4

A mathematical basis for simulation of seriatable data

I. Herzog*

I. Scollar†

4.1 Introduction

Seriation methods are probably one of the most published topics in the literature on computer applications in archaeology. At the time of writing, the authors have collected 117 references to the subject (included in the handbook distributed with the Bonn Seriation and Clustering Package), and there are probably a lot more. Only a very few workers have attempted to simulate the processes which lead to seriatable data, and to test these simulations having known sequences and properties with existing methods (Graham *et al.* 1976; Wilkinson 1974; Doran & Hodson 1975, p. 267–284). Some archaeologists have tested one or another method on real data whose dating is externally determined, e.g. through the physical geometry of the site or known chronological order rather than through find associations (Hodson 1968, Eggert *et al.* 1980). But a simulation technique which is founded on reasonable statistical premises about populations and type production can give greater insight into the limitations of a given method. In addition, simulated data is excellent for testing programs with regard to execution time and storage, since it is much easier to simulate a large quantity of data than to enter real information.

Simulation has also been used in other fields to model real processes. If the fit of the simulated model to real data is very good, it is then assumed that the parameters of the simulation also apply to the real data. As archaeological recovery can not be modeled by simple random sampling (there are systematic effects as well), it seems dangerous to use seriation simulation in this way, but there is nothing to prevent the user from doing so at his own risks and perils. In this paper, we describe some of the statistical and mathematical bases for the program SERSIM which will be distributed with Version 3.2 of the Bonn Seriation and Clustering Package (I. & I. 1987) with seriation algorithm

* Rheinisches Amt für Bodendenkmalpflege,
Colmantstr. 14,
D 5300 Bonn 1,
West Germany

† Rheinisches Amt für Bodendenkmalpflege,
Colmantstr. 14,
D 5300 Bonn 1,
West Germany

following Ihm, (Ihm 1982). Results of tests made with data produced by SERSIM will be described in the near future.

The following mathematical outline of the simulation of a feature complex is based partly on the simulation concept of (Graham *et al.* 1976). The new design offers greater variability in the choices of the type lifetime distribution function and of the population function. It has a more realistic concept of the type production function, and it replaces the confusing 'richness' parameter of that paper by the mean number of incidences per feature.

With the new program presence/absence as well as abundance data can be simulated. Also, the user may choose to simulate two feature groups which have only a certain number of types in common. Each feature complex simulation has three phases: First the types then the features and finally the incidences are generated.

4.2 Creating the types

Each type is produced only during its lifetime interval. It is assumed that the centres of these lifetime intervals are equally distributed in the period during which the feature complex was in use, plus a small margin. The type production function is a curve which has values zero at both lifetime interval border points. It is assumed that the curve increases rapidly to a maximum and then gently decreases to zero, the famous 'battleship' curve of the seriation literature.

Given the lifetime interval $[a, b]$ and the location of the maximum (c, d) with c less than $(a - b)/2$, how can we construct a curve which fulfills the properties mentioned above? Mathematically the properties are:

- the curve is monotone increasing in the interval $[a, c]$ from $(a, 0)$ to (c, d)
- the curve is monotone decreasing in the interval $[c, b]$ from (c, d) to $(b, 0)$
- the curve is continuous differentiable
- the only point within the interval with vanishing first derivative is c

It may not be evident that the type production curve should be continuously differentiable if the type is a pot for example, since only a certain number of pots are produced every day and a decrease in production is reflected by a lower number of total pots per day. But during the long time periods we are modelling, these differences level out and the trend remains of a slowly changing demand bringing the pot of our example in and out of fashion.

Our initial idea for modelling the type production function was to construct a cubic polynomial which is uniquely defined by the three points $(a, 0)$, (c, d), $(b, 0)$ and the condition that the first derivative in c should vanish. But if c is chosen close to a, the third zero of the cubic polynomial lies within the interval $[a, c]$ and so the function would violate at least two of the properties desired for the type production function. If a polynomial of degree four is constructed with the same conditions and the additional property that the first derivative of the curve vanishes in b, the same problem occurs. We also tried a cubic spline, but this led in some cases to an undesired maximum greater than d in the interval $[a, c]$.

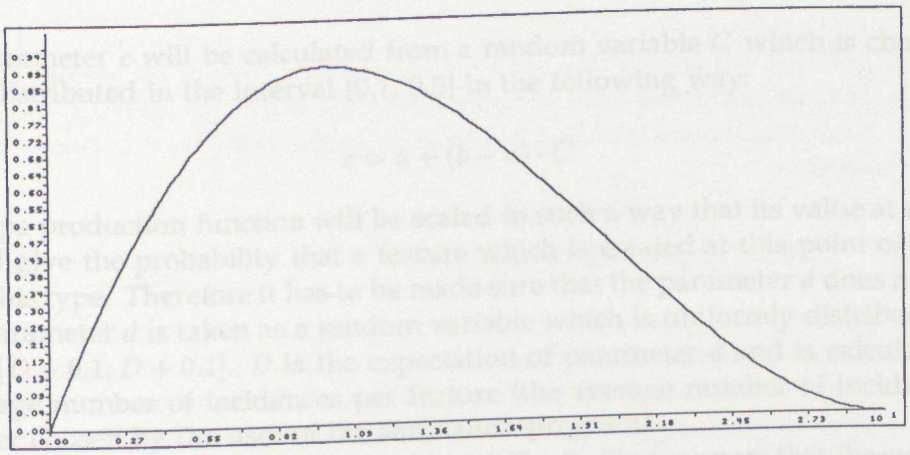

Figure 4.1: A type production curve with a lifetime of thirty years with the maximum after ten years of type production. The curve is scaled in such a way that the maximum production is at 0.9.

Finally we decided to use a curve which is partly a second degree and partly a third degree polynomial. It fulfills the above mentioned properties and its first derivative vanishes in b. This curve is given by the following formulas:

$$f(x) = \begin{cases} 0 & \text{if } x \leq a \text{ or } x \geq b \\ (x-a)(sx+t) & \text{if } a < x \leq c \\ (x-b)^2(\sigma x+\tau) & \text{if } c < x < b \end{cases}$$

with

$$s = -\frac{d}{(c-a)^2}$$
$$t = -\frac{d(a-2c)}{(c-a)^2}$$
$$\sigma = \frac{2d}{(c-b)^3}$$
$$\tau = d\frac{3c-b}{(c-b)^3}$$

There are two minor disadvantages in this method: It is not possible to construct a symmetric curve; and the second derivative in c is not continuous. In our experiments with this type production form, the resulting curves always appeared smooth as shown in Figs. 4.1 and 4.2.

The values a, b, c, d are random variables for each type. The random variables a and b are calculated from two other random variables, the mean type date and the type lifetime. As mentioned above, the mean type dates are equally distributed in the time interval of feature complex use plus a small margin.

Most archaeologists have no intuitive ideas about type lifetime distributions. So we offer an extremely flexible concept, allowing the archaeologist to experiment with three parameters governing the lifetime distribution function. It is assumed that the density of this function is piecewise linear, increasing in the time interval from 0 to t_1, constant from t_1 to t_2, and decreasing from t_2 to t_3. It is permissible to set date t_1 equal to t_2.

The expectation of a random variable with this distribution function is given by (with A = maximum of density function):

$$\frac{A}{6}(-t_1^2 + t_2^2 + t_3^2 + t_2 t_3) \quad \text{with} \quad A = \frac{2}{-t_1+t_2+t_3}$$

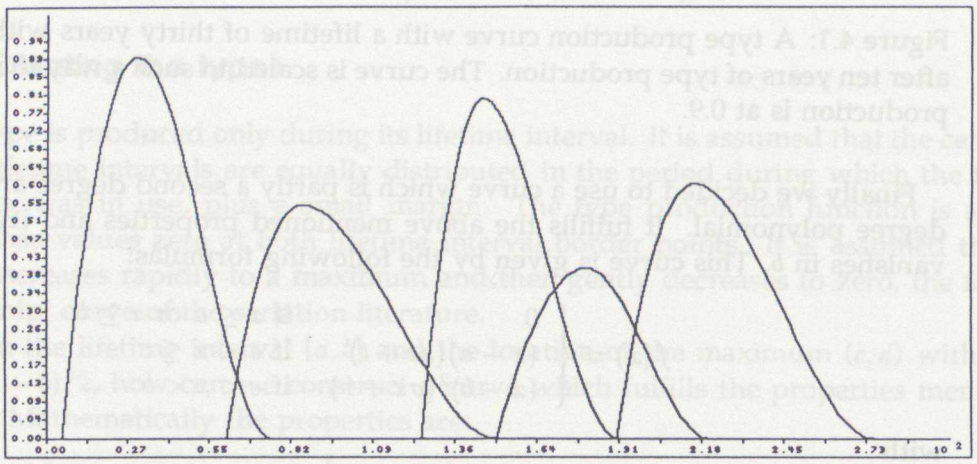

Figure 4.2: Five overlapping type production curves with different lifetimes and maximum heights distributed over a period of three hundred years. The lifetimes vary between 65 and 90 years, the mean is 76 years. The maxima are scaled so that they are in the range 0.4 ... 0.9.

The parameter c will be calculated from a random variable C which is chosen to be equally distributed in the interval [0.1, 0.5] in the following way:

$$c = a + (b - a) \cdot C$$

The type production function will be scaled in such a way that its value at a point of time will give the probability that a feature which is created at this point of time will contain this type. Therefore it has to be made sure that the parameter d does not exceed 1. The parameter d is taken as a random variable which is uniformly distributed in the interval $[D - 0.1, D + 0.1]$. D is the expectation of parameter d and is calculated from the average number of incidences per feature (the average number of incidences is a parameter chosen by the user of the simulation program).

A strict calculation of the value D is rather difficult. First we note that the expectation of the number of incidences at a given date T is equal to the sum of function values of all type production functions at this date. So if we integrate all type production functions over the period of feature complex use, sum these integrals up and divide the sum by the length of the feature complex usage period, we should get the expectation of the average number of incidences per feature. The full integral of a type production function with parameters a, b, c, d is given by:

$$\tfrac{2}{3}d(c - a) + \tfrac{1}{2}d(b - c) = d(\tfrac{1}{6}c - \tfrac{2}{3}a + \tfrac{1}{2}b)$$

There are also types whose production starts before the period of feature complex use and others whose production continues after the feature complex period. So we need the integrals over the beginning or ending segment of the type production function too. We get:

if $c_1 < c$:

$$\int_a^{c_1} f(x)dx = \frac{d(c_1 - a)^2}{3(c - a)^2}(3c - c_1 - 2a)$$

if $b_1 > c$:

$$\int_{b_1}^{b} f(x)dx = \frac{d(b_1 - b)^3}{2(c - b)^3}(b + b_1 - 2c)$$

The other segment integrals can be easily calculated by subtracting the above segment integrals from the full integral. The important thing to note is that all integrals (segment and full) are dependent on the factor d. This allows for the following way to calculate the expectation D of d: First, we set the factor d to one and evaluate all the integrals over the feature complex period and sum them up to a value we may call Intsum. By dividing Intsum by the length of the feature complex period we get the average number of incidences per feature, if d is to be one. But we want to calculate the expectation D of d from the average number of incidences as requested by the user of the simulation program. This is done by solving the following equation for D:

$$\frac{\text{Intsum}}{\text{FeatureComplexPeriodLength}} D = \text{AverageNumberOfIncidences}$$

Now it has to be made sure that D is neither greater than 0.9 nor less than 0.1 (because D is the expectation of a uniformly distributed random variable in the range

$[D - 0.1, D + 0.1]$ and this random variable may never be greater than 1 or less than 0). If D as calculated from the above formula is greater than 0.9, the program will take $D = 0.9$ and issue a warning to the user. In the same way D becoming less than 0.1 is avoided.

4.3 Creating graves or features

In general, grave or feature dates are not equally distributed within the time interval of feature complex use. We note that each grave has a well-defined date but this may not be so for features like pits or buildings which have been used for a certain period of time or contain earlier material. This difference is ignored in the following. In a stricter setup we should create a feature lifetime distribution function and develop a model as to how the number of objects which are thrown into the pit varies with time.

The density of a feature date distribution is in most cases proportional to the corresponding population function (the number of people living at a site at a given moment). Therefore the user of the simulation program will be asked to model the population function. There are some exceptions to the rule, for example in cases of epidemics and wars. In these cases it is useful to remember that the name population function has only chosen for convenience and that the proper name should be something like 'function reflecting the number of deaths depending on time' in the case of graves, for example.

Three ways are offered by the program to model the population function:

1. population remains constant

2. population increases or decreases linearly: User enters the population number at beginning date and end date of the feature complex.

3. population has at least one (local) minimum or maximum: User enters the population number at beginning date and end date of the feature complex and all minima and maxima in this time interval.

The third option is the only non-trivial one. In this case we have to find a curve which has the extreme values given by the user (and only these) which is smooth and as simple as possible. Our solution is to make up the population function with third and second degree polynomials so that the first derivative of the function is continuous.

This means in detail: Let t_1 be the beginning date of the feature complex, T_n the ending date, and $T_2, \ldots T_{n-1}$ the locations of the extrema. Let P_i be the population number at date $T_i, i = 1..n$. Then the population curve is constructed for each interval $[T_i, T_{i+1}]$:

1. If one of T_i, T_{i+1} is a boundary point and not an extremum, the local population curve is a second degree polynomial. There is either an extremum in T_i or in T_{i+1}. If the extremum is in T_i we get for the population function:

$$f(x) = ax^2 + bx + c \qquad x \in [T_i, T_{i+1}]$$
$$\text{with} \qquad a = \frac{P_{i+1} - P_i}{(T_{i+1} - T_i)^2} \quad b = -2aT_i \qquad c = P_i + aT_i^2$$

$[D - 0.1, D + 0.1]$ and this random variable may never be greater than 1 or less than 0). If D as calculated from the above formula is greater than 0.9, the program will take $D = 0.9$ and issue a warning to the user. In the same way D becoming less than 0.1 is avoided.

4.3 Creating graves or features

In general, grave or feature dates are not equally distributed within the time interval of feature complex use. We note that each grave has a well-defined date but this may not be so for features like pits or buildings which have been used for a certain period of time or contain earlier material. This difference is ignored in the following. In a stricter setup we should create a feature lifetime distribution function and develop a model as to how the number of objects which are thrown into the pit varies with time.

The density of a feature date distribution is in most cases proportional to the corresponding population function (the number of people living at a site at a given moment). Therefore the user of the simulation program will be asked to model the population function. There are some exceptions to the rule, for example in cases of epidemics and wars. In these cases it is useful to remember that the name population function has only chosen for convenience and that the proper name should be something like 'function reflecting the number of deaths depending on time' in the case of graves, for example.

Three ways are offered by the program to model the population function:

1. population remains constant

2. population increases or decreases linearly: User enters the population number at beginning date and end date of the feature complex.

3. population has at least one (local) minimum or maximum: User enters the population number at beginning date and end date of the feature complex and all minima and maxima in this time interval.

The third option is the only non-trivial one. In this case we have to find a curve which has the extreme values given by the user (and only these) which is smooth and as simple as possible. Our solution is to make up the population function with third and second degree polynomials so that the first derivative of the function is continuous.

This means in detail: Let t_1 be the beginning date of the feature complex, T_n the ending date, and $T_2, \ldots T_{n-1}$ the locations of the extrema. Let P_i be the population number at date $T_i, i = 1..n$. Then the population curve is constructed for each interval $[T_i, T_{i+1}]$:

1. If one of T_i, T_{i+1} is a boundary point and not an extremum, the local population curve is a second degree polynomial. There is either an extremum in T_i or in T_{i+1}. If the extremum is in T_i we get for the population function:

$$f(x) = ax^2 + bx + c \qquad x \in [T_i, T_{i+1}]$$
$$\text{with} \qquad a = \frac{P_{i+1} - P_i}{(T_{i+1} - T_i)^2} \quad b = -2aT_i \qquad c = P_i + aT_i^2$$

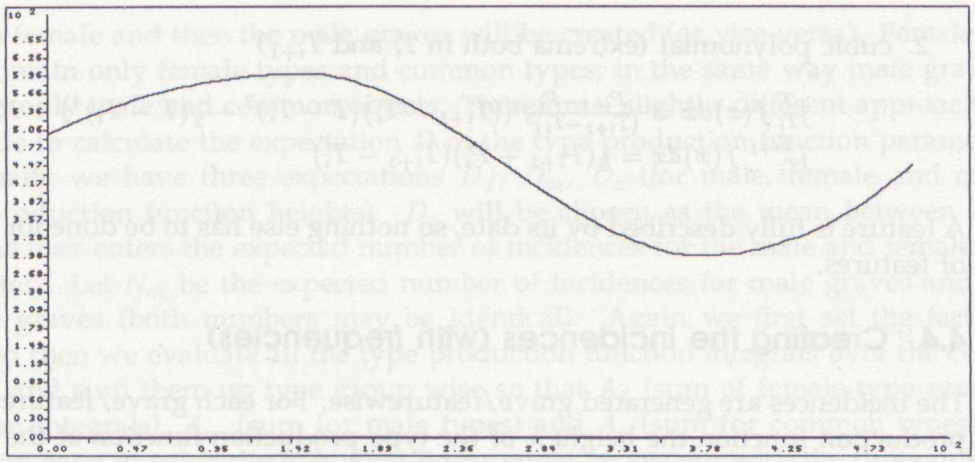

Figure 4.3: A population function showing a period of 500 years. At first the population number is 500, after 150 years a maximum of 600 people is reached and 230 years later the population number decreased to a minimum of 300 people and finally the population increases again to 450 people.

If the extremum is in T_{i+1} we get the symmetric result:

$$f(x) = ax^2 + bx + c \qquad x \in [T_i, T_{i+1}]$$
$$\text{with} \qquad a = \frac{P_i - P_{i+1}}{(T_i - T_{i+1})^2} \quad b = -2aT_{i+1} \quad c = P_{i+1} + aT_{i+1} + ^2$$

2. If both dates T_i and T_{i+1} are extrema, the local population curve is a third degree polynomial:

$$f(x) = ax^3 + bx^2 + cx + d \qquad x \in [T_i, T_{i+1}]$$
$$\text{with} \qquad a = \frac{2(P_{i+1} - P_i)}{(T_i - T_{i+1})^3} \quad b = -\frac{3}{2}a(T_i + T_{i+1})$$

This construction ensures that the first derivative vanishes at the extrema and that the curve is monotonic between two extrema. An example is given in Fig. 4.3.

The population function has to be normalised so that it may become the density function of the feature date distribution function. So we have to calculate the integral of the population function and to divide the P_is by this value. We also need the integrals at different points in time for generating random variables with the help of the distribution function (see section: Implementing the simulation concepts).

The integrals are given by:

1. second degree polynomial: if an extremum is in T_i:

$$\int_{T_i}^{T} f(x)dx = \frac{(P_{i+1} - P_i)}{3(T_{i+1} - T_i)^2}(T - T_i)^3 + P_i(T - T_i)$$
$$\int_{T_i}^{T_{i+1}} f(x)dx = \frac{1}{3}(P_{i+1} + 2P_i)(T_{i+1} - T_i)$$

if an extremum is in T_{i+1}:

$$\int_{T_i}^{T} f(x)dx = \frac{(P_i - P_{i+1})(T - T_i)^2}{(T_{i+1} - T_i)}\left(\frac{T - T_i}{3(T_{i+1} - T_i)} - 1\right) + P_i(T - T_i)$$
$$\int_{T_i}^{T_{i+1}} f(x)dx = \frac{1}{3}(P_i + 2P_{i+1})(T_{i+1} - T_i)$$

2. cubic polynomial (extrema both in T_i and T_{i+1})

$$\int_{T_i}^{T} f(x)dx = \frac{P_{i+1}-P_i}{(T_{i+1}-T_i)^3}\left((T_{i+1}-T_i)(T-T_i)^3 - \tfrac{1}{2}(T-T_i)^4\right) + P_i(T-T_i)$$
$$\int_{T_i}^{T_{i+1}} f(x)dx = \tfrac{1}{2}(P_{i+1}+P_i)(T_{i+1}-T_i)$$

A feature is fully described by its date, so nothing else has to be done for the generation of features.

4.4 Creating the incidences (with frequencies)

The incidences are generated grave/featurewise. For each grave/feature and each type production function the height h of the type production function at the grave/feature date is evaluated.

If a cemetery with presence/absence data is created, the height h of the type production function at the grave date gives the probability that the type is in the grave. It is easy to generate a random variable which is 1 with probability h and 0 with probability $1-h$, if a decent random number generator is available.

If a feature complex with abundance data is created, things are slightly more difficult: The user has entered the mean number of objects of one type in one feature (M). This number is now multiplied by the height h and divided by the mean height of all type production functions (H). The resulting value becomes the expectation for the number of objects of this type in the current feature. The distribution of the number of objects of a type in a feature is discrete with positive probabilities at non-negative integers.

We create such a distribution by mixing the Zero-one distribution or Alternative distribution, (Rektorys 1969) with the binomial distribution in such a way that the resulting distribution is approximately symmetric around its expectation (in the range $1 \ldots n$). We get:

$$P(X=0) = 1-h$$
$$P(X=j+1) = h\binom{n}{j}p^j(1-p)^{n-j} \qquad j = 0..n$$

with $E(X) = h(np+1) = h \cdot \frac{M}{H}$

and $n = 2 \cdot \left(\frac{M}{H} - 1\right)$

$p = \frac{1}{n}\left(\frac{M}{H} - 1\right)$

Note that the parameters p and n are constant for the whole feature complex.

4.5 Modification for two grave or feature groups

The concept described above can be modified for two grave or feature groups, for example if female and male graves are present which have only a certain percentage of types in common. The wording of this example will be kept in the following note about two feature groups.

We have three sorts of types: 'female' types, 'male' types and common types. The number of male and female graves, the number of male, female and common types will be determined by the user. Nothing will change in the generation of graves and types. Only the first types generated will be male, the next types will be female and the last type group will contain the common types; and also the first graves generated

will be female and then the male graves will be created (or vice-versa). Female graves may contain only female types and common types; in the same way male graves will contain only male and common objects. Therefore a slightly different approach has to be made to calculate the expectation D of the type production function parameter d.

Actually we have three expectations D_f, D_m, D_c (for male, female and common type production function heights). D_c will be chosen as the mean between D_f and D_m The user enters the expected number of incidences for the male and female graves separately. Let N_m be the expected number of incidences for male graves and N_f for female graves (both numbers may be identical). Again we first set the factor d to one and then we evaluate all the type production function integrals over the cemetery period and sum them up type group wise so that A_f (sum of female type production function integrals), A_m (sum for male types) and A_c (sum for common types) result. Now we have to solve the following linear equation system with the three unknowns D_f, D_m and D_c:

$$\begin{aligned} D_f A_f + D_c A_c &= N_f \cdot L \\ D_m A_m + D_c A_c &= N_m \cdot L \\ D_c &= \tfrac{1}{2}(D_m + D_f) \end{aligned}$$

with L=length of cemetery period.

The solution of this equation system is given by:

$$\begin{aligned} D_f &= \tfrac{L}{det}\left((A_m + \tfrac{1}{2}A_c)N_f - \tfrac{1}{2}A_c N_m\right) \\ D_m &= \tfrac{L}{det}\left(-\tfrac{1}{2}A_c N_f + (A_f + \tfrac{1}{2}A_c)N_m\right) \\ \text{with } det &= A_f A_m + \tfrac{1}{2}A_c(A_f + A_m) > 0 \end{aligned}$$

The incidences are generated in the same way as for one feature group, except that female features may contain only female and common types and that male features have male and common types only.

4.6 Implementing the simulation concept

It is difficult to create random variables with a computer when their density function is as complex as the normalized population function with several extrema. The general concept for generating random variables with density is to create a uniformly distributed random variable and then the inverse of the distribution function at this value (Knuth 1969, p. 102–103).

This concept can be applied to the type lifetime distribution function. The distribution function is derived from the density function by integration:

$$F(t) = \begin{cases} \frac{A}{2t_1}t^2 & \text{if } 0 \le t < t_1 \\ A(t - \frac{1}{2}t_1) & \text{if } t_1 \le t < t_2 \\ 1 - \frac{A}{2(t_3 - t_2)}(t_3 - t)^2 & \text{if } t_2 \le t \le t_3 \end{cases}$$

The inverse of the type lifetime distribution function is easily calculated and has the following form:

$$F^{-1}(x) = \begin{cases} +\sqrt{\frac{2t_1}{A}x} & \text{if } 0 \le x < \frac{A}{2}t_1 \\ \frac{1}{A}x + \frac{1}{2}t_1 & \text{if } \frac{A}{2}t_1 \le x < At_2 - \frac{A}{2}t_1 \\ t_3 - \sqrt{\frac{2}{A}(t_3 - t_2)(1 - x)} & \text{if } At_2 - \frac{A}{2}t_1 \le x \le 1 \end{cases}$$

So with a decent random number generator the type lifetimes can be easily generated. The population distribution function is partly a third and partly a fourth degree polynomial and it is not easy to invert these polynomials in a numerically stable way. So we decided to evaluate the distribution function at a large set of points and to approximate it by piecewise linear functions which can be easily inverted. For construction of a grave or feature date, a uniformly distributed random variable is created; via table lookup the corresponding date interval is identified, the distribution function is approximated by a linear function in this interval and then this linear function is inverted at the value of the uniformly distributed random variable.

References

DORAN, J. & F. R. HODSON 1975. *Mathematics and computers in Archaeology.* Edinburgh University Press, Edinburgh.

EGGERT, M., S. KURZ, & H. P. WOTZKA 1980. "Historische Realität und archäologische Datierung", *Prähistorische Zeitschrift*, 55: 110–145.

GRAHAM, I., P. GALLOWAY, & I. SCOLLAR 1976. "Model Studies in Computer Seriation", *Journal of Archaeological Science*, 3: 1–30.

HODSON, F. R. 1968. *The La Tène Cemetery at Münsingen-Rain*, Acta Bernensia V. Stämpfli, Berne.

I., H. & SCOLLAR I. 1987. "Ein 'Werkzeugkasten' für Seriation und Clusteranalyse", *Archäologisches Korrespondenzblatt*, 17: 273–279.

IHM, P. 1982. "Ein einfacher Algorithmus zur Bestimmung des dominanten Eigenvektorpaares bei einer Korrespondenzanalyse". *in Studien zur Klassifikation*, 10, pp. 54–57. Indeks Verlag, Frankfurt.

KNUTH, D. E. 1969. *The Art of Computer Programming, Vol. 2: Seminumerical Algorithms.* Addison-Wesley Publishing Company, Reading Massachusetts.

REKTORYS, K., (ed.) 1969. *Survey of Applicable Mathematics*, 1255. The M.I.T. Press, Cambridge, Massachusetts.

WILKINSON, E. M. 1974. "Techniques of data analysis: seriation theory", *Archaeo-Physika*, 5: 3–142.