1

Introduction

This essay aims to bring out some of the distinctive features and special problems of statistical inference on spatial processes. Realistic spatial stochastic processes are so far removed from the classical domain of statistical theory (sequences of independent, identically distributed observations) that they can provide a rather severe test of classical methods. Although much of the literature has been very negative about the problem, a few methods have emerged in this field which have spread to many other complex statistical problems. There is a sense in which spatial problems are currently the test bed for ideas in inference on complex stochastic systems.

Our definition of 'spatial process' is wide. It certainly includes all the areas of the author's monograph (Ripley, 1981), as well as more recent problems in image processing and analysis. Digital images are recorded as a set of observations (black/white, greylevel, colour ...) on a square or hexagonal lattice. As such, they differ only in scale from other spatial phenomena which are sampled on a regular grid. Now the difference in scale *is* important, but it has become clear that it is fruitful to regard imaging problems from the viewpoint of spatial statistics, and this has been done quite extensively within the last five years.

Much of our consideration depends only on geometrical aspects of spatial patterns and processes. Two of the fundamental difficulties of spatial processes are the lack of any causal ordering of the observations, and the simple observation that

$$\frac{\text{number of 'neighbours' at distance} \leqslant t}{t} \to \infty$$

as $t \to \infty$. Both difficulties divide the study of spatial processes from that of time series. In the 1950s there was an optimistic view that time series

methods could be extended simply to spatial processes (e.g. Whittle, 1954). This was accepted for a long time, and the error in the assumption (although rather simple to demonstrate) was forcefully rebutted only by Guyon's (1982) careful treatment of the spatial problem. From the perspective of 1986 the chasm appears to be between

classical and time series methods

on the one hand, and

spatial and complex system methods

on the other.

The difficulties of inference for spatial processes can conveniently be grouped under a small number of headings.

(a) Edges

Consider $\{1, \ldots, n\}^d \subset \mathbb{Z}^d$. This contains $N = n^d$ points, of which $[n^d - (n-2)^d]$ are 'outside' points, in that they have fewer than $2d$ neighbours. The number of outside points $N_0 \sim 2dn^{d-1} = 2dN^{1-(1/d)}$ and so

$$\frac{N_0}{N} \sim 2dN^{-1/d}$$

Suppose we had independent observations at each point of the lattice, and computed some measure of the similarity of the observation at each point to those of its $2d$ neighbours. We would expect the statistic so computed to converge in distribution at rate $N^{-1/2}$, by the central limit theorem. The problem of spatial processes is that for $d \geqslant 2$ the error term due to the outside points is of at least the same order as the statistical fluctuation term. Far from being asymptotically negligible, edge effects often dominate the asymptotic distribution, and almost always dominate problems of typical size.

(b) Which way to infinity?

In classical problems and in time series it is totally clear how to embed the problem uniquely in a series of problems so as to perform asymptotic calculations. This is not at all the case in spatial problems. Consider an $m \times n$ portion of $\mathbb{Z}^2$. Clearly either m or n should tend to infinity, but very little can be concluded about the ratio m/n. The reader could be forgiven for assuming that this might be immaterial, and in some problems it is. However, there are problems (Ripley, 1982, p. 252) in which the asymptotic distribution has a mean depending on the asymptotic value of m/n!

These problems have a further twist for spatial point processes. Suppose we observe n points irregularly distributed within a study region E. One

way to embed this in a sequence of problems is to let $n \to \infty$ and keep E fixed. This will have the effect of increasing the interaction between the points, and the edge effects will be O(1). Perhaps a more natural approach is to suppose that the pattern extends throughout space but was only observed within E. Then an asymptotic sequence of problems will be to take a sequence of 'windows' E increasing to $\mathbb{R}^d$. This keeps the interaction rate fixed but diminishes the edge effects. The ways in which these different asymptotic regimes give different limit results are studied in chapters 3 and 4.

(c) Long-range dependence

In time series analysis short-range dependence is the norm and special models are needed to demonstrate long-range dependence. In spatial problems long-range dependence appears inevitable. This is best illustrated by a few examples.

(i) Moving average contouring methods. Suppose we wish to interpolate a continuous surface $Z(\ )$, observed at n points $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$. A very simple way to do so is to use a weighted average

$$\hat{Z}(\mathbf{x}) = \sum_1^n \lambda_i(\mathbf{x}) Z(\mathbf{x}_i), \qquad \Sigma \lambda_i(\mathbf{x}) = 1$$

with

$$\lambda_i(\mathbf{x}) \propto w(d(\mathbf{x}, \mathbf{x}_i))$$

Clearly the fitted surface will interpolate if $w(r) \to \infty$ as $r \to 0$. We would also like $\hat{Z}(\ )$ to be differentiable, not least so as to be able to contour it without difficulty. This needs $r/w(r) \to 0$ as $r \to 0$ (Ripley, 1981, p. 36). On the other hand, let us consider large r. We would expect $O(r^{d-1})\Delta r$ points at distances r to $r + \Delta r$ away from $\mathbf{x}$, and these contribute $O(r^{d-1}w(r))\Delta r$ to the total of the weights. Thus unless

$$\int_1^\infty r^{d-1} w(r) dr < \infty$$

then $\hat{Z}(\ )$ will not be a local average but will depend entirely on the size of the study region.

This algorithm has been used quite widely in contouring packages for twenty years. One common choice of $w(r)$ was r^{-2}, which gives a smooth (C^2) interpolated surface, but with long-range dependence in $\mathbb{R}^2$.

(ii) Boundary conditions. The important distinction between a process defined only on the study region E and one defined throughout $\mathbb{Z}^d$ or $\mathbb{R}^d$ but observed in E has often been overlooked. The difference is one of

boundary conditions. For example, consider a simple conditional autoregression (Ripley, 1981, p. 88), a Gaussian process with

$$\mathrm{E}(Z_i \mid \text{rest}) = \alpha \sum_{j \text{ nhbr of } i} Z_j$$

For the process defined only for a finite part of the lattice the outside points will have fewer neighbours, but the effect is identical to setting $Z_i \equiv 0$ for $i \notin E$. Again, consider a point process of centres of discs randomly packed within E. If the process extends throughout space the discs will be forced away from the edges of E by discs whose centres lie outside E and so are unobserved. Both the proportion of edge sites *and* the geometry of the problem mean that the differences in the boundary conditions have an effect throughout E and not just near the edges.

(iii) Lattice processes. Consider again the conditional autoregression. This has direct dependence only between neighbouring sites. On $\{\ldots, -1, 0, 1, \ldots\}$ the connection between points n time units apart is through all intermediate points, and so the correlation decays as ρ^n. In $\mathbb{Z}^d$, $d > 1$, this is no longer the case. There are many paths between points n steps apart on the lattice, and the number of paths increases with n. This can balance the decay along each individual path, so that correlations decay quite slowly with n. It also ensures that a process which is defined on a grid can appear quite isotropic at medium and large scales. Besag (1981) gives some detailed calculations to support these assertions.

This remark is connected with the phenomenon of *phase transition* in statistical physics. Consider the simple model on the lattice $\mathbb{Z}^d$ which takes values ± 1 at each point. Let

$$\mathrm{P}(Z_{ij} = +1 \mid \text{all other values}) = \beta \sum_{\text{nhbrs}} Z_{rs}$$

where there are $2d$ neighbours. Then on $\mathbb{Z}$ the process has characteristics continuous in β. For $d \geqslant 2$ there is a *critical point* β_c at which many characteristics have a discontinuity. In particular, for $\beta > \beta_c$ the correlation between values at sites distance n apart does not decrease to zero with n, but converges to some positive constant, and realizations almost surely have infinite patches of $+1$ and of -1. For further details of this process see Pickard (1987).

(d) Geometry of likelihoods

There are essentially two problems with likelihood inference for spatial processes. The best known is computational. There are problems in which it is impossible to write the likelihood in a simple closed form, and

others in which the form is simple but the combinatorial terms involved are prohibitive. We give two examples from spatial point processes.

(i) Cluster processes. Suppose we have m parent points distributed independently and uniformly within E. Around each parent there is a random number n (with distribution $f(n)$) of daughter points, each of which is independently distributed about the parent with a uniformly distributed orientation and radial distance p.d.f. $P_R(\ )$. Then the p.d.f. of the n observed daughter points is

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \sum f(n_i) \prod_{j=1}^{n_i} P_R(\|\mathbf{x}_{ij} - \mathbf{y}_i\|)$$

where the sum is over the assignment of the observed daughter points to the parents $\mathbf{y}_1, \ldots, \mathbf{y}_m$. Since m and the location of the parents are usually unknown, this expression will need to be averaged over $\mathbf{y}_i$ uniform in E and over m. The combinatorial sum is prohibitive even for moderate n: there are m^n terms!

(ii) Gibbsian point processes are a way to model interactions between points, for instance to allow spacing out of birds' nests. One class of models gives a p.d.f. to n points as

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \frac{1}{Z} \prod_{i<j} h(d(\mathbf{x}_i, \mathbf{x}_j))$$

where

$$Z = \int_{E^n} \prod_{i<j} h(d(\mathbf{x}_i, \mathbf{x}_j)) \, \Pi \, d\mathbf{x}_i$$

To perform likelihood inference we need to know Z as a function of the parameters in h. This is a problem in which some progress has been made both in approximating Z and in estimating Z (Monte Carlo inference), discussed further in chapter 4.

There is, however, a more fundamental problem. Until recently it has been widely assumed that likelihood methods are in some sense near-optimal in spatial problems. (About the only published dissenting voice is Ripley, 1984b.) There is no theoretical basis for this belief. Classical theorems on strong consistency and best asymptotic normality apply to sequences of independent identically distributed (i.i.d.) random variables. The work of Mann and Wald extended these to time series, but in a context in which there is a sequence of i.i.d. innovations. There are efficiency results for spatial processes (e.g. Mardia and Marshall, 1984) but these depend on embedding the problem in an asymptotic sequence of

almost independent copies, so it is no surprise that the classical results are obtained. As we saw in (b) 'which way to infinity?', this particular asymptotic formulation need not provide useful guidance for even moderately sized problems.

The author has been suspicious of these results for some time, but only the very recent examples presented in chapter 2 demonstrated the scale of the problem. It appears likely that similar difficulties in fact occur in many other spatial contexts.

(e) Stationarity

Some assumption of stationarity plays a crucial role in virtually all of spatial statistics. Despite the fact that this is almost axiomatic it has often been argued against by users. Statistical inference is impossible without *some* stationarity assumption. Most spatial problems have only one data set and replication has to be attained from stationarity. It is this that the users misunderstand; they inconsistently argue against stationarity and simultaneously compute average characteristics of their data sets. If stationarity is false then these average characteristics have no meaning!

There is still a worthwhile debate to be had on what stationarity assumption should be made. Perhaps the easiest way to explain the difficulties is to draw parallels with time series analysis. There departures from stationarity are usually trends in mean level. (Trends in variance are studied occasionally.) These trends are usually of a simple form such as a smooth increase and/or seasonal variation, and are typically removed by differencing, perhaps after an instantaneous transformation of the data (say to log scale). Some differences are:

(i) Spatial differencing is not comprehensive. Künsch (1987) shows that the class of differenced autoregressions in $\mathbb{R}^d$, $d \geqslant 2$, is not a natural generalization of autoregressions, and some important limits of spatial autoregressions are omitted. This difficulty, like several others, stems from the lack of factorizability of polynomials over $\mathbb{C}^d$ for $d \geqslant 2$.

(ii) The class of possible departures is very large. Suppose we look at a spatial point pattern and decide it is not stationary. We could have noticed

trend in intensity from top to bottom
trend in any other direction
'banding' from periodic variation in intensity in any direction
'patchiness' on one of many scales

at least. In fact, we will usually notice something even in simulations from stationary processes! The problem is essentially that of multiple com-

parisons. We are implicitly applying tens or even hundreds of significance tests to the pattern we see, and reporting the most significant.

(iii) Stationarity can be under translations and/or rotations. As well as the complications of having two separate ideas, *homogeneity* (stationarity under translations) and *isotropy* (stationarity under rotations about a fixed point, with homogeneity, about any point), there is a structural problem. It is rather convenient that with stationarity under translations the group is isomorphic to the underlying space ($\mathbb{Z}^d$ or $\mathbb{R}^d$). The trivial observation that this is untrue for the group of rigid motions in $\mathbb{R}^d$ has nontrivial consequences.

One attempt to overcome these problems is the *intrinsic hypothesis of* Matheron (1973), which applies to a real-valued surface in $\mathbb{R}^d$ (and hence specializes to lattice processes). A generalized increment of order k is $\Sigma\lambda_i Z(\mathbf{x}_i)$, the weights λ_i satisfying

$$\Sigma\lambda_i x_{i1}^{a_1} \dots x_{id}^{a_d} = 0 \qquad \forall \alpha_i \geqslant 0, \quad \Sigma\alpha_i \leqslant k$$

where $\mathbf{x}_i = (x_{i1}, \dots, x_{id})$. The hypothesis is that all generalized increments processes of order k are stationary. The increment of order k filters out polynomials of order k, so a process can have a polynomial trend and still satisfy the intrinsic hypothesis. This avoids the problem of polynomial trends by differencing them away.

Another approach, currently favoured by the author, is to model trends as another layer of stochastic variation. We then have the stationary model

$$Z(\mathbf{x}) = Z_m(\mathbf{x}) + Z_l(\mathbf{x})$$

where Z_l has short-range dependence but Z_m has long-range dependence and so its realizations resemble trends. This doubly stochastic approach has the advantage that we do not need to specify what trends can occur, and unlike polynomial trend surfaces, it extrapolates safely.

(f) Discretization

Like many other applications of stochastic processes, the non-point-process part of spatial statistics is concerned with sampled or aggregated data. So is time series, but there are two important differences. First, even when the underlying continuous process is isotropic, the sampled version will not be. The option of modelling the continuous phenomenon and basing inference on the sampled or aggregated data is usually computationally prohibitive. Easy models on a rectangular lattice mostly do not give even approximately isotropic realizations. This problem is mainly recognized, then ignored.

The second problem is that whereas time series are usually sampled regularly, spatial phenomena are not. (Even the varying lengths of months and quarters in time series are usually ignored.)

The above catalogue of problems may give a rather bleak impression, but this would be incorrect. It is intended rather to show why spatial problems are different and challenging. The rest of this essay shows how some of these challenges have been addressed.

2

Likelihood analysis for spatial Gaussian processes

The framework used in this chapter was developed by the author (Ripley, 1981, chapter 4) to provide a formal statistical framework for the ideas of Matheron and his school of 'geostatistiques'. The ideas provide a generalization to spatial problems of the Wiener–Kolmogorov theory of prediction in time series, and provide a flexible framework for smoothing and interpolation of spatial surfaces.

We suppose that a surface $Z(\mathbf{x})$ is defined for $\mathbf{x} \in X \subset \mathbb{R}^d$. This could be topographic height over a geographical region or porosity in a (three-dimensional) oil reserve. The surface is assumed to be smooth, at least continuous and preferably differentiable. The most tractable model is a spatial Gaussian process. This is defined by the mean function

$$m(\mathbf{x}) = EZ(\mathbf{x})$$

and covariance function

$$c(\mathbf{x}, \mathbf{y}) = \text{cov}[Z(\mathbf{x}), Z(\mathbf{y})]$$

plus joint normality of the finite-dimensional distributions. Let us parameterize the mean function by a spatial regression model as

$$m(\mathbf{x}) = f(\mathbf{x})^T \beta$$

Now suppose the surface is observed at $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and we wish to predict the surface elsewhere. The minimum mean square error unbiased predictor $\hat{Z}(\mathbf{x})$ is given by

$$\hat{Z}(\mathbf{x}) = y^T k(\mathbf{x}) + f(\mathbf{x})^T \hat{\beta}$$

where

$$K = [c(\mathbf{x}_i, \mathbf{x}_j)] \qquad k(\mathbf{x}) = [c(\mathbf{x}, \mathbf{x}_i)]$$

$$F = \begin{bmatrix} f(\mathbf{x}_1)^T \\ \vdots \\ f(\mathbf{x}_n)^T \end{bmatrix} \qquad \mathbf{Z} = \begin{bmatrix} Z(\mathbf{x}_1) \\ \vdots \\ Z(\mathbf{x}_n) \end{bmatrix}$$

LL^T is the Cholesky decomposition of K, so L is lower triangular.

$L^{-1}\mathbf{Z} = L^{-1}\mathbf{F}\hat{\beta}$ gives $\hat{\beta}$.

$L(L^T y) = [Z(\mathbf{x}_i) - f(\mathbf{x}_i)^T\hat{\beta}]$ gives y.

This is a computationally stable form of 'universal kriging'. Further,

$$\text{var}[Z(\mathbf{x}) - \hat{Z}(\mathbf{x})] = c(\mathbf{x}, \mathbf{x}) - \|\mathbf{e}\|^2 + \|\mathbf{g}\|^2 \tag{1}$$

$$L\mathbf{e} = k(\mathbf{x})$$

$$R^T\mathbf{g} = f(\mathbf{x}) - (L^{-1}F)^T\mathbf{e}$$

R is the orthogonal reduction of $L^{-1}F$

This is a model-based approach which is both its strength and its weakness. The freedom to choose $c(\ ,\ )$ gives great flexibility to the technique. Conversely, c is never known and must be estimated from the data. This is done in *ad hoc* ways by the geostatistiques school. Since examples in Ripley (1981, chapter 4) and Warnes (1986) show that rather small changes in c can give rise to large changes in the fitted surface $\hat{Z}(\ )$, it is clear that the prediction variance given by (1) can very seriously underestimate the true prediction uncertainty. This is analogous to using a regression equation for prediction whilst ignoring the variability of the regression coefficients.

Spatial Gaussian processes have also been used as models for the error term of a spatial regression, particularly in geography and for agricultural field trials (Ripley, 1981, chapter 5; Cook and Pocock, 1983; Besag and Kempton, 1986). The model is again

$$Z(\mathbf{x}) = m(\mathbf{x}) + \varepsilon$$

$$m(\mathbf{x}) = f(\mathbf{x})^T\beta \tag{2}$$

but no prediction is involved, so $\mathbf{x}$ is restricted to n sites $\mathbf{x}_1, \ldots, \mathbf{x}_n$. An example might be to explain health variables on small geographical units by environmental factors. The errors $(\varepsilon_1, \ldots, \varepsilon_n)$ are assumed to be spatial autocorrelated from other environmental factors not explicitly included in the regression (2). Thus

$$\boldsymbol{\varepsilon} \sim \text{MVN}(\mathbf{0}, K)$$

where MVN denotes a multivariate normal distribution. Again $c(\ ,\ )$ and hence K is unknown and has to be fitted from the data. A parametric form $c(\mathbf{x}, \mathbf{y}; \theta)$ is assumed, and in this field maximum likelihood has been proposed for the estimator of (β, θ). The log likelihood is

$$L(\beta, \theta; \mathbf{Z}) = \text{const} - \tfrac{1}{2}[\ln|K_\theta| + (\mathbf{Z} - F\beta)^T K_\theta^{-1}(\mathbf{Z} - F\beta)] \tag{3}$$

We ignore the constant from now on. The MLE of β minimizes the quadratic form, and so is the generalized least squares estimate $\hat{\beta}$ given by

$$L^{-1}\mathbf{Z} = L^{-1}F\hat{\beta}$$