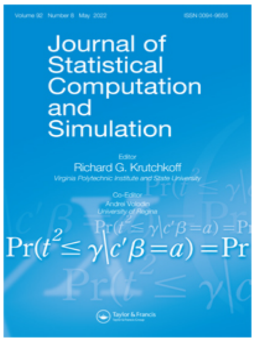

Journal of Statistical Computation and Simulation

Richard G. Krutchkoff

Taylor & Francis
Taylor & Francis Group

Journal of Statistical Computation and Simulation

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/gscs20

Simulation methods for squared Euclidean and Mahalanobis type distances for multivariate data and their application in assessing the uncertainty in hierarchical clustering

Efthymia Nikita & Panos Nikitas

To cite this article: Efthymia Nikita & Panos Nikitas (2022): Simulation methods for squared Euclidean and Mahalanobis type distances for multivariate data and their application in assessing the uncertainty in hierarchical clustering, Journal of Statistical Computation and Simulation, DOI: 10.1080/00949655.2022.2030338

To link to this article: https://doi.org/10.1080/00949655.2022.2030338

Published online: 02 Feb 2022.

JOURNAL OF STATISTICAL COMPUTATION AND SIMULATION
https://doi.org/10.1080/00949655.2022.2030338

Simulation methods for squared Euclidean and Mahalanobis type distances for multivariate data and their application in assessing the uncertainty in hierarchical clustering

Efthymia Nikita [a] and Panos Nikitas[b]

[a]Science and Technology in Archaeology and Culture Research Centre, The Cyprus Institute, Aglantzia, Cyprus; [b]Department of Chemistry, Aristotle University of Thessaloniki, Thessaloniki, Greece

ABSTRACT

This paper extends the Suzuki and Shimodora method [Suzuki R, Shimodora H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics. 2006;22:1540–1542] for assessing uncertainty in hierarchical cluster analysis to multivariate datasets and examines the reliability of the simulated cluster probabilities in relation to the simulation method adopted. The extension is applied to squared Euclidean and Mahalanobis-type distances and employs three simulation methods, the Monte-Carlo and bootstrap methods, and a new proposed method, the distance distribution. The distance distribution method is very fast and gives satisfactory predictions for the distance, its standard deviation, skewness, and kurtosis. The performance of the Monte-Carlo method is equally satisfactory, whereas the bootstrap method usually gives acceptable predictions only for the distance. The distance distribution and Monte-Carlo methods give similar cluster probabilities. The bootstrap probabilities are not very different; thus this method can be used in datasets of unknown distance distribution.

ARTICLE HISTORY
Received 12 July 2021
Accepted 13 January 2022

KEYWORDS
Biodistance estimation; cluster analysis; Euclidean distance; Mahalanobis distance

Introduction

Hierarchical cluster analysis (HCA) [1–3] is a popular iterative technique for classifying objects based on their (dis)similarities. Initially, each data point is considered as an individual cluster and then, at each iteration, based on a (dis)similarity measure, clusters merge with other clusters until one cluster is formed. There are several similarity/dissimilarity measures; the former quantify the similarity, while the latter quantify the difference between two objects. The output is a tree-like diagram, a dendrogram. Thus, a dendrogram visualizes the hierarchical relationship between objects and the presence of clusters in a dataset. If, in addition, a dendrogram can assess the uncertainty in the clustering process, then it enables the estimation of the probability of formation of the various clusters and, therefore, their stability.

CONTACT Efthymia Nikita ✉ e.nikita@cyi.ac.cy Science and Technology in Archaeology and Culture Research Centre, The Cyprus Institute, Aglantzia, Nicosia 2121, Cyprus

Supplemental data for this article can be accessed here. https://doi.org/10.1080/00949655.2022.2030338

Suzuki and Shimodora [4] proposed a bootstrap algorithm to assess uncertainty in HCA, which is freely available via the pvclust package of the R language. In general, in a simulation approach, the cluster probabilities in hierarchical cluster analysis may be approximated from the percentage of the appearance of the various dendrogram patterns in a large simulated dataset. Therefore, the basic steps to assess uncertainty in the dendrogram of hierarchical cluster analysis are the following. Given a dataset of samples, first all pairwise distances among the samples are estimated and then a large dataset of simulated distances is generated. Subsequently, all the dendrograms of the original and simulated distances are estimated and the number that each pattern in the original dendrogram appears in the simulated distances is counted. The uncertainty in a pattern is assessed from the percentage of the appearance of this pattern in the simulated dataset, which gives the simulated probability of the formation of this pattern. Therefore, the assessment of uncertainty in HCA requires the creation of simulated distances.

The simplest approach to generate simulated distances is via the Monte-Carlo or the bootstrap methods. In these methods, we first generate simulated samples and then these samples are used to compute simulated distances. The Monte-Carlo method presumes that we know the underlying statistical distributions in the data, whereas bootstrapping is free from distributional assumptions. However, both usually require long computational time. In the present paper we propose a new, very fast, approach that directly simulates the distance distribution, provided that this distribution is known. Thus, it can be used to generate simulated distances without the creation of artificial datasets. The performance of these three simulation methods, i.e. the Monte-Carlo (MC), the bootstrap (B), and the distance distribution (DD) methods, is tested using squared Euclidean and Mahalanobis type distances. Special attention is paid to expressions of Euclidean or Mahalanobis distances that are unbiased estimators of population divergence since such distances are of primary interest in many biological/anthropological studies. Finally, the Suzuki-Shimodora [4] approach to assess uncertainty in HCA is extended to multivariate datasets and the impact of the performance of a simulation method on the accuracy of the estimated cluster probabilities is examined.

Theoretical part

Simulation methods for distances and their application in HCA

The simulation of a distance measure, d, between two samples generates a bunch of artificial distances that have the same or a similar distribution as the distribution of d. Thus, the simulated distances should have at least the first four moments, mean, variance, skewness, and kurtosis, as close as possible to those of the original distance. Therefore, to assess the performance of a simulation method, we should compare the simulated first four moments to the corresponding moments calculated on the original dataset. However, at this point we should clarify the following. Due to the asymmetry of the distance distribution, the distance is not equal to its first moment. Therefore, we should either transform the distance to its first moment or vice versa. Here, we have chosen the latter approach and the relationships used to connect a certain distance to its first moment are presented and discussed below.

As already mentioned in the Introduction, in the present paper we examined three simulation methods; the conventional Monte-Carlo and bootstrap methods, and a new proposed simulation based on the distance distribution. Consider g multivariate samples, i.e. samples consisting of r variables. We assume that these samples come from multivariate normal or binary populations depending on whether the samples consist of continuous or binary data. The squared Euclidean and Mahalanobis type distances that can be calculated between pairs of samples can be expressed in matrix notation as a quadratic form [5], i.e. in the form $x^T Ax$, where x is a column vector of r random variables, A is an r-dimensional symmetric matrix, and T denotes the transpose matrix. Note that when a quadratic form is used to express a squared Euclidean or Mahalanobis type distance, the values of vector x are not sample values. If the dataset consists of continuous variables, the values of x are the differences in the mean values per variable between two samples. For datasets of binaries, the values of x can be the differences in the proportions or proper transformations of the proportions per variable between two samples. A basic property of a quadratic form is that its expected value is given from [5]:

$$E[x^T Ax] = \mu^T A\mu + \mathrm{tr}(A\Sigma) \tag{1}$$

where $\mu = E[x]$ and $\Sigma = \mathrm{Cov}[x]$ are the expected value and covariance matrix of x, respectively, and tr denotes the trace of a matrix.

The conventional Monte-Carlo and bootstrapping methods can be performed straightforwardly. In the Monte-Carlo method, based on the mean values per variable and sample and the corresponding sample variances or the pooled covariance matrix of all samples, we generate the populations from which the original samples have been drawn. Then samples are randomly drawn from the populations, pairwise distances are calculated and the first four moments of their distribution are estimated. In bootstrapping, the simulated samples are produced by resampling with replacement. Thus, there is no need to generate simulated populations; the simulated distances are computed directly on the simulated samples. Finally, in the simulation based on the distance distribution, for the simulated distances we adopted the following relationship:

$$d(\mathrm{sim.DD}) = E[d] + s(X - \mu)/\sigma \tag{2}$$

where $E[d]$ is the expected value of the original distance measure d, s is the standard deviation of d, X is a random value generated from the distance distribution and μ and σ are the mean value and the standard deviation of this distribution. Equation (2) has been adopted since, due to the term $\mathrm{s}(X-\mu)/\sigma$, the simulated distances, d(sim.DD), follow the distance distribution with standard deviation s, whereas the term $E[d]$ shifts the expected simulated value to the $E[d]$ value. Note that for the application of this approach, the distance distribution and the standard deviation s of d are needed.

In the subsections below the most important squared Euclidean and Mahalanobis type distances and their unbiased estimators for population divergence between multivariate samples are presented. For each of them the specific expression of Equation (2) is developed. The variance, skewness, and kurtosis of the distances discussed here are presented as supplementary material in the file: *Supplement-1-Moments*. These are needed for testing the performance of the simulations as well as for the application of the DD method via Equation (2).

As soon as a distance measure is properly simulated, we can extend the Suzuki-Shimodora [4] method to assess uncertainty in HCA in a multivariate dataset. Figure 1 shows schematically the proposed approach. We first estimate the selected distance measure d, its standard deviation (sd), skewness (sk), and kurtosis (ku) in the original dataset and apply HCA to determine the dendrogram visualizing the clusters in this dataset. Next, using one or more of the MC, DD, and B simulation methods, we create simulated distances ($\sim$ 5000) and from their distribution we estimate the simulated d, sd, sk, and ku values. Now by comparing the original d, sd, sk, and ku values to the simulated ones, using either plots or quantitative measures such as the Root Mean Square Percentage Error (RMSPE) defined below, the most appropriate simulation method can be selected, and its simulated distances can be used to estimate cluster probabilities. In case the distance distribution is unknown, or the dataset does not fulfil the requirements that are necessary for a reliable estimation of the distance moments, this comparison does not provide any useful information and the cluster probabilities should be based on the bootstrap method.

Squared Euclidean and corrected squared Euclidean distances

Consider two multivariate samples, originating from two populations. Each sample consists of r continuous variables and n_1 and n_2 observations, respectively. If missing values are included, the number of observations per sample and variable i will be n_{1i} and n_{2i}. Between the centroids of these two samples, the squared Euclidean distance (ED) is defined from:

$$\mathrm{ED} = (\bar{x}_1 - \bar{x}_2)^T(\bar{x}_1 - \bar{x}_2) \tag{3}$$

where $\bar{x}_1, \bar{x}_2$ are the vectors of the mean values of samples 1 and 2, respectively. According to Equation (1), the expected value of ED is equal to the squared Euclidean distance between the populations, $\mathrm{ED}_p = (\mu_1 - \mu_2)^T(\mu_1 - \mu_2)$, where μ_1, μ_2 are the vectors of the mean values of populations 1 and 2, respectively, plus the trace $\mathrm{tr}(\mathrm{Cov}[\bar{x}_1 - \bar{x}_2])$, which is given by:

$$\mathrm{tr}(\mathrm{Cov}[\bar{x}_1 - \bar{x}_2]) = \sum_{\mathrm{i}=1}^{\mathrm{r}} \mathrm{Var}[\bar{x}_{1i}] + \mathrm{Var}[\bar{x}_{2i}] = \sum_{i=1}^{\mathrm{r}} \frac{\sigma_{1i}^2}{n_{1i}} + \frac{\sigma_{2i}^2}{n_{2i}} \tag{4}$$

since $\mathrm{Var}[\bar{x}_{1i}] = \sigma_{1i}^2/n_{1i}$ and $\mathrm{Var}[\bar{x}_{2i}] = \sigma_{2i}^2/n_{2i}$, where $\sigma_{1i}^2, \sigma_{2i}^2$ are the variances of variable i in populations 1 and 2. Note that in the simulation process the squared Euclidean distance between the populations, ED_p, is practically equal to the squared Euclidean distance calculated in the original dataset, ED, because for the generation of populations we used the unbiased estimators mean values and variances of the original samples and, thus, these quantities are practically common in the samples and the corresponding populations. Therefore, when applying the Monte-Carlo or the Bootstrap method, the estimated first distance moment of the simulated values, E[ED(sim.MC/B)], can be used to estimate the simulated squared Euclidean distance, ED(sim), by subtracting from E[ED(sim.MC/B)] the trace given from Equation (4). That is:

$$\mathrm{ED(sim)} = E[\mathrm{ED(sim.MC/B)}] - \mathrm{tr}(\mathrm{Cov}[\bar{x}_1 - \bar{x}_2] \tag{5}$$

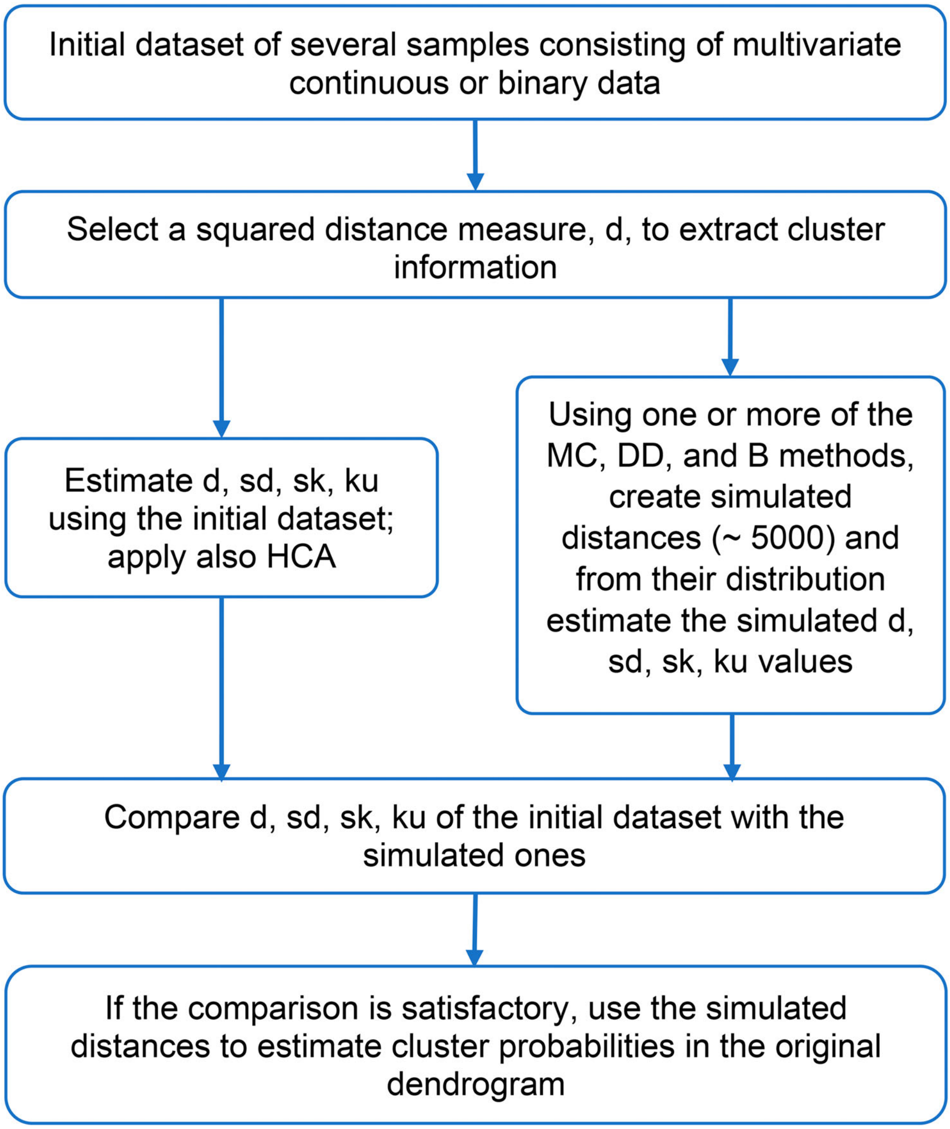
Initial dataset of several samples consisting of multivariate continuous or binary data
Select a squared distance measure, d, to extract cluster information
Estimate d, sd, sk, ku using the initial dataset; apply also HCA
Using one or more of the MC, DD, and B methods, create simulated distances (~ 5000) and from their distribution estimate the simulated d, sd, sk, ku values
Compare d, sd, sk, ku of the initial dataset with the simulated ones
If the comparison is satisfactory, use the simulated distances to estimate cluster probabilities in the original dendrogram

Figure 1. Schematic representation of the steps followed for the estimation of cluster probabilities in multivariate datasets.

At this point, we should clarify that although the squared Euclidean distance does not account for correlations among the variables, it can be applied to correlated data. However, the expressions for the distance moments presented in *Supplement-1-Moments* are strictly valid when the variables are independent. This limitation should be taken into consideration in the calculation of the variance, skewness, and kurtosis as well as in the implementation of the MC method.

To apply the DD method, Equation (2), we may take into account that $\bar{x}$ follows at least asymptotically the normal distribution with mean value and variance equal to μ and σ^2/n, respectively, where μ, σ are the mean value and the standard deviation of the population distribution and n is the sample size. Therefore, the difference $\bar{x}_{1i} - \bar{x}_{2i}$ also follows the normal distribution with mean value and variance equal to $\mu_{1i} - \mu_{2i}$ and $\sigma_{1i}^2/n_{1i} + \sigma_{2i}^2/n_{2i}$, respectively, which yields:

$$\frac{(\bar{x}_{1i} - \bar{x}_{2i})^2}{\sigma_{1i}^2/n_{1i} + \sigma_{2i}^2/n_{2i}} \sim \chi^2(1, \lambda_i) \tag{6}$$

where $\chi^2(1, \lambda_i)$ is the noncentral chi-square distribution with 1 degree of freedom and λ_i is the noncentrality parameter given by:

$$\lambda_i = \frac{(\mu_{1i} - \mu_{2i})^2}{\sigma_{1i}^2/n_{1i} + \sigma_{2i}^2/n_{2i}} \tag{7}$$

Therefore, for the application of Equation (2), we may use the following expression under the assumption that variables are independent:

$$\text{ED(sim.DD)} = \text{ED} + \text{tr}(\text{Cov}[\bar{x}_1 - \bar{x}_2]) + \sum_{i=1}^{r} s_i\{\, \chi^2(1{,}\lambda_i) - 1 - \lambda_i\}/\sqrt{2(1 + 2\lambda_i)} \tag{8}$$

where ED is the squared Euclidean distance in the original dataset and s_i is the standard deviation of the squared Euclidean distance when it is computed on variable i. It can be estimated from the square root of term i of Equation (S1–18) in *Supplement-1-Moments*. For the simulated squared Euclidean distance according to the DD method, we can use again Equation (5) after replacing the first moment obtained from Monte-Carlo or bootstrapping by that obtained from the data generated from Equation (8). This easily arises from Equation (8) if we calculate the expected value of both sides. Note that in Equation (8), $\chi^2(1, \lambda_i)$ denotes a random value generated from the noncentral chi-squared distribution with 1 degree of freedom and λ_i noncentrality parameter. For this distribution, the mean value is equal to $1+\lambda_i$ and the standard deviation is $\sqrt{2(1 + 2\lambda_i)}$. When a random variable is multiplied by a constant, then both its mean and standard deviation are multiplied by this constant and for this reason the quantity $\sigma_{1i}^2/n_{1i} + \sigma_{2i}^2/n_{2i}$ does not appear in Equation (8).

The Euclidean distance is a distance with a wide range of applications. However, it is not an unbiased estimator of population divergence since its expected value is not equal to the Euclidean distance between populations. From the treatment presented above, this property can be incorporated into the Euclidean distance if it is corrected as follows:

$$\text{cED} = (\bar{x}_1 - \bar{x}_2)^T(\bar{x}_1 - \bar{x}_2) - \text{tr}(\text{Cov}[\bar{x}_1 - \bar{x}_2]) \tag{9}$$

where the trace is still given by Equation (4). For simulated cED distances based on the MC, B or DD methods, we may simply use the expression:

$$\text{cED(sim)} = \text{ED(sim)} - \text{tr}(\text{Cov}[\bar{x}_1 - \bar{x}_2]) \tag{10}$$

Finally, we should clarify that in many expressions above the population quantities $\mu_{1i}, \mu_{2i}, \sigma_{1i}^2, \sigma_{2i}^2$ appear. If these quantities are unknown, they may be approximated from the corresponding sample mean and variance since the sample mean and variance are unbiased estimators of the corresponding population statistics. The same holds for all relevant expressions presented in this section.

Squared Mahalanobis and corrected squared Mahalanobis distances

The Euclidean distances allow for missing values, but they do not consider differences in the dispersion of the sample points in space. A first approach to account for the distribution of sample points is to normalize the ED, cED distances. There are several ways to normalize an Euclidean distance. However, a better distance measure that assumes an anisotropic Gaussian distribution of the sample points is the Mahalanobis distance (MD).

For the definition of the Mahalanobis distance, consider g multivariate normal populations with a common covariance matrix, $C = C_1 = C_2 = \ldots = C_g$. Then the Mahalanobis distance between populations 1 and 2 is defined from [6–8]:

$$\mathrm{MD_{pop}} = (\mu_1 - \mu_2)^T C^{-1} (\mu_1 - \mu_2) \tag{11}$$

This expression can be used to estimate the corresponding sample Mahalanobis distance provided that μ_1, μ_2 are replaced by $\bar{x}_1, \bar{x}_2$ and the matrix C is known. If C is unknown, it may be replaced by the pooled covariance matrix S, which is an unbiased estimator of C. To distinguish these two expressions of sample Mahalanobis distance, we will denote them by MD1 when C is known, and MD2 when S is used.

The major disadvantage of the sample Mahalanobis distances, MD1 and MD2, in biological and anthropological studies is that they are not unbiased estimators of population divergence. For the MD1 we can define an unbiased estimator from:

$$\mathrm{cMD1} = \mathrm{MD1} - \mathrm{rf} = (\bar{x}_1 - \bar{x}_2)^T C^{-1} (\bar{x}_1 - \bar{x}_2) - \mathrm{rf} \tag{12}$$

where $f = 1/n_1 + 1/n_2$. Equation (12) arises if we apply Equation (1) to MD1 and take into account that $\mathrm{tr}(\mathrm{A}\Sigma) = \mathrm{tr}(C^{-1}\Sigma) = \mathrm{rf}$ because Σ is equal to Cf (see Equation 6–22 in ref. [6]).

Note that for the MD1 and cMD1 the covariances among the variables are taken into account in the calculations. For this reason, in the Monte-Carlo method the simulated multivariate normal populations should be generated based on the mean values per variable and sample and the pooled covariance matrix of the original dataset. As in the case of ED, when applying the Monte-Carlo or the Bootstrap method, the estimated first distance moment is used to compute the simulated MD1 from:

$$\mathrm{MD1(sim)} = \mathrm{E[MD1(sim.MC/B)]} - \mathrm{rf} \tag{13}$$

For the simulated cMD1 distances, we can use the relationship:

$$\mathrm{cMD1(sim)} = \mathrm{MD1(sim)} - \mathrm{rf} \tag{14}$$

For simulations using the DD method, we can make use of the following theorem of a quadratic form [5]: If x is a multivariate normal random vector with a mean vector μ and a covariance matrix Σ, i.e. $x \sim N_r(\mu, \Sigma)$, and$A\Sigma$ is idempotent, then the quadratic form $x^T A x$ follows the noncentral chi-square distribution, $\chi^2(p, \lambda)$, with $p = \mathrm{rank}\,(A\Sigma)$ degrees of freedom and noncentrality parameter $\lambda = \mu^T A \mu$. In the present case, this theorem yields:

$$\mathrm{MD1}/f \sim \chi^2(r, \lambda) \tag{15}$$

where $\lambda = (\mu_1 - \mu_2)^T C^{-1}(\mu_1 - \mu_2)/f$ since $A = C^{-1}/f$, $\Sigma = \mathrm{C}f$ and $A\Sigma = 1$ is idempotent. Therefore, for simulations based on the distance distribution, we may use the following expression:

$$\mathrm{MD1(sim.DD)} = \mathrm{MD1} + \mathrm{rf} + s(\chi^2(r, \lambda) - r - \lambda)/\sqrt{2(1 + 2\lambda)} \tag{16}$$

where MD1 is the original distance and *s* is the standard deviation of MD1. The latter may be estimated from the square root of Equation (S1-22) in *Supplement-1-Moments*. For the corresponding simulated cMD1 distances, we can use Equation (14).

The MD2 distance does not follow the chi-square distribution. It has been shown that the distribution of MD2 is related to the non-central F statistic [9–11] (for details see *Supplement-1-Moments*). Its unbiased estimator may be expressed as [11,12]:

$$\mathrm{cMD2} = q \cdot \mathrm{MD2} - \mathrm{rf} \tag{17}$$

where $q = (N - g - r - 1)/(N - g)$. The Monte-Carlo and bootstrap simulations can be performed such as in the case of the MD1. However, taking into account Equation (S1-29), the simulated MD2 and cMD2 are estimated from:

$$\mathrm{MD2(sim)} = q \cdot E[\mathrm{MD2(sim.MC/B)}] - \mathrm{rf} \tag{18}$$

$$\mathrm{cMD2(sim)} = q \cdot E[\mathrm{cMD2(sim.MC/B)}] - \mathrm{rf} \tag{19}$$

For simulations based on the DD method, we can use Equation (2) where now X is a random value generated from the noncentral F distribution with $v_1 = r$ and $v_2 = N\text{-}g\text{-}r{+}1$ degrees of freedom and λ the noncentrality parameter given from $\lambda = \mathrm{MD2}/f$ [9–11]. In particular, the expressions that should be used are:

$$\mathrm{MD2(sim.DD)} = (MD2 + \mathrm{rf})/q + s_{\mathrm{MD2}}(F_{v_1,v_2}(r, \lambda) - \mu)/\sigma \tag{20}$$

$$\mathrm{cMD2(sim.DD)} = (\mathrm{cMD2} + \mathrm{rf})/q + s_{\mathrm{cMD2}}(F_{v_1,v_2}(r, \lambda) - \mu)/\sigma \tag{21}$$

where s_{MD2}, s_{cMD2} can be estimated via Equation (S1–28) and μ and σ are the mean value and the standard deviation of the F distribution given from Equation (S1–26) and the square root of Equation (S1–27), respectively. The simulated MD2 and cMD2 via the DD method may be estimated from Equations (18), (19), where the expected values computed from the MC and B methods are replaced by those computed from the DD method.

Mean measure of divergence

The mean measure of divergence (MMD) is a distance exclusively for binary data. This measure was introduced by statistician C.A.B. Smith for use by Grewal [13] and it may be expressed as [14–16]:

$$\mathrm{MMD} = \frac{1}{r}\sum_{i=1}^{r}\{(\theta_{1i} - \theta_{2i})^2 - \frac{1}{n_{1i}} - \frac{1}{n_{2i}}\} \tag{22}$$

where n_{1i} is the number of individuals from sample 1 in which the presence of trait i is examined, n_{2i} is the number of individuals from sample 2 in which the presence of trait i

is examined, and θ_{1i} and θ_{2i} denote the transformed frequencies of each trait per sample. There are various methods for the estimation of the transformed frequencies, such as the arcsine transformations suggested by Freeman-Tukey or Anscombe [17]. A complete theory of this distance measure is presented by Sjøvold [14], where it is shown that the MMD is an unbiased estimator for the squared difference between the angular transformations of the population incidences provided that there are no traits (variables) that exhibit a particularly high (>0.95) or low (<0.05) frequency within one or more samples. For this reason, such traits are removed before the calculation of the MMD. The MMD is, in fact, a type of the cED distance, where its value is averaged by division by r, the mean sample values $\bar{x}_{1i}, \bar{x}_{2i}$ are replaced by the transformed frequencies θ_{1i}, θ_{2i} and the trace term is given by the sum of variances:

$$\mathrm{Var} = \frac{1}{r}\sum_{i=1}^{r}\{\frac{1}{n_{1i}} + \frac{1}{n_{2i}}\} \tag{23}$$

For the application of the Monte-Carlo method, populations of binary data can be created using multivariate binary variates of known marginal probabilities. For the marginal probabilities, we may use the φ values of the original samples, since φ is an unbiased estimator, or we may approximately use their adjustment $(\varphi + 3/(8n))/(1 + 3/(4n))$ when the Anscombe transformation is used. However, we found that these two approaches lead to similar results and, therefore, for simulations we may use either of them. Since the MMD is a type of the cED, the algorithm used for Monte-Carlo and bootstrapping simulations for the cED can be also adopted for the MMD.

For simulations using the DD method, we can take into account that θ follows the normal distribution with mean value and variance equal to Θ (i.e. the value of θ in the population) and 1/n, respectively, and that in the MMD all r traits are assumed to be independent [14]. Therefore, Equation (8) is transformed to the following expression:

$$\mathrm{MMD(sim.DD)} = \mathrm{MMD} + \mathrm{Var} + \sum_{i=1}^{r} s_i\{\chi^2(1, \lambda_i) - 1 - \lambda_i\}/\sqrt{2(1 + 2\lambda_i)} \tag{24}$$

where s_i is given by the square root of Equation (S1-31) and the noncentrality parameter λ_i can be estimated from:

$$\lambda_i = \frac{(\Theta_{1i} - \Theta_{2i})^2}{1/n_{1i} + 1/n_{2i}} \tag{25}$$

where the population parameter Θ can be approximated from the corresponding sample parameter θ. From Equation (24) we readily obtain that the simulated MMD via the DD method may be calculated from:

$$\mathrm{MMD(sim)} = \mathrm{E[MMD(sim.DD)]} - \mathrm{Var} \tag{26}$$

Corrected squared Euclidean distance for binary data

If the corrected Euclidean distance is applied directly to the proportions of the original binary data, we obtain the distance [18]:

$$\mathrm{UMD} = \sum_{i=1}^{r} \{(\varphi_{1i} - \varphi_{2i})^2 - \frac{p_{1i}(1-p_{1i})}{n_{1i}} - \frac{p_{2i}(1-p_{2i})}{n_{2i}}\} \tag{27}$$

where φ_{1i} is the proportion of variable i of sample 1 having the trait in question and, similarly, φ_{2i} is the corresponding proportion for sample 2. Note that in the case of binary data we have $E[\varphi] = p$ and $\mathrm{Var}(\varphi) = p(1\text{-}p)/n$, where p is the proportion of the population having the trait in question. It is seen that the distance is similar to the MMD with the trace term given by:

$$\mathrm{Var} = \sum_{i=1}^{r} \left\{ \frac{p_{1i}(1-p_{1i})}{n_{1i}} + \frac{p_{2i}(1-p_{2i})}{n_{2i}} \right\} \tag{28}$$

For the simulated distances based on the Monte-Carlo method or bootstrapping, we can proceed as in the MMD. Since the UMD is an unbiased estimator of population divergence in the entire range of φ values, i.e. from $\varphi = 0$ to $\varphi = 1$ [18], for the creation of populations of binary data, we can use as marginal probabilities the sample φ values. For simulations based on the distance distribution, Equation (24) becomes:

$$\mathrm{UMD(sim.DD)} = \mathrm{UMD} + \mathrm{Var} + \sum_{i=1}^{r} s_i\{\chi^2(1,\lambda_i) - 1 - \lambda_i\}/\sqrt{2(1+2\lambda_i)} \tag{29}$$

where s_i is given by the square root of Equation (S1–34) and λ_i can be estimated from:

$$\lambda_i = \frac{(p_{1i} - p_{2i})^2}{p_{1i}(1-p_{1i})/n_{1i} + p_{2i}(1-p_{2i})/n_{2i}} \tag{30}$$

using in the calculations φ in place of p.

Table S1 of the file *Supplement-2-Tables* summarizes the conditions used for the simulation study of each distance measure.

Materials and methods

Materials

To test the performance of the simulations and the uncertainty in hierarchical clustering in relation to the distances studied in this paper, real and simulated datasets were used. The real datasets were obtained from the literature and, in particular, from the following three sources: (a) The dataset of Egyptian skulls that can be found in Thomson and Randall-MacIver [19], and in online domains such as https://www3.nd.edu/∼busiforc/handouts/Data%20and%20Stories/regression/egyptian%20skull%20development/EgyptianSkulls.html or from the file ‘skulls’ of the R package HSAUR; (b) the Howells Craniometric Data Set, compiled by Dr. William Howells [20–23] and made available online by Dr Benjamin Auerbach at https://web.utk.edu/∼auerbach/

HOWL.htm; and (c) the Ossenberg database of cranial nonmetric traits [24], freely available online at: http://library.queensu.ca/webdoc/ssdc/cntd.

The first dataset (EG1) consists of 150 cases of four measurements of male Egyptian skulls from 5 different time periods, where thirty skulls are measured from each time period. To create a dataset with different sample sizes, we have selected the first 5, 10, 15, 20, and 30 cases of EG1 and named this dataset EG2.

The William Howells Craniometric Data Set consists of 82 measurements (variables) obtained from 2524 human crania from 30 populations. From this dataset, we have removed 11 variables with missing values and selected the male subset of 12 Pacific populations (dataset WH1). These populations comprise the Moriori, Maori South, Maori North, Mokapu, Easter Island, Ainu, Guam, Japan North, Japan South, Atayal, Hainan, and Philippines. Geographically, the first five groups are classified as 'Pacific' (P) and the latter seven as 'Western Pacific' (WP). This dataset consists of 484 individuals and has a q value equal to 0.85. The q value affects the shape of the distribution of the Mahalanobis distances, hence it is interesting to examine the effect of this parameter on the simulation of the Mahalanobis distances MD2 and cMD2. Note that the q value in all datasets of Egyptian skulls (EG1-EG2) is greater than 0.9, thus these datasets cannot be used per se for the examination of the abovementioned effect. For this reason, we have created a subset of the WH1 dataset with a significantly low value of q as follows. We selected the first 9 samples of WH1 and from each sample we selected the first 5–30 individuals. The created dataset, WH2, consisted of 130 individuals with $q = 0.4$.

The Nancy Ossenberg database includes cranial nonmetric traits from the Arctic and Northwestern North America as well as Northeast Asia, Eurasia, Africa, and the South Pacific [24]. From this database, we created a dataset of male individuals consisting of 11 samples of 574 individuals, 5 from Africa with 195 individuals and 6 from Eurasia with 379 individuals. From the Eurasia populations, the samples Bavaria, Czechoslovakia, Europe, France, Germany, and Russia have been removed because of their very small size. The number of cranial traits in the Ossenberg database is 69. However, 15 of them were eliminated because they were scored on an ordinal scale from 0 to 2 or 3. In addition, we removed from the dataset another 10 variables due to many missing values. From this dataset (NO1) of 44 traits, two subsets with 31 (NO2) and 22 (NO3) traits were also examined after data editing to remove traits with $p \leq 0.01$ and $p \leq 0.02$, respectively. Finally, we created a large dataset (NO4) consisting of 44 traits and 3838 male individuals from all over the world grouped into 5 samples: Africa (AF), Northeast Asia (AS), Eurasia (EU), Native America and Greenland (NAG), and South Pacific (SP). This dataset was also examined after data editing to remove intercorrelated traits, a process that resulted in a dataset of 21 variables (NO5).

The simulated datasets were created to follow the multivariate normal distribution with mean values those of the real dataset EG1 and containing correlated and uncorrelated continuous variables. For correlated variables, the covariance matrix of the dataset EG1 was used. In particular, we created four datasets, SIM1 to SIM4. In the first two, SIM1, SIM2, the variables are independent and sample sizes are equal to those of the EG1 and EG2 datasets, respectively. In the last two, SIM3, SIM4, the variables are correlated and the sample sizes are equal to the corresponding sizes of SIM1, SIM2 and EG1, EG2, i.e. (30, 30, 30, 30, 30) and (5, 10, 15, 20, 30) cases, respectively. With these simulated data we can test how the proposed methods work, i.e. if there are deviations and their extent when we work on real data in relation to data that follow the 'ideal' multivariate normal distribution.

At this point, we should clarify the following. The first four datasets consist of continuous variables that measure the linear distance between several cranial landmarks. These data are analysed almost exclusively using the Mahalanobis distance corrected for small sample sizes [11,25]. For getting statistical inference, such as statistical significance [11] and confidence intervals [10], this distance presumes that the data follow the multivariate normal distribution. Therefore, these datasets as well as all the simulated SIM1 to SIM4 datasets are appropriate for testing the MD1, cMD1, MD2, and cMD2 distances. The EG1 and EG2 datasets consist of small samples with just 4 variables. For this reason, the statistically significant correlations among the variables are very limited and, therefore, they can be also used, along with the SIM1 and SIM2 datasets, to analyse the ED and cED distances. Finally, the binary datasets NO1 to NO5 include samples with edited or non-edited frequencies or intercorrelations designed to test the MMD and UMD. For the selection, properties, and biological importance of the traits/variables in binary datasets, see [14].

Details of the datasets used in the current study are presented as Supplementary Material in Table S2 of the file *Supplement-2-Tables*. The complete datasets are available from the authors upon request.

Software

For each distance measure presented in this study three homemade functions have been written in R to implement the MC, B and DD simulation methods. The functions generate simulated data that are used to give: (a) sample, (population), and simulated sample distances; (b) standard deviations and values of skewness and kurtosis of the sample distances estimated from the theory and the distribution of the simulated sample distances; (c) cluster probabilities assessing the uncertainty in the dendrogram; (c) several plots among the various distances, plots of standard deviations, skewness, and kurtosis values, and dendrograms of the original and simulated sample distances. A function that implements the Pearson distribution system via the PearsonDS library of R has also been written to plot the distribution density of a distance from its moments, i.e. from the mean, variance, skewness, and kurtosis. All software material, along with detailed instructions, is presented as supplementary material in *Supplement-3-Instructions&Code*.

The default number of iterations used to estimate simulated distances was 5000 for all simulation methods. In the MC method the 5000 samples were randomly drawn from populations with size equal to 100,000 each. To examine the number of iterations on the performance of the simulation methods, the 5000 iterations were increased to 50,000 in certain simulations using the B and DD methods. In similar simulations with the MC method, the iterations were increased to 10,000 without increase in the population size.

Finally, hierarchical clustering was performed using the *hclust()* function from the base *stats* package of R. For the cluster agglomeration method (linkage method), the Ward minimum variance method was selected, i.e. the method that minimizes the total within-cluster variance.

Results and discussion

To test the performance of the simulation methods, we may examine the comparison plots between original and simulated data of pairwise squared distances (d), their standard

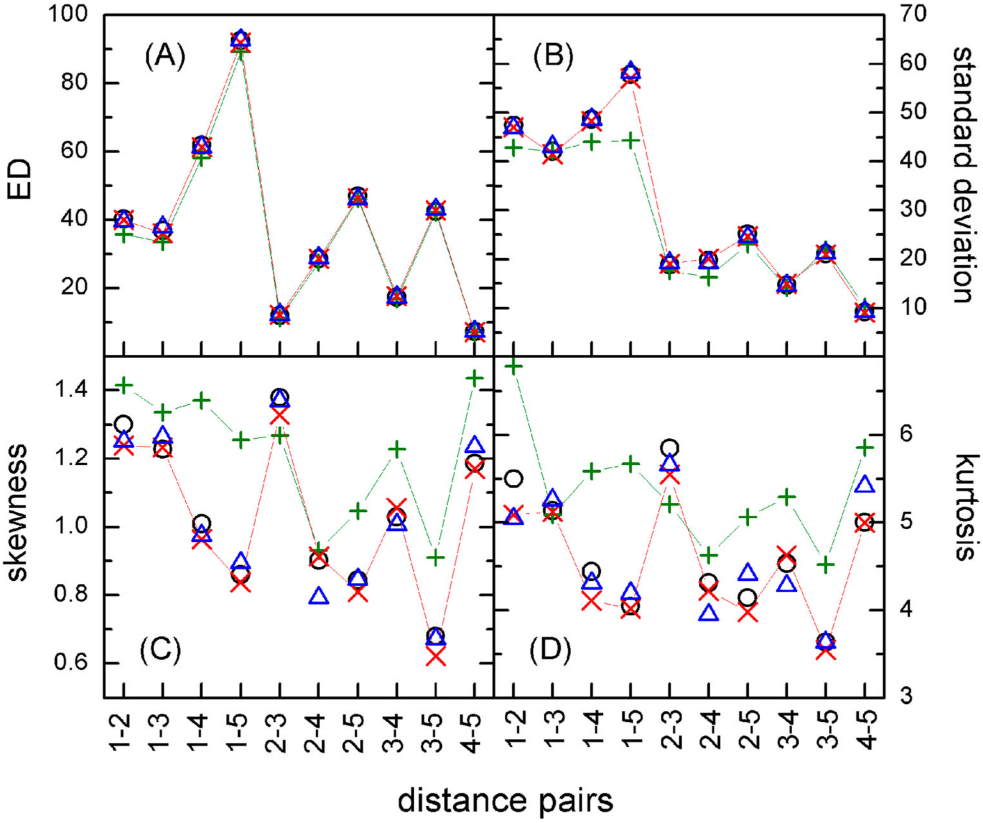

Figure 2. Comparison plots between original and simulated ED distances estimated on the EG2 dataset (A), their standard deviations (B), the skewness (C), and the kurtosis (D) of their distribution. The original data are shown with (o) and the simulated data using MC, DD, and B with (×), (△), and (+), respectively.

deviations (sd), the corresponding skewness (sk) and kurtosis (ku) of their distribution density curves. Two examples of such plots are given in Figures 2 and 3, which show plots between original and simulated ED and MD2 distances and their sd, sk, ku plots estimated on the EG2 dataset for the ED and the WH1 dataset for the MD2.

A quantitative comparison can be carried out via a scoring measure for prediction models, such as the Root Mean Square Percentage Error (RMSPE) defined from:

$$\text{RMSPE} = 100\sqrt{\frac{\sum_{j=1}^{N}(y_{jo}-y_{js})^2}{N(y_{o,\max}-y_{o,\min})^2}} \tag{31}$$

where y_{jo} is the jth original value of the quantity being tested, y_{js} is the corresponding quantity obtained from simulated data, $y_{o,\max}$, $\mathrm{y}_{\mathrm{o,min}}$ are the maximum and minimum values of y_{jo}, and N is the number of y_{jo} values. In case of kurtosis, the range $y_{o,\max} - y_{o,\min}$ was replaced by the corresponding mean value because there were distances where this difference was very close to zero yielding misleading RMSPE values. Table 1 presents RMSPE values for d, sd, sk, and ku that correspond to the ED, MD1, MD2, MMD and UMD when

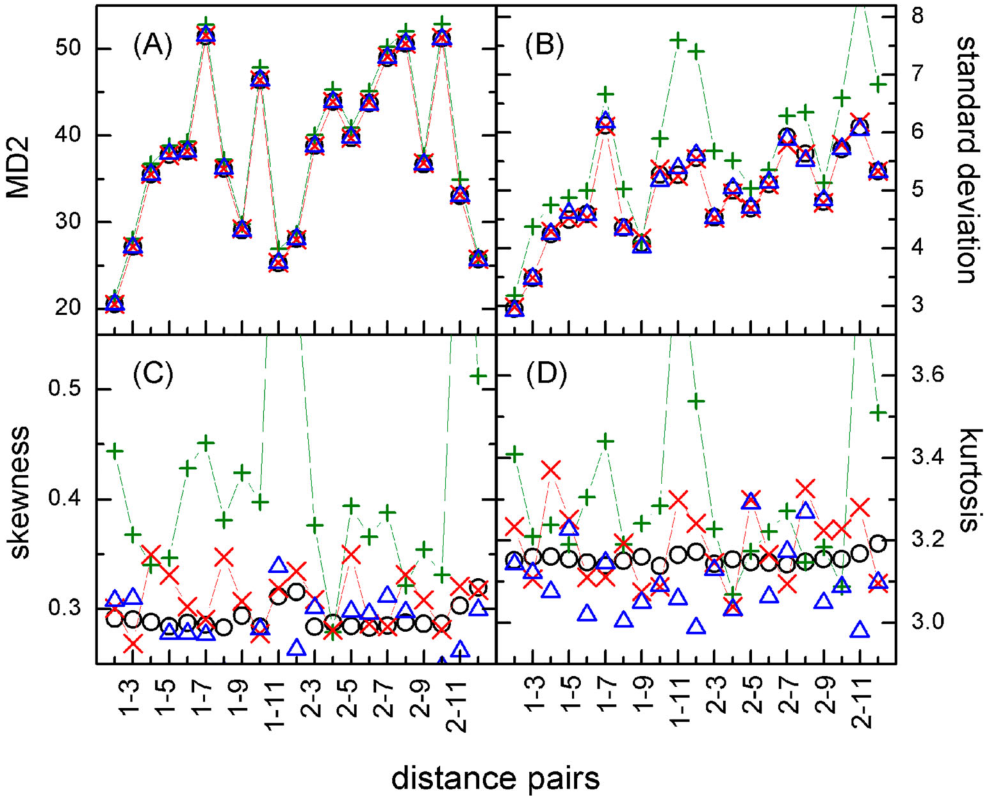

Figure 3. Comparison plots between original and simulated MD2 distances estimated on the WH1 dataset (A), their standard deviations (B), the skewness (C), and the kurtosis (D) of their distribution. The original data are shown with (o) and the simulated data using MC, DD, and B with (×), (△), and (+), respectively.

estimated using the simulation methods MC, B and DD. The corresponding RMSPE values of cED, cMD1, cMD2 are identical or almost identical to those of ED, MD1, and MD2, respectively, and for this reason they have not been included in the table. Note that in this table, as in all other tables, EG1-2 indicates results estimated from the mean of those obtained from datasets EG1 and EG2, and the same holds for SIM1-2 and SIM3-4.

From the comparison plots and the RMSPE values, we observe that the MC and DD simulation methods outperform B. Thus, the RMSPE values for MC and DD tend to converge with each other, while the corresponding RMSPE values are much higher for the bootstrap method. A significant improvement in the performance of the bootstrap method is observed only in NO5, i.e. a binary dataset with nearly independent traits. Note that the performance of the MC and DD methods can be improved if we increase the number of the simulated distances. For example, if we increase this number from 5000 to 50,000, the RMSPE values when using the MD1 distance measure with DD are reduced from 0.43, 0.083, 0.075 to 0.16, 0.028, 0.021 in the datasets EG1-2, WH1, WH2, respectively, i.e. RMSPE on average is reduced by one third. In contrast, the RMSPE values for B remain practically unaffected by the increase in the number of the simulated distances. Finally, we observe that the performance of the simulation methods on the simulated datasets is

Table 1. RMSPE for the squared distances (d), standard deviations (sd), skewness (sk) and kurtosis (ku) of the ED, MD1, MD2, MMD and UMD when estimated using the MC, DD, and B methods.

Distance/dataset	d(MC)	d(DD)	d(B)	sd(MC)	sd(DD)	sd(B)	sk(MC)	sk(DD)	sk(B)	ku(MC)	ku(DD)	ku(B)
ED/EG1-2	0.51	**0.43**	1.51	**0.72**	0.87	7.24	77.63	**5.95**	22.68	11.64	**5.33**	15.06
ED/SIM1-2	0.48	**0.36**	1.42	**0.72**	1.12	7.75	4.71	**4.54**	13.44	**5.67**	6.9	18.38
ED/SIM3-4	0.51	**0.34**	1.16	1.06	**0.86**	7.77	7.15	**5.58**	16.85	6.63	**4.87**	13.65
MD1/EG1-2	**0.43**	**0.43**	1.4	**0.61**	0.9	8.3	7.84	**6.91**	31.68	7.97	**7.17**	16.04
MD1/SIM1-2	**0.35**	0.45	3.16	**0.57**	0.84	13.75	6.82	**5.89**	29.21	6.33	**5.4**	25.6
MD1/SIM3-4	**0.41**	0.45	2.15	**0.93**	1.28	5.5	7.43	**6.74**	20	**5.26**	5.66	8.95
MD1/ WH1	0.09	**0.08**	0.81	**0.64**	0.67	12.27	23.3	**20.07**	199.15	3.06	**2.69**	22.18
MD1/ WH2	0.11	**0.07**	1	0.89	**0.77**	26.23	**28.39**	33.22	556.25	**2.67**	2.71	46.11
MD2/EG1-2	0.67	**0.52**	2.35	1.28	**1.16**	8.79	**8.22**	16.36	44.01	**6.84**	16.69	21.82
MD2/SIM1-2	0.61	**0.54**	3.33	1.26	**0.49**	13.74	**6.1**	10.67	29.75	**5.38**	11.95	23.04
MD2/SIM3-4	0.59	**0.48**	2.49	1.51	**0.87**	9.04	11.78	**9.99**	36.13	7.01	**6.47**	10.72
MD2/ WH1	0.13	**0.11**	2.17	**0.85**	0.95	14.09	**11.33**	13.6	80.76	3.4	**3.04**	11.78
MD2/ WH2	0.2	**0.19**	**	**0.82**	1.14	**	9.17	**5.31**	**	10.11	**6.15**	**
MMD/NO3	2.16	**0.24**	1.22	2.96	**0.84**	16.07	15.48	**12.22**	79.99	4.05	**2.92**	16.4
MMD/NO4	3.8	**0.15**	1.26	7.52	**0.82**	16.47	33.1	**15.3**	41.27	4.56	**1.65**	2.48
MMD/NO5	2.69	**0.3**	1.51	7.84	**1.32**	10.87	27.45	**15.36**	15.59	**1.75**	2.51	3.4
UMD/NO1	0.33	**0.24**	1.23	1.77	**0.55**	28.3	19.44	**8.58**	67.97	7.33	**3.68**	35.12
UMD/NO2	0.36	**0.21**	0.61	1.31	**0.8**	17.78	22.58	**10.81**	37.52	7.17	**3.99**	17.77
UMD/NO3	0.35	**0.22**	0.45	0.89	**0.67**	17.16	15.93	**10.83**	60.09	5.3	**4.16**	16.48
UMD/NO4	1.08	**0.2**	0.44	1.85	**1.12**	27.61	27.46	**26.09**	50.74	**1.95**	2.04	3.49
UMD/NO5	0.81	**0.41**	0.42	**0.78**	0.83	3.91	**15.59**	20.96	26.32	**1.64**	4.76	3.16

Note: Values in bold show minimum values of RMSPE per distance, standard deviation, skewness, and kurtosis, whereas underlined values have been estimated as the RMSPE of kurtosis.
**Computational problems; system is computationally singular.

practically the same as that observed in the corresponding real datasets, indicating that the selected real datasets fulfil the preconditions necessary for the application of the distances under study or the violation of certain conditions does not have a significant effect on the results.

To examine the effect of the discrepancy between theoretical and simulated moments on the distribution density curve of a distance, we can use the Pearson distribution system via the PearsonDS library of R. As an example, Figure 4 shows distance distribution density plots obtained from the MD1, MD2, and MMD when applied on pairs of samples of the EG2, WH1, and NO5 datasets, respectively. These distances were selected because they show relatively large differences between theoretical and simulated moment values. It is seen that the MC and DD methods describe very satisfactorily the distribution density curves of the MD1, MD2, whereas the bootstrap method shows great deviations from the expected behaviour. In the NO5 dataset, both MC and B show deviations from the expected behaviour.

To sum up, the MC and DD simulation methods are much better than the bootstrap method. The performance of the MC and DD is very similar and increases with the increase of the number of simulated distances. The DD method is much faster than the MC and B methods and therefore, if necessary, we can easily increase the number of the simulated distances. For example, if we apply the simulation methods with the MMD to the NO3 and NO4 datasets and the MD1, MD2 to the WH1 dataset, then on average the DD method is 100 times faster than MC and 1100 times faster than the B method. In this example, the DD is 2500 times faster than the B method when it is used to apply the MMD to the NO4 dataset and 350 times faster than the MC when it applies the MD1 to the WH1 dataset. On the other hand, the MC method can give information on whether a distance measure is or

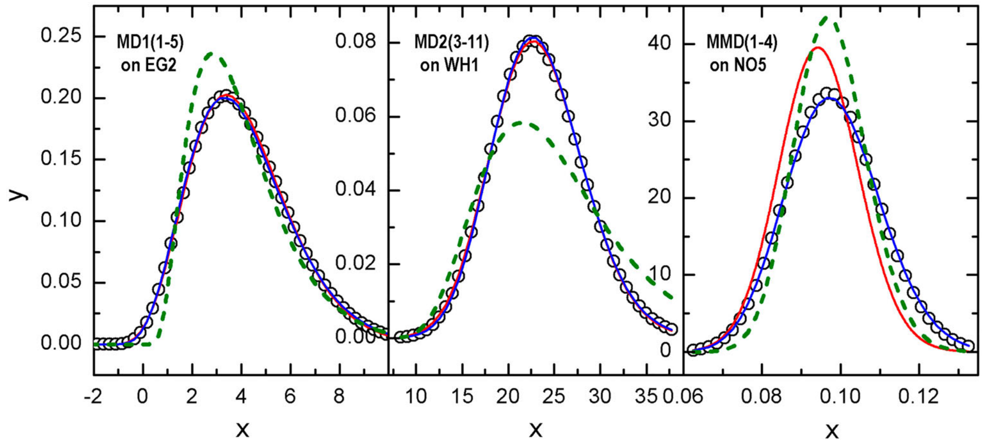

Figure 4. Distance distribution density plots obtained from the MD1, MD2, and MMD when applied on the pairs of samples 1 and 5 of the EG2, 3 and 11 of the WH1, and 1 and 4 of the NO5 datasets. The plots have been calculated using the dpearson function of the PearsonDS library of R by means of the mean, variance, skewness, and kurtosis of each distance (o) and the corresponding simulated quantities using MC (—), DD (—), and B (- - -).

can be corrected to become an unbiased estimator of population divergence. This is very useful information, especially for the MMD, since this distance measure is valid within a certain range of frequencies, usually between 0.1 and 0.9. To meet this requirement, a data editing procedure is usually adopted to remove variables with low/high frequencies. However, when we improve the MMD to behave as an unbiased estimator, a large reduction of variables may be needed. For example, in our datasets we need to reduce the number of traits from 44 in dataset NO1 to 22 in NO3 in order to make the MMD behave as an unbiased estimator. A similar great loss of traits occurs when we remove intercorrelated variables. For example, from 44 traits in NO4, we ended up with 21 in NO5. In such cases a loss of useful information regarding population affinity cannot be excluded.

Figures 6 and 7 as well as the Figures given in *Supplement-4-Dendrograms* show selected dendrograms along with the cluster probabilities obtained from the application of the R functions. We observe that, overall, the estimated cluster probabilities from the three simulation methods do not exhibit significant differences. This is also seen in Figure 5, which shows cluster probabilities estimated by means of the MC, DD, and B methods when using the cMD2 over all relevant datasets and the MMD over NO3, NO4, and NO5. If we take into account that in most cases the bootstrap method does not simulate satisfactorily the distance distribution, this is a rather unexpected result, indicating the tolerance of the computed cluster probabilities over the performance of the simulation method used in these calculations.

To test quantitatively the impact of the simulation method in assessing the uncertainty in HCA, we examined the differences between cluster probabilities estimated from the three simulation methods. Table 2 shows the mean and maximum percent of absolute difference between cluster probabilities estimated from the MC and DD methods, and Table 3 presents differences in the cluster probabilities between the MC and B simulation methods. It is seen that the MC and DD methods, which exhibit practically the same simulation

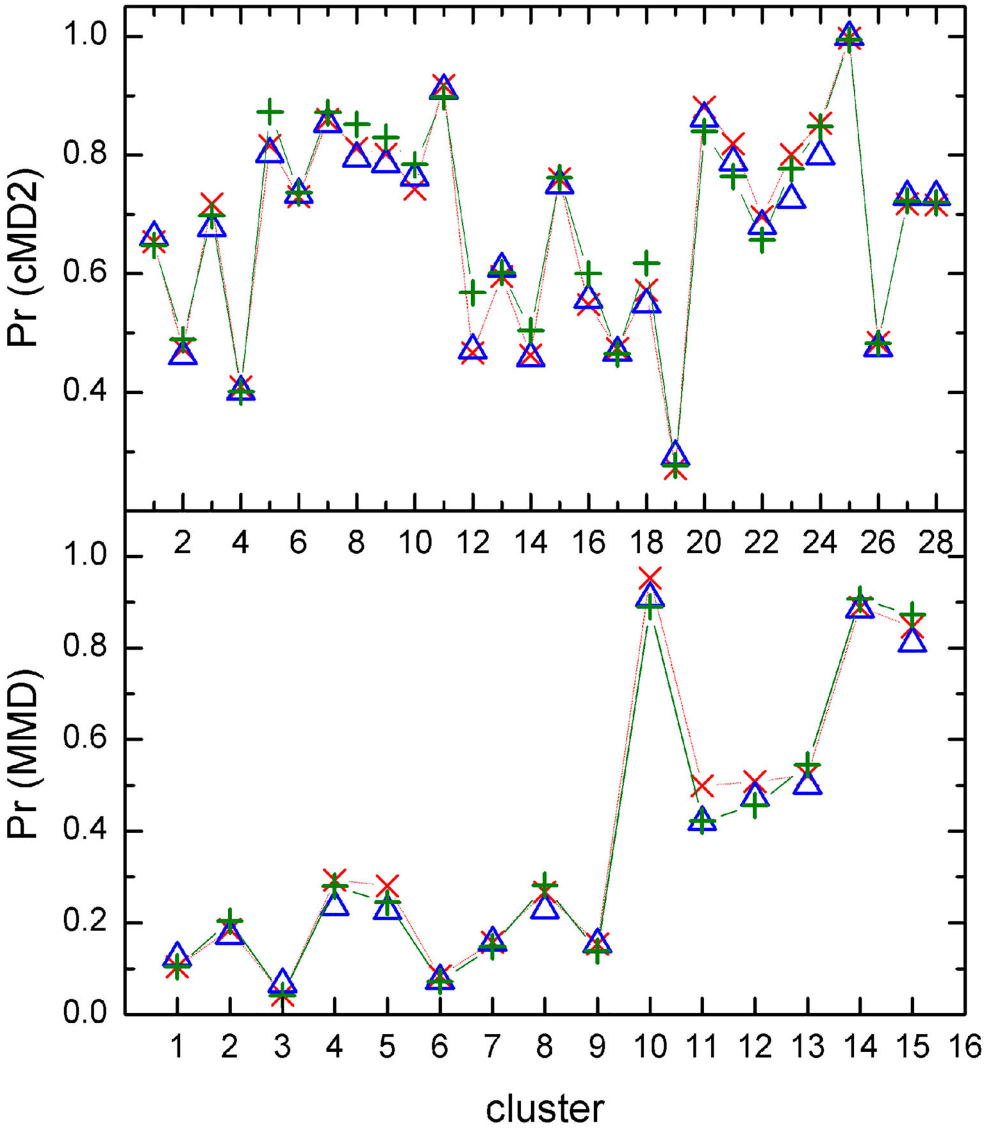

Figure 5. Cluster probabilities when using the cMD2 over all relevant datasets and the MMD over NO3, NO4, and NO5 estimated by means of MC (×), DD (△), and B (+) simulations.

performance, give similar cluster probabilities. On average, the mean and maximum differences between the cluster probabilities obtained from these methods are 1.7% and 3.6%, respectively. Note that these differences are not significantly affected from the number of iterations used, i.e. the observed differences are not significantly reduced if we increase the number of iterations, due to the random selection of the simulated distances. In line with the patterns seen in Figure 5, Table 3 shows that bootstrapping by simple resampling with replacement does not give as bad results as expected from the comparisons using distance moments. On average, the mean and maximum differences between MC and B are

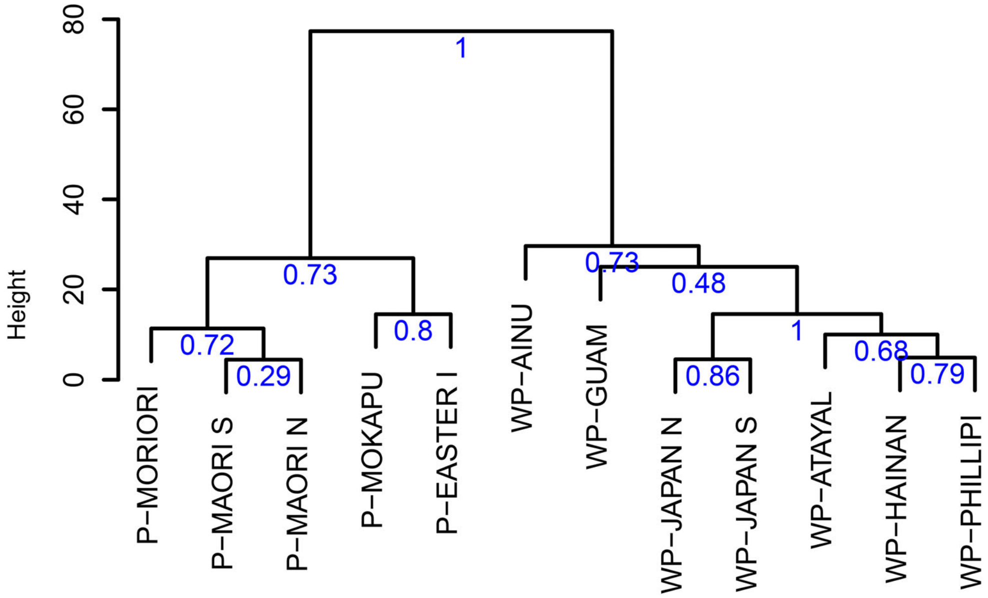

Figure 6. Ward's dendrogram obtained from the cMD2 when applied to the dataset WH1 and cluster probabilities estimated using the DD method.

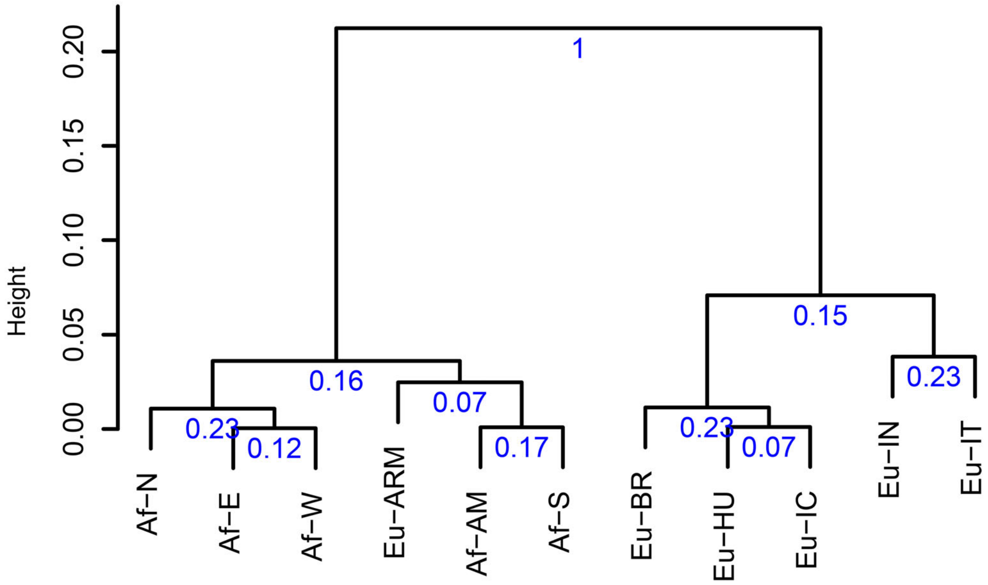

Figure 7. Ward's dendrogram obtained from the MMD when applied to the dataset NO3 and cluster probabilities estimated using the DD method.

2.9% and 6.2%, respectively. Noting also that the maximum percent difference in the cluster probabilities between the MC and B methods is 11% and that between MC and DD is 8%, it becomes clear that the estimated probabilities do not depend much on the simulation method. This is an interesting result since it shows that the bootstrap method can be

Table 2. Mean and maximum percent difference between cluster probabilities estimated from Monte-Carlo simulations and simulations based on the DD method.

Distance	Dataset	Mean	Max	Distance	Dataset	Mean	Max
ED	EG1-2	2.08	4.5	MD2	SIM3-4	1.25	2.8
cED	EG1-2	1.3	3.3	cMD2	SIM3-4	1.73	4.1
MD1	EG1-2	0.97	2.7	MD1	WH1	1.87	7.2
cMD1	EG1-2	0.7	1.7	cMD1	WH1	2.31	7.4
MD2	EG1-2	0.78	2.4	MD2	WH1	1.88	5.2
cMD2	EG1-2	1.05	1.6	cMD2	WH1	2.53	7.6
ED	SIM1-2	1.18	2.6	MD1	WH2	1.99	3.4
cED	SIM1-2	0.95	1.9	cMD1	WH2	2.76	4.4
MD1	SIM1-2	0.82	2	MD2	WH2	3.06	5.3
cMD1	SIM1-2	0.57	0.9	cMD2	WH2	1.49	2.7
MD2	SIM1-2	1.23	2.3	MMD	NO3	2.44	5.8
cMD2	SIM1-2	1.03	2.4	UMD	NO3	2.43	5.7
ED	SIM3-4	1.5	4.3	MMD	NO4	5.27	7.9
cED	SIM3-4	1.35	2.4	UMD	NO4	1	1.7
MD1	SIM3-4	1.15	2.1	MMD	NO5	2.1	3.6
cMD1	SIM3-4	1.08	1.9	UMD	NO5	1.23	1.9

Table 3. Mean and maximum percent difference between cluster probabilities estimated from Monte-Carlo simulations and simulations based on bootstrapping.

Distance	Dataset	Mean	Max	Distance	Dataset	Mean	Max
ED	EG1-2	1.5	3.6	cMD1	SIM3-4	3.2	7.7
cED	EG1-2	2.12	4	MD2	SIM3-4	3.02	5.4
MD1	EG1-2	2.23	6.2	cMD2	SIM3-4	2.52	5.2
cMD1	EG1-2	1.97	4.3	MD1	WH1	1.55	3.8
MD2	EG1-2	1.95	4.9	cMD1	WH1	1.77	4.4
cMD2	EG1-2	2.03	4.1	MD2	WH1	2.53	5.8
ED	SIM1-2	3.22	9	cMD2	WH1	1.83	5.5
cED	SIM1-2	3.18	8.5	MD1	WH2	2.66	5.5
MD1	SIM1-2	4.47	9.3	cMD1	WH2	4.59	9.1
cMD1	SIM1-2	5.02	11.1	MMD	NO3	1.36	3.7
MD2	SIM1-2	3.55	8.1	UMD	NO3	3.27	8.3
cMD2	SIM1-2	4	10.1	MMD	NO4	6.4	7.7
ED	SIM3-4	3.62	7.2	UMD	NO4	4.47	5.3
cED	SIM3-4	3.83	6.6	MMD	NO5	2.27	2.7
MD1	SIM3-4	3.7	7.8	UMD	NO5	0.63	1

a useful alternative to the MC and DD methods in datasets that follow unknown distance distributions.

Finally, Table 4 gives selected values of the mean and minimum percent of cluster probabilities estimated from simulations based on the DD method. It is evident that these probabilities for a certain dataset and distance measure determine the reliability of the information about the clusters appearing in the dendrogram. For example, in Figure 6 we observe that the populations from North Japan, South Japan, Atayal, Hainan, and Philippines form a cluster with a probability 1.0. Note that this probability is not related to a specific configuration of the sub-clusters formed by these populations. It is also interesting to observe that this cluster appears in the dendrograms of all Mahalanobis distances, MD1, cMD1, MD2, and cMD2, with very high probabilities ranging from 0.98 to 1. Therefore, the craniometric data clearly indicate the affinity of these populations. In contrast, Figure 7 shows that all the cluster probabilities when using the MMD are particularly low and, therefore, this dendrogram cannot be used to draw conclusions for the existence of

Table 4. Mean and minimum percent of cluster probabilities estimated from simulations based on the DD method.

Distance	Dataset	Mean	Min	Distance	Dataset	Mean	Min
ED	EG1-2	60.8	24.8	MD1	WH2	76.71	55.5
cED	EG1-2	58.28	30.8	cMD1	WH2	75.14	59.3
MD1	EG1-2	67.12	38.6	MD2	WH2	43.37	24.7
cMD1	EG1-2	65.03	46.4	cMD2	WH2	35.86	24.6
MD2	EG1-2	66.17	38.4	MMD	NO3	15.94	6.6
cMD2	EG1-2	63.97	45.8	UMD	NO3	21.92	5
MD1	WH1	78.28	46.9	MMD	NO4	60.07	42
cMD1	WH1	75.18	29.3	UMD	NO4	97.77	97.3
MD2	WH1	74.38	39.6	MMD	NO5	73.13	49.9
cMD2	WH1	70.81	29.2	UMD	NO5	42.07	38.3

clusters of populations. The use of the UMD in place of the MMD increases these probabilities but they are still too low to give reliable information about the existence of clusters. From Table 4 and the cluster probabilities shown in the dendrograms of Figures 6, 7 and in *Supplement-4-Dendrograms,* we note that metric data are more informative than binary data with one exception: The binary data can give reliable information only if we use very large samples, such as those of datasets NO4 and NO5 (Table 4).

Conclusions

The assessment of the uncertainty in hierarchical clustering is an essential step to draw reliable conclusions about the existence of clusters in a dataset. This assessment can be performed via the computation of simulated probabilities for the appearance of the various dendrogram patterns. In the examples used in the present study, these probabilities were high enough when using craniometric data, allowing the determination of clusters of populations of close affinity with certainty. In contrast, the use of binary datasets with cranial traits can give useful cluster information only if the samples are of particularly large size.

The estimation of reliable simulated cluster probabilities presumes reliably simulated distances since these probabilities arise from the percentage of the appearance of the various dendrogram patterns in a large simulated dataset of distances. In the present study, we have examined three simulation methods: the conventional Monte-Carlo method and the bootstrapping by simple resampling with replacement, and a new proposed method, the distance distribution method. It is shown that the MC and DD simulation methods perform similarly and give very satisfactory predictions for the first four distance moments, provided that the data fulfil the requirements that are necessary for a reliable estimation of the distance moments. In addition, the performance of these two methods can be further improved if we increase the number of the simulated distances, which can be easily carried out by the DD method since this method is very fast. In contrast, the bootstrap method used in the present study gives rather acceptable predictions of the first distance moment, but it fails to simulate satisfactorily the standard deviation, skewness, and kurtosis.

The MC and DD methods give very similar cluster probabilities. On average, the mean and maximum differences between the cluster probabilities obtained from these methods are 1.7% and 3.6%, respectively. The bootstrap probabilities are not very different from those obtained from the MC and DD methods, despite the rather poor simulation performance of this method. This result shows that the bootstrap method can be used in datasets that exhibit strong deviations from the distributions examined in the present study or datasets of unknown distance distribution.

Acknowledgements

The authors would like to thank two anonymous reviewers for their valuable input.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

EN's contribution to this work was co-funded by the European Regional Development Fund and the Republic of Cyprus through the Research and Innovation Foundation [Project: EXCELLENCE/1216/0023]. This project has also received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement No 811068.

ORCID

Efthymia Nikita http://orcid.org/0000-0003-2094-5047

References

[1] Frank IE, Todeschini R. The data analysis handbook. Vol. 14, 1st ed Amsterdam: Elsevier; 1994.
[2] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. 2nd ed. New York (NY): Springer; 2009.
[3] Kaufman L, Rousseeuw PJ. Finding groups in data: an introduction to cluster analysis. 1st ed. New York (NY): John Wiley; 1990.
[4] Suzuki R, Shimodora H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics. 2006;22:1540–1542.
[5] Mathai AM, Provost SB. Quadratic forms in random variables. Boca Raton: CRC Press; 1992.
[6] Johnson RA, Wichern DW. Applied multivariate statistical analysis. 6th ed. New Jersey: Prentice Hall; 2007.
[7] Mardia KV, Kent JT, Bibby JM. Multivariate analysis. San Diego (CA): Academic Press; 1995.
[8] Mclachlan GJ. Mahalanobis distance. Resonance. 1999;4:20–26.
[9] Rao CR. Linear statistical inference and its applications. 2nd ed. New York (NY): Wiley; 2001.
[10] Reiser B. Confidence intervals for the Mahalanobis distance. Commun Stat Simul. 2001;30:37–45.
[11] Sjøvold T. Some notes on the distribution and certain modifications of mahalanobis' generalized distance (D2). J Hum Evol. 1975;4:549–558.
[12] van Vark GN. Some statistical procedures for the investigation of prehistoric human skeletal material [dissertation]. Groningen: University of Groningen; 1970.
[13] Grewal MS. The rate of genetic divergence in the C57BL strain of mice. Genet Res. 1962;3:226–237.
[14] Sjøvold T. Non–metrical divergence between skeletal populations. The theoretical foundation and biological importance of C.A.B. Smith's mean measure of divergence. OSSA. 1977;4(suppl.):1–133.
[15] de Souza P, Houghton P. The mean measure of divergence and the use of non-metric data in the estimation of biological distances. J Archaeol Sci. 1977;4:163–169.

[16] Irish JD. The mean measure of divergence: its utility in model-free and model-bound analyses relative to the Mahalanobis D^2 distance for nonmetric traits. Am J Hum Biol. 2010;22:378–395.
[17] Edward F, Harris EF, Sjøvold T. Calculation of Smith's mean measure of divergence for intergroup comparisons using nonmetric data. Dent Anthropol J. 2004;7:83–93.
[18] Nikita E, Nikitas P. Measures of divergence for binary data used in biodistance studies. Archaeol Anthropol Sci. 2021;13:40.
[19] Thomson A, Randall-MacIver D. The ancient races of the thebaid. Oxford: Clarendon Press; 1905.
[20] Howells WW. Cranial variation in man: a study by multivariate analysis of patterns of difference among recent human populations. Cambridge: Harvard University Press; 1973.
[21] Howells WW. Skull shapes and the map: craniometric analyses in the dispersion of modern homo. Cambridge: Harvard University Press; 1989.
[22] Howells WW. Who's who in skulls: ethnic identification of crania from measurements. Cambridge: Harvard University Press; 1995.
[23] Howells WW. Howells' craniometric data on the internet. Am J Phys Anthropol. 1996;101:441–442.
[24] Ossenberg NS. Brief communication: cranial nonmetric trait database on the internet. Am J Phys Anthropol. 2013;152:551–553.
[25] Nikita E. Osteoarchaeology: a guide to the macroscopic study of human skeletal remains. San Diego (CA): Academic Press; 2017.