INTERACTIVE SPATIAL ANALYSIS ON A MICRO

J.G.B. Haigh† and M.A. Kelly††

School of Archaeological Sciences, University of Bradford,
Bradford, West Yorkshire, BD7 1DP.

1. Introduction

The majority of archaeological observations contain a strong subjective element, such as the assessment of style, purpose or shape. Consequently there are only limited opportunities for the objective application of quantitative methods to archaeological data. One such opportunity lies in the analysis of the distribution of some class of object over an appropriate region. Even locational data of this type have some subjective elements: since the objects will have finite size, some way must be defined for specifying the coordinates of the "centre" of an object, and it is necessary to decide precisely which objects belong to the chosen class. Nevertheless, provided that the class is a distinctive one and that the objects within it are small compared with their spatial separation, the subjective elements may be considered insignificant.

The spatial analysis of distributions of "point" objects is thus an attractive proposition for quantitative archaeology, and the use of nearest neighbour analysis has been discussed at some length by Hodder and Orton (1976). Unfortunately, contradictory effects may mask each other in the calculation of a single nearest neighbour statistic, and highly structured data may fail to show any significant deviation from random. The authors therefore decided to investigate the possibility of using a number of distances to higher order neighbours, in order that various structural aspects may be revealed in different ways (Kelly, 1986).

2. Limitations of classical nearest neighbour analysis

Nearest neighbour analysis was originally developed for ecological application (Clark and Evans, 1954) and was soon extended for higher-order neighbours (Thompson, 1956). In ecology, the techniques are usually applied to a large population of points, which can be considered to have been produced by a uniform stochastic process and from which comparatively small samples may be taken. With these considerations it is reasonable to assume that the classical formulae may be used to estimate the expected mean and variance for the neighbour distances.

† Also attached to the School of Mathematical Sciences.
†† Currently at Bradford University Software Services Ltd.

In archaeology, however, the underlying stochastic processes are likely to be much more complicated and the populations to be much smaller. It is then desirable to evaluate the neighbour distances for every point in the population, in order that the "sample" size may be sufficiently large to give a statistically significant result, and that a global description of the distribution may be obtained. Under those circumstances the unmodified classical formulae may no longer be applied. The first reason is that each point appears both as a sample point and as a neighbour to other sample points, resulting in dependence among the neighbour distances; the appropriate modification for this effect is fairly well established. The second reason is that a small population must have a finite boundary, and that points near the boundary must be expected to have larger neighbour distances than those in the interior; the modifications for edge effects are much larger than those for interdependence among the distances.

There have been several attempts to find the appropriate modifications, both by purely analytical means and with the aid of simulation. In general, the analytical attempts have not been very successful, but Donnelly (1978) has published an interesting set of results based on simulation. Unfortunately Donnelly established his formulae only for nearest neighbour distances, and the authors have set out to establish corresponding formulae for higher-order neighbours.

3. New corrections for k-th order neighbours

By a combination of empirical analysis and extended simulation, the authors have concluded that the following modified formulae apply for the mean k-th neighbour distance ρ_k and for the variance of the mean:

$$\tilde{E}(\rho_k) = c_k \sqrt{\frac{A + bLE(\rho_k)}{n - 1}}$$

$$\tilde{var}(\bar{\rho}_k) = 1.029\ var(\bar{\rho}_k) + \frac{\sqrt{A}\, gL}{n(n-1)^{\frac{3}{2}}}$$

where
$$\left.\begin{aligned} E(\rho_k) &= c_k \sqrt{\frac{A}{n-1}} \\ var(\bar{\rho}_k) &= \frac{A}{n(n-1)}\left(\frac{k}{\pi} - c_k^2\right) \end{aligned}\right\} \text{ the classical formulae}$$

$$c_k = \frac{1.3.5. \ldots .(2k-1)}{2^k (k-1)!}$$

$$b = 0.3934 - 0.0425896\ \ln(k)\ \exp(-0.1803368k)$$

$$g = 0.03059k^{1.357},$$

and there are n points having a uniformly random distribution over an area A enclosed by a boundary of length L.

The authors intend to describe the derivation of these formulae elsewhere, and for present purposes it is sufficient to give reasons for accepting the reliability of the formulae. The first reason is that the mean and variance are correct in the classical limit for large values of n, with an appropriate correction in the variance for the effects of interdependence of distances between neighbours. Secondly, in the case of nearest neighbour distances when k = 1, the formulae are in very good agreement with those of Donnelly (1978), although account has to be taken of Donnelly's expressions being in terms of n, without allowing for the fact that a given point has only (n - 1) neighbours. Thirdly, the mean neighbour distances and their variances have been calculated for a large number and variety of random distributions and have been found to agree with the formulae to within statistical variation.

Donnelly (1978) explains that his expressions are likely to be valid for regions of comparatively simple geometry, but are unlikely to work for regions with complicated boundaries. The authors have found a fairly sharp deviation from the formulae and have attempted to quantify the effect, in terms of the distance of the centroid of the region from the nearest point on the boundary. If this distance is less than twice the mean k-th neighbour distance, then the formulae are not reliable for that value of k. This rule may be rather biassed against crescent-shaped or toroidal areas, but works very well for a wide variety of common shapes. The authors' formulae are likely to be reliable over a wider range of shapes than those of Donnelly.

The variance of the mean increases very rapidly with k, and it is not easy to fix the formula for it with absolute certainty. Even for the nearest neighbour distance, k = 1, there is some uncertainty in the choice of parameters in the variance formula. Donnelly quotes his parameters with great confidence, but that confidence does not appear entirely reflected in the curves which he presents as evidence for his choice.

4. Implementation on the micro

The main concern of this paper is not with the formulae themselves, but with the presentation of the data which are to be analysed in terms of them. The ideal way of obtaining data is on the basis of a computerised database, from which a file of coordinates may be obtained automatically. Indeed, the authors have carried out spatial analyses of sites in North Yorkshire, based upon searches of the County's Sites and Monuments Record.

In most cases, however, the data are likely to be obtained from traditional archaeological sources, namely maps or plans on which the objects are represented as symbols. To determine coordinates from such sources, using ordinary measuring instruments, is laborious, time-consuming, and probably not particularly reliable. The authors have attempted to provide convenient and reliable means of transferring spatial information from a plan to a computer, of subsequently managing and analysing the data, and of presenting the results in a useful form.

The computer used for the work consists of a Research Machines RML 380Z computer, which is interfaced through its RS232 port to a Hewlett-Packard 7475A plotter and a Graphtec KD4030 digitising pad. Communication from the digitising pad is achieved by means of a Y-cable arrangement at the plotter's interface. This set of hardware has been collected primarily for the purpose of analysing archaeological air photos, as an extension of the method described by Chamberlain and Haigh (1982); without the digitiser it is also used for contouring the results of geophysical survey (Kelly and Haigh, 1984).

The bulk of the software has been written in Microsoft FORTRAN 80, but a number of machine code subroutines have been added to the standard libarary. These subroutines include access to high-resolution graphics (provided by Research Machines Limited), direct display of text on the screen, direct response from the key-board, access to CP/M primitives, handshake with the plotter, and input from the digitiser.

The runtime system is menu-driven, with the user having a free choice over the sequence of operations, although he would normally be expected to use a group of input commands, followed by analysis of the accumulated data, followed by a group of output commands. In principle, the system works with just one menu, with control returning to the menu after each operation. In practice, there are too many operations to be displayed at one time, and the commands have been fairly arbitrarily divided into two menus, one largely containing input operations and the other largely output operations.

5. Facilities in interactive software

The fundamental operation of the system is the input of coordinates from the plan through the digitiser. A series of preliminary operations allows the user to set the scale and the axes. As the coordinates are received from the digitiser, corresponding points are displayed on the screen, so that the user can check that the complete set of data has been fed in correctly. As a further aid in checking the data, it has been arranged that one button on the digitising cursor should cause a screen cursor to be displayed; this provides a "mouse"-like facility, and enables the user to examine the correspondence between the digitised input and the screen display.

At any point, the user may request that the coordinates so far accumulated be saved as a data file. He may then start to accumulate a new set. Data from existing files or from the digitiser may be combined in any way that the user finds convenient. Because the data files are written in a very clear and simple format, it is also possible to introduce data from other sources.

The analysis requires that a boundary should be defined for the region over which the data are distributed. In fact, the program makes provision for the introduction of two boundaries. The outer boundary is intended to be the one used in the natural geographical description of the region, so that the final plan output by the plotter may be readily recognisable; the inner boundary is intended

to be the one actually used for analytical purposes. The outer boundary may be stored on a file in the same manner as the actual data; no provision has been made for the storage of the inner boundary, since this is usually of quite a simple form and may easily be recreated from the digitiser. Either boundary may readily be replaced, the outer boundary from either the digitiser or a previously stored data file, the inner boundary from the digitiser alone. The replacement of the inner boundary is particularly important, since it is necessary to ensure that results are not sensitive to the detailed choice of boundary.

Once all the data and the appropriate inner boundary are stored in the computer, the user may request that the analysis be made. The computer then sorts out the data points that lie within the inner boundary, works out the nearer-neighbour distances from each of them (normally up to k = 20), and calculates the observed mean distances over the set. Having calculated the length L of the inner boundary, the area A within it, and the coordinates of the centroid, it then uses the formulae of section 3 to calculate the expected mean distances for a random distribution, and the variances of the expected means. The difference between the observed and expected means is expressed as a number of standard deviations (in effect a value of Student's t-statistic), and a warning is given when the limit of reliability, described in section 3, is exceeded. The results are normally output on a small printer, connected to the parallel port of the RML 380Z.

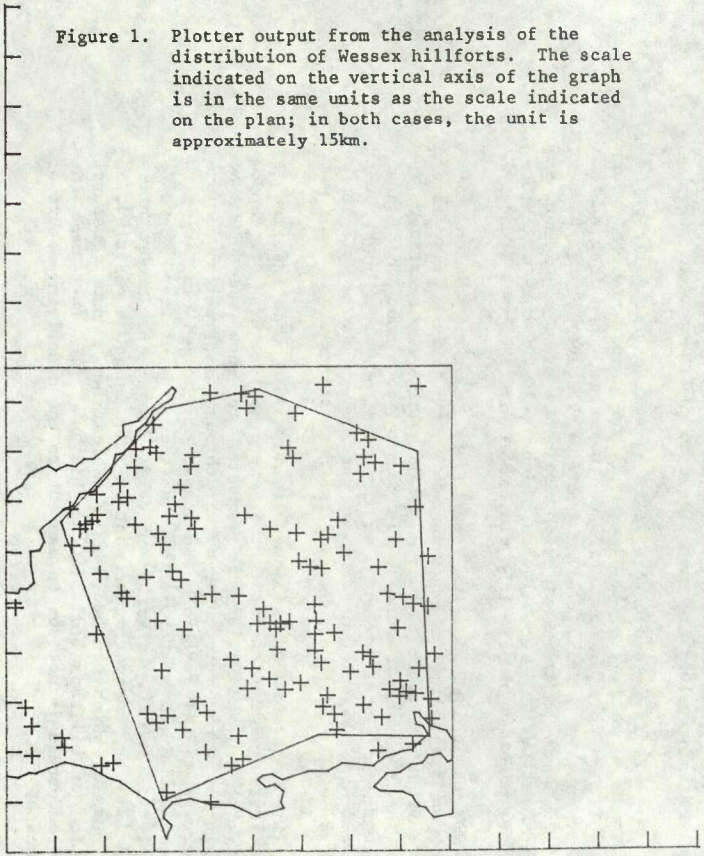
The more important form of output is through the plotter. It is intended that this should include a plan showing the distribution of the data, together with the associated boundaries and scale, and a graph of the observed mean k-th neighbour distances in comparison with the expected values and the expected standard deviation. The whole presentation should be labelled with suitable captions, as near self-explanatory as possible, so that it is virtually ready for immediate publication. A more detailed explanation of one particular output is given in section 6.

6. The Wessex hillforts: an example of plotted output

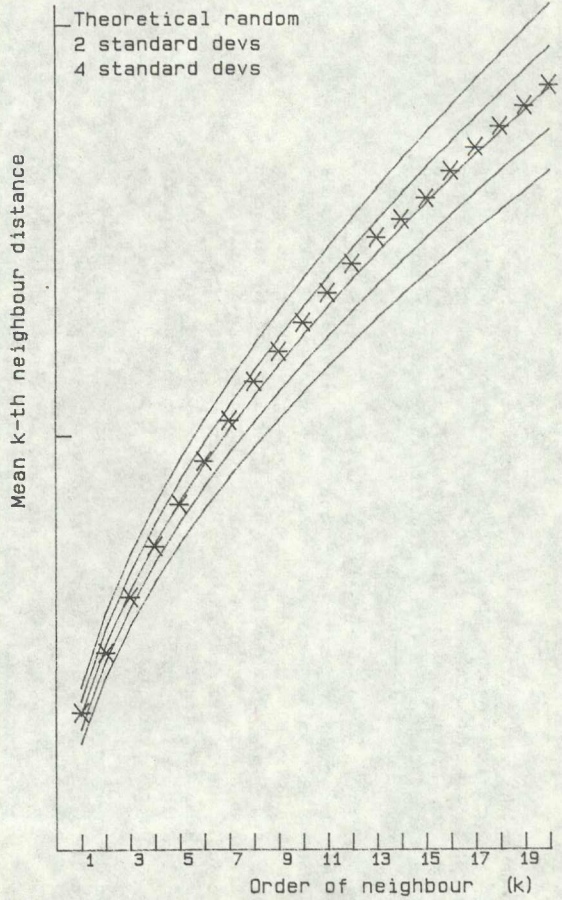
Figure 1 shows the output produced from the data for one of the classic examples of spatial analysis, the distribution of Wessex hillforts described by Hodder and Orton (1976). The shape of the coastline around the Wessex region has been digitised and stored, together with rectangular edges to complete the enclosure, as the outer boundary. When plotted out it has produced the characteristic shapes of the Bristol Channel and of the South Coast, including Portland Bill and the Isle of Wight, seen in the plan on the left. The digitised locations of the hillforts are shown against this outline.

Also drawn on the plan is the inner boundary, which is chosen to be a simple shape enclosing the majority of the hillforts. Only the points within the inner boundary are included in the analysis and, when the diagram is plotted in colour, this is made clear by plotting points outside the inner boundary in a lighter colour. The three

Figure 1. Plotter output from the analysis of the distribution of Wessex hillforts. The scale indicated on the vertical axis of the graph is in the same units as the scale indicated on the plan; in both cases, the unit is approximately 15km.



Map of Wessex
Main section of
Wessex univallate hillforts.



k-th NEIGHBOUR ANALYSIS

captions at the bottom of the plan are stored with the three sets of data. The first, "Map of Wessex", is associated with the outer boundary; the second, "Main section of", is associated with the inner boundary; the third, "Wessex univallate hillforts", is associated with the main data file. With appropriate choice of captions, each diagram can be labelled unambiguously.

The mean k-th neighbour distances are shown as asterisks on the graph at the right of figure 1, and are compared with graphs representing the values expected from the formulae of section 3 applied to the area within the inner boundary. Curves indicating values 2 and 4 standard deviations on either side of the expected mean are also shown. Again, when colour plotting is available, the theoretical curves are shown in colours which contrast with the points representing the observed values. The units indicated on the vertical scale of the graph are the same as those shown on the plan; in this instance, each unit is approximately 15km.

It is seen that, in the case of the Wessex hillforts, there is remarkably good agreement between the observed and expected values for the mean neighbour distances. In no case is the difference between the values greater than two standard deviations, and hence no observed value individually need be considered of statistical significance. On the other hand, the values in the block between $k = 7$ and $k = 13$ lie consistently above the expected central values, while the remaining values lie much closer to the theoretical curve.

A little care is needed before reaching a conclusion, since the results may be sensitive to the choice of the inner boundary. For instance, if the inner boundary is made too large, then the results will almost certainly lie below the theoretical curve for all values of k. This would be taken as strong evidence of a contagious or clustered distribution, a conclusion which would be correct, since all the data points would lie in a cluster within the inner boundary.

In the present instance, the results are remarkably insensitive to any reasonable choice of inner boundary. In almost every case, the middle group of values is slightly high, while the remainder lie close to the theoretical curve. It may be concluded that these results define a global property of the data set as a whole, not merely something associated with a particular choice of inner boundary, or with a limited subset of the data points.

At first impression, it is somewhat surprising that the results show so little deviation from the expected pattern of random behaviour. Human activities are expected to show some form of discernible pattern or structure, either through dispersal to achieve maximum access to the available landscape, or through clustering to achieve concentrations of population. Neither type of trend seems to show to a statistically significant degree among the Wessex hillforts. It may be, however, that what is seen is not a distribution of locations for which man had a free choice, but rather a natural distribution of hilltops suitable for his use. In that case the apparent randomness is not due to man but to the underlying nature of the landscape.

Although no individual value is significant, the fact that the mean k-th neighbour distances in the range k = 7 to k = 13 are rather larger than expected may have some significance, and suggests that the data are clustered in groups of around seven points. Careful examination of the plan indicates that this is indeed a reasonable conclusion. Whether this slight trend towards clusters of consistent size is the result of man's choice or of the underlying geological pattern is a question which this form of spatial analysis is not designed to answer.

7. Conclusions

The hardware and software configuration described in sections 4 and 5 provides a reliable and convenient means of getting access to a sophisticated form of spatial analysis. The formulae set out in section 3 appear to give a good prediction of the values expected from a random distribution over the appropriate region, and section 6 demonstrates that they are capable of giving worthwhile results. A particularly important feature of the interactive system is the ability to vary the inner boundary, which defines the region to be analysed. This provides a check on the reliability and consistency of the final results, and ensures that they are not significantly dependent on the choice of inner boundary.

The system comes close to the limit of what can be achieved on a Z80-based microcomputer. The calculation of the mean nearer-neighbour distances is quite complicated and, for a reasonably large data set, occupies several minutes of computer time. Clearly this is unsatisfactory for an interactive system. Besides the mathematical calculation, the system also includes a number of other large sub-routines, to control the screen graphics, the digitiser, the plotter, and the menu generators. With space for 300 data points, 100 points on the outer boundary and 100 on the inner boundary, the total program size is 36 k bytes, which is close to the limit when allowance is made for the operating system and for a co-resident linker.

Personal computers in the range currently available are capable of overcoming these limitations. They have memories sufficiently large to contain an almost indefinite number of data points. Furthermore, they can be fitted with arithmetic coprocessors which are capable of reducing the 'number crunching' calculation of neighbour distances to a matter of a few seconds. Thus current technology offers the prospect of considerable enhancement to the existing system, in terms both of interactive convenience and of offering a wider range of statistical utilities.

References

Chamberlain, M.P. and Haigh, J.G.B. (1982). 'A microcomputer system for practical photogrammetry', Computer Applications in Archaeology 1982, 142-149, University of Birmingham. ISBN 0 7044 0627 6.

Clark, P.J. and Evans, F.C. (1954). 'Distance to nearest neighbour as a measure of spatial relationships in populations', Ecology 35, 445-453.

Donnelly, K.P. (1978). 'Simulations to determine the variance and edge effect of total nearest neighbour distance', in Hodder, I.R. 'Simulation studies in archaeology', Cambridge University Press. ISBN 0 521 22025 4.

Hodder, I.R. and Orton, C.R. (1976). 'Spatial analysis in archaeology', Cambridge University Press.

Kelly, M.A. (1986), Thesis to be submitted in part fulfilment of Ph.D. regulations, University of Bradford.

Kelly, M.A. and Haigh, J.G.B. (1984). 'Automatic data logging for resistance surveying and subsequent data-processing', Computer Applications in Archaeology 1984, 161-169, University of Birmingham. ISBN 0 7044 0731 0.

Thompson, H.R. (1956). 'Distribution of distance to n-th neighbours in a population of randomly distributed individuals', Ecology 37, 391-394.