Jeffers, J. N. R.. 1983 Analysis of spatial data. In: Fuller, R. M., (ed.) Ecological mapping from ground, air and space. NERC/ITE, 16-21. (ITE Symposium, 10).

# ANALYSIS OF SPATIAL DATA

J N R JEFFERS
*Institute of Terrestrial Ecology, Merlewood Research Station, Grange-over-Sands*

**ABSTRACT**

Methods of analysis of spatial data are defined partly by the design of the experiment or survey from which the data are derived and partly by the nature of the data themselves. A decision table for the type of analysis is suggested, together with a range of analysis techniques appropriate to various types of data.

## INTRODUCTION

It is commonly assumed that the appropriate output of any scheme of ecological mapping from the ground, air or space is a map of some kind. Much of this symposium is, therefore, primarily concerned with problems of cartography, mapping or the presentation of spatial information in some graphical representation. However, the data from which such representations are derived may also be submitted to statistical analysis of various kinds, whether these data consist of attributes, categories or discrete or continuous variables. A further assumption seems to be that, where analysis is to be done, it will follow, rather than precede, the cartographic representation. In the worst of all possible cases, the analysis will actually be done on data derived from the cartographic representations themselves, rather than from the primary sources from which the cartographic presentation was itself derived. In this paper, I will challenge this assumption. It is my belief that analysis should precede cartographic representation, and that the analysis should be based on the primary data from which the cartographic representation is derived.

I do not want to dwell on the point, in this paper, but there are considerable dangers in the sheer sophistication of the techniques now available for the cartographic representation of spatial data. The principal danger is that of subjectivity. Given the enhanced flexibility of modern cartography, with its ability to stretch, tilt, colour and slice images, it is obvious that an infinity of combined images is now available to the well-trained cartographer. How does the cartographer select the final image from the infinity of possible combinations? Given the flexibility of modern instrumentation, it seems likely that he will manipulate the various elements until he achieves a representation which satisfies either his aesthetic taste or his preconceptions about the problem with which he is concerned. At this point, he may well cease any further experimentation with the variables that he is able to change in the representation. The resulting cartographic presentation may, therefore, be no more than a subjective expression of the cartographer's interpretation of the problem. What is more, even if this process is used to generate hypotheses, all of the data will already have been used in the generation of the hypothesis, thus leaving no 'new' data for the explicit testing of that hypothesis. Modern cartography does not, in this way, meet the criteria of the scientific philosophy by providing a falsifiable hypothesis. This is, of course, a somewhat extreme view of what cartographers do, and I have expressed it in this extreme way primarily to show the dangers of unequivocal acceptance of the sophistication which is available to modern technology.

A more important point about the use of cartographic information is that graphical and cartographic representation of spatial data frequently, if not inevitably, involves a disassociation between the objectives of the investigation, the methods by which the data were collected, and the methods by which those data are interpreted. As I shall emphasize later in the paper, the greatest dangers lie in this disassociation. The main message of my paper, therefore, is that analysis should always precede the cartographic presentation, and that the analysis should be defined by the objectives of the investigation, and by the method by which the data have been collected. Cartography and any form of visual presentation of spatial data would then be a secondary process, which necessarily follows analysis.

My assertion is dependent upon an important distinction between the 2 principal theories of data. Elsewhere (Jeffers 1981), I have emphasized that there are 2 quite distinct theories about the ways in which data can be used. First, there is what I call the 'accounting' theory of data which insists that data can be used in any way that seems appropriate to the user, in much the same way that records of transactions in company accounts are used to demonstrate various facets of the profitability, assets, and investment of a commercial or industrial company. The forms of presentation of such data take little or no account of the units in which the data were originally collected, the aggregations made at various stages in the accounting process, or the relationships of the data to other variables within the management and administration of the company. In contrast, the 'statistical' theory of data stresses the dependence of methods of analysis on the objectives of the investigation and the methods which were used to collect the data. These methods include such important points as sampling design, sampling units, definition of the attributes or variables collected within the investigation, and, particularly, the presence or absence of randomization, and any constraints that may have been placed on that randomization. Lakhani, earlier in this symposium, has already dealt with the essentials of sampling techniques, experimental designs and statistical methodology. Under this statistical theory of data, the kinds of analysis which can be made of spatial data are limited by the objectives of the investigation, and by the decisions made during the design and collection of the data. We may, therefore, make some considerable headway by a careful analysis of the strategies to be used in various kinds of environmental studies resulting in the collection of spatial data.

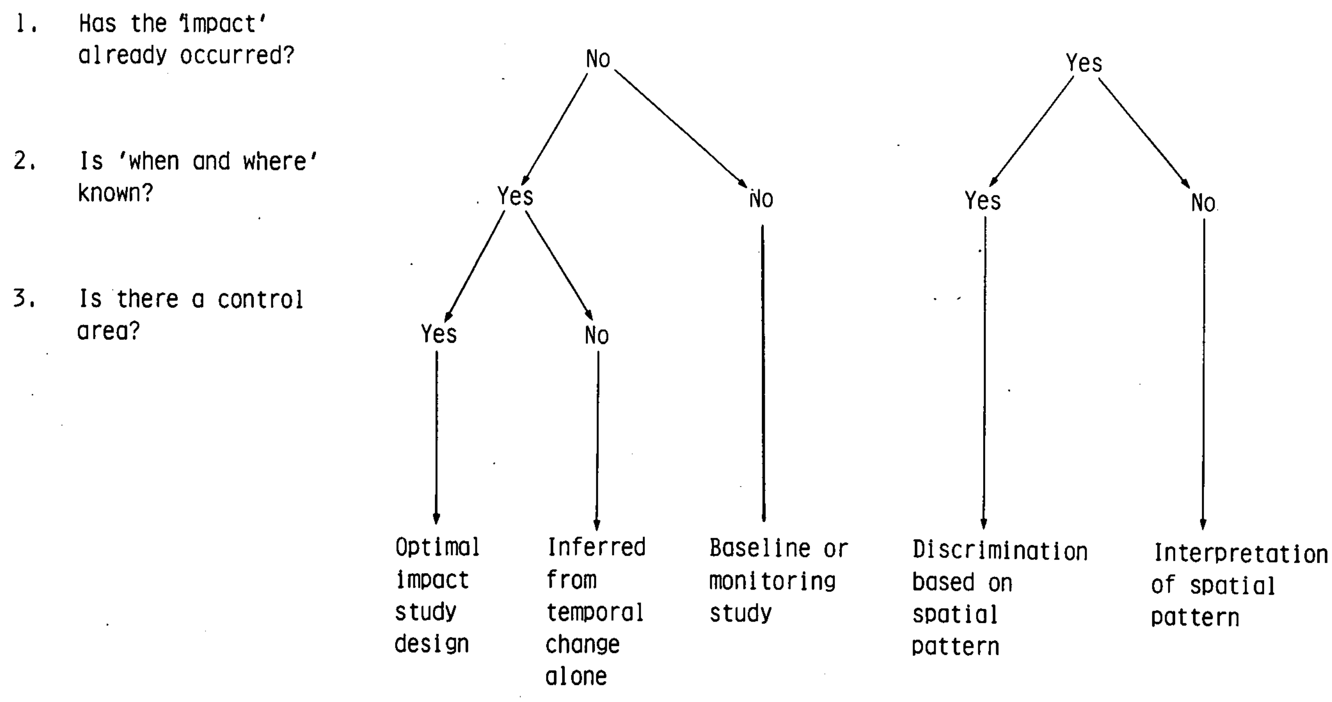


*Figure 1* *A decision key to determine the main strategies in studying environmental impacts (after Green 1979)*

One of the most interesting classifications of the principal types of environmental study has been given by Green (1979) as a decision key related to the existence or non-existence of some 'impact', knowledge of when and where any such impact may have occurred, and the presence or absence of a control area. The main sequence of categories is given in Figure 1. From this Figure, it can be seen that the main strategies are determined quite simply by a sequence of yes/no answers to the 3 principal questions. If the 'impact' has already occurred, the principal division is made on whether or not the time and location of the impact occurrence is known. Similarly, if the impact has not occurred, and it is not known when or where such an impact is likely to occur, little more than baseline monitoring studies can be attempted. If the timing and place of the impact is known, then the determining factor is the presence or absence of a control area. In the paragraphs that follow, I will suggest methods of analysis for each of the 5 principal strategies defined by this simple decision table.

## METHODS OF ANALYSIS

### *Interpretation of a spatial pattern*

Where a supposed impact has already occurred, but it is not known when or where such an impact has taken place, the analysis is reduced to the interpretation of an existing spatial pattern. There are a large number of available techniques, but they can, broadly, be grouped into 3 categories, namely a) ordination, b) multidimensional scaling, c) cluster analysis. Ordination, or reification, is an attempt to attach some biological or physical meaning to mathematical expressions of spatial variation (Mather & Openshaw 1974). Techniques such as principal component analysis, factor analysis, principal co-ordinate analysis, reciprocal averaging, and indicator species analysis all have possible applications, but considerable skill and experience are necessary if any of the methods of analysis are to give a true picture of the physical meaning of the data. Indeed, if mathematical expressions of this sort have any obvious physical meaning, the result must be attributed to a lucky chance, or to the fact that the data have a strongly marked structure that shows up in the analysis. Even in the latter case, quite small sampling fluctuations can upset the interpretation (Harris 1975).

Multidimensional scaling is closely related to the techniques of ordination and has the aim of producing a visual representation with the smallest number of dimensions and which distorts the pattern of spatial representation as little as possible. In applications to spatial analysis, however, the technique is frequently based on a circular argument, the satisfactoriness of the representation being related to subjective interpretations of the displayed patterns. Perhaps the best known multidimensional scaling procedure is that of Sheppard, further developed by Kruskal (1964). The effectiveness of the method has been demonstrated by a remarkable reconstruction of the map of the departments of France by Kendall (1971). An entirely different technique (not involving multidimensional scaling) has since been devised by Kendall (1974) explicitly for the solution of problems such as this, which arise in a natural way in the context of historical geography. An earlier attempt was made by Geary (1954) through his development of the contiguity ratio.

### *Discrimination based on spatial patterns*

The purpose of discriminant analysis is to investigate the relationship between a known grouping of data and the variables recorded (Harris 1975). The outcome of the analysis is an allocation rule, often, but not necessarily, associated with the values taken by a discriminant function. This rule can then be used for assigning to their appropriate groups the members of subsequent samples. With any allocation rule, there is associated a certain probability of miscalculation, and it is useful to be able to estimate this probability, at least roughly, and, of course, to choose the rule to make it as small as possible. The discriminant function is an extension of the linear models of the analysis of variance and of the multivariate analysis of variance, with a one-way classification. The disadvantage of the technique is that the statistical significance of the differences between the impact and non-impact areas is judged against variation within the areas. The absence of a control area precludes the possibility of a genuine error term in the analysis.

*Baseline or monitoring studies*

Baseline or monitoring studies precede the intended impact, and are intended to provide a base against which some future changes may be recorded. A wide range of techniques exist for such analyses, but one of the most popular, and, with some precautions, the most useful, is that of trend surface analysis (Unwin 1975). Provided that the appropriate mathematical functions are used to describe the trend surface, with adequate provision for discontinuities, the assumptions of this method are not too difficult to meet. The alternative method of automatic contouring, however, while apparently less demanding, has some extremely uncomfortable assumptions which are seldom justified in practical work. Nevertheless, automatic contouring remains a topic of great interest to mathematicians, an interest which is probably in inverse proportion to its applicability to real-life situations.

*Impact inferred from temporal change alone*

The absence of a control area in any spatial analysis which precedes the occurrence of a particular impact places severe constraints upon the kinds of analysis which can be undertaken. Typically, the analysis depends upon a comparison at 2 or more times spanning the period before and after the imposition of the change. Various methods of time series analysis are therefore appropriate, but the measure of experimental and survey error is inadequate because of the absence of the control (Findley 1978). An alternative approach is through the use of Markov models, where the transition probabilities before and after the introduction of change can be assessed (Collins *et al.* 1974). Such models are, however, somewhat restricted in their assumptions, although they have the advantage of a certain transparency in these assumptions.

*Optimal impact study design*

The optimal impact study design, where one or more control areas exist, allows for the use of 2-way analysis of variance or multiple analysis of variance. A wide range of possible models exist, depending upon the particular design of the control areas. In especially favourable conditions, the use of factorial designs is possible, thus enabling many factors to be studied, and to be separated from environmental variation not associated with the supposed impact.

*Special problems*

In the application of appropriate methods of analysis for the basic strategies of spatial analysis, there are some particular warnings to be heeded.

1. Pay careful attention to the methods of sampling used in the collection of data. These methods will largely define the methods of analysis which are appropriate.

2. The use of systematic sampling introduces special problems, particularly in the estimation of sampling errors.

3. If random sampling has been used, make sure that any constraints placed on the randomization are known and understood, as they help to define the appropriate method of analysis.

4. Be especially careful of data from a 'data bank', as these data will almost certainly have become disassociated from their methods of collection.

5. It is not necessary to use all the data. Subsamples of the data set will provide a check on the repeatability of the results of the analysis.

6. Check carefully any assumption made by the methods of analysis that you intend to use.

7. Be especially cautious about using any method which is claimed to be 'the method'!

## CONCLUSIONS

The analysis of spatial data follows closely the basic principles for the statistical analysis of any set of data, but has the added complication of the presence of spatial autocorrelation. Nevertheless, appropriate methods of analysis exist, and these are defined by the objectives of the investigation and by the methods used to collect the data. Three useful texts dealing with many of the problems touched on in this short paper are given by Cliff and Ord (1973) and Haggett *et al.* (1977a, b).

## REFERENCES

CLIFF, A.D. & ORD, J.K. 1973. *Spatial autocorrelation.* London: Pion.

COLLINS, L., DREWETT, R. & FERGUSON, R. 1974. Markov models in geography. *Statistician*, **23**, 179-210.

CORMACK, R.M. 1971. A review of classification. *Jl R. statist. Soc. A*, **134**, 321-367.

FINDLEY, D.F. ed. 1978. *Applied time series analysis.* London: Academic Press.

GEARY, R.C. 1954. The contiguity ratio and statistical mapping. *Statistician*, **5**, 115-145.

GREEN, R.H. 1979. *Sampling design and statistical methods for environmental biologists.* New York: Wiley.

HAGGETT, P., CLIFF, A.D. & FREY, A. 1977a. *Locational analysis in human geography. 1. Locational models.* London: Edward Arnold.

HAGGETT, P., CLIFF, A.D. & FREY, A. 1977b. *Locational analysis in human geography. 2. Locational methods.* London: Edward Arnold.

HARRIS, R.J.C. ed. 1975. *A primer of multivariate statistics.* London: Academic Press.

JEFFERS, J.N.R. 1981. The development of models in urban and regional planning. *Nature & Resour.*, **17**, 14-19.

KENDALL, D.G. 1971. Construction of maps from "odd bits of information". *Nature, Lond.*, **231**, 158-159.

KENDALL, D.G. 1974. Data-analytic problems in archaeology and history. *Proc. Euro. Meeting of Statisticians, 2nd, Budapest.*

KRUSKAL, J.B. 1964. Multidimensional scaling by optimizing goodness-of-fit to a non-metric hypothesis. *Psychometrika*, **29**, 1-27.

MATHER, P.M. & OPENSHAW, S. 1974. Multivariate methods and geographical data. *Statistician*, **23**, 283-308.

UNWIN, D. 1975. *An introduction to trend surface analysis*. (Concepts and techniques in modern geography no. 5). Norwich: Geo Abstracts.