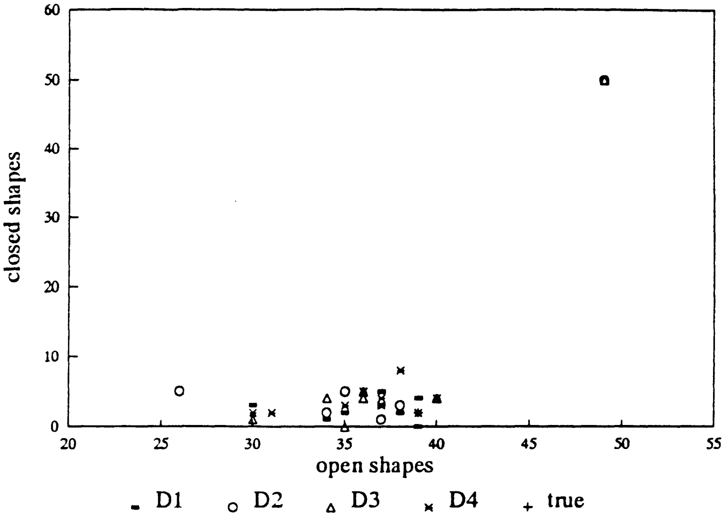

Figure 2.

Cluster results for each index with 25% missing values.

D_2 and D_3 completely fail in detecting the existing clusters, always giving rise to a single big cluster containing all the units. When the proportion of missing values decreases D_1, D_2, D_3, D_4 give similar results. All samples studied gave results similar to the ones shown for one of them in pictures 1 and 2 (only the count of the units of one cluster is shown, the one for the other cluster may be determined as difference); to avoid congestion on the pictures only a small subset of the simulation results has been represented.

REFERENCES

[1] **Beale, E.M.** and **Little, R.J.A.** (1975). "Missing values in multivariate analysis". *Journal of the Royal Statistical Society*, **B 37**, 129–145.

[2] **Brothwell, D.R.** and **Krzanowski, W.J.** (1974). "Evidence of biological differences between early British populations from Neolithic to Medieval times as revealed by eleven commonly available cranial vault measurements". *Journal of Archaeological Sciences*, **1**, 249–260.

[3] **Capitanio, A., Mignani, S.** and **Montanari, A.** (1992). "Le elaborazioni statistiche". In: S. Santoro Bianchi (ed.) *Castel Raimondo. Scavi 1988–90.* II *I materiali.* L'Erma di Bretschneider, Roma (in press).

[4] **Doran, J.** and **Hodson, F.** (1975). *Mathematics and computers in archaeology.* Edinburgh University Press. Edinburgh.

[5] **Gordon, A.D.** (1981). *Classification.* Chapman and Hall. London.

[6] **Gower, J.C.** (1971a). "A general coefficient of similarity and some of its properties". *Biometrics*, **27**, 857–872.

[7] **Gower, J.C.** (1971b). "Statistical methods of comparing different multivariate analyses of the same data". In: *Mathematics in the archaeological and historical sciences.* (F.R. Hodson, D.G. Kendall, P. Tatu eds.). University Press. Edinburgh.

[8] **Gower, J.C.** and **Legendre, P.** (1986). "Metric and euclidean properties of dissimilarity coefficients". *Journal of Classification*, **3**, 5–48.

[9] **Guermandi, M.P.** (1990). "La ceramica grezza. Analisi computerizzata e classificazione: problemi di metodo". *Antichità Altoadriatiche*, **XXXVI**.

[10] **Krzanowski, W.J.** (1988). *Principles of multivariate analysis.* Clarendon Press. Oxford.

[11] **Rizzi, A.** (1985). *Analisi dei dati.* Nis. Roma.

[12] **Seber, G.A.F.** *Multivariate observations.* John Wiley and Sons. New York.

[13] **Shennan, S.** (1988). *Quantifying archaeology.* Edinburgh University Press. Bristol.

[14] **Sneath, P.H.A.** and **Sokal, R.R.** (1973). *Numerical taxonomy.* Freeman and Co. San Francisco.

[15] **Wishart, D.** (1978). "Treatment of missing values in cluster analysis". In: *Compstat 1978* (C.A. Corsten, J. Hermans eds.). Physica-Verlag. Wien.