13
Spatial interrelationships analysis and its simple statistical tools
Germà Wünsch
Secció d'Arqueologia, Institució Milà i Fontanals (CSIC), c/Egipciaques, 15, 08001 Barcelona (SPAIN)

13.1 Introduction
One of the main aspects of our approach to the study of the spatial articulation of the archaeological record is the gradual introduction of quantitative methods (Wünsch 1991, 1992a, 1992b). With this approach we hope to attain objectivity in the descriptive treatment of the information contained in the archaeological record (Wünsch 1989a, 1989b, 1991). Statistical processing must be considered only a tool that makes the handling of this information easier and that enables its description, ordering and, especially, its hierarchising in terms of differential significance.

In order to avoid the limitations brought about by the absence of previous theoretical thought we have tried to carry out the introduction of quantitative tools within the framework of global reflection that characterizes our proposal for Spatial Interrelationships Analysis (ANITES in Spanish). This proposal tries to elicit meaningful information that could complement the reconstruction of the organizational strategies of hunter-gatherer communities (Figure 13.1).

From this simple reflection we have laid the controlled application of several statistical techniques, by creating a design for computerised statistical data processing. Here we present a basic tool aimed at contrasting explanatory hypotheses about the management of social space, food and goods distribution, etc. (Figure 13.2). In preceding works we carried out a previous evaluation of the budgets and of the application conditions of the selected statistical tests that was further complemented by relevant modifications in order to suit the processing of archaeological data at the spatial level (Wünsch 1989a, 1989b).

Succinctly, the design demands simple restrictive validation requirements for the application of the whole set of tests that imply: a) analysis of a representative archaeological unit of a socioeconomic stage (occupational level) that is well delimited and relevant; b) processing, exclusively, those analytically informative categories on organizational grounds; and lastly, c) adjusting the selection of categories to the minimal effective group or the essential limitations for correct operation of the statistical tests.

Figure 13.1: Diagram synthesising the main aspects of the theoretical design of Spatial Interrelationships Analysis.

Figure 13.2: The operational logic stages of data processing within the Spatial Interrelationships Analysis (ANITES) framework.
We begin with a proposal of complementarity between the various tests that enable us to deal with two different types of data recordings: a) those localized within a regular grid of cells (Dacey 1973; Whallon 1973a, 1973b); b) those located individually within a delimited space (Ebdon 1982; Hodder & Orton 1990).

In devising the statistical processing we integrate several interesting features. Firstly, the seminal computerisation of the chaining of statistical tests is aimed at maximising the speed of data processing. We need to emphasize that we have given priority to "simplicity" as the selection criterion for the statistical techniques. We intend to avoid excessive complexity in the elaboration and subsequent reading of the results of the statistical tests which have hindered its management and have restricted its handling. Therefore, our design of statistical processing is based on the rejection of any unnecessary technical sophistication and consequently has developed into a tendency towards the priority usage of tests that are easily interpreted and handled.

Another interesting aspect is that of our defense of the proposal of complementarity between several tests which operate on different types of recorded data. This is an essential procedure for the assessment of the spatial articulation of the archaeological record, and it enables us to integrate: a) the material remains recorded in three dimensions during the excavation; b) the remains registered only on the basis of the excavation grid, and c) the data relative to qualitative categories resulting from technical analysis. The design of the statistical processing can be described from its integration into an articulation of linked analytical stages (Figure 13.3):

1. Thought about the information relative to the socioeconomic dynamics that is to be analysed and archaeologically represented by a delimited occupational level. It ought to enable the first contact with the features of the just mentioned dynamics, especially in relation to the activities that are being carried out and to the working processes applying. Moreover, it must facilitate the assessment of the incidence of overall factors such as the seasonality, the period of occupation or the postdepositional processes. To lead the analysis it is important to state preliminary working hypotheses.

2. Observation of plans and profiles in order to have the first sight of the occupational level. Special attention should be drawn to distribution plans for categories and to the cells frequency tables. This previous observational process enables us tentatively to locate possible differential concentration nuclei, reiterate locations of certain elements or areas with no material remains, and hence it allows us to evaluate the initial working hypotheses.

3. Regarding the established hypotheses, an accurate selection of the quantitative, as well as qualitative categories to be processed should be carried out. Moreover, a parallel preparation of the different tables and/or files for the statistical processing:

4. Determination of distribution and arrangement patterns of every selected quantitative category. Obtaining the category groups (AGP) and the differential concentration nuclei (NUC). First evaluation of the interrelationships among the groups of the different categories (INT) and determination of the preliminary spatial interrelationships (INTES).

5. Search for significant element associations (AES). The introduction of new qualitative categories that increase the informative potential must be given priority along the complementary lines of interest.

6. Use of the results provided by the different tests in order to make an overall evaluation of the spatial interrelationships. The isolation of spatial interrelationships is the source for the delimitation and location of internal areas, synthesized in the spatial interrelationship patterns (PIE).

In summary, the application of the statistical processing is the starting point in ANITES for the contrast and elaboration of hypotheses related to the organisational strategies. In order to make the operational scheme of this statistical processing design more simple and understandable it will be useful to outline in brief the features of the selected tests in the analytical stages.

![Figure 13.3: The main elements and selected tests of the statistical data processing that generates the source information for the implementation of the Spatial Interrelationship Analysis.](image-url)
13.2 Data processing for three-dimensional located points

We begin with the individualized location of the points within a delimited space corresponding to the studied area (Ebdon 1982; Hodder & Orton 1990). With these original data we create files for every selected category. Every record must contain at least the identification or number of the object, as well as the three-dimensional coordinates and an analytical reference which act as selection filter for these points according to the evaluated working hypotheses. This system of data recording enables us to process the material remains recorded in three dimensions during the excavation, namely lithic and faunal remains.

Processing this type of data implies carrying out the following four linked stages for every selected category:

a) Determine the observed distribution pattern of points using nearest-neighbour analysis. This operates from located points providing a measure of divergence between the observed distribution and the random distribution calculated from the distances between points and their nearest-neighbours, independent of direction (Clark & Evans 1954; Dacey 1960, 1973; Ebdon 1982; Hodder & Orton 1990). To overcome the limitations of this test, such as the boundary effect of size variation in the study area, several authors have proposed remodellings (Pielou 1961; Pinder & WHiterick 1975; Vincent 1976; Haworth & Vincent 1976; Pinder 1978; Aplin 1979; Pinder et. al. 1979).

We have developed a remodelling of the test to enable the joint processing of the three dimensions x, y, z. Following Pinder (1978), before the test we carry out a simulation to obtain the parameters required in the formulae, and also to minimize the boundary influence. Therefore, we would like exclusively to outline the resulting operative procedure which consists of comparing the mean of the observed minimal distances:

\[ \bar{r}_{A} = \frac{\sum r}{N} \]

with the expected distance in a random distribution of the points (that is where \( \text{vol} \) is equal to the studied volume and n is equal to the total number of points):

\[ \bar{r}_{E} = C(\sqrt{\text{vol}/n-1}) \]

where \( C = 0.498107243 + 0.018592611(\sqrt{1000}/n-1) \)

Here, \( C \) is a reduction coefficient obtained from the simulation program. In this way we obtain the R index which indicates the divergence of the observed distribution from randomness. If the distribution is random, R is 1, for maximal clustering R is 0 and, for maximal dispersion the upper value of R is 2.1491.

\[ R = \frac{\bar{r}_{A}}{\bar{r}_{E}} \]

To determine the significance of the R index, we can use the two-tailed critical values table by formulating a simple equation, keeping in mind the total number of points (n):

\[ c = \frac{\bar{r}_{A} - \bar{r}_{E}}{\sigma_{E}} \]

In this case \( \sigma_{E} \) is the standard error of the mean distance from the nearest-neighbour in a random distribution, and is calculated as:

\[ \sigma_{E} = V(\bar{r}_{E}) \]

where \( V = (-0.04478945 + 0.07231321)(\sqrt{1000}/n-1) \)

In the equation, \( V \) is a variation coefficient from the simulation program. We have to take into account that this is a descriptive analysis and therefore it only provides information about the point distribution and not about its arrangement within the studied space. When interpreting the distribution pattern, the results of the proof applications indicate the existence of an inverse proportionality between the number of effectives and the R coefficient (according to which a reduced number of effectives can give a high R coefficient and vice versa). Moreover, there is a certain influence between the R coefficient and the number of groups, whereas there is no direct influence between the number of points and the number of groups (Wünsch 1989b).

b) In the case of a non-random pattern the different groups should be divided by calculating the rupture critical distance which is determined as the sum of the mean and the standard deviation of the distances from the nearest-neighbours:

\[ d_{\text{crit}} = \bar{d} + \sigma(1.65) \]

The union of points determines the groups by drawing circles with a radius (or diameter if we want to weight the areas) equal to the critical distance, and so we have curvilinear boundaries and maximal areas.

c) In order to analyze every obtained group, we propose synthesis of the dispersion pattern by calculating the gravity center and the direction of the axes of the standard deviational ellipse (Raine 1978; Ebdon 1982). The ellipse is located in the centre of the point distribution within each group, with the main axis in the direction of maximal dispersion and the minor axis in the direction of minimal dispersion. The calculations are done in four steps:

1. Centre the coordinate system by translating the original coordinates into new ones centred on the centre of gravity:
   \[ x' = x - \bar{x} \]
   \[ y' = y - \bar{y} \]

2. Calculate the rotation angle of the new axes \( x' \) and \( y' \) with respect to the original axes \( x \) and \( y \):
   \[ \tan \theta = \frac{(\sum x^2 - \sum y^2) + \sqrt{((\sum x^2 - \sum y^2)^2 + 4(\sum x'y')^2)}}{2\sum x'y'} \]

3. Calculate standard deviation along \( x \) axis of ellipse:
   \[ \sigma_x = \sqrt{\frac{((\sum x^2)(\cos \theta)^2 - 2(\sum x'y')\sin \theta \cos \theta + (\sum x'^2)(\sin \theta)^2)}}{n} \]
4. Calculate standard deviation along y axis of ellipse:
\[ \sigma_y = \sqrt{\frac{(\Sigma y^2)(\sin \theta)^2 + 2(\Sigma x'y')(\sin \theta \cos \theta) + (\Sigma x'^2)(\cos \theta)^2}{n}} \]

The lengths of the ellipse axes are calculated as \((2\sigma_x\text{ and } 2\sigma_y)\), respectively. The angle obtained with the first equation is that between the \(y'\) axis and the \(y\) axis of the ellipse, and is measured clockwise. If the result is negative, the sign is ignored when looking for the angle value in the tangent table, but subsequently the result is subtracted from 90°.

d) Determine the interrelationships among the groups of selected categories in terms of the spatial overlap of the maximal areas delimited before. To delve into the analysis of each established interrelationship we can compare the distributions two by two, or otherwise apply the correspondence analysis to evaluate the relation between the new areas and the implied categories.

13.3 Data processing for cells frequency tables

From a regular grid of cells that divides the studied area, the cell frequency tables are made for every selected category (Greig-Smith 1952; Whallon 1973a, 1973b; Mead 1974; Johnson 1984). This system of data recording enables the processing of the material remains not recorded in three dimensions during the excavation, however there are interesting analytical categories. The processing of this type of data supposes the realization of three linked stages for every selected category:

a) Determine the distribution pattern of observed frequencies. In order to assess the randomness or non-randomness we use a simple statistic resulting from the division of the variance by the mean, and thus it enables us to have the first characterization.

\[ \frac{\sigma^2}{\bar{x}} \]

If a map has a random pattern the expected value of this ratio is close to 1, a minor result indicates a scattered distribution and a major one a clustered distribution. To find out whether the observed ratio differs greatly from randomness, the dispersion index is compared to the chi-squared distribution and a major one a clustered distribution. To the already defined symbols we add: \(n\) which is the number of studied areas and \(J\) the number of common boundaries. We also take account of the variable \(x\), the value of an area, the mean of all values of this variable, as well as the values of two contiguous zones \(x_i, x_p\).

\[ I = \frac{n\sum (x_i - \bar{x})(x_j - \bar{x})}{J\sum (x - \bar{x})^2} \]

The result obtained by applying Moran's coefficient does not have a value by itself when describing the spatial autocorrelation of a variable or category. To evaluate the significance of \(I\) a normal standardized variable must be created. For this purpose the expected value of \(I\) and the standard deviation in terms of a randomization procedure (for we cannot accept the null hypothesis which greatly complicates the calculations. The equation of the expected \(I\) value according to the randomized null hypothesis is:

\[ E_i = -\frac{1}{n-1} \]

The equation of the standard deviation is fairly complicated due to the adoption of the randomized null hypothesis which greatly complicates the calculations.

\[ \sigma_I = \sqrt{\frac{nJ(n^2 + 3 - 3n) + 3J^2(n^2 - n) + k^2 J(n^2 - n) + 6J^2 - 2n\Sigma E^2}{J(n-1)(n-2)(n-3)}} \]

To the already defined symbols we add: \(\Sigma L\) which is the number of areas with that an area has common boundaries and the sum of its squares (\(\Sigma L^2\)). Furthermore we also include the calculation of \(k\) (the kurtosis of the \(x\) variable) and hence the standard deviation should be calculated in the first place:

\[ \text{Kurtosis} = \frac{\sum (x - \bar{x})^4}{n\sigma^4} \]

Finally, the observed value \(I\) can be normalized with a simple equation:

\[ z_i = \frac{I - E_I}{\sigma_I} \]

In order to evaluate the significance of the obtained values, we must refer to the two-tailed critical values table for a standardized normal variable. It is preferable to apply a two-tailed test because the concrete devitional
direction in relation to the randomness is unknown to us. The expected value of the coefficient for the random arrangement (absence of spatial autocorrelation) is close to 0, whereas negative values indicate a scattered arrangement, and positive ones a clustered arrangement.

13.4 Determination of significant element associations (AES)

This is an important complementary stage for the study of the spatial interrelationships. We propose a double approach to the AES which enables us to build new hypotheses:

a) On the one hand, the application of association or similarity tests that operate in the terms of data recorded in presence/absence tables, comparing the selected categories two by two. Above all, the implementation of these tests enables to use both qualitative and quantitative categories. The most adequate tests are:

- The Chi-squared test (2*2). It operates from contingency tables that shows the observed occurrence of selected categories (Dacey 1973).

\[ \chi^2 = \frac{n(ad - bc) - n/2)^2}{(a+b)(c+d)(a+c)(b+d)} \]

- The significance of the results is found in the critical values table with one degree of freedom. The null hypothesis sets forth the independence of the different observations, and consequently the absence of spatial association between the two analyzed categories.

- Jaccard's I coefficient (Wünsch 1989a, 1989b) also operates by contrasting the selected categories two by two from a contingency table. Its peculiarity lies in the fact that the d cell is unused in the calculation, that is, the mutual absences are not taken into account. With this we intend to avoid, or at least diminish, the eventual distortion of the cells with observed frequency equal to 0, that can form the majority when working on a wide area and with a clustered distribution pattern.

\[ I = \frac{a}{a+b+c} \]

The significance of the observed differences is carried out by means of the Chi-squared statistic (2*2) with one degree of freedom. The significant results can be summarised graphically by dendrograms. There is a slight inconvenience: when the d cell of the table is not taken into account, this coefficient bans the evaluation if the obtained value deviates significantly from that expected in the null hypothesis, and hence we must appeal to the chi-squared statistic.

- The phi coefficient. As in the two preceding cases, this also operates from a contingency table (Driver 1961). It is of great interest because, in a way, it takes the features of the two previous tests.

\[ \phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} \]

The result lies between +1, the maximal direct association, and -1, the the maximal inverse association (exclusion); values close to 0 indicate absence of association. We can calculate whether the obtained value deviates significantly from that expected from the null hypothesis if it is translated into chi-squared results with one degree of freedom by means of a simple formula (in which N represents the total sum).

\[ \chi^2 = (\phi^2)N \]

b) On the other hand there is a complementary application of the lien table to evaluate the information contained in a contingency table (Laplace 1979-80) and/or of the correspondence analysis which works upon qualitative variables or categories and can operate with reduced effectives (Cibois 1983; Volle 1981; Lesage 1991; Mora & Roca 1991). The great advantage of correspondence analysis regarding the previous tests is the possibility of correlating all types here considered (variables or categories and observations or individuals).

With the application of correspondence analysis we try to carry out a factorial decomposition which could facilitate the interpretation of the initial complexity of a contingency table with several rows and columns. The obtained factors allow us to discriminate the different individuals and, therefore, to establish the eventual attractions or associations. However, we cannot forget that we are dealing with a descriptive analysis that only offers us an overall information synthesis or resume. Moreover, the reduction of information implies an essential relative loss in order to attain an understanding of the whole. We are interested in using correspondence analysis to determine significant associations between the selected categories and the cells or areas used as a source of the effectives count from which we elaborate the contingency table. Its major potential lies, of course, in its implementation upon tables with large numbers of rows and/or columns, in that we need to synthesize the data in order to eliminate the irrelevant information. Nevertheless, we have to outline that in the searching for the significant element associations (AES), our aim is not to replace the association tests presented earlier by the correspondence analysis because we consider that they provide us with information about different aspects, and they can complement each other perfectly well.

13.5 Joint evaluation of spatial interrelationships in the archaeological record

The determination of the spatial interrelationships formed by significant associations of diagnostic elements or by differential concentrations of a given category is finally based on the complementarity between the results obtained in the different analytic stages. This supposes the joint evaluation of the inter-relationships between the groups (INT), the differential concentration nuclei (NUC) and the significant elements associations (AES).
Bearing in mind that this is a quantitative tool which must enable the objectivity of the delimitation and location criteria of the internal areas, in the search frame of the spatial interrelationship patterns (PIE), we want to conclude by outlining two aspects:

- On the one hand, it is of the utmost importance to potentiate the introduction of new qualitative categories in the statistical data processing. These categories must be obtained in terms of the constant search for new diagnostic elements of the working processes that are distinctive of the different subsistence and maintenance activities carried out in the settlements.

- On the other hand, initially it seems necessary to articulate operational research schemes aimed at contrasting specific organizational hypotheses. This supposes the realization of more restricted and controlled applications of the statistical process. The aim is to carry out initial controlled applications upon ethnoarchaeological records (Wünsch, 1993) that enable us to set to proof the operativity of the designed quantitative tool. For this reason we need to reduce, within the possible, the wide range of involved variables without this leading to a loss of relevant information. Furthermore, the previous thought about the different models that can be evaluated when relating some given working processes or activities with specific spatial interrelationships of the different elements that form the archaeological record, should be enhanced.

References


WUNSCH, G. 1989b. 'La organización interna de los asentamientos de comunidades cazadoras-recolectoras: el análisis de las interrelaciones espaciales de los elementos arqueológicos', Trabajos de Prehistoria, 46: 13–33.


