QÜESTIIÓ, vol. 22, 3, p. 529-538, 1998

THE USEFULNESS OF DISCRIMINATION BASED ON DISTANCES ON HUMAN EVOLUTION

C. ARENAS*
D. TURBÓN**
Universitat de Barcelona

The reconstruction of human history from the fossil record often runs up against incomplete or differential preservation of specimens. In anthropological studies a large number of variables are usually taken and missing values can be a problem. Here we analyze three population samples of extinct aborigines from Tierra del Fuego. The first sample, with sex and ethnic group known, is used to compare the step-wise discriminant analysis and the discriminant analysis based on distances. With the second sample a first approach to the assignation of poorly documented specimens in relation to sex or ethnic group is presented here by comparing the results from the two discriminant methods. A third sample of skulls with ethnic group and sex unknown is used to illustrate the advantages of distance-based discriminant analysis to solve the problem of allocating individuals when some values are missing.

Keywords: Distance discrimination; step-wise discrimination; missing values; Tierra del Fuego aborigines.

* Facultat de Biologia. Departament d'Estadística. Universitat de Barcelona. Av. Diagonal, 645. 08028 Barcelona.
** Departament de Biologia Animal (Antropologia). Facultat de Biologia. Universitat de Barcelona. Av. Diagonal, 645. 08028 Barcelona.

– Received February 1997.
– Accepted April 1998.

1. INTRODUCTION

One important aim in anthropology is to reconstruct extinct human groups, and thus to find relationships between them, to analyse differences and similarities with other extinct groups or with modern groups and to establish a possible common origin. For these reasons it is essential that the material studied should be accurately identified. Moreover accurate identification is necessary in order to allocate problematic individuals. However, the anthropological reconstruction of extinct human groups from archaeological sites is usually conditioned by the state of preservation of the remains, usually skull and long bones, and the statistical treatment often has to deal with a variable set of missing values. Another source of difficulty comes from the evolutionary context. For instance, the size of the bones could be problematic when determining sex of remains belonging to neighbouring ethnic groups, as some females from an hypothetically robust group could be erroneously classified as males from another group. This is the case when dealing with skulls of aborigines from Tierra del Fuego (particularly the Ona, pedestrian hunters-gatherers in Isla Grande) which show great osteological robusticity, all of them probably corresponding to a Paleoindian stock (Lahr 1995). This great robusticity prevents the distinction between the Ona female skulls and male skulls of the sea-canoe aborigines Yamana and Alakaluf. All these aborigines were decimated upon contact with Europeans, leading to their virtual extinction between the turn of the nineteenth century and the early twentieth century. In particular, the Ona were moved away from their original land, where they mixed with other ethnic groups.

This study is a tentative classification of a number of Fueguian skulls from different European and American museums and collections (Turbón 1995). Some cases are of uncertain attribution because of their physical displacement, complicated by the difficulty of discriminating the robust Ona females from the sea-canoe males. Sometimes more than one possible identification is given or contemporany anthropologists contradict former identifications.

Our skulls were further classified in three groups depending on previous identification. One with completely identified skulls (ethnic group and sex); another with no sure sex identification, and a third group with poor ethnic and sex identification. The aim of this study is to clarify the identification of these last two groups. A discriminant analysis is proposed, but as usual in the analysis of measurements of human skulls the following difficulties arise. When a broken skull was found, only some measurements could be taken. Thus, the data were incomplete and then several choices are available to compute the missing values. In what follows, some of the more usual solutions found in the literature are commented. One choice is to remove all cases for which the data are incomplete, which often reduces the number of samples dramatically and could exclude particular cases of critical importance for the analysis. A second choice is the replacement of missing values, either by the group mean or by values

obtained through multiple regression, which is not satisfactory in our case since some subgroups contain only a few specimens. Another possibility is the suppression of the variables for which a large number of values are missing, which would involve a significant reduction of information. Furthermore, the longer skulls could be prone to damage or poor preservation (Rao 1989) and a distinction must be made between measurements taken on well preserved skulls and those from damaged skulls. In this study we work under the hypothesis that the distribution of missing values is random. Finally, other difficulties can arise when classical discriminant analysis is applied. For example, if there many variables classical step-wise discriminant analysis is appropriate. However, the step-wise forward method with the F criterion for including a variable requires the data to be normally distributed. Furthermore the variables selected by this method are not always optimal (see Mc Cabe 1975). Moreover it is possible that the new individuals to be classified present missing values in the variables selected (sometimes in all of them), so correct allocation is not possible. In order to avoid some of these problems, this study uses the discriminant analysis based on distances introduced by Cuadras (1989). First a brief description of the method is presented and some interesting properties are discussed. After describing our data, a discriminant analysis and the assignation of some skulls with problems in the sex or ethnic group identification are performed first using the step-wise discriminant analysis, and then using the distance-based method. The results given by these methods are compared and we discuss some of their advantages and disadvantages.

2. MATERIAL AND METHOD

The material studied consists of 162 Fueguian skulls belonging to three ethnic groups: Yamana, Alakaluf and Ona (Table 1). A total of 65 biometrical traits were measured following W.W. Howells' technique (Howells 1973, 1989). These measurements (see Howells 1973 for details) are useful in the identification of the sex and ethnic group. This material is classified in three groups. Sample S_1 contains skulls with sure sex and ethnic group identification; sample S_2 contains skulls with sure ethnic group identification but doubtful sex identification; sample S_3 contains skulls with doubtful sex and ethnic group identification.

The distance-based (DB) discriminant analysis was introduced by Cuadras (1989) and it has recently been explained in detail (Cuadras 1989,1991,1992; Arenas *et al.* 1994). Its goal and rule of classification may be briefly summarised as follows.

Given some groups $\prod_i$ $(i = 1, \ldots, k)$ and a selected distance function $\delta(\cdot,\cdot)$ between individuals, then the rule of classification for a new individual x is,:

«allocate x to $\quad (i = 1, \ldots, k) \quad$ if and only if $\quad f_i(x) = \min\{f_1(x), \ldots, f_k(x)\}$ »,

where

$$f_i(x) = \frac{1}{n_i} \sum_{l=1}^{n_i} \delta_{li}^2 - \frac{1}{2n_i^2} \sum_{l,j=1}^{n_i} \delta_{lji}^2$$

and n_i, the sample size of group $\prod_i$; δ_{lji}^2 the square distance between objects l and j of group $\prod_i$ and δ_{li}^2 the square distance between object l of group $\prod_i$ and the new object x.

Cuadras *et al.* (1997a) proved that each $f_i(x)$ can be interpreted as the proximity of x to $\prod_i$. Thus the DB rule assigns an individual to the nearest group (see also Cuadras *et al.* 1997b). It can also be showed that it is equivalent to the linear discriminant rule, the quadratic discriminant rule or the euclidean discriminant rule, if an appropriate distance function is taken. Furthermore, as it is based on a distance, it can be applied to binary, qualitative or mixed variables by using a suitable distance.

It is clear that the results of the distance-based discriminant analysis depend on the distance selected. In this study Gower's distance (Gower, 1971) was chosen. This distance is obtained by assigning a score $0 \leqslant s_{ijk} \leqslant 1$ and a weight w_{ijk} for variable k.

The expression of this distance is given by $d_{ij} = 1 - \dfrac{\sum_k s_{ijk}\, w_{ijk}}{\sum_k w_{ijk}}$ where for continuous variables $s_{ijk} = \Sigma\left(1 - \left|x_{ik} - x_{jk}\right| / G_k\right)$, G_k is the range of the kth continuous variable.

For qualitative or binary variables s_{ijk} is 1 for matches between states and 0 for mismatches. The weight w_{ijk} is set to 1 when a comparison is considered valid for variable k and to 0 when the value of variable k is unknown for one or both observational units. As proved in Montanari (1994) it is a suitable distance for the treatment of data with missing values because it seems to be the least biased and reproduces the original cluster structure.

Summarising, distance-based discrimination has the following advantages:

- It works with mixed variables.
- It allows to work with a large number of variables.
- It can deal with missing values.
- It allocates new individuals with missing values.
- It does not need calculation of any inverse-matrix, so it is robust to the problem of ill-conditioned covariance matrices.
- It does not need any hypothesis about the distribution (normality) of data.

For all these reasons, a distance-based discriminant analysis might be preferable to classical linear discrimination or step-wise discriminant analysis in some cases. In this study, a comparison between the DB-method and the step-wise forward method using the F criterion is carried out. For calculating the probability of miss-classification the leave-one-out method is used. The analysis with the step-wise method is performed using the BMDP package. The DB method is implemented in the package of multivariate analysis Multicua (Arenas *et al.* 1991, 1993, 1998). A version for a large number of data was written by F. Oliva in SAS/IML.

3. RESULTS

In our data (Table 1), 75% of the skulls measured had a variable number of missing values, affecting 68% of the 65 variables observed. As mentioned above, the skulls were classified in three different subsamples, S_1 = completely known; S_2 = no sure sex identification and S_3 = ethnic group and sex unknown.

Table 1. *Description of the data: number of skulls for males (M) and females (F)*

	S_1		S_2	S_3
	M	F		
Yamana	31	22	11	
Alakaluf	11	10	2	
Ona	19	6	32	
Total	99		45	18

First we consider the data of sample S_1. The results of three DB discriminant analysis on the three groups of S_1 compared with those derived from a classical step-wise discriminant analysis are shown in Table 2. Table 3 shows the sex assignation given by the DB model and the step-wise model for the skulls of group S_2. The results of a new discriminant analysis when both samples S_1 and S_2 (with the final assignation) are put together are presented in Table 4. Finally using all the skulls of group S_1 an analysis is made in order to assign individuals of the S_3 group. The results of the discriminant analysis and the assignations are given in Tables 5 and 6 respectively.

In the first analysis (Table 2) a higher percentage of correct classification is obtained by the step-wise method, confirming the well known efficiency of this procedure. However, when we try to assign individuals of S_2, the advantages of the DB-discriminant analysis are clear (Table 3). With the step-wise discriminant analysis, some skulls cannot be allocated because they have missing values in the variables selected. From these results, it is clear that although the classical step-wise discriminant analysis initially gives better results, the DB-discriminant analysis is more useful for new

assignations. Furthermore, if the variables selected by the step-wise discriminant analysis are removed in order to work only with variables for which the elements of S_2 have no missing values, the assignation for all skulls is then possible although the probability of misclassification becomes greater (0.06 for Yamana; 0.2 for Ona). A new discriminant analysis is performed when samples S_1 and S_2 are put together. As Table 4 shows, with the step-wise method some skulls of known sex and ethnic group (from S_1) are now incorrectly classified. Finally when a discriminant analysis is performed with the skulls of group S_1 (see Table 5) the step-wise method , as before, gives a better classification than the DB rule, but again problems arise when skulls of S_3 are assigned (see Table 6). In this case using the second assignation, it is impossible to assign the skulls of S_3. These skulls present missing values in the variables selected by the step-wise discriminant analysis. If the variables selected by the step-wise are removed then the probability of bad classification (0.42) by the step-wise discriminant analysis becomes greater than the probability of bad classification using the DB-discriminant analysis (0.232).

Table 2. *Results of a DB-discriminant analysis and classical step-wise analysis on S_1*

DB-discriminant analysis						Step-wise discriminant analysis					
Matrix of misclassification						Matrix of misclassification					
		M	F	Prob. misclassif.	number variables			M	F	Prob. misclassif.	variables selected
Yamana	M	27	4	0.132	65	Yamana	M	31	0	0	8
	F	3	19				F	0	22		
Alakaluf	M	10	1	0.095	65	Alakaluf	M	11	0	0	10
	F	1	9				F	0	10		
Ona	M	17	2	0.28	65	Ona	M	18	1	0.04	2
	F	5	1				F	0	6		

Table 3. *Results of the assignation of skulls from group S_2*

Assignation according to the Museum (initial assignation) and biometrical assignation (DB and step-wise). ?= have missing values in the selected variables and no assignation is possible.

	Initial assignation	DB assignation	Step-wise assignation
Yamana	4M	3M 1F	0M 3F 1?
	7F	7F	1M 6F
Alakaluf	1M	1M	1F
	1F	1F	1M
Ona	23M	18M 5F	4M 17F 2?
	9F	3M 6F	5M 4F

Table 4. *Results of a DB-discriminant analysis and classical step-wise analysis on S_1 and S_2*

DB-discriminant analysis						Step-wise discriminant analysis					
Matrix of misclassification						Matrix of misclassification					
		M	F	Prob. misclassif.	number variables			M	F	Prob. misclassif.	variables selected
Yamana	M	30	4[1]	0.125	65	Yamana	M	30	2[4]	0.06	4
	F	4[2]	26				F	2[4]	29		
Alakaluf	M	11	1[1]	0.087	65	Alakaluf	M	10	2[4]	0.174	22
	F	1[1]	10				F	2[4]	9		
Ona	M	33	7[3]	0.210	65	Ona	M	25	3[4]	0.091	6
	F	5[3]	12				F	2[4]	25		

1 the same incorrectly assigned skulls as in the first analysis (Table 2)
2 three of the incorrectly assigned skulls in the first analysis (Table 2) and the skull initially assigned as M and finally assigned as F.
3 skulls of sample S_2.
4 skulls with ethnic group and sex known (from S_1) that now are incorrectly classified.

Table 5. *Results of a DB-discriminant analysis and classical step-wise analysis on S_1 group.*

DB-discriminant analysis Matrix of misclassification							Step-wise discriminant analysis Matrix of misclassification						
	1	2	3	4	5	6		1	2	3	4	5	6
1	21	3	3	0	4	0	1	29	0	0	1	1	0
2	3	16	0	3	0	0	2	0	22	0	0	0	0
3	2	0	7	1	1	0	3	0	0	9	1	1	0
4	0	1	1	8	0	0	4	0	0	1	9	0	0
5	4	0	3	0	10	2	5	1	0	1	0	16	1
6	0	1	1	0	3	1	6	0	0	0	0	0	6

1=Yamana M; 2=Yamana F;3= Alakaluf M; 4=Alakaluf F; 5=Ona M; 6=Ona F.

Prob. misclassification	number variables	Prob. misclassification	variables selected
0.361	65	0.08	6

Table 6. *Results of the assignation of skulls from group S_3*

Assignation according to the Museum (initial assignation) and biometrical assignation (DB and step-wise). ?= have missing values in the selected variables and no assignation is possible.

Initial assignation	DB assignation	Step-wise assignation
18 skulls	10 are reconfirmed 8 change	?

4. CONCLUSIONS

Palaeontological studies based on quantitative variables often deal with heterogeneous samples that do not belong to empirical biological populations and include incomplete data sets. Furthermore, isolated specimens are usually considered in the comparative analyses establishing phylogenetic relationships. The DB discriminant analysis could be a valuable statistical tool as it works with morphological distances and avoids the missing values problem at the same time. This is particularly useful when substitution of missing values is impossible or inadvisable if a step-wise analysis method were initially chosen, which in this case is actually more efficient than the DB method. Whether substitution of missing values, a combination of both techniques, or direct application of the DB rule is the right choice depends on the information sought, since the three choices respectively present as many advantages as disadvantages. However the above results indicate that it seems that in order to make new assignations, it is better to use the DB-discriminant analysis than a classical step-wise discriminant analysis when there are missing values. So it is clear that the DB-discriminant analysis has some advantages when some values are missing. A summary of some advantages and disadvantages of the DB discriminant method with respect to the classical step-wise method is presented bellow. The DB-rule can use qualitative, quantitative, binary or mixed variables without any transformation. The step-wise method uses quantitative variables, and can also use qualitative variables although a codification as binary variables is needed. If the number of variables is large with respect to the number of individuals, the DB-rule can work with all of them. The step-wise method has to select some of them and this selection is not always optimal. The step-wise method usually gives better allocation for predetermined groups than the DB-rule. With the step-wise method if the new individual to allocate has missing values in the selected variables, assignation is not possible. The DB-rule deals with missing values and can allocate individuals with values of this kind.

5. ACKNOWLEDGEMENTS

We thank Prof. Cuadras for his helpful suggestions and R. Rycroft (S.A.L. University of Barcelona) for the English correction. This work has been supported in part by grants from Spanish DGCYT (PB93-0021 and PB96-1004) and the Generalitat de Catalunya (GRC96-UB2506 and SGR97-00183).

6. REFERENCES

[1] **Arenas, C.** and **Bernal, M.** (1994). «Multivariate approach to the classification of *genus Dianthus L. (Caryophyllaceae)*». *Selected topics on stochastic modelling*, R. Gutierrez y M.J. Valderrama Editores, World Scientific, Singapore.

[2] **Arenas,C.; Cuadras, C.M.** and **Fortiana, J.** (1991). *Multicua. Paquete no estandard de análisis multivariante, versión 0.75*. Publicacions del Departament d'Estadística, n° 4, Barcelona, Spain.

[3] **Arenas, C.; Cuadras, C.M.** and **Fortiana, J.** (1993). *Multicua. Paquete no estandard de análisis multivariante, versión 0.77*. Publicacions del Departament d'Estadística, n° 4, Barcelona, Spain.

[4] **Arenas, C.; Cuadras, C.M.** and **Fortiana, J.** (1998). *Multicua. Paquete no estandard de análisis multivariante, versión 0.77 ampliada*. Publicacions del Departament d'Estadística (nueva colección), n° 1, Barcelona, Spain.

[5] **Cuadras, C.M.** (1989). «Distance analysis in discrimination and classification using both continuous and categorical variables». *Statistical Data Analysis and Inference*. (In: Y. Dodge, ed.) Elsevier North Holland, Amsterdam pp. 459–473.

[6] **Cuadras, C.M.** (1991). «A distance approach to Discriminant Analysis and its Properties». *Universitat de Barcelona, mathematics preprint series*, n° 90.

[7] **Cuadras, C.M.** (1992). «Some examples of distance based discrimination». *Biometrical Letters*, **29 (1)**, 3–20.

[8] **Cuadras, C.M.; Fortiana, J.** and **Oliva, F.** (1997a). «The proximity of an individual to a population with applications to discriminant analysis». *Journal of Classification*, **14**, 117–136.

[9] **Cuadras, C.M.; Atkinson, R.A.** and **Fortiana, J.** (1997b). «Probability densities from distances and discriminant analysis». *Statistics & Probabilities Letters*, **33**, 405–411.

[10] **Gower, J.C.** (1971). «A general coefficient of similarity and some of its properties». *Biometrics*, **27**, 857–874.

[11] **Howells, W.W.** (1973). *Craneal Variation in Man. A study by Multivariate Analysis of Patterns of Difference Among Recent Human Populations*. Papers of the Peabody Museum Harvard University, vol. 67, 259 pp.

[12] **Howells, W.W.** (1989). *Skull Shapes and the Map. Craniometric Analysis in the Dispersion of Modern Homo*. Papers of the Peabody Museum Harvard University, vol.79, 189 pp.

[13] **Lahr, M.M.** (1995). «Patterns of modern human diversification. Implications for Amerindian origins». *Yearbook of Physical Anthropology*, **38**, 163–198.

[14] **Mc Cabe, G.P., Jr.** (1975). «Computations for variable selection in discriminant analysis». *Technometrics*, **17**, 103–109.

[15] **Montanari, A.** and **Mignari, S.** (1994). «Notes on the bias of dissimilarity indices for incomplete data sets: the case of archaelogical classifications». *Qüestiió*, **18**, 39–49.

[16] **Rao, C.R.** (1989). *Statistics and Truth*. International Co-oper. Pub. House, Fairland, USA.

[17] **Turbón, D.** (1995). «Antropología de los aborígenes fueguinos». Estévez J. and Vila A. Eds.: *Encuentros en los conchales fueguinos*. C.S.I.C. Madrid, 275–289.