A method for chronological ordering of archeological sites

By Annamária G. Varga

1. Introduction

The chronological ordering of archeological material is an important question of the archeological investigation. For the solution of this problem, besides the classical archeological methods, various methods using tools of natural sciences and mathematics are known.

In this paper we are going to describe a mathematical method based on the theory of regression. This theory gives a natural approach to the problem of chronological ordering. By the aid of this theory we are able to decide in which cases the chronological order obtained by the method of Brainerd-Robinson [1] and by similar methods can be accepted. The idea of the application of the theory of regression was given by an analysis of the methods of Brainerd-Robinson and Dempsey-Baumhoff [2].

2. Prerequisites

The purpose of this section is to summarize concepts and to state results which are familiar to mathematicians but not to archeologists and which will be used in what follows. Whenever the word 'set' is used it will be interpreted to mean a subset of a given set which will be denoted by S. If x is an element of S, and E is a subset of S, the notation $x \in E$ means, that x belongs to E; the negation of this assertion, i.e. the statement that x does not belong to E, will be denoted by $x \notin E$. If E and F are subsets of S, the notation $E \subset F$ means that E is a subset of F i.e. that every point of E belongs to F. Two sets E and F are called equal if and only if they contain exactly the same elements or, equivalently, if and only if $E \subset F$ and $F \subset E$.

If $P(x)$ is a proposition concerning x then the symbol $\{x\colon P(x)\}$ denotes the set of those elements x for which the proposition $P(x)$ is true. In general the brace notation $\{...\}$ will be reserved for the formation of sets. Thus for instance if x and y are elements then $\{x, y\}$ denotes the set whose only elements are x and y.

If $\mathbf{E}$ is any set of subsets of S, the set of all points of S which belong to *at least one* set of $\mathbf{E}$ is called the *union* of the sets of $\mathbf{E}$; it will be denoted by $\cup \mathbf{E}$ or $\cup \{E\colon E \in \mathbf{E}\}$. For the union of a special set of sets various special notations are used. If for instance $\mathbf{E} = \{E_1, E_2, ..., E_n\}$, then $\cup \mathbf{E}$ is denoted also by $E_1 \cup E_2 \cup ...$ $... \cup E_n$ or $\bigcup\limits_{i=1}^{n} E_i$.

If E any set of subsets of S, the set of all elements of S which belong to *every set* of E is called the *intersection* of the sets of E; it will be denoted by $\bigcap \mathbf{E}$ or $\bigcap\{E\colon E\in \mathbf{E}\}$.

Two sets E and F are called *disjoint* if they have no elements in common. A disjoint set is a set E of sets such that every two distinct sets of E are disjoint.

If E and F are subsets of S, the *difference* between E and F, denoted by $E-F$, is the set of all elements of E which do not belong to F. The symmetric difference of two sets E and F, denoted by, $E\triangle F$ is defined by $E\triangle F=(E-F)\cup(F-E)$. It is the set of all elements which belong to one and only one of E and F.

Let R be any set whose elements are called, for suggestivity, points. If to each pair x, y of elements of R a non-negative real number, denoted by $\varrho(x, y)$ and called the distance of x and y, is attached such that

(1) if $x=y$ then $\varrho(x,y)=0$,
(2) if $\varrho(x,y)=0$ then $x=y$,
(3) $\varrho(x,y)=\varrho(y,x)$,
(4) for each three elements x, y, z of R

$$\varrho(x,y)\leqq \varrho(x,z)+\varrho(z,y),$$

the resulting "space" M is called a *metric space* over the groundset R with *metric* ϱ.

A function ϱ which satisfies (1), (3), (4) only, is called a *pseudo-metric* and the resulting space is called a *pseudo-metric space M* over the groundset R with pseudometric ϱ.

Let M be a pseudo-metric space and let D be the family of all sets $G_x=\{y\in M\colon \varrho(x,y)=0\}$. If $u\in G_x$ and $v\in G_y$ then

$$\varrho(u,v)\leqq \varrho(u,x)+\varrho(x,y)+\varrho(y,v) \doteq \varrho(x,y).$$

Consequently, since in this case it is also true that $x\in G_u$ and $y\in G_v$, $\varrho(u,v)=\varrho(x,y)$. Let A and B be two members of D and let $\tau(A,B)$ be equal to $\varrho(x,y)$ for every x in A and for every y in B. Thus D with the function $\tau(A,B)$ is a metric space. In the sequel we shall call the set D with $\tau(A,B)$ the metric space induced by the pseudometric space M. A set N is called a subset of a metric space M provided N is a subset of the groundset R of M and the distance of any two points x, y of N is the same as their distance in M. If N and L are subsets of two metric space M and Q, respectively, we say N is congruent to L provided there exists a one-to-one distance-preserving correspondence between the points of N and the points of L; that is for every pair x, y of points of N $\varrho(x,y)=\varrho'(x',y')$, where x', y' are the points of L that correspond, respectively, to points x, y of N and ϱ, ϱ' denote the distance in N and L, respectively.

A subset N of a metric space M is congruently imbeddable in a metric space Q provided there is a subset L of Q such that N is congruent to L.

We shall apply in the sequel the theory of regression. We need the linear regression. For our purposes it is necessary to know only the following. We consider n points $(x_1,y_1), (x_2,y_2), \dots, (x_n,y_n)$ in the plane. It is convenient to write the equation of the straight line which we fit to these n points in the from

$$y'=a+b(x-\bar{x}), \tag{1}$$

where $\bar{x}$ is the arithmetic mean of $x_1, x_2, \ldots, x_n$; b is the slope of this line and a is the y intercept on the line $x=\bar{x}$. The y intercept on the y axis is $a-b\bar{x}$. The problem is to determine the parameters a and b so that the sum of the squares

$$\sum_{i=1}^{n} (y_i - y_i')^2$$

will be a minimum. When y' is replaced by its value as given by (1), it becomes clear that this sum is a function of a and b only. If this function is denoted by $F(a, b)$ then

$$F(a, b) = \sum_{i=1}^{n} [y_i - a - b(x_i - \bar{x})]^2.$$

If this function is to have a minimum value, it is necessary that its partial derivates vanish there; hence, a and b must satisfy the equations

$$\frac{\partial F}{\partial a} = \sum_{i=1}^{n} 2[y_i - a - b(x_i - \bar{x})][-1] = 0,$$

$$\frac{\partial F}{\partial b} = \sum_{i=1}^{n} 2[y_i - a - b(x_i - \bar{x})][-x_i - \bar{x}] = 0.$$

When the summations are performed term by term and the sums that involve y_i are transposed, these equations assume the form

$$an + b \sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} y_i$$

$$a \sum_{i=1}^{n} (x_i - \bar{x}) + b \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} (x_i - \bar{x}) y_i.$$

Since $\sum_{i=1}^{n} (x_i - \bar{x}) = 0$, the solution of these equations is given by

$$a = \bar{y} \quad \text{and} \quad b = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

These values when inserted in (1) yield the line $y' - \bar{y} = b(x_i - \bar{x})$ which is usually called the regression line.

If we write

$$S_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}) y_i$$

and

$$S_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2,$$

then we may write

$$b = \frac{S_{xy}}{S_x^2} = \frac{S_{xy}}{S_x S_y} \cdot \frac{S_y}{S_x} = r \frac{S_y}{S_x}.$$

Here $r = \frac{S_{xy}}{S_x S_y}$ is called the correlation coefficient. The value of r must satisfy the inequality $-1 \leqq r \leqq 1$. The value of r will be equal to ± 1 if and only if, the points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ lie on the regression line.

3. The archeological bases of the methods of Brainerd-Robinson and Dempsey-Baumhoff

Let us assume that we compare n sites. We denote by S_i $(i=1, 2, \ldots, n)$ the set of the objects of i-th site and by T_i $(i=1, 2, \ldots, n)$ the set of the types of the i-th site. Put $S = \bigcup_{i=1}^{n} S_i$ and $T = \bigcup_{i=1}^{n} T_i$. The number A_{KS_i} means the precentage of the objects of type K belonging to the i-th site. The correlation between site i and site j according to Brainerd and Robinson is defined by

$$X_{ij} = 200 - \sum_{K \in T} |A_{KS_i} - A_{KS_j}|. \tag{2}$$

This may be written in the following from

$$X_{ij} = 200 - \sum_{K \in T_i - T_j} A_{KS_i} - \sum_{K \in T_j - T_i} A_{KS_j} - \sum_{K \in T_i \cap T_j} |A_{KS_i} - A_{KS_j}|.$$

From this one can easily seen that the method of Brainard-Robinson is based on the following principle. If two sites have types in essentially different percentages or if there are types which belong to one of the two sites but absent other site then the two sites originate from different times.

If $T_i = T_j$ i.e. the i-th and the j-th sites have the same types then in the above formula the first and second sums are equal to zero. Thus the agremeent between the i-th and j-th sites is determined by third sum. If the desagreement is small between i-th and j-th sites then the members of the foregoing sum $\left(\sum_{K \in T_i \cap T_j} |A_{KS_i} - A_{KS_j}| \right)$ are also small. This is the only case, according to Brainerd-Robinson's method, the two sites are of an age.

This means that the percentage of each type is approximately the same in the both sites.

Even if the site S_i contains essentially more objects than the site S_j, consequently, the site S_i has a greater number of the objects of the type K than the site S_j. The point of view of archeology this is such a requirement regarding to two sites which only rarely holds.

The element X_{ij} of the matrix used in Dempsey-Baumhoff's method is given by formula

$$X_{ij} = N - \sum_{K \in T_i \Delta T_j} 1, \tag{3}$$

where N means the number of all types belonging to sites S_i $(i=1, 2, \ldots, n)$. According to this formula the method of Dempsey-Baumhoff is based on the following principle. If two sites have the same types than both sites are of an age. However there exist such types which belong to one of the two sites only then they originate from different times.

These principles show that the two methods are essentially different. Later we shall return this question and we shall formulate the difference between these methods in the language of mathematics.

4. The mathematical analysis of the methods of Brainard-Robinson and Dempsey-Baumhoff

The first method assigns to each pair (S_i, S_j) $(i, j = 1, 2, \dots, n)$ of sites the number given by formula (2), the second one assigns the number given by formula (3).

Let us correspond to each pair (S_i, S_j) either the number

$$\sum_{K \in T_i \triangle T_j} |A_{KS_i} - A_{KS_j}| + \sum_{K \in T_i \cap T} |A_{KS_i} - A_{KS_j}| \tag{4}$$

or the number

$$\sum_{K \in T_i \triangle T_j} 1. \tag{5}$$

For the sake of brevity, let us denote the number (4) by $r(S_i, S_j)$ and the number (5) by $\varrho(S_i, S_j)$, respectively. The function corresponding to the first method is $200 - r(S_i, S_j)$ and the function corresponding to the second one is $N - \varrho(S_i, S_j)$. It is clear that the determination of chronological order we may use the function $r(S_i, S_j)$ instead of $200 - r(S_i, S_j)$ in the case of the first method and the function $\varrho(S_i, S_j)$ instead of $N - \varrho(S_i, S_j)$ in the case of the second one.

We shall prove that the functions r and ϱ satisfy the

$$\varrho(S_i, S_j) \leqq \varrho(S_i, S_k) + \varrho(S_k, S_j) \tag{6}$$

and

$$r(S_i, S_j) \leqq r(S_i, S_k) + r(S_k, S_j) \tag{7}$$

inequalities, respectively.

First we prove the inequality (6). Let us correspond to each subset L of the set T the number of the element of L (that is the number of types contained in L) which we denote by $\mu(L)$. The domain of the function $\mu(L)$ is the set $P(T)$ of all subsets of T and its values are non-negativ numbers. If L and M are disjoint subsets of T then

$$\mu(L \cup M) = \mu(L) + \mu(M),$$

i.e. the function $\mu(L)$ is additive.

The function $\varrho(S_i, S_j)$ can be given with the aid of function $\mu(L)$ as follows

$$\varrho(S_i, S_j) = \mu(T_i \triangle T_j).$$

Thus the inequality (6) obviously follows from the additivity of μ.

After this we are going to prove the inequality (7). This may be rewritten in the following form

$$\sum_{K \in T_i \cup T_j} |A_{KS_i} - A_{KS_j}| \leqq \sum_{K \in T_i \cup T_k} |A_{KS_i} - A_{KS_k}| + \sum_{K \in T_k \cup T_j} |A_{KS_k} - A_{KS_j}|. \tag{8}$$

Now the left-hand side of (8) in detail is

$$\begin{aligned}(9)\qquad &\sum_{K\in T_i-(T_j\cup T_k)} A_{KS_i}+\sum_{K\in T_j-(T_i\cup T_k)} A_{KS_j}+\sum_{K\in(T_i\cap T_j)-T_k} |A_{KS_i}-A_{KS_j}|+\\ &+\sum_{K\in(T_j\cap T_k)-T_i} A_{KS_j}+\sum_{K\in(T_i\cap T_k)-T_j} A_{KS_i}+\sum_{K\in T_i\cap T_j\cap T_k} |A_{KS_i}-A_{KS_j}|\end{aligned}$$

and the right-hand side of (8)

$$\begin{aligned}&\sum_{K\in T_i-(T_j\cup T_k)} A_{KS_i}+\sum_{K\in T_k-(T_i\cup T_j)} A_{KS_k}+\sum_{K\in(T_i\cap T_j)-T_k} A_{KS_i}+\\ (10)\qquad &+\sum_{K\in(T_i\cap T_k)-T_j} |A_{KS_i}-A_{KS_j}|+\sum_{K\in(T_j\cap T_k)-T_i} A_{KS_k}+\sum_{K\in T_i\cap T_j\cap T_k} |A_{KS_i}-A_{KS_k}|+\\ &+\sum_{K\in T_k-(T_i\cap T_j)} A_{KS_k}+\sum_{K\in T_j-(T_i\cup T_k)} A_{KS_j}+\sum_{K\in(T_i\cap T_j)-T_k} A_{KS_j}+\\ &+\sum_{K\in(T_i\cap T_k)-T_j} A_{KS_k}+\sum_{K\in(T_j\cap T_k)-T_i} |A_{KS_k}-A_{KS_j}|+\sum_{K\in T_i\cap T_j\cap T_k} |A_{KS_k}-A_{KS_j}|.\end{aligned}$$

We omit from (9) and (10) the members occuring in the both (9) and (10).

By the application of the triangle inequality we get

$$(11)\qquad \sum_{K\in T_i\cap T_j\cap T_k} |A_{KS_i}-A_{KS_j}| \leqq \sum_{K\in T_i\cap T_j\cap T_k} |A_{KS_i}-A_{KS_k}|+\sum_{K\in T_i\cap T_j\cap T_k} |A_{KS_k}-A_{KS_j}|;$$

$$(12)\qquad \sum_{K\in(T_i\cap T_j)-T_k} |A_{KS_i}-A_{KS_j}| \leqq \sum_{K\in(T_i\cap T_j)-T_k} A_{KS_i}+\sum_{K\in(T_i\cap T_j)-T_k} A_{KS_j};$$

$$(13)\qquad \sum_{K\in(T_i\cap T_k)-T_j} A_{KS_i} \leqq \sum_{K\in(T_i\cap T_k)-T_j} |A_{KS_i}-A_{KS_k}|+\sum_{K\in(T_i\cap T_k)-T_j} A_{KS_k};$$

$$(14)\qquad \sum_{K\in(T_j\cap T_k)-T_i} A_{KS_j} \leqq \sum_{K\in(T_j\cap T_k)-T_i} |A_{KS_k}-A_{KS_j}|+\sum_{K\in(T_j\cap T_k)-T_i} A_{KS_k}.$$

The left-hand sides of (11), (12), (13), (14) add up the left-hand side of the (8) and similarly the right-hand sides of (11), (12), (13), (14) add up the right-hand side of the (8), disregarding the omitted members and the sum

$$2\sum_{K\in T_k-(T_i\cup T_j)} A_{KS_k}.$$

From this we can infer that the inequality (8) and automatically the inequality (7) holds..

5. The application of the regression theory to the chronological seriation

From the foregoing it can be easily seen that the function $r(S_i, S_j)$ in the method of Brainerd—Robinson and the function $\varrho(S_i, S_j)$ in the method of Dempsey-Baumhoff determine each a pseudometric space. In the prerequisites it was shown that a pseudo-metric induces a metric on the set of all sets $G_i=\{S_j: \varrho(S_i, S_j)\}=0$. Thus we may assume, with no loss of generality, that the function $r(S_i, S_j)$ and $\varrho(S_i, S_j)$ are metrics. Arises the question what kind of a metric are induced by the function r and ϱ in the set of the sites. Are they similar to the metric of the straight line or euclidean plane. Precisely, they are whether or not congruently imbeddable in the euclidean plane. It may happen that the imbedding is not possible.

Namely, let us consider, four sites A, B, C, D. Assume that each of the sites have the same types: I, J, K, L, M, N, P, Q. Assume moreover, that in the site A the type I occurs in percentage 25, the type M in percentage 45, and the other types occur in percentages $5-5$; in the site B the type J occurs in percentage 25, the type N in percentage 45 and the other types in percentages $5-5$; in the site C the type K occurs in percentage 25, the P in percentage 45 and the other types occur in percentages $5-5$; and finally in the site D the type L occurs in the percentage 25, the type Q in percentage 45 and the other types occur in percentages $5-5$.

By the method of Brainerd-Robinson

$$r(A, B)=r(A, C)=r(A, D)=r(B, C)=r(B, D)=r(C, D)=120,$$

i.e. the distance of each pair of the four sites is the same. Since we cannot find in the plane four distinct points such that any pair of them has the same non-zero distance, the metric space determined by the set $\{A, B, C, D\}$ and the metric r is not congruently imbeddable in the plane. We may make a similar example in the case of the method of Dempsey-Baumhoff. It is easy to see that in such cases neither the Brainerd-Robinson's method nor Dempsey-Baumhoff's method cannot give a chronological order.

In the reality, however, such cases occur only when we commit an error in the preparation of the archeological material or in our calculations. After a new examination we may find the trouble.

We have seen the difference between principles on which the methods of Brainerd-Robinson and Dempsey-Baumhoff are based. This may the right time to straighten out the different in another way. Arises the question that the metric space induced by the sites and the metric $\varrho(S_i, S_j)$ can be congruently imbeddable in the metric space induced by the sites and the metric $r(S_i, S_j)$. In general this is not possible. Consequently, the chronological orders obtained by the two methods are not the same, because both methods determine the chronological order comparing the sizes of the distances of the sites.

In order to establish the chronological seriation we need at least demand that the metric space induced by the set of sites and the function r or ϱ be congruently imbeddable in the plane. But in this case it is reasonable to apply the theory of regression. First we must decide that the metric space induced by the function $r(S_i, S_j)$ on the set of the sites are imbeddable whether or not in the plane. If this is not possible then we must examine preliminary analyses particularly the isolation of the types.

It is known various methods to decide the possibility of the imbedding. We may use the following general theorem [3].

An arbitrary metric space S with metric r is congruently imbeddable in euclidean n-dimensional space if and only if (i) S contains an $t+1$-tuple $p_0, p_1, \dots, p_t$ $(t \leqq n)$ such that the determinant

$$D(p_0, p_1, \dots, p_k) = \begin{vmatrix} 0 & 1 & 1 & . & . & . & 1 \\ 1 & 0 & r^2(p_0 p_1) & . & . & . & r^2(p_0 p_k) \\ 1 & r^2(p_1 p_0) & 0 & . & . & . & r^2(p_1 p_k) \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ 1 & r^2(p_k p_0) & . & . & . & . & 0 \end{vmatrix}$$

where $k=1, 2, \dots, t$, has the sign of $(-1)^{k+1}$, (ii) for every pair (x, y) of points of S the determinants $D(p_0, p_1, \dots, p_t, x)$, $D(p_0, p_1, \dots, p_t, y)$, $D(p_0, p_1, \dots, p_t, x, y)$ vanish. We use this theorem in the case of $n=2$. Since each set of three points of a metric space is congruently contained in the euclidean plane, we must verify that for any four points p_0, p_1, p_2, p_3 of the metric space of the sites the determinant

$$D(p_0, p_1, p_2, p_3) = \begin{vmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & r^2(p_0p_1) & r^2(p_0p_2) & r^2(p_0p_3) \\ 1 & r^2(p_0p_1) & 0 & r^2(p_1p_2) & r^2(p_1p_3) \\ 1 & r^2(p_0p_2) & r^2(p_1p_2) & 0 & r^2(p_2p_3) \\ 1 & r^2(p_0p_3) & r^2(p_1p_3) & r^2(p_2p_3) & 0 \end{vmatrix}$$

vanish. If the metric space determined by the set of the sites and the metric r is imbeddable in the plane then we do imbedding (for example graphically). After this we compute the coordinates of the points of the plane corresponding to the sites and with the aid of the theory of regression the regression line to points corresponding in the plane to the sites. The correlation coefficient shows the position of the points which represent the sites in the plane, relative to the regression line. If correlation coefficient is equal to $+1$ or -1 then every point lies on the regression line. If the correlation coefficient differs from ± 1 then there exist points do not lie on the regression line. If the correlation coefficient is close to ± 1 then the distances of the points from the regression line which are outside of the regression line are small.

Inasmuch as each point is on the regression line, we consider the position of the points on this line as the chronological order of the sites. Otherwise we project the points onto the regression line perpendicularly, and we consider the position of the images as the chronological order. Thus the reliability of the chronological seriation depend upon the value of the correlation coefficient. If the correlation coefficient is close $+1$ or -1 then the chronological seriation is satisfactory. The advantage of this method is that we can control simultaneously the preciseness of the preliminary analyses.

References

[1] Brainerd, G. W., The place of chronological ordering in archeological analysis, *Amer. Antiquity,* v. 16, 1951, pp. 301—313.

Robinson, W. S., A method for chronological deposits, *Amer. Antiquity,* v. 16, 1951, pp. 293—301.

[2] Dempsey, P. & M. Baumhoff, The statistical use of artifact distributions to establish chronological sequence, *Amer. Antiquity,* v. 28, 1963, pp. 496—509.

[3] Blumenthal, L. M., *Theory and applications of distance geometry,* Oxford, 1953, p. 104.

(Received March 23, 1970)