Przegląd Antropologiczny – Anthropological Review • Vol 62, pp. 75–83 (1999)

https://doi.org/10.18778/1898-6773.62.09

Analysis of morphological differences between prehistoric populations using a non-hierarchic method of data clustering

Arkadiusz Sołtysiak[1], *Piotr Jaskulski*[2]

ABSTRACT Presented in this paper non-hierarchic method of cluster analysis allows studying of morphological variability in big populations on the basis of individual characteristic of specimens. Test taken by the authors have confirmed correct results of the proposed algorithm and showed the role of proper preparation of data, choice of appropriate distance measure and importance of the process of interpretation the results of clustering.

Arkadiusz Sołtysiak, Piotr Jaskulski, 1999; *Przegląd Antropologiczny – Anthropological Review*, vol. 62, Poznań 1999, pp. 75–83, figs 2. ISBN 83-86969-44-X, ISSN 0033-2003

Introduction

One of the main problems concerning anthropologists who deal with prehistoric populations is attempt to reconstruct history of specified human group (i.e., its "time and space"). Such reconstructions are conducted by the means of analyzing biological differences between populations, which consider both synchronic (by paralleling contemporary neighbor groups) and diachronic (by comparing populations, which occupied the same territory in different periods of time) comparison. Those studies uses various analytical methods suitable for checking inter-populational differences, but we can say that morphological analyses have the longest tradition. Other methods consider blood-groups as well as molecular structures. One of most popular methods preferred by biologists in the last twenty years is a construction of phylogenetic trees and similarity dendrograms, based on analysis of nuclear or (more frequently) mitochondrial DNA. Those methods, considering recent progress of genetic science, are also used for evaluating the history of human migrations and explaining modern intra-specific differentiation. However it should be stated that analysis on the DNA level are seldom useful for studying fossil material. Main causes for that are the technical difficulties and also high costs of such analysis [MAYS 1998]. For that reason study of morphological variability can be of great importance, considering its comparatively easier to conduct methods and lack of advanced laboratory technology. Of course we realize that there are some problems one has to be conscious about:

[1] Department of Historical Anthropology
Warsaw University, Krakowskie Przedmieście 26/28
Warszawa 64
[2] Department of Computer Science
State Archaeological Museum
Długa 52, 00-950 Warszawa

1. Necessity of realizing the difference between biological population and so-called "fossil population";

2. Data fragmentariness;

3. Necessity of proper selection of characteristics (which often depend rather on their availability than optimization);

4. Necessity of adapting the method of numerical analysis for the specified problem.

Fossil finds are only a kind of "film frame" showing the moment from the evolution process of specified population, due to fact that composition of "fossil population" strongly depends on chance. Researcher has to select characteristics by their extent of heritability and environmental sensibility. The statistical analyses in multi-characteristic space have also some limitations: too many characteristics (or dimensions in multi-characteristic space) lead to blur intra-populational difference, but on the other hand more characteristics increase credibility of results. The researcher's goal should then be taking the intermediate stand between those two inclinations.

One should also have in mind that random composition of the fossil population has to result in rejecting all difference analyses methods based on the calculating and comparing of means and other central measures. Series of skeletons used in analyses can actually originate from a number of biological populations. For that reason we assumed that it is necessary to introduce a method, which allows both: to regard individual characteristics, and to analyze numerous series of objects (even thousands of individuals). Then we assumed that the method suitable for that should be cluster analysis, popular among anthropologists who are engaged in studying biological variability [HENKE 1991, LYNCH 1989]. Most often they use dendrograms (i.e., hierarchical methods of clustering), and previously so-called Czekanowski's diagrams. However both of these methods do not comply with the second requirement we have presented – they allow analyzing series of no more than 100-200 objects. Our attention has been attracted by an algorithm of complete and non-hierarchical data clustering [OVERALL & KLETT 1972, MAREK & NOWOROL 1987] in which closest pairs of objects form kernels of succeeding clusters. Following objects are linked to these shells under condition that their distance to the gravity center of a cluster is by specified constant less than mean distance to the objects not belonging to the cluster. That results in distinguishing some clusters in the multi–characteristic space. We can call them as "morphological groups". The objects within these groups resemble each other more than objects belonging to other groups. The division is complete – every object must be assigned to one of the groups (in extreme case a group could include only one object) and no single object can belong to more than one group.

Description of algorithm

To begin the process of clustering one need to calculate matrix of distances, using selected measure of distance or similarity. Next we are searching for a pair of most similar objects in the matrix; i.e., the objects of the lowest distance measure or highest similarity measure. In the original algorithm these two objects compose a kernel of the first cluster, while in our modification only one of them is selected. Then the mean distance of all other objects to this selected one is calculated. If the ratio of the distance measure between first pair of objects to the average distance to other objects is lesser than assumed constant, the first cluster is created and includes two closest objects. In other case, only the first selected object remains in the cluster and we already know that each object will be clustered separately, i.e., in individual cluster. The aim of such modification is to avoid the creation of clusters containing "mess"; i.e., pairs of objects distant from each other, only for the reason that their distance coefficient is lesser than this of other objects not yet classified.

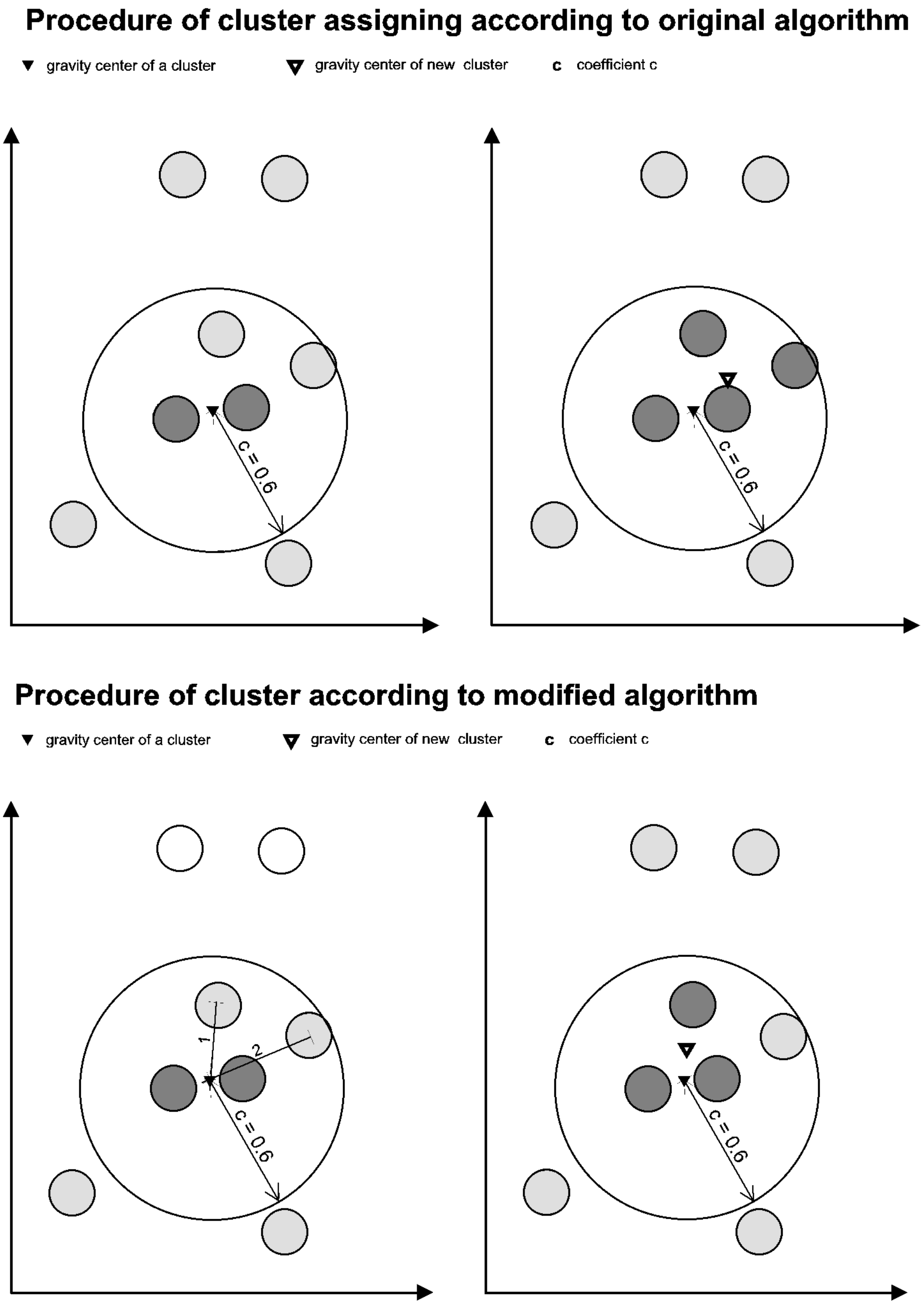
The next step is to calculate the mean distance between each object not assigned to the cluster and the gravity center of the cluster (this is mean distance of an object outside the cluster to each object assigned to the cluster separately) and next the mean distance between each object not assigned to the cluster and all other not assigned objects. The coefficient *c* is calculated – the ratio between these two means. At the beginning of the analysis, a threshold constant value of this coefficient has to be assumed. The recommended value is 0.6. In original algorithm all objects, for which the coefficient *c* is lesser than threshold value, are assigned to the cluster and this step is repeated until the cluster fills up, i.e., until the moment when there is no object outside the cluster, which fulfil the condition. In our modification only the object with lowest coefficient *c* is assigned to the cluster, of course if this value is lesser than the constant. Next the whole procedure starts again and this iteration endures until the cluster fills up. The purpose of such modification is to obtain more dense clusters – two objects very different from each other but both with average distance from the gravity center of the cluster cannot be assigned to the cluster simultaneously. In the modified algorithm only closer one will be assigned, the gravity center will move in its direction and the further one can be then distant enough to be rejected in the next step (Fig. 1). It considerably prolongs the time of analysis but the results, in our opinion, are more correct. When the cluster fills up, we search for next two closest objects from those, which were not assigned to a cluster. The procedure of assigning new objects is repeated. The calculated distances in this case are: mean distance to objects assigned to new cluster and mean distance to other objects, including these assigned to previous clusters. Finally, if there are no objects outside clusters, the procedure finishes and we obtain a division of series into a number of clusters.

Handy determinant of the level of division can be the following coefficient:

$$\Omega_p = \frac{2}{MN(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \tau_{pij}$$

where: *N* denotes the number of clusters, *M* – the number of variables (characteristics), while τ_{pij} – the number of differences in means between clusters *i* and *j* for specified characteristics, at the *p*–level of significance. This coefficient can assume values from 0 to 1. The value equal 1 denotes that the division is perfect, while the value equal 0 – that the clusters do not differ from each other.

Comparing two clusters with at least two objects assigned to each of them, Gosset's t–test can be used, which, although not recommended for too short



Procedure of cluster assigning according to original algorithm

▼ gravity center of a cluster ▽ gravity center of new cluster **c** coefficient c

Procedure of cluster according to modified algorithm

▼ gravity center of a cluster ▽ gravity center of new cluster **c** coefficient c

Fig. 1. Comparison of original and modified procedure of clustering in two-dimensional space

series, can be sufficient for diagnostic purposes. If we have one cluster of at least two objects and the second including only one object, the significance of the difference can be checked by standardization of characteristics of this single object to the mean and standard deviation for the characteristics of more numerous cluster. With an assumption of normal distribution (which can be tested, of course) we can determine whether this single object occurs adequately far away from the mean of more numerous cluster. Most difficult situation one meets when two single objects must be compared. In that case the only solution is the standardization of their characteristics with the means and standard deviations of whole series. Then, assuming the normal distribution, one can determine if they are sufficiently far away from each other on two wings of distribution to assume the difference between them is significant.

The concurrence of two cases of division of the same series cannot be determined with Pearson's χ^2, Cramer's V, nor Góralski's r_p, because one must expect many cells with frequencies lesser than 5 and seldom close to 0. For that reason we propose the following measure of concurrence, taking the values between 0 and 1:

$$r_m = \left(\sum_{i=1}^{N} x_{i,max} - \frac{\left(\sum x_i\right) - x_{i,max}}{M-1} \right) * \\ * \left(\sum_{j=1}^{M} x_{max,j} - \frac{\left(\sum x_j\right) - x_{max,j}}{N-1} \right) \Bigg/ \\ \Bigg/ \left(\sum_{i=1}^{N} \sum_{j=1}^{M} x_{i,j} \right)^2$$

where: $x_{i,max}$ and $x_{max,j}$ are maximal frequencies in a row and a column of a matrix (which is rectangular for the reason that the numbers of clusters can differ in different divisions), N is the number of columns while M – the number of rows or inversely. Value r_m equal 0 means that there is no concurrence between the results of two divisions, while the value equal 1 denotes that two divisions are identical. We propose the following partitions for the coefficient: lack of concurrence (0.0–0.1), weak concurrence (0.1–0.3), mean concurrence (0.3–0.5), strong concurrence (0.5–0.8), very strong concurrence (0.8–1.0).

The difference between modified and original algorithms concerns two characteristics. First, two closest objects, which are not assigned to any cluster, do not form the new cluster automatically – before the algorithm determines whether they are actually resembling each other. It prevents the appearance of "mess" clusters. Secondly, when new objects are assigned to a cluster, at each step only the closest object is taken into account, instead of all the objects for which the ratio of means distances is lesser than a threshold value. Also for that reason the clusters are more compact.

Some effects of these differences between two versions of algorithm were tested on 34 series of skulls excerpted from the HOWELLS' [1989] collection. It appeared that the modified algorithm is more rigorous but also more time-consuming. The simulation gives us the information that the coefficient $\Omega_{0.05}$ for modified algorithm was higher in 19 cases for 33 (58%) and the coefficient $\Omega_{0.001}$ in 17 cases for 33 (52%).

Selection of characteristics

Optimal selection of characteristics (one of key elements of analysis) can be obtained by the calculation of a matrix of their correlation in order to choose these, which are the most independent. In case there are too many most independent indices, these with lesser variability range should be excluded. They cannot be too numerous and selected ones must describe analyzed objects in best possible way – e.g., the morphology of skull in case of anthropological researches. What seems especially important is regarding of the descriptive features of skulls. This topic, however, exceeds limits of the present paper and will be discussed in separate article. Also the standardization and regard of sex differences must be included.

Standardization

Objects belonging to the compared series are described with few characteristics, i.e., indices of continuous value or cranioscopic descriptions expressed in ordinal or nominal variables. The indices used in anthropology are typified by different means, ranges and standard deviations. It implies that during the analysis they influence the result in various degrees (there is an assumed input weight, not defined by a researcher). The indices of higher mean and standard deviation have greater influence on the result than these of lower ones. Of course, we can state that height–lenght index is less important than the cranial index, but such statement needs expressed motive. The difference in weights should be a result of such motive and not of the natural difference in dimension. For that reason it seems to be clear that the standardization is necessary, while definition of weights for specified indices can take place later, e.g. if someone assumes that the measures of cranial and facial parts of skull are of equal importance, two cranial indices can have the weight of $^1/_2$ and three facial indices – the weight of $^1/_3$.

One should also regard the differences originating in sex dimorphism [PIONTEK 1985]. The preliminary simulation points out that height-length index is sex independent. Also three facial indices – Kollman’s, Virchov’s and morphological one – are independent or slightly dependent. Other examined indices are depended on sex: females have average shorter skulls, broader noses, higher orbits, and lower height-breadth index. For that reason, the standardization for females and males should be performed separately, with separate parameters of general population. Only in case of specified facial indices and the height-length index the unified parameters can be used.

Distance or similarity coefficient

A series of simulations revealed that surely the worst measure of distance for discussed algorithm and for anthropological purposes is Tschebyshev’s coefficient. Also the Euclidean and Czekanowski’s distances should be rejected for they results in distinguishing too many clusters. Amongst other measures the inversion of angular distance proves to be correct in case of complete data set and average square or cube of Euclidean distance (so-called Henzel’s distance) if the series contains broken objects, under condition that they are used for standardized data [LARO 1998, ZAKRZEWSKA 1987].

Interpretation of results

The distinguishing of clusters and test of the quality of division is only an introduction to the succeeding interpretation of results. The next stage of analysis is the calculation of frequencies of the morphological groups in the source series (so–called “fossil populations”). It reflects pretty well the internal structure of these populations (decidedly better than simply by the mean values of characteristics with standard deviations). Since even in case of very great data sets the number of groups rarely exceeds twenty, such frequencies are easy to present in form of box–and–whiskers diagrams, which, besides the frequency itself, shows also the level of confidence (Fig 2). It is very important because most frequently we deal with the series of various numbers. Because each object has been unequivocally assigned to one of morphological groups, there is also a possibility to count up other frequencies – e.g., by taking in account the geographical distribution setting the series from one region together, and so on. The diagram of frequencies allows us to compare “fossil populations” using Czekanowski’s diagraphical method or the method of dendrograms, in which the frequencies of specified morphological groups serve as the characteristics of populations.

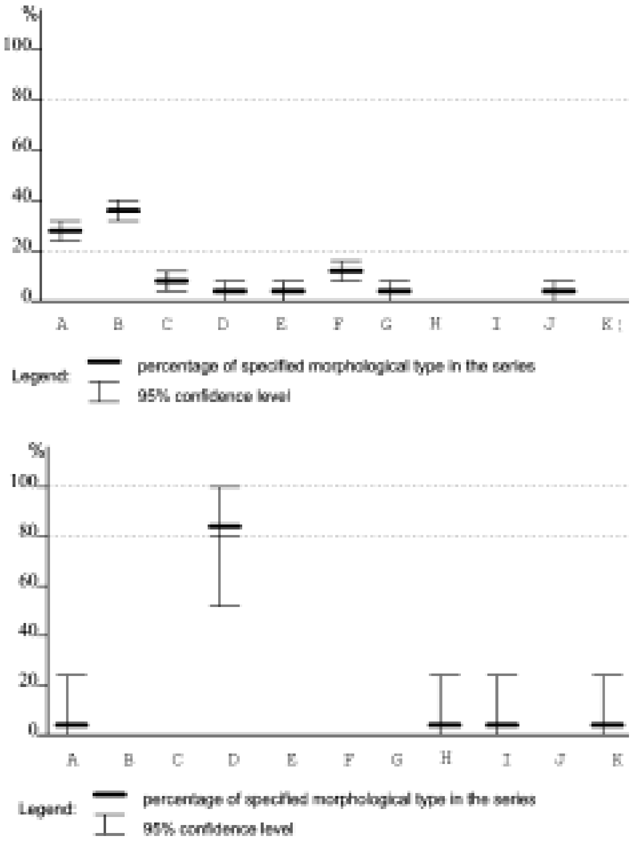


Fig. 2. Sample bar diagram for frequencies

Tools

Computer program was originally written in Pascal (DOS operating system, limitation to 600 objects), then the version for Windows 95 operating system came into existence in Visual Basic 5.0. The main technical problem concerns the time–consuming feature of the method (the clustering of 2200 objects defined with 5 characteristics lasts exactly one hour on the computer Pentium II 266 MHz). However, it seems that this feature can be acceptable, especially in the confrontation with very fast development of computer technology.

Conclusions

There are three main conclusions. Data analysis by the use of discussed algorithm allows us to examine very long series of objects (e.g., skulls) treated in individual way. The series of test analyses indicates the key importance of preliminary stages of analysis – data preparation, selection of distance measure. It seems that the flexibility of the method can probably lead to its use not only in craniometric researches and not only in anthropology.

References

HENKE W., 1991, *Biological distances in late Pleistocene and early Holocene human populations in Europe*, Variability and Evolution, **1**, 39-46

HOWELLS W.W, 1989, *Skull shapes and the map: Craniometric analyses in the dispersion of modern Homo*, Papers of the Peabody Museum, 79

LARO P.M., 1998, *Dobór współczynników podobieństwa w badaniach zbiorowisk roślinnych*, [in:] *Metody numeryczne w badaniach struktury i funkcjonowania szaty roślinnej*, E. Kazimierczak (ed.), Wyd. UMK, Toruń, 119 - 131

LYNCH M., 1989, *Phylogenetic hypotheses under the assumption of neutral quantitative-genetic variation*, Evolution, **43**, 1-17

MAREK T., C. NOWOROL, 1987, *Zarys analizy skupień. Niehierarchiczne i hierarchiczne techniki skupiania*, [in:] *Wielozmiennowe modele statystyczne w badaniach psychologicznych*, J. Brzeziński (ed.), PWN, Poznań, 184-204

MAYS S., 1998, *The archaeology of Human Bones*, Routledge, London

OVERALL J. E., C. J. KLETT, 1972, *Applied multivariate analysis*, McGraw Hill, New York

PIONTEK J, 1985, *Biologia populacji pradziejowych*, Wyd. Nauk. UAM, Poznań

ZAKRZEWSKA T., 1987, *O miarach podobieństwa obiektów i cech przydatnych w psychologicznych zastosowaniach analizy skupień*, [in:] *Wielozmiennowe modele statystyczne w badaniach psychologicznych*, J. Brzeziński (ed.), PWN, Poznań, 205-259

Streszczenie

Prowadzone przez nas prace dotyczą metod analizy zmienności fenotypowej populacji ludzkich za pomocą algorytmów grupowania danych. Problem bioróżnorodności i klasyfikacji jest ostatnio szeroko dyskutowany, zwłaszcza w świetle najnowszych osiągnięć genetyki. Analizy na poziomie DNA są jednak rzadko możliwe w przypadku materiałów kopalnych. Po pierwsze ze względu na trudności techniczne, po wtóre na znaczne koszty. Dlatego istotne znaczenie mogą mieć badania zmienności morfologicznej, łatwiejsze do przeprowadzenia i nie wymagające stosowania zaawansowanych technik laboratoryjnych.

Analizując opisane w literaturze algorytmy grupowania danych, zwróciliśmy uwagę na niehierarchiczny algorytm grupowania zupełnego, ze względu na brak ograniczających założeń wstępnych (dotyczących wiedzy na temat struktury analizowanej populacji i jej wielkości – ograniczeniem są tu jedynie możliwości sprzętu komputerowego). Oprogramowanie napisane zostało w języku Visual Basic 5.0 dla systemu Windows 95. Do testów praktycznych użyto serii danych kraniometrycznych z łatwo dostępnej serii Howellsa (2500 czaszek z różnych populacji). W procedurze grupującej najbliższe sobie pary obiektów tworzą jądra kolejnych grup, a następnie dołączane są do nich kolejne obiekty pod warunkiem, że ich odległość od punktu ciężkości grupy jest mniejsza o określony współczynnik od średniej odległości do obiektów pozostających poza skupieniem. W wyniku procesu grupowania wyróżnione zostają pewne skupie-

nia w przestrzeni wielocechowej, które można określić jako „grupy morfologiczne". Obiekty wewnątrz grup są do siebie bardziej podobne niż do obiektów należących do innych grup. Podział jest zupełny – każdy obiekt zostaje przydzielony do jakiejś grupy (w skrajnym przypadku jest to grupa składająca się z jednego obiektu), żaden nie należy do więcej niż jednej z nich. Wydzielenie grup i sprawdzenie jakości podziału jest jedynie wstępem do dalszej interpretacji wyników. Kolejnym etapem analizy jest obliczenie frekwencji występowania poszczególnych grup morfologicznych w seriach źródłowych („populacjach kopalnych"), co dość dobrze odzwierciedla wewnętrzną strukturę tych populacji (z pewnością lepiej niż średnie wartości cech z odchyleniami standardowymi).

Prezentowany algorytm pozwala na badanie dużych serii indywidualnie traktowanych obiektów (np. czaszek). Serie przeprowadzonych analiz wskazują na kluczową rolę wstępnych etapów analizy – przygotowania danych, wyboru funkcji odległości. Wydaje się, że elastyczność metody pozwoli na zastosowanie jej nie tylko w badaniach kraniometrycznych i nie tylko w antropologii.