16

Combining stratigraphic information and finds

Irmela Herzog

(Rheinisches Amt für Bodendenkmalpflege, Bonn, Germany)

16.1. Introduction

In 1980 Orton discussed how to include information about finds in a Harris diagram. One suggestion was to include symbols for the presence or absence of finds in the diagram. This is quite easy if only four find types are considered as in Orton's example. Since the presence of a type is more significant than its absence, it appears to be sensible to include symbols solely for the presence of types. An example which was created by a new version of my program HARRIS (Herzog & Scollar 1991, Herzog 1993) is given in Fig. 16.1.

Another presentation of finds is as a table resulting from seriation. All the strata on a single level in the Harris diagram may be merged and presented as one feature in the table. Alternatively, the strata within a level could be ordered arbitrarily and all strata containing finds shown in the table. The latter presentation was added to the HARRIS program.

But finds help us with dating a stratum, especially if a stratum belongs to a so-called floating sequence. As Orton (1980) mentions, the Harris diagram is based on the relation "is older than". The set of strata and the relations form a partially ordered set. There are many ways to create a linear ordering from a partially ordered set and the number of Harris diagrams which can be drawn based on this set structure is quite large as Fig. 16.2 illustrates.

In this example, there are ten strata with numbers 11 to 20 each of which can be moved within a range of ten stratum levels, indicated by the strata with numbers 0 to 9 on the left hand side of the diagram. One may now describe the configuration of strata 11 to 20 by a ten digit number, the first digit is the position of stratum 11 (which is 4 in the above example), the second digit gives the position of stratum 12 and so on. Every number with ten or less digits corresponds to a stratum configuration in this example, so that the total number of configurations is 10^{10}. This is only a lower bound for the number of configurations, as the diagram may also be drawn so that all strata 11 to 20 appear earlier than strata 0 to 9, (or vice versa), or intermediate levels may be created so that, for example, stratum 11 is positioned on a level between strata 3 and 4.

The aim is now to choose the one which reflects best the find information out of the millions of possibilities. But even if the layout of a Harris diagram with a PC takes only 1 second, with 10^{10} layouts the PC user will have to wait 317 years until all layouts are created. So this naïve approach is not feasible. But there is another problem: What exactly is the criterion which can be used by the computer to decide which diagram is best? At the beginning of my study, I had the idea that this criterion could well be the canonical correlation coefficient which is maximised in correspondence analysis (Greenacre 1984), but I was not sure. So I decided to work with simulated data first. A diagram is considered optimal if it reflects the simulated chronological sequence of these strata.

16.2. Simulating stratigraphic data sets with finds

The finds content of the strata and their date was simulated with the help of an already existing program (Herzog & Scollar 1988) called SERSIM. The only disadvantage of this program is that it is not possible to create truly contemporary strata. So for the subsequent simulation program HARRISIM, which creates the stratigraphic relations, contemporary strata are consecutive strata in the SERSIM sequence.

The HARRISIM simulation program creates an ideal situation without errors of observation of the above, below or contemporary relations, without complex archaeological strata like outlines of pits and without any redeposition of finds if finds are considered as well.

It is very difficult to describe three-dimensional strata mathematically. To simplify the theory, each stratum is considered to consist of small cubes or bricks which are adjacent to each other. If the cubes are small enough, one may use this simplification to build a model of an actual site with very little error. In my model, I consider a rectangular area of N x M cubes which corresponds to the excavation area. If a single profile is to be considered, N or M may be set to 1.

The first cube of the first stratum is given an arbitrary position within this rectangle, then the stratum grows cube by cube in arbitrary directions. All the other strata follow. If a new stratum is created, the probability that its first cube is at the lowest position within the rectangle is highest. Additionally, the stratum tends to grow in the direction of its lowest cube neighbour. This simulation technique tends to create complex stratum shapes like hot wax poured into cold water. An example view of the bricks during the formation process can be seen in Fig. 16.3. Each pattern corresponds to a different stratum on the surface.

If a cube of a stratum overlies a cube of another stratum this relationship is stored in the relationship data base used by the HARRIS program. In general, quite a few redundant relations are created. If five strata are contemporary, five non-overlying positions for the first cubes of the strata are determined, and the cube growing process ensures that the cubes of these strata will not overlie each other.

By increasing the area of the rectangular model excavation area and holding the cube numbers per stratum constant, more strata will be needed to cover the whole area and the probability that contiguous strata will overlap will

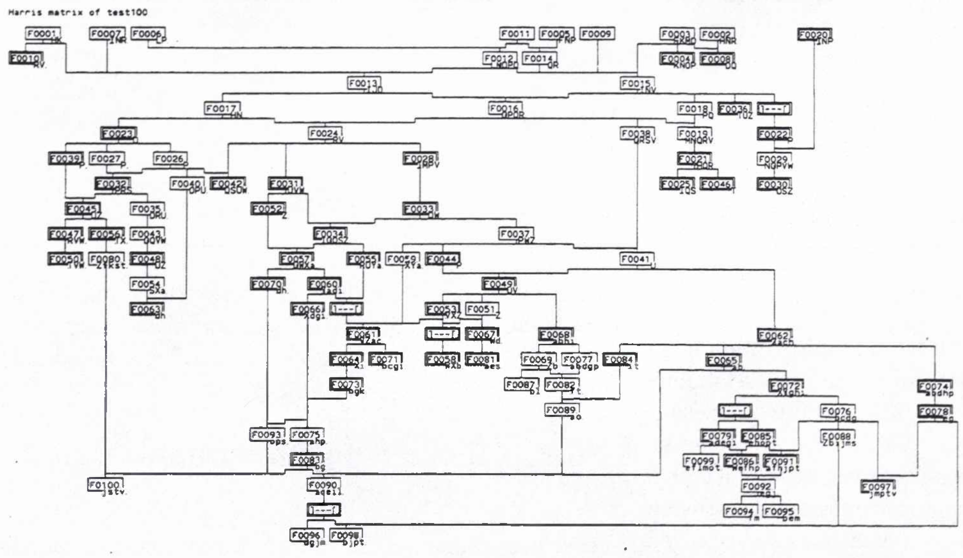

Figure 16.1: A simulated Harris data set with 100 strata. The presence of a find type is indicated by a letter in the lower right corner of the stratum box. The chronological order of the find type indicators is reflected by the find identifiers, capital letter finds are later than lower case letter finds, the finds within each letter group are sorted alphabetically. Similar examples of labelling a Harris diagram are given by Harris and Reece.

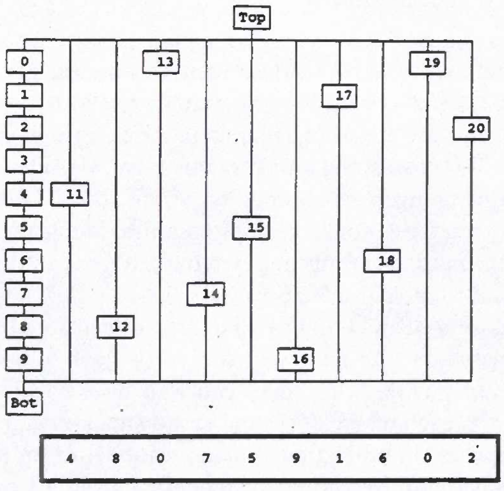

Figure 16.2: Example to show that the Harris matrix is very large, in general. To each level-assignment of strata 11 to 20 corresponds a 10 digit number, in this example the number is 4807591602.

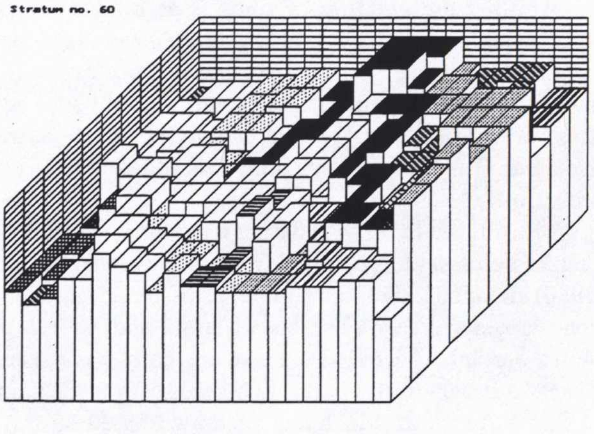

Figure 16.3: The surface of the excavation area after sixty strata have been simulated. Each stratum consists of bricks, the mean number of bricks per stratum is 80. The bricks belonging to one stratum have the same surface pattern.

decrease. The corresponding effect in the Harris diagram is that fewer levels are needed to show the strata and their relationships. So it may well happen, contrary to most published Harris diagrams, that the diagram is a lot larger in breadth than in depth (see Fig. 16.1).

The example presented in Fig. 16.1 will be the model for testing the methods suggested in this paper. It is called TEST100, because 100 strata were simulated, 88 of which contain at least two different find types. Some pain was taken to ensure that neither Harris analysis nor seriation leads to an optimal result for this data set. According to the simulation parameters, the 100 strata were created within 50 years, the mean lifetime of types was set to 30 years, 3.5 find types are present on average in each stratum. The Harris diagram created by program HARRIS arranges the strata on 21 different levels, the 41st and the 80th stratum in chronological sequence are on the same level. The sequence calculated by seriation shown in Fig. 16.4 can be improved as well. For example the 29th, 10th and 40th stratum appear consecutively. Layer numbers of the simulated layers indicate the ideal ordering. In this example they have been compressed into 21 levels by the HARRIS program. The problem then remains as one of how best to expand this model to something approaching a "correct" solution.

How can the goodness of fit of the relative chronology and the simulated sequence be measured? My first idea was to use Kendall's τ. But W. Vach, of Freiburg University drew my attention to the fact that this sequence correlation coefficient tends to favour Harris diagrams with very few levels. He suggested using the number of concordant and discordant pairs when comparing the simulated and the calculated sequence. With a simulated chronological sequence of n strata s_i with dates d_i given, f the function which assigns each stratum its position in relative chronology, whenever $d_i < d_j$ and $f(s_i) < f(s_j)$, the number of concordant pairs is increased, if $f(s_i)>f(s_j)$, the number of discordant pairs is incremented. So if the simulated sequence equals

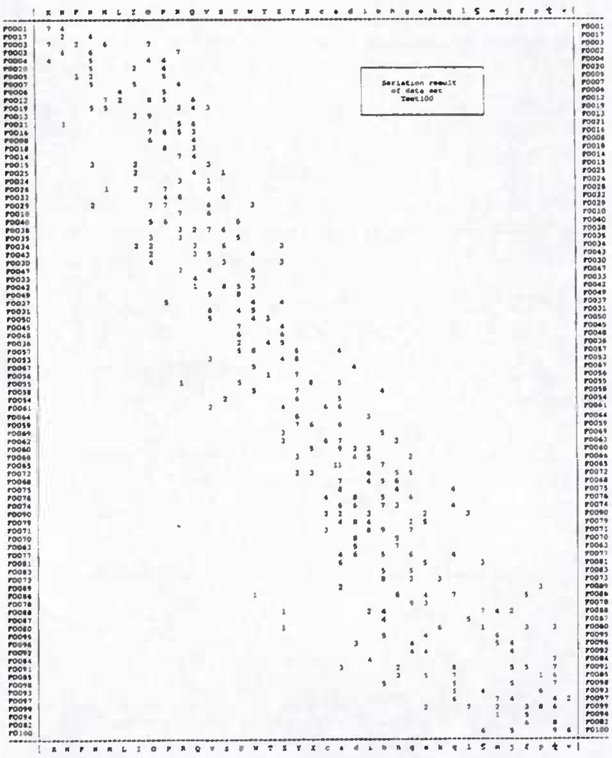

Figure 16.4: Seriation result of the simulated data set TEST100. The mean type lifetime is quite long compared to the stratum creation time so that the seriation result is not optimal. The simulated chronological order of strata is F0001, F0002, F0003 and so on, the ideal order for types is A..Z, a..z.

the calculated sequence, the number of concordant pairs is n(n − 1)/2, the number of discordant pairs is 0. With seriation, the function f assigns each stratum its score, in the Harris diagram each stratum is assigned the level it is on. So with the Harris diagram, pairs on the same level are not comparable and add neither to the concordant nor to the discordant count. In order to be able to create an ordering for the types with the help of the Harris diagram, the corresponding function f assigns each type the average level of this type in the diagram. Table 16.1 presents the results of the comparison between calculated and simulated sequence for data set TEST100.

16.3. A simple algorithm to combine stratigraphy and seriation results

When I first started to think about the problem of combining stratigraphy and finds information I planned to solve this problem via some maximisation technique used in discrete mathematics like simulated annealing (Press, Flannery, Teukolsky & Vetterling 1986). The value to be maximised is the canonical correlation coefficient, but the difficulty is to determine the procedure which generates small changes in the configuration. To move a stratum one layer up or down seems to be an obvious choice for this procedure, but not all strata can be moved upwards any further, and if they are moved downwards, there may be a domino effect in that hundreds of other strata are moved downwards, too.

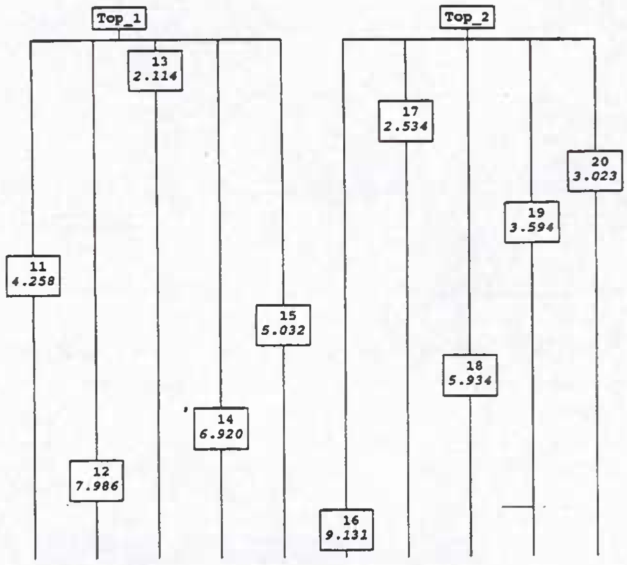

Figure 16.5: This diagram illustrates the simple combination of Harris diagram and seriation results. The stratum boxes show the seriation scores in italics. If several strata are candidates for a new level (in this example level 2), they are sorted according to the seriation scores.

So it is probable that this method is feasible but is likely to be very time consuming.

Another approach is to put those strata on one level which are sufficiently similar according to some measure of similarity and to start the next level with the stratum which is most similar to the previous level´s strata. This method is quite similar to single-linkage clustering and would have all the known disadvantages of this method.

The method suggested in this paper uses the seriation sequence if Harris analysis does not determine the strata sequence. The algorithm is an extension of the level assignment procedure which is well known in graph theory and which was introduced into stratigraphic analysis by Magnar Dalland (1984). All strata which have no later relations are placed in a candidate stack. The stratum with the lowest seriation score is placed on the first level. All later relations which end at this stratum are removed from consideration, and all strata which therefore have no later relations are placed in the candidate stack as well. The stratum with the lowest seriation score is placed on the second level, and so on, until the level of each stratum is determined. Fig. 16.5 gives an example for this procedure.

In practice there are some minor difficulties which must be overcome before implementing this procedure. The seriation algorithm creates only a roughly chronological sequence of the strata, but the direction of the sequence is not determined, i.e. the earliest strata may come first contrary to the requirements of the algorithm. There are two ways to solve this problem: Either Kendall's τ is calculated comparing Harris and seriation sequence *a posteriori*, and if it is negative the seriation sequence is turned upside down. Alternatively, the seriation algorithm is started with the sequence as calculated for the Harris diagram (without taking finds into account), the reciprocal averaging process

TEST100	*Strata Total*	*Kendall's tau*	*Number Concord. Pairs*	*Number Discord. Pairs*	*Type Total*	*Kendall's tau*	*Number Concord. Pairs*	*Number Discord Pairs*	*Canon. Corr.Coef.*
Ideal	98	1.00000	4753	0	38	1.00000	703	0	0.9260
Seriation	88	0.86677	3573	255	37	0.85586	619	47	0.9715
Harris	98	0.89476	4303	239	38	0.80085	633	70	0.9026

Table 16.1: The goodness of fit of seriation and Harris matrix analysis compared to the simulated sequence for data set TEST100. The closer an algorithm´s results are to the ideal values, the better the method. On the left hand side are the results for strata, on the right hand side the results for types. Additionally, the canonical correlation coefficient was calculated which is amazingly small in the ideal case compared to the coefficient reached by seriation.

TEST100	*Stratum Total*	*Kendall's tau*	*Number Concord. pairs*	*Number Discord. pairs*	*Type Total*	*Kendall's tau*	*Number Concord. pairs*	*Number Discord. pairs*	*Canon. Corr. Coeff.*
Ideal	98	1.00000	4753	0	38	1.00000	703	0	0.9260
Seriation	88	0.86677	3573	255	37	0.85586	619	47	0.9715
Harris	98	0.89476	4303	239	38	0.80085	633	70	0.9026
Combination 1	98	0.89471	4499	250	38	0.85491	652	51	0.9576
Combination 2	88	0.88871	3615	213	37	0.86186	620	46	0.9561
Comb./Var. (10)	98	0.92395	4427	175	38	0.83215	644	59	0.9469

Table 16.2: The goodness of fit of the Combination and the Combination with variance method for data set TEST100.

TEST400	*Stratum Total*	*Kendall's tau*	*Number Concord. pairs*	*Number Discord. pairs*	*Type Total*	*Kendall's tau*	*Number Concord. pairs.*	*Number Discord. pairs*	*Canon. Corr. Coeff.*
Ideal	387	1.00000	74691	0	94	1.00000	4371	0	0.9816
Seriation	345	0.94937	57836	1502	91	0.94969	3992	103	0.9974
Harris	387	0.88366	68778	4248	94	0.93959	4238	132	0.9765
Combination	387	0.96062	73194	1470	94	0.95287	4268	103	0.9923
Comb./ Var.(16)	387	0.96729	72243	1201	94	0.95469	4271	99	0.9926

Table 16.3: The table showing the results of analysing data set TEST400.

tries to refine this sequence iteratively and will converge to the solution which is closest to the Harris diagram. Another problem is contemporary strata which are to be put on the same level, regardless of any differences in their seriation scores. Therefore, these strata are assigned the average score of all the strata with which they are contemporary. If strata without finds (and therefore without seriation scores) appear in the candidate list, they are positioned on the level which is created next.

Having programmed this algorithm I wanted to test it with my model data set TEST400, which is a simulated data set with 400 strata. TEST400 was my model data set before I started to think about presenting the results in a publication, i.e. before I started worrying about how to put one data set on a single sheet of paper. It took my 386-20 computer more than an hour to compute the layout for this diagram, which is about ten times more than a normal layout. Additionally, the storage requirements on disk were tenfold compared to normal diagram generation because the diagram was a lot longer than before. For exact figures see Table 16.4.

I felt that the computation time and the final size of the diagram were not in favour of this method. But, if the results are presented as a seriation table, neither computation time nor amount of paper required is high, so the new version of the HARRIS program supports this output form. The result for data set TEST100 is given in Fig. 16.6. The goodness of fit of this method and the one which is presented next will be discussed in the following section.

TEST400	*Levels*	*Size in kb*	*Minutes*
Harris	45	182	7
Combination	393	2088	60
Comb./Var(16)	72	348	11

Table 16.4: Time and storage requirements for laying out the data set TEST400. Note that the combination method needs less time than any other method, if only seriation table output is generated.

16.4. Combining stratigraphy and finds using the variance method, results and evaluation

The results of the method discussed above were not bad, but the disadvantage is that no Harris diagram can be produced sensibly with this procedure. So I thought of putting those strata on one level which are reasonably similar. Similarity was to be defined with the help of the seriation scores. Unfortunately, these scores are not equally distributed even if the time difference between the strata is equal. But large gaps in stratum scores normally correspond to large gaps in chronology. In a variation of the method discussed in the previous section, first all forward variances of all strata are calculated within a certain range, say 10; i.e. for a stratum, a variance calculation is made taking into account the next 10 strata scores in the seriation sequence. All candidates in the stack are allowed to enter the current level if the scores are within twice the standard deviation of the stratum with the lowest score. The range is user-defined, and should be increased with large data sets. This method requires only little more effort than normal diagram layout if the range is properly chosen. Fig. 16.7 shows the results of this method for data set TEST100. The program generated 30 different stratum levels, with the variance range set to 10. Table 16.2 indicates which method comes closest to the simulated sequence.

Combination 1 in Table 16.2 means that all strata containing a find are considered (as in Harris analysis), Combination 2 is restricted on strata with at least two different find types (as in seriation). This way it is easy to compare Combination 2 and seriation and Combination 1 and Harris analysis. The results show that the combination algorithm is superior to the two separate methods. With the variance method there are fewer stratum pairs which are comparable, therefore a decrease in the number of concordant pairs is noted, but on the other hand the number of discordant pairs was decreased by approximately the same amount, so that this solution is not inferior to the simple combination method. The explanation of the phenomenon is, that with strata which are close together in time, the probability of putting these strata into the wrong sequence is quite high. If these strata are made non-comparable, the number of discordant pairs is likely to decrease.

The final column in Table 16.2 gives the canonical correlation coefficients of the different results. It is amazing to see that the Combination1 coefficient is higher than the Combination with variance coefficient, though it was just shown that both solutions are of equal quality. But the table shows too, that the ideal scores, the scaled simulated dates, form a significantly lower canonical correlation coefficient than the methods which take the seriation sequence into account. Therefore, it remains to be shown whether optimising the canonical correlation coefficient will lead to better results.

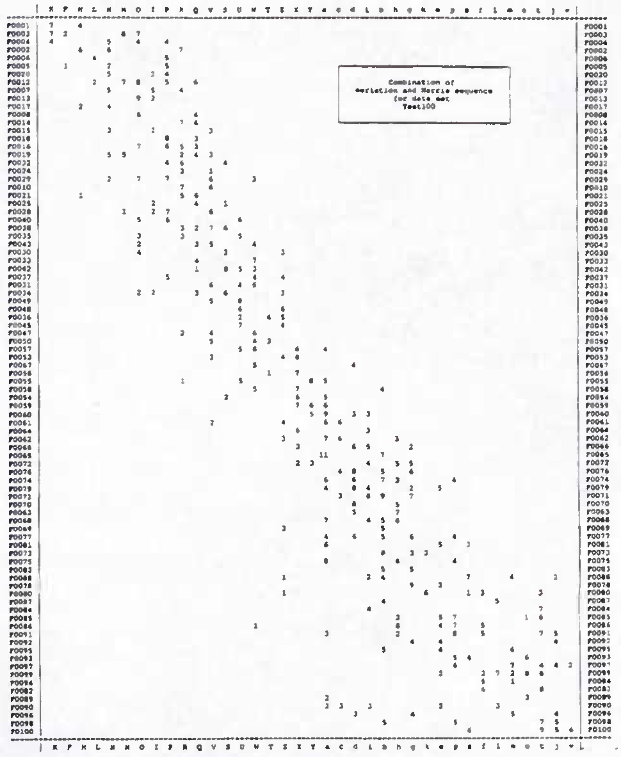

Figure 16.6: The result of the simple combination method in seriation table form for model data set TEST100.

The TEST400 data set has a more accurate seriation result than TEST100, because here the time difference between oldest and youngest stratum is 100 years, the mean type lifetime 22 years, so that the production times of the types do not overlap as much as in TEST100. The initial Harris diagram consists of 45 levels, so there are about nine strata to each level whereas with TEST100 there are five strata on a level on the average. Even with good conditions for seriation as in this example, these results can be further improved as is shown in Table 16.3. The variance method, this time with a range parameter of 16, reaches nearly the quality of the simple combination method.

Table 16.4 shows that the penalty paid when using the simple combination method for Harris diagram layout is about ten times the normal effort in time and space, whereas with the variance method the amount of time and space which are required still remains less than twice the normal effort.

16.5. Conclusion

Two methods are presented to permit combining stratigraphic information and the results of seriation. With the help of simulation experiments, it was shown that these

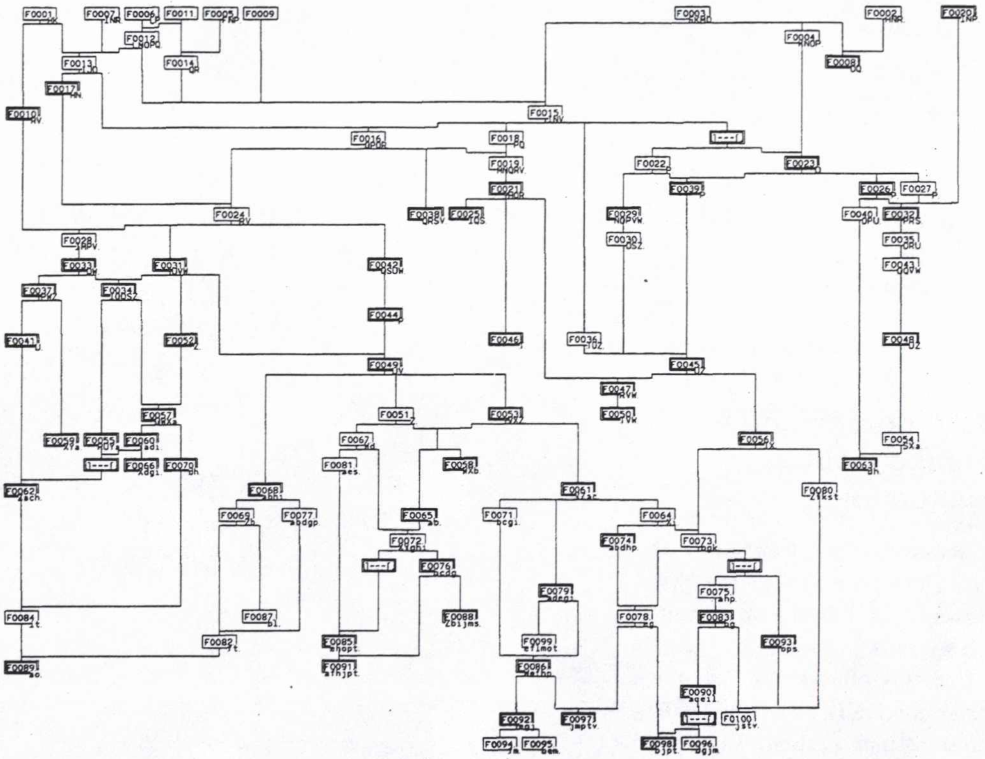

Figure 16.7: The result of the combination with variance method for data set TEST100. For variance calculation, the ten next neighbours with higher seriation scores were considered. The number of levels was increased from 21 to 30.

methods are superior to seriation and Harris diagram analysis calculated separately. Real data tend to contain more errors which are often not randomly distributed. Gerrard (1993) discusses several ways how the artefactual content of a stratum may become corrupted and presents a method to detect these corruptions. In this paper it was assumed that all the stratigraphic relations are without errors. In practice it will be necessary to define some kind of procedure which will show major differences between observed stratigraphic relations and calculated seriation sequences, so that errors in the excavation records concerning the stratigraphic relations may be detected as well.

Acknowledgements

I would like to thank Johanna Banck, Landesdenkmalamt Baden-Württemberg, for suggesting putting in symbols to indicate artefacts in the Harris diagram. She is actually working on quite a different problem, analysing cloth layers and their substances of an ancient bed. This may show that stratigraphic analysis may see some applications in many fields in the future. Additionally, I would like to thank Werner Vach for his suggestions, Frank Siegmund and Clive Bridger for encouragement and Irwin Scollar for correcting my English.

Bibliography

DALLAND, M. 1984. "A procedure for use in stratigraphical analysis", *Scottish Archaeological Review* 3(2): 116–127.

GERRARD, R. H. 1993. "Beyond crossmends: stratigraphic analysis and the content of historic artefact assemblages on urban sites", *in* Harris, E. C., M. Brown III & G. Brown (eds), *Practices of archaeological stratigraphy*, pp. 229–249. Academic Press, London.

GREENACRE, M. J. 1984. *Theory and applications of correspondence analysis*. Academic Press, London.

HERZOG, I. & I. SCOLLAR 1988. "A mathematical basis for simulation of seriatable data", *in* Rahtz, S. (ed.), *Computer and Quantitative Methods in Archaeology 1988*, pp. 53–62. British Archaeological Reports International Series 446(ii). Oxford.

HERZOG, I. & I. SCOLLAR 1991. "A new graph theoretic oriented program for Harris matrix analysis", *in* Lockyear, K. & S. Rahtz (eds), *Computer applications and quantitative methods in archaeology 1990*, pp. 53–59. British Archaeological Reports International Series 565, Oxford.

HERZOG, I. 1993. "Computer-aided Harris matrix generation", *in* Harris, E. C. , M. Brown III and G. Brown (eds), *Practices of archaeological stratigraphy*, pp. 201–217. Academic Press, London.

ORTON, C. 1980. *Mathematics in Archaeology*. Collins, London.

PRESS, W. H., B. P. FLANNERY, S. A. TEUKOLSKY & W. T. VETTERLING 1986. *Numerical Recipes: The art of scientific computing*. Cambridge University Press, Cambridge.

I. Herzog
Rheinisches Amt für Bodendenkmalpflege
Endenicher Str. 133
Germany W-53115 Bonn.