Spatial Statistics

by
Henning Omre
Department of mathematical sciences
NTNU
January 2003

1 Introduction

The intention of this brief report is to present the broad lines of the course Spatial Statistics in a consistent notation.

Statistical practice has three major dimensions:

- explorative statistics, also termed data analysis, covers evaluation of the data material under a minimum of model assumptions.
- predictive statistics, covers forecasting and prediction of non-observed outcomes of the random variable, and quantification of the associated precision. The model assumptions are pragmatic in the sense that it is the quality of the forecasts and predictions which is in focus.
- confirmative statistics, covers the evaluation of underlying effects of the data material. These effects will frequently appear as parameters in the model used. Model choice is hence crucial.

Spatial statistics is also practiced along these dimensions.

1.1 Spatial Variable

The spatial variables are usually multivariate with geographical reference, and they are denoted by,

$$\{z(x); x \in D\}$$

with $z(.)$ vector of interest, x geographical reference, and D a geographical domain in two or three dimensions.

The spatial variables are divided into three types:

- continuous variables, where $r(.) \in R$ takes real values and is continuous almost everywhere in D. Examples are: terrain height, depth to geological layer, distribution of air pollution and temperature distribution.
- mosaic variables, where $l(.) \in \{l_1, \ldots, l_k\}$ are class assignments, and there is a large degree of symmetry between the classes. This entails that no class can be considered a background class. Examples are: classes of area type, rock types in geology and human cells.

- event variables, where $z(.)$ is reparametrized to a list of events $(s_1, \ldots, s_n)$, for example geometrical objects which are located on a background m_0. Examples are: trees in a forrest, geological faults in a reservoir and sperms in plasma.

2 Stochastic Model - Prior Model

In order to give the spatial variable a probabilistic interpretation, a stochastic model has to be defined. It is usually denoted random field, contrary to random process for time series and random variable for the non-reference case.

The random field is denoted by,

$$\{Z(x); x \in D\}$$

with possible realizations, or outcomes, $\{z_1(x); x\epsilon D\}, \ldots, \{z_s(x); x\epsilon D\}$ where one of these might be the true field one wishes to explore. Hence one establishes some kind of super population concept.

The various variable types takes the following notation:

- random continuous fields - $\{R(x); x \in D\}$
- random mosaic fields - $\{L(x); x \in D\}$
- random event field - parametrized by $\{M_0, S_1, \ldots, S_n\} = \{M_0, (X_1, T_1, M_1), \ldots, (X_n, T_n, M_n)\}$ where M_0 is the background class, $X_.$ is the reference location, $T_.$ is object geometry, $M_.$ is class type, and n is number of objects.

A random field is probabilistically fully specified by its probability density function (pdf),

$$\{Z(x); x \in D\} \rightsquigarrow f_Z(z)$$

which then will be defined on a infinite dimensional vector z with all its complications.

In order to treat these pdfs, one has to define parametric models in one form or the other. Denote the associated model parameters by Θ, and note that these might be functions of the reference variable x.

Philosophically there are two approaches to this:

- frequentistic approach - assumes that there exists a true model with a unique true parameter value θ. The model is known to us, but the parameter value is completely unknown.

  Known stochastic model,

  $$f_Z(z; \theta)$$

  with θ true but unknown parameter value.

- Bayesian approach - assumes that the model is subjectively chosen with the stochasticity introduced through a stochastic interpretation of the parameter Θ. This parameter is assigned a prior pdf, $f_\Theta(\theta)$.

  Known stochastic model form,

$$f_Z(z) = \int f_{Z|\Theta}(z|\theta) f_\Theta(\theta) d\theta$$

  with Θ being a stochastically defined parameter with prior pdf $f_\Theta(\theta)$.

NOTE ! The notation in geostatistics and image analysis is different. In geostatistics the variable of interest is defined as $Z(.)$, as here. In image analysis, the parameter θ is usually the variable of interest. This is only a notational choice and the two disciplines are in many ways very similar.

3 Observations - Likelihood Model

There are a variety of ways to observe spatial variables. Firstly, one may average over a support volume. Secondly, preferential sampling is made since one uses previously collected information in the data acquisition procedure. Note further that the observations will be dependent due to spatial dependence in the spatial variable.

There are two fundamentally different ways of sampling:

- observations in different locations in one particular realization $\{z(x); x \in D\}$.
- repeated observations in the same locations over several realizations $\{z_1(x); x \in D\}, \ldots, \{z_n(x); x \in D\}$

In the following only the first sampling procedure will be considered.

Assume that the realization under study is $\{z(x); x \in D\}$, and the observations are denoted by,

$$o = g(\{z(x); x \in D\}) + u$$

where o is a vector of observations, $g(.)$ is a vector valued function of the realization under study representing the data acquisition procedure, and u is the observation error.

Examples of $g(.)$, also termed the forward or transfer function, are:

- indicator variables in location $(x_1, \ldots, x_n)$ which make $o = (z(x_1) + u_1, \ldots, z(x_n) + u_n)$ and correspond to observations in the location $(x_1, \ldots, x_n)$ with associated observations error.
- a convolution function around location $(x_1, \ldots, x_n)$ which corresponds to sampling as an average in a small area around the specified locations.

- a function representing fluid flow through $z(.)$, which corresponds to fluid flow testing through the matrix.

In order to integrate these observations with the stochastic model, a probabilistic formalism is required. Hence a stochastic notation is introduced:

$$[O|\{Z(x) = z(x); x \in D\}] = g(\{z(x); x \in D\}) + U$$

By assigning a pdf to the error term U, the likelihood model is defined,

$$f_{O|Z}(o|z) = f_{O|Z,\Theta}(o|z, \theta)$$

due to conditional independence between O and Θ when Z is given.

NOTE ! In image analysis with the variable of interest being Θ, the observations will be termed $Z(.)$.

4 Conditional Stochastic Model - Postorior Model

Prediction of characteristics of $\{Z(x); x \in D\}$ must be based on the conditional model, which in the Bayesian setting will be:

$$f_{Z|O}(z|o) = [f_O(o)]^{-1} f_{O|Z}(o|z) f_Z(z) = const \times \int f_{O|Z}(o|z) f_{Z|\Theta}(z|\theta) f_\Theta(\theta) d\theta$$

This posterior pdf can only in very particular cases be evaluated analytically, usually it must be explored by Markov chain Monte Carlo simulation.

5 Parameter Estimation

Understanding of the underlying effects can be obtained by studying estimates of the parameters of the model, θ.

Two approaches to parameter estimation exists:

- frequentist approach by maximum likelihood estimates (MLE):

$$\hat{\theta} = argmax_\theta\{f_O(o; \theta)\} = argmax_\theta\{\int f_{O|Z}(o|z) f_Z(z; \theta) dz\}$$

- Bayesian approach by maximum posterior estimates (MAP):

$$\tilde{\theta} = argmax_\theta\{f_{\Theta|O}(\theta|o)\} = argmax_\theta\{\int f_{O|Z}(o|z) f_{Z|\Theta}(z|\theta) f_\Theta(\theta) dz\}$$

Only under very particular models can these estimates be determined analytically. In the majority of cases numerical optimization must be used.

6 Functions of Random Fields

Frequently, focus is on functions of random fields,

$$w = w(\{z(x); x \in D\})$$

where w is an arbitrary function of the spatial variable. Examples are: maximum of the variable in D, the proportion of D having the variable exceeding z_0 and the probability for having contact between two locations in a binary random field.

Under the prior model, one has:

$$W = w(\{Z(x); x \in D\}) \rightsquigarrow f_W(w)$$

where $f_W(.)$ usually must be determined by simulation from $f_Z(.)$.

Under the posterior model, one has:

$$[W|O = o] = w(\{Z(x); x \in D|O = o\}) = w(\{[Z(x)|O = o]; x \in D\}) \rightsquigarrow f_{W|O}(w|o)$$

where $f_{W|O}(.|o)$ usually must be determined by simulation from $f_{Z|O}(.|o)$.

Best predictions can be defined by:

$$E\{W|O = o\}$$
$$MAP\{W|O = o\}$$

Prediction precision can be quantified by:

$$Var\{W|O = o\}$$

.