Statistical Analysis of Non-lattice Data

JULIAN BESAG[†], *University of Liverpool and Princeton University*

A Markovian approach to the specification of spatial stochastic interaction for irregularly distributed data points is reviewed. Three specific methods of statistical analysis are proposed; the first two are generally applicable whilst the third relates only to "normally" distributed variables. Some reservations are expressed and the need for practical investigations is emphasized.

1. Introduction

In rather formal terms, the situation with which this paper is concerned may be described as follows. We are given a fixed system of n sites, labelled by the first n positive integers, and an associated vector x of observations, $x_1, \ldots, x_n$, which, in turn, is presumed to be a realization of a vector X of (dependent) random variables, $X_1, \ldots, X_n$. In practice, the sites may represent points or regions in space and the random variables may be either continuous or discrete. The main statistical objectives are the following: firstly, to provide a means of using the available concomitant information, particularly the configuration of the sites, to attach a plausible probability distribution to the random vector X; secondly, to estimate any unknown parameters in the distribution from the realization x; thirdly, where possible, to quantify the extent of disagreement between hypothesis and observation.

Consider the following example. Cliff and Ord (1973, Section 6.4) examine data on the yield of wheat in hundredweights per acre for each county in England during 1936 (Yule and Kendall 1968, p. 311). They find that when the data are regressed against a particular measure of productivity for each county (Kendall 1939, pp. 25–9), there is evidence of high positive correlation between the residuals from adjacent counties. This leads them to suggest "productivity alone is not sufficient to account for the spatial variation in wheat yields in the English counties, and that additional variables such as rainfall or soil type should be considered". It is clear that such complications will arise in many geographical contexts. If local measurements of the offending variables are available, then the regression can be reformulated in an appropriate manner, but often the relevant information is not at hand and some attempt must be made to incorporate spatial stochastic interaction into the analysis.

[†] Now at University of Durham.

One possible means of formulating spatial schemes, as suggested in an earlier paper (Besag 1974a, is through the adoption of a *conditional probability* approach. This requires the user to specify the conditional distribution of each random variable X_i, given the values x_j at all the remaining sites $j \neq i$. The alleged advantage of this tactic is that it breaks down the original problem into a series of simpler sub-problems since, instead of having to do battle with all n random variables simultaneously, we may attack each X_i in turn. This approach is summarized below: note that for ease of presentation, the term "conditional distribution" will be used to indicate that the conditional distribution of a particular variate, *given all other site values*, is being considered.

The simplest assumption which can be made is that the conditional distribution at each site is independent of all other site values. This, of course, corresponds to the classical notion of statistical independence. However, we are here primarily interested in departures from this assumption which arise under the influence of unknown local variations in environmental conditions. This suggests that as a plausible first approximation, we might assume that the conditional distribution at site i depends only upon the values at those sites which are, *in some sense*, in the proximity of site i. This is the essence of local (or *Markov*) statistical dependence in the spatial context. Formally, site j $(\neq i)$ is called a *neighbour* of site i if and only if the conditional distribution at site i depends upon the value at site j. A system of sites, each with specified neighbours, is called a *graph*. The formulation of conditional probability schemes may therefore be considered in two parts: firstly, the choice of a suitable graph and, secondly, the selection of appropriate conditional distributions, consistent with that graph.

It is contended that, in practical applications, the choice of neighbours for each site should be implemented on intuitive grounds. It must be held in mind that the formulation is not intended to suggest a direct causal relationship between the random variables but merely aims to take account of local variations in extraneous conditions—what one might call "third-party" dependence. It follows that the choice of neighbours will largely be subjective. However, as a rule of thumb, it is likely that, in physical geography, geometrical contiguity of sites will be all-important whilst, in human applications, other measures, such as accessibility between sites, may also play their parts. As a rough guide, one might anticipate an end-product averaging around six neighbours per site in typical geographical contexts.

It is with the other half of the formulation that the conditional probability approach runs into trouble. For it transpires that, given a particular graph, the choice of a valid conditional distribution at each site is subject to some extremely severe and unobvious consistency conditions. Indeed, these remained largely unidentified until the arrival in 1971 of an unpublished paper by Hammersley and Clifford. (For a detailed account, including a very short proof of the relevant theorem, the interested reader may refer to Besag(1974a).In section 2.2, we shall mention the impact of the consistency conditions when the site variables are "normally" distributed but, this apart, the general problem of specifying valid conditional distributions will not be discussed in any detail in the present paper.

Instead, we shall, in section 2, concentrate upon the Gaussian situation, examining particular aspects of conditional probability schemes and the rival simultaneous autoregressions. Not only does the normality assumption generate the most easily handled distribution theory but also the resulting schemes probably constitute those of most interest to geographers. In section 3, some methods of statistical analysis for locally dependent schemes are described. Two of these are applicable to non-normal as well as to normal schemes and will be discussed in a general setting, since this involves no additional difficulty in exposition. However, before finishing this section, it may be useful to give an outline description of the Hammersley–Clifford theorem and to follow this by some cautionary remarks concerning the conditional probability approach as a whole.

The Hammersley–Clifford theorem is formulated in terms of the *cliques* generated by a given graph. Here, a clique is defined as any set of sites which either consists of a single site or else in which every site is a neighbour of every other site in the set. Roughly what the theorem says is that, for discrete random variables, the joint probability distribution of $X_1, \ldots, X_n$ must be a product of functions, one function corresponding to each clique. For continuous random variables, the analogous result holds for the joint density function of $X_1, \ldots, X_n$. In either case, it is then a simple matter to obtain the allowable forms of the conditional distributions for the site variables.

Although it is being suggested in this account that the conditional probability approach can provide an intuitively appealing and tractable means of analysing certain types of spatial phenomena, it must be admitted that there still remain many criticisms and many unanswered questions in regard to practical applications. First and foremost, there is, so far as I am aware, no empirical evidence yet available to support the use of spatial Markov schemes. At best, such schemes should perhaps be viewed as mimics of reality rather than models. Thus, I have specifically avoided terms such as "model" and "process" in the present paper, since their usage might have suggested considerations of causality. Indeed, the very notion of placing intuitive interpretations on conditional probability statements, in purely spatial settings, is viewed dubiously by some researchers and should not be accepted uncritically. For some further comments, see Whittle (1963, and in discussion of Besag 1974a). Finally, I have tended, in the present paper, to use the term "local dependence" rather than "spatial Markovity". This is in an attempt to avoid confusion with the (non-standard) definition of spatial Markovity (in terms of simultaneous autoregressions) which has been adopted in some of the geographical literature; see, for example, Cliff and Ord (1973). Fuller details are given in section 2.3 of the paper.

2. Some Multivariate Normal Schemes

2.1. *Introduction*

There may be many practical situations where it is reasonable, as a first approximation, to suppose that the site variables $X_1, \ldots, X_n$ have a multivariate normal distribution. To make any progress with such an assumption, it is necessary to postulate a plausible structure for the mean vector μ and the dispersion matrix V of the distribution. The specification of μ is likely to be made upon classical grounds and, most commonly, will involve the use of a linear function of a few unknown parameters. On the other hand, the description of V may present some difficulty, if the intention is to escape from the classical assumption of independence (that is, $V=\Lambda$, say, where Λ is a diagonal matrix of variances). Various proposals might be made. For example, the conditional probability approach, with which this paper is primarily concerned, leads to the so-called *auto-normal* prescription. This may be interpreted as a generalized version of the Markovian regular lattice schemes of Lévy (1948). A second approach is based upon the regular lattice schemes of Whittle (1954) and, in turn, leads to the specification of *simultaneous normal autoregressions*. Unfortunately, there has sometimes been confusion between these two proposals and we therefore examine their differences in section 2.3. A third approach, which may be particularly relevant when sampling a continuum at a number of point sites, is to attack V directly by assuming that the covariance between X_i and X_j depends only upon their distance apart (in the isotropic case) and then specifying an appropriate correlation function. This technique may be discussed by other contributors and will not be considered further in the present paper.

2.2. *The Auto-normal Prescription*

The aim here, as with all conditional probability formulations, is to isolate each random variable in turn for specific consideration. In particular suppose that the conditional distribution of X_i is normal with conditional mean

$$E(X_i \mid x_j, j \neq i) = \mu_i + \sum_{j=1}^{n} \beta_{i,j}(x_j - \mu_j) \tag{2.1}$$

where $\beta_{i,i}=0$, and conditional variance,

$$\text{var}\,(X_i\,|\,x_j, j\neq i)=\sigma_i^2 \tag{2.2}$$

As might be expected from the discussion in section 1, certain consistency conditions must be imposed upon the coefficients in order for the formulation to be valid. Indeed, it can be shown that the dispersion matrix V for the scheme is given by

$$V=B^{-1}\Lambda \tag{2.3}$$

where Λ is the $n\times n$ diagonal matrix of conditional variances σ_i^2 and B is the $n\times n$ matrix with diagonal elements unity and off-diagonal (i,j) element $-\beta_{i,j}$. Now since any dispersion matrix must be both symmetric and positive-definite, the same holds true of $\Lambda^{-1}B$; the symmetry condition implies that (2.1) and (2.2) are valid only if

$$\beta_{i,j}\sigma_j^2=\beta_{j,i}\sigma_i^2 \tag{2.4}$$

for all i and j, whilst positive-definiteness can, in general, only be checked once the coefficients are known numerically. Since, conversely, any given dispersion matrix V determines unique values for B and Λ, it follows that the class of all valid auto-normal schemes is, in fact, equivalent to that of all multivariate normal schemes. Thus, the auto-normal formulation does not imply any inherent simplification of the parameter space. Nevertheless, the hope is that, by considering each site in turn and invoking the concept of local dependence, it should be possible to reduce the number of unknowns, firstly by setting $\beta_{i,j}=0$ for sufficiently "remote" sites, as described in section 1, and secondly by postulating plausible relationships between the remaining coefficients. This will be discussed further in section 2.4.

2.3. *Simultaneous Normal Autoregressions*

Alternatively, suppose that the structure of the random variables is specified by the n simultaneous equations,

$$X_i=\mu_i+\sum_{j=1}^{n}\beta_{i,j}(X_j-\mu_j)+Z_i \tag{2.5}$$

where $\beta_{i,i}=0$ and $Z_1,\ldots,Z_n$ is a set of independent normal variates, with Z_i having mean zero and variance σ_i^2. It can easily be shown that the class of all such simultaneous autoregressions again generates the class of all multivariate normal distributions but that now,

$$V=B^{-1}\Lambda(B^{-1})' \tag{2.6}$$

where Λ and B are defined as before. Thus, the validity of (2.5) requires only that B is non-singular. (This fact can result in identification problems since, in general, more than one value of B will generate a given dispersion matrix V.) Again the aim is to attach a plausible structure to the coefficients. Indeed, it is likely that, in any given situation, the formal rules employed to reduce the number of unknown parameters will be similar (or even the same), whichever of the two approaches is adopted. However, it is crucial to recognize that the same rules will lead to *different* covariance structures, as can be seen from equations (2.3) and (2.6). That is, the two approaches lead to distinct schemes. For example, taking expectations in equation (2.5), conditional upon the values $x_j (j \neq i)$, does *not* produce the result (2.1). The reason is that Z_i is correlated with the X_j in (2.5) and, although $E(Z_i)=0$, the conditional expectation $E(Z_i \mid x_j, j \neq i)$ is, in general non-zero. As a specific example, it may be useful to resurrect the simple rectangular lattice illustration of Besag (1974, section 5.2.2) with sites now conveniently labelled by integer pairs (r, s). Thus, it can be shown that, if $\Lambda = \sigma^2 I$, the simultaneous zero-mean normal autoregression,

$$X_{r,s} = \beta_1 X_{r-1,s} + \gamma_1 X_{r+1,s} + \beta_2 X_{r,s-1} + \gamma_2 X_{r,s+1} + Z_{r,s}$$

yields conditional moments,

$$\begin{aligned} E(X_{r,s} \mid \text{all other site values}) = \omega\{&(\beta_1+\gamma_1)(x_{r-1,s}+x_{r+1,s}) \\ &+(\beta_2+\gamma_2)(x_{r,s-1}+x_{r,s+1}) - (\beta_1\gamma_2+\gamma_1\beta_2)(x_{r-1,s-1}+x_{r+1,s+1}) \\ &-(\beta_1\beta_2+\gamma_1\gamma_2)(x_{r-1,s+1}+x_{r+1,s-1}) - \beta_1\gamma_1(x_{r-2,s}+x_{r+2,s}) \\ &-\beta_2\gamma_2(x_{r,s-2}+x_{r,s+2})\} \end{aligned}$$

and

$$\operatorname{var}(X_{r,s} \mid \text{all other site values}) = \omega\sigma^2$$

where $\omega = (1+\beta_1^2+\gamma_1^2+\beta_2^2+\gamma_2^2)^{-1}$. Note how the symmetry requirement is automatically satisfied when the conditional expectation structure is evaluated, without any prior constraints being placed upon β_1, β_2, γ_1, and γ_2.

The difference between the two approaches has not always been recognized in the literature on spatial schemes and this has sometimes resulted in confusion. Although previous warnings have been given, especially for lattice schemes (Bartlett 1971a, 1974; Besag 1972a, 1974a; Moran 1973), my impression is that the matter still requires publicity. Of course, the preceding discussion does not tell us *which* (if either!) of the two approaches should be used in practice; indeed, this must, for the moment, remain largely a matter of personal preference. For myself, I find that the conditional probability approach, as a whole, is one which contains considerable intuitive appeal, though it must, of course, be remembered not to interpret the conditioning in a causal sense but rather in the sense of information flow. Incidentally, it also happens that the ordinary method of least squares provides a consistent technique of parameter estimation for (local) auto-normal schemes, whereas it is well-known that the method fails for simultaneous autoregressions (Whittle 1954; Mead 1967, 1971; Cliff and Ord 1973; Hepple 1974; Ord 1975) and other means of estimation *must* be sought. We shall discuss this fully in section 3.3. For some opposing views, see, for example, Whittle (1954, 1963 and in discussion of Besag 1974a).

2.4. *Choice of Auto-normal Scheme*

We should now examine the problem of reducing the number of unknown parameters in given geographical situations. It has already been suggested that, whichever of the two approaches in sections 2.2 and 2.3 is adopted, the choice of relationships between the coefficients is likely to look much the same. However, the rationale and the end products are different and my personal prejudice therefore leads me to discuss the problem in terms of the conditional probability viewpoint, although, formally, the reduction rules will be very similar to those used for simultaneous autoregressions by Mead (1971), Cliff and Ord (1973) and Ord (1975).

The choice of a structure for the μ_i can usually be administered separately, and should not present undue difficulties. The simplest case is that of an invariant unconditional mean; that is, $\mu_i = \mu$, say, for all i. However, this situation is presumably of rather limited geographical relevance and it may usually be more appropriate to set up a classical linear representation of the μ_i, typically written as

$$\boldsymbol{\mu} = \boldsymbol{D}\boldsymbol{\theta} \tag{2.7}$$

where $\boldsymbol{D}$ is an $n \times p$ matrix of explanatory data and $\boldsymbol{\theta}$ is the associated $p \times 1$ vector of unknown parameters. Thus, in physical geography, such a specification might represent a trend surface analysis, with the ith row of $\boldsymbol{D}$ carrying (powers of) the locational coordinates of site i, whilst, in human applications, the ith row of $\boldsymbol{D}$ might relate μ_i to the values of certain economic indicators in site (region) i.

In contrast, the problem of reducing the number of unknown interaction parameters is non-classical and, if we intend to interpret the schemes, demands the ability to translate experience and intuition into bald statements of conditional probability. Whether this is a viable proposition remains to be seen. The first step must be to select a restricted set of neighbours for each site so that the scheme becomes one of local dependence, as described in section 1. Then $\beta_{i,j}$ is set equal to zero unless sites i and j are neighbours of each other; that is, unless the pair (i, j) forms a clique. Incidentally, the fact that sites i and j are not neighbours does not of course imply that the random variables X_i and X_j are independent but merely that the conditional distribution of X_i, given all other values, does not depend upon the value obtaining at site j. This is the analogous situation to that encountered with classical Markov chains: whilst no two variates are generally independent, any two are conditionally independent given the value at some intermediate time point.

Having chosen the neighbours for each site, the second step is to postulate plausible relationships between the non-zero interaction coefficients. It is suggested that this can often be done by making use of the same measures as those employed to produce the reduced neighbourhood structure, or graph, of the system. In particular, quantities such as the common boundary length, $l_{i,j}$, of contiguous regions i and j, and the Euclidean or generalized distance, $d_{i,j}$, between their appropriately defined centres may be used. Incidentally, Voronyi polygons, as suggested previously by Brown (1965), Mead (1971), Besag (1974a) and Ord (1975), provide one simple method of attaching conceptual regions to point sites. The eventual aim is to relate the non-zero interaction coefficients to perhaps one or two unknown parameters, at the same time ensuring that the symmetry condition (2.4) is satisfied. To reduce the amount of handwaving, we consider some specific examples. These are based upon suggestions made for simultaneous normal autoregressions by Mead (1971), by Cliff and Ord (1973) and by Ord (1975). We assume that $l_{i,j}(=l_{j,i})$ and $d_{i,j}(=d_{j,i})$ are defined as above and that sites i and j are deemed to be neighbours if and only if $l_{i,j}>0$. The perimeter of the site i boundary is denoted by l_i, whilst ϕ and σ represent unknown parameters. In these terms, we consider six possible specifications of the non-zero $\beta_{i,j}$. These are generated by cases (i) and (ii) below, according to the choices $\kappa=0$, 1, and 2, respectively:

(i) $\beta_{i,j}=d_{i,j}^{-\kappa}\phi$ with $\sigma_i^2=\sigma^2$

(ii) $\beta_{i,j}=(l_{i,j}/l_i)\,d_{i,j}^{-\kappa}\phi$ with $\sigma_i^2=\sigma^2/l_i$

The forms of the respective conditional variances are clearly necessary, from equation (2.4), unless the system as a whole can be split up into two or more independent subsystems. However, the requirement $\sigma_i^2=\sigma^2/l_i$ in (ii) looks a little worrying, since the implication is that the larger the boundary, the smaller the conditional variance. We therefore return to the form of $\beta_{i,j}$ in (ii) and see that the term $l_{i,j}/l_i$ is included as a weighting coefficient. As such, it may be suggested that the scheme is only appropriate when X_i represents a *normalized* version of absolute "yield", such as yield per unit area. In that case, a conditional variance which declines with increase in area would seem not only reasonable but actually desirable.

Whilst this final argument needs to be looked at critically, it is nevertheless hoped that the suggestions embodied in the section as a whole give some idea of how auto-normal schemes can be formulated in practice, especially for regional systems. Whether such an approach will help to explain any geographical phenomena remains to be seen. The proof of the pudding will be in its eating.

Before moving on to methods of Monte Carlo simulation, we remark that, with a choice such as (ii), it is convenient for subsequent statistical analysis, though not necessary, to rescale the X_i so that the scheme has invariant conditional variances. Thus, for (ii), the transformation, $Y_i = l_i^{1/2} X_i$, yields

$$E(Y_i \mid y_j, j \neq i) = \bar{\mu}_i + \sum_{j=1}^{n} \beta_{i,j}(y_j - \bar{\mu}_j)$$

where $\bar{\mu}_i = l_i^{1/2}\mu_i$, and $\mathrm{var}(Y_i \mid y_j, j \neq i) = \sigma^2$, for all i.

2.5. *Monte Carlo Simulation*

In section 3, we shall be discussing some methods of statistical analysis for conditional probability schemes. Since the sampling properties of the techniques, beyond consistency, are largely unknown and likely to be analytically intractable, it would be useful to carry out Monte Carlo simulation studies, where feasible. For discrete random variables, no direct methods of simulation have yet been found. In principle, it is possible to set up a discrete time, spatial-temporal Markov chain which yields as its stationary temporal limit the required spatial distribution. The simulation procedure is to consider the sites cyclically and, at each stage, to amend or leave unaltered the particular site value in question, according to a probability distribution whose elements depend upon the current values at neighbouring sites. For further details, see Hammersley and Handscomb (1964, Chapter 9); however, the technique is unlikely to be particularly helpful in many other than binary situations and the Markov chain itself has no practical interpretation.

On the other hand, a direct approach *is* available for the Gaussian schemes of sections 2.2 and 2.3. For example, consider the simultaneous autoregression of equation (2.5). Re-written in matrix form this becomes,

$$\boldsymbol{B}(\boldsymbol{X} - \boldsymbol{\mu}) = \boldsymbol{Z} \tag{2.8}$$

where $\boldsymbol{B}$ is defined as before and $\boldsymbol{Z}$ denotes a vector of independent normal variates, Z_i having mean zero and variance σ_i^2. To obtain a realization $\boldsymbol{x}$ of the scheme, it is merely necessary to generate a random sample $\boldsymbol{e} = (e_1, \ldots, e_n)'$ from the standard normal distribution and then to use the transformation,

$$\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{B}^{-1}\boldsymbol{\Lambda}^{1/2}\boldsymbol{e}$$

where $\boldsymbol{\Lambda}^{1/2}$ is the diagonal matrix with (i, i) element equal to σ_i.

For the auto-normal scheme of section 2.2, the procedure is less obvious but is computationally equivalent. Thus, given $\boldsymbol{B}$ and $\boldsymbol{\Lambda}$, suppose that $\boldsymbol{C}$

denotes any $n \times n$ matrix such that $C'C = \Lambda^{-1}B$. Then, if e again denotes a random sample from the standard normal distribution, the transformation,

$$x = \mu + C^{-1}e$$

produces the desired result. This follows since the sampling distribution of x has mean μ and dispersion matrix $C^{-1}(C^{-1})' = B^{-1}\Lambda$, as required. At first sight, the numerical evaluation of a suitable C may appear difficult; however, remembering that $\Lambda^{-1}B$ must be symmetric and positive-definite, we can use the fact that the standard library routine for inverting such a matrix consists of finding the unique upper (or lower) *triangular* matrix C such that $C'C = \Lambda^{-1}B$ and then inverting C (which is, of course, trivial) to obtain $B^{-1} = C^{-1}(C^{-1})'$. It follows that an appropriate matrix C and its inverse C^{-1} are readily available and that the computational effort required to simulate auto-normal schemes is no greater than that needed for simultaneous autoregressions.

3. Statistical Analysis of Conditional Probability Schemes

3.1. *Introduction*

The two principal techniques of analysis in the classical theory of statistics are the methods of maximum likelihood and of least squares, respectively. Unfortunately, neither of these appears applicable in the present context, unless the variates are blessed with a multivariate normal distribution. On the one hand, due to the occurrence of a grotesque normalizing function, the likelihood is generally intractable both to analytical and computational progress; on the other, it turns out that linear formulations are generally inappropriate unless the variates are normally distributed (Besag 1972a). Even for the auto-normal schemes, the implementation of maximum likelihood estimation is not straightforward. For such reasons, it is necessary to devise other general methods of estimation, two of which will be described in sections 3.2 and 3.3. The first involves the use of a "coding technique" to generate a relatively simple *conditional* likelihood function for the scheme. Parameter estimates are then obtained by maximizing this quantity in the usual way. However, the coding technique necessarily ignores a substantial proportion of the information which should be available from the sample. The second method of estimation, which involves the maximization of an intuitively plausible *pseudo*-likelihood function, is intended to partially overcome this deficiency. Oddly enough, its current justification is in terms of the coding technique itself. Each of the methods can easily be implemented numerically. Finally, in section 3.4, maximum-likelihood estimation for auto-normal schemes is described. The development is algebraically similar to the work of Mead (1971) and Ord (1975) on simultaneous normal autoregressions, although the schemes are, of course, non-equivalent (see section 2.3). It is emphasized that none of the three techniques described in the sequel has yet been used in practice for irregularly distributed sites. They should not be viewed uncritically!

In accordance with earlier discussion, we assume henceforth that all schemes are locally dependent (that is, each site only has a limited number of neighbours) and that the conditional distribution at site i is fully specified in terms of a vector ψ consisting of a few unknown parameters. No assumption of stationarity or homogeneity is made. Given a realization x, we shall, for discrete random variables, use $p_i(\psi)$ to denote the conditional probability of observing x_i at site i, given the values at all other sites. For continuous variates, we shall use $p_i(\psi)$ to denote the corresponding probability density. The primary objective is to obtain a reasonable estimate of ψ from the realization x.

3.2. *The Coding Technique*

Although the coding technique was first introduced in the context of regular lattice systems (Besag 1972b, 1974a), such restriction is unnecessary and will not be assumed here. The first step is to divide up the n sites, S_n, into two groups, the one of "dependents" $S_{n,0}$, the other of "conditioners". In particular, let $S_{n,0}$ denote a subset of the sites $i=1, \ldots, n$, chosen in such a way that no two members of $S_{n,0}$ are neighbours. Assign the "colour" black to each site in $S_{n,0}$ and white to the remainder. It is then evident, by the very essence of the conditional probability formulation, that the set of black-site variates, given the values at the white sites, are mutually independent. Hence, the corresponding *conditional* likelihood is obtained by multiplying together those terms $p_i(\psi)$ for which $i \in S_{n,0}$. That is, the corresponding conditional log-likelihood function is given by

$$L_{n,0}(\psi) = \sum_{i \in S_{n,0}} \ln p_i(\psi) \tag{3.1}$$

An estimate $\overset{\circ}{\psi}$ of ψ may then be found by maximizing $L_{n,0}$ with respect to ψ in the usual way. Note that this is a task of classical proportions since the need to evaluate any awkward normalizing functions has been obviated. It is therefore possible to carry out the maximization of (3.1) by plugging into a standard computer algorithm.

Furthermore, provided the number of blackened sites is not too small, *classical* maximum-likelihood theory (Kendall and Stuart 1967, Chapters 18 and 24) can be used to obtain the approximate conditional distribution of $\overset{\circ}{\psi}$ and to construct conditional likelihood-ratio tests for examining the adequacy of more restrictive schemes. Since the efficiency of the technique clearly depends upon the number of blackened sites, $S_{n,0}$ will naturally be chosen to contain as many sites as possible. Such a choice will not necessarily be unique. For the simplest lattice schemes, the proportion of blackened sites may be as high as 50 per cent but, in typical (non-lattice) geographical applications, one might expect a value nearer 30 per cent.

Experience with the coding technique has thus far been extremely limited. Two very simple lattice examples are considered numerically in Besag (1974a), whilst Besag and Moran (1975) discuss the efficiency of the technique for basic auto-normal lattice schemes. These investigations are unfortunately of little direct relevance in non-experimental situations. In fact, as a general rule, the coding technique would seem a somewhat inefficient procedure in the way that the white-site terms $p_i(\psi)$ are entirely ignored once $S_{n,0}$ has been generated. We therefore consider an alternative proposition.

3.3. *A Pseudo-likelihood Technique*

Given the previous set-up, perhaps the most naive approach to the estimation of the unknown parameters in the terms $p_i(\psi)$ would be to take that vector $\tilde{\psi}$ which maximizes the quantity

$$L_n(\tilde{\psi}) = \sum_{i=1}^{n} \ln p_i(\psi) \qquad (3.2)$$

with respect to ψ. Of course, L_n is not the true log-likelihood function for the sample (except in the trivial case of complete independence) and yet its maximization, especially in view of the coding technique, would seem to present an intuitively plausible method of estimation. That this intuition can be given a theoretical foundation will be seen later on. Note that although the pseudo-likelihood technique confers the advantages that no coding is required and that all the p-functions are used in the maximization process, it does have the drawback that no sampling properties of the estimates are yet known. Nevertheless, one supposes that the method will, on the whole, produce better point estimates of the parameters than does the coding technique.

Although maximum pseudo-likelihood estimation is intended to have fairly widespread applicability, it is of special interest to see how it fares with the auto-normal schemes of section 2. In particular, we suppose that

$$p_i(\psi) = (2\pi\sigma^2)^{-1/2} \exp\left[-\tfrac{1}{2}\sigma^{-2}\left\{x_i - \mu_i - \sum_{j=1}^{n} \beta_{i,j}(x_j - \mu_j)\right\}^2\right] \qquad (3.3)$$

so that, possibly following suitable rescaling, X_i has conditional variance σ^2 for each i. Observing the proposals in section 2.4, we also assume that $\boldsymbol{\mu} = \boldsymbol{D\theta}$ and that $\beta_{i,j} = h_{i,j}\phi$, where $\boldsymbol{D}$ and $\boldsymbol{H} \equiv \{h_{i,j}\}$ are known $n \times p$ and $n \times n$ matrices, respectively; thus $\boldsymbol{B} = \boldsymbol{I} - \boldsymbol{H}\phi$ and $\boldsymbol{\theta}$, ϕ and σ together constitute ψ. It can be seen immediately from equations (3.2) and (3.3) that the maximum pseudo-likelihood estimates $\tilde{\boldsymbol{\theta}}$ and $\tilde{\phi}$ are obtained by minimizing

$$\Omega = \sum_{i=1}^{n} \left\{ x_i - \mu_i - \sum_{j=1}^{n} \beta_{i,j}(x_j - \mu_j) \right\}^2$$

$$= (\boldsymbol{x} - \boldsymbol{D\theta})' \boldsymbol{B}^2 (\boldsymbol{x} - \boldsymbol{D\theta}) \tag{3.4}$$

with respect to $\boldsymbol{\theta}$ and ϕ. That is, the technique reduces to the ordinary method of least squares. Also, it follows from equation (3.4) that

$$\tilde{\boldsymbol{\theta}} = (\boldsymbol{D}'\tilde{\boldsymbol{B}}^2\boldsymbol{D})^{-1}\, \boldsymbol{D}'\tilde{\boldsymbol{B}}^2\boldsymbol{x} \tag{3.5}$$

where $\tilde{\boldsymbol{B}} = \boldsymbol{I} - \boldsymbol{H}\tilde{\phi}$, and that

$$\tilde{\phi} = (\boldsymbol{x} - \boldsymbol{D}\tilde{\boldsymbol{\theta}})'\, \boldsymbol{H}(\boldsymbol{x} - \boldsymbol{D}\tilde{\boldsymbol{\theta}}) / \{(\boldsymbol{x} - \boldsymbol{D}\tilde{\boldsymbol{\theta}})'\, \boldsymbol{H}^2(\boldsymbol{x} - \boldsymbol{D}\tilde{\boldsymbol{\theta}})\} \tag{3.6}$$

This suggests finding $\tilde{\boldsymbol{\theta}}$ and $\tilde{\phi}$, in practice, by using an iterative procedure of successive approximations, usually initiated by taking $\tilde{\phi} = 0$. An estimate of σ^2 is then obtained from the residual sum of squares; that is,

$$\tilde{\sigma}^2 = n^{-1}(\boldsymbol{x} - \boldsymbol{D}\tilde{\boldsymbol{\theta}})'\, \tilde{\boldsymbol{B}}^2(\boldsymbol{x} - \boldsymbol{D}\tilde{\boldsymbol{\theta}}) \tag{3.7}$$

The use of least squares estimators for auto-normal schemes may cause surprise amongst the advocates of simultaneous normal autoregressions. For it is well known (Whittle 1954; Mead 1967, 1971; Cliff and Ord 1973; Hepple 1974; Ord 1975) that, in the latter context, the use of ordinary least squares methods leads to inconsistent parameter estimates. As an example, suppose

$$X_i = \phi \sum_{j=1}^{n} h_{i,j} X_j + Z_i \tag{3.8}$$

for $i = 1, \ldots, n$, where the $h_{i,j}$ are known coefficients, with $h_{i,i} = 0$. If the Z_i are independent normal variates, each with mean zero and variance σ^2, we have a special case of the simultaneous autoregression (2.5). The ordinary least squares estimator of ϕ is then

$$\tilde{\phi} = \sum_{i=1}^{n} \sum_{j=1}^{n} h_{i,j} X_i X_j \Bigg/ \left\{ \sum_{i=1}^{n} \left(\sum_{j=1}^{n} h_{i,j} X_j \right)^2 \right\} \tag{3.9}$$

and hence,

$$\tilde{\phi} - \phi = \left\{ \sum_{i=1}^{n} Z_i \left(\sum_{j=1}^{n} h_{i,j} X_j \right) \right\} \Bigg/ \left\{ \sum_{i=1}^{n} \left(\sum_{j=1}^{n} h_{i,j} X_j \right)^2 \right\} \tag{3.10}$$

However, for the simultaneous autoregression, it follows from section 2.3 that Z_i is correlated with the X_j in equation (3.8) and therefore we have no reason to expect the right-hand side of equation (3.10) to tend to zero, in probability, as n tends (conceptually) to infinity. On the other hand, consider the analogous auto-normal scheme, for which X_i has a conditional variance σ^2 and conditional mean,

$$E(X_i \mid x_j, j \neq i) = \phi \sum_{j=1}^{n} h_{i,j} x_j$$

with $h_{i,i} \equiv 0$ and $h_{j,i} \equiv h_{i,j}$. The least squares estimator of ϕ is again given by (3.9) and if we now *define* a set of Z_i by (3.8), equation (3.10) will still hold. Moreover, although the Z_i themselves now form a *dependent* set, each Z_i is *independent* of the X_j, for $j \neq i$, and the numerator in (3.10) has mean zero. It follows that $\tilde{\phi} \rightarrow \phi$, in probability, as $n \rightarrow \infty$. Thus, in the first case, $\tilde{\phi}$ is quite inappropriate as an estimator of ϕ whilst, in the second, where ϕ of course has a different interpretation, $\tilde{\phi}$ is at least consistent.

We now return to the general problem of providing some mathematical justification for maximum pseudo-likelihood estimators. The only property which will be established is that of consistency and, as such, we shall have to admit a conceptual passage of n to the infinite limit. How relevant this is to spatial applications, where n is usually fixed, is a matter for debate; for example, the imagination palls at the thought of extending the counties of Eire to an infinite set! Nevertheless, the property of consistency might be thought of as a minimal statistical requirement. We shall sketch its proof; a rigorous treatment would, for example, require some consideration of the system boundary as n increases.

Thus, we hypothesize a system consisting of a denumerably infinite collection of sites, labelled by the positive integers in any convenient manner. We assume that no site has more than ν neighbours (ν finite) and then choose $\nu+1$ "colours", called $c_0, \ldots, c_\nu$, respectively. It follows that each site in the infinite system can be assigned one of these colours in such a way that, firstly, no two neighbours have like colours and, secondly, the number of sites coloured c_k tends to infinity as $n \rightarrow \infty$, for each k. In particular, consider the first n sites and let $S_{n,k}$ denote those of colour c_k. Each fixed k then generates a coding technique which uses $S_{n,k}$ as "dependents" (see section 3.2) and is based upon the conditional log-likelihood

$$L_{n,k}(\psi) = \sum_{i \in S_{n,k}} \ln p_i(\psi)$$

In this way, we can obtain $\nu+1$ consistent estimators of ψ, based upon $L_{n,0}, \ldots, L_{n,\nu}$, respectively. Moreover, if L_n denotes the logarithm of the n-site pseudo-likelihood, as defined in equation (3.2), then

$$L_n(\psi) = L_{n,0}(\psi) + \ldots + L_{n,\nu}(\psi)$$

and, under suitable regularity conditions, it follows that the maximization of $L_n(\psi)$ must also lead to a consistent estimator of ψ. Indeed, $\tilde{\psi}$ can be thought of as a weighted average of coding estimators.

3.4. *Maximum-likelihood Estimation for Auto-normal Schemes*

Suppose we reconsider the auto-normal scheme discussed in section 3.3. That is, in the terminology of section 2, we assume that $\boldsymbol{\Lambda}=\boldsymbol{I}\sigma^2$, that $\boldsymbol{\mu}=\boldsymbol{D\theta}$ and that $\boldsymbol{B}=\boldsymbol{I}-\boldsymbol{H}\phi$, where $\boldsymbol{D}$ and $\boldsymbol{H}$ are known matrices and $\boldsymbol{\theta}$, ϕ and σ are unknown parameters. This leads to the dispersion matrix $\boldsymbol{V}=\boldsymbol{B}^{-1}\sigma^2$ (cf. equation (2.3)) and hence a log-likelihood function given by

$$\mathscr{L}_n(\boldsymbol{\theta}, \phi, \sigma)= -\tfrac{1}{2}n \ln (2\pi\sigma^2)+\tfrac{1}{2} \ln |\boldsymbol{B}|-\tfrac{1}{2}\sigma^{-2}(\boldsymbol{x}-\boldsymbol{D\theta})^1 \boldsymbol{B}(\boldsymbol{x}-\boldsymbol{D\theta})$$

Performing the appropriate differentiations and using circumflexes to denote maximum-likelihood estimates, it is easily shown that

$$\hat{\boldsymbol{\theta}}=(\boldsymbol{D}'\hat{\boldsymbol{B}}\boldsymbol{D})^{-1}\,\boldsymbol{D}'\hat{\boldsymbol{B}}\boldsymbol{x} \qquad (3.12)$$

and

$$\hat{\sigma}^2=n^{-1}(\boldsymbol{x}-\boldsymbol{D}\hat{\boldsymbol{\theta}})'\,\hat{\boldsymbol{B}}(\boldsymbol{x}-\boldsymbol{D}\hat{\boldsymbol{\theta}}) \qquad (3.13)$$

where $\hat{\boldsymbol{B}}=\boldsymbol{I}-\boldsymbol{H}\hat{\phi}$. Substituting back into (3.11), it follows that $\hat{\phi}$ is obtained by minimizing

$$-n^{-1} \ln |\boldsymbol{B}|+\ln [n^{-1}\boldsymbol{x}'\boldsymbol{B}\{\boldsymbol{I}-\boldsymbol{D}(\boldsymbol{D}'\boldsymbol{B}\boldsymbol{D})^{-1}\boldsymbol{D}'\boldsymbol{B}\}\,\boldsymbol{x}] \qquad (3.14)$$

with respect to ϕ (cf. section 6.3 in Besag (1974b) where some expressions have a missing n^{-1}).

Computationally, the only problem which arises in minimizing (3.14) iteratively occurs in the evaluation of the determinant $|\boldsymbol{B}|$ at each stage. For auto-normal schemes, in general, where $\boldsymbol{B}$ depends upon several unknown parameters, this obstacle is likely to be prohibitive but when, as here, $\boldsymbol{B}=\boldsymbol{I}-\boldsymbol{H}\phi$, we can implement a device suggested by Ord (1975) for use with simultaneous autoregressions. In particular, suppose that $\boldsymbol{H}$ has eigenvalues $\xi_1, \ldots, \xi_n$; then

$$|\boldsymbol{B}|= \prod_{i=1}^{n} (1-\xi_i\phi)$$

This implies that, once the eigenvalues of $\boldsymbol{H}$ have been found, $\ln |\boldsymbol{B}|$ can easily be evaluated for any given value of ϕ.

Finally, it is of interest to compare the least-squares estimates (3.5) and (3.7) with those obtained by maximum likelihood, (3.12) and (3.13). The apparent contradictions can be resolved by proving that for any auto-normal scheme, with constant conditional variance σ^2,

$$E\{(\boldsymbol{X}-\boldsymbol{\mu})'\,\boldsymbol{A}\boldsymbol{B}(\boldsymbol{X}-\boldsymbol{\mu})\}=\sigma^2 \operatorname{tr}(\boldsymbol{A})$$

where $\boldsymbol{A}$ is an arbitrary $n\times n$ matrix and $\operatorname{tr}(\boldsymbol{A})$ denotes the trace of $\boldsymbol{A}$ (that is, the sum of its diagonal elements). Then it follows, for example, that

$$E\{(\boldsymbol{X}-\boldsymbol{\mu})'\,\boldsymbol{B}^2(\boldsymbol{X}-\boldsymbol{\mu})\}=E\{(\boldsymbol{X}-\boldsymbol{\mu})'\,\boldsymbol{B}(\boldsymbol{X}-\boldsymbol{\mu})\}=n\sigma^2$$

since $\operatorname{tr}(\boldsymbol{B})=n$.

4. Final Remarks

The previous sections have reflected some personal views concerning spatial Markovity and its statistical analysis, my hope being that the techniques will be of some interest in quantitative geography. It should be borne in mind that the Markov approach is still in its infancy, with almost all the theoretical developments appearing in the last five years or so—indeed, one or two items have made their débuts here. However, whilst there is therefore much scope for further mathematical excursions, I believe that the immediate requirement is for some practical analyses to be undertaken. This will hopefully provide a much-needed preliminary assessment of the relevance, if any, of spatial Markovity to geographical problems.

Finally, although this paper is concerned with the statistical analysis of purely spatial data, it would be unrealistic to make no mention at all of the spatial–temporal framework which is often intimately bound up with such considerations. Firstly, spatial systems are usually generated as instantaneous cross-sections of spatial–temporal processes, and occasionally it is possible to link the two theoretically (see, for example, Bartlett 1971b, 1974; Besag 1972a, 1974b; Whittle 1962). Secondly, it may be that data are actually available at two (or more) points in time and that the problem is one of relating the instantaneous realizations statistically. In such cases, it is sometimes plausible to use a classical spatial–temporal autoregression (see, for example, Bartlett's (1974) analysis of contagion in hop-plants); however, this will only be relevant if the process truly has only a discrete-time mechanism. For if the process is developing continuously through time and is observed at time points t and $t+1$, say, there will, in general, be no reason to suppose that $X_i(t+1)$ and $X_j(t+1)$ are independent, given (Xt). It would then be unreasonable to set up a simple spatial–temporal model which did not allow for spatial stochastic interaction (contradicting the isolated statement on spatial–temporal models in Besag (1974a)). Such problems may be considered elsewhere; but first the data analysis!

REFERENCES

BARTLETT, M. S. (1971a). Two-dimensional nearest-neighbour systems and thei ecological applications. *Statistical Ecology*, **1**, 179–94. Pennsylvania State University Press.

BARTLETT, M. S. (1971b). Physical nearest-neighbour models and non-linear time series. *Journal of Applied Probability*, **8**, 222–32.

BARTLETT, M. S. (1974). The statistical analysis of spatial pattern. *Advances in Applied Probability*, **6**, 336–58.

BESAG, J. E. (1972a). Nearest-neighbour systems and the autologistic model for binary data. *Journal of the Royal Statistical Society B*, **34**, 75–83.

BESAG, J. E. (1972b). On the statistical analysis of nearest-neighbour systems. *Proceedings of the 9th European Meeting of Statisticians, Budapest*, pp. 101–5.

BESAG, J. E. (1974a). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society*, **36**, 192–235.

BESAG, J. E. (1974b). On spatial–temporal models and Markov fields. *Proceedings of the 10th European Meeting of Statisticians, Prague* (to appear).

BESAG, J. E. and MORAN, P. A. P. (1975). On the estimation and testing of spatial interaction in Gaussian lattice processes. *Biometrika*, **62**.

BROWN, G. S. (1965). Point density in stems per acre. *New Zealand Forestry Service Research Note*, **38**, 1–11.

CLIFF, A. D. and ORD, J. K. (1973). *Spatial Autocorrelation*. London: Pion.

HAMMERSLEY, J. M. and HANDSCOMB, D. C. (1964). *Monte Carlo Methods*. London: Methuen.

HEPPLE, L. W. (1974). The impact of stochastic process theory upon spatial analysis in human geography. In *Progress in Geography*, **5**, 91–142. London: Arnold.

KENDALL, M. G. (1939). The geographical distribution of crop productivity in England (with discussion). *Journal of the Royal Statistical Society*, **102**, 21–48.

KENDALL, M. G. and STUART, A. (1967). *The Advanced Theory of Statistics*, **2**. London: Griffin.

LÉVY, P. (1948). Chaînes doubles de Markoff et fonctions aléatoires de deux variables. *Comptes Rendus de L'Academie des Sciences, Paris*, **226**, 53–5.

MATERN, B. (1960). Spatial variation. Stochastic models and their application to some problems in forest surveys and other sampling investigations. *Meddelanden Fran Statens Skogsforskningsinstitut*, **49**, 1–144.

MEAD, R. (1967). A mathematical model for the estimation of inter-plant competition. *Biometrics*, **23**, 189–205.

MEAD, R. (1971). Models for interplant competition in irregularly distributed populations. In *Statistical Ecology*, **2**, 13–22. Pennsylvania State University Press.

MORAN, P. A. P. (1973). A Gaussian Markovian process on a square lattice. *Journal of Applied Probability*, **10**, 54–62.

ORD, J. K. (1975). Estimation methods for models of spatial interaction. *Journal of the American Statistical Society* (to appear).

WHITTLE, P. (1954). On stationary processes in the plane. *Biometrika*, **41**, 434–49.

WHITTLE, P. (1962). Topographic correlation, power-law covariance functions, and diffusion. *Biometrika*, **49**, 305–14.

WHITTLE, P. (1963). Stochastic processes in several dimensions. *Bulletin of the International Statistical Institute*, **40**, 974–94.

YULE, G. U. and KENDALL, M. G. (1968). *Introduction to the Theory of Statistics*. London: Griffin.