

# CONTINUOUS SPATIAL DATA ANALYSIS

## 1. Overview of Spatial Stochastic Processes

The key difference between *continuous spatial data* and *point patterns* is that there is now assumed to be a meaningful value,  $Y(s)$ , at *every* location,  $s$ , in the region of interest. For example,  $Y(s)$  might be the temperature at  $s$  or the level of air pollution at  $s$ . We shall consider a number of illustrative examples in the next section. But before doing so, it is convenient to outline the basic analytical framework to be used throughout this part of the NOTEBOOK.

If the region of interest is again denoted by  $R$ , and if the value,  $Y(s)$ , at each location,  $s \in R$  is treated as a random variable, then the collection of random variables

$$(1.1.1) \quad \{Y(s) : s \in R\}$$

is designated as a *spatial stochastic process* on  $R$  (also called a *random field* on  $R$ ). It should be clear from the outset that such (uncountably) infinite collections of random variables cannot be analyzed in any meaningful way without making a number of strong assumptions. We shall make these assumptions explicit as we proceed.

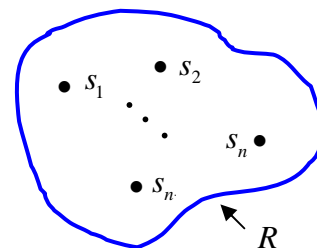
Observe next that there is a clear parallel between spatial stochastic processes and *temporal stochastic processes*,

$$(1.1.2) \quad \{Y(t) : t \in T\}$$

where the set,  $T$ , is some continuous (possibly unbounded) interval of time. In many respects, the only substantive difference between (1.1) and (1.2) is the *dimension* of the underlying domain. Hence it is not surprising that most of the assumptions and analytical methods to be employed here have their roots in *time series analysis*. One key difference that should be mentioned here is that time is naturally ordered (from “past” to “present” to “future”) whereas physical space generally has no preferred directions. This will have a number of important consequences that will be discussed as we proceed.

### 1.1 Standard Notation

The key to studying the infinite collections of random variables such as (1.1) is of course to take finite samples of  $Y(s)$  values, and attempt to draw inferences on the basis of this information. To do so, we shall employ the following standard notation. For any given set of *sample locations*,  $\{s_i : i = 1, \dots, n\} \subset R$  (as in Figure 1.1), let the *random vector*:



**Fig.1.1. Sample Locations**

$$(1.1.3) \quad Y = \begin{bmatrix} Y(s_1) \\ \vdots \\ Y(s_n) \end{bmatrix} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$$

represent the possible list of values that may be observed at these locations. Note that (following standard matrix conventions) we always take vectors to be *column* vectors unless otherwise stated. The second representation in (1.3) will usually be used when the specific locations of these samples are not relevant. Note also that it is often more convenient to write vectors in *transpose* form as  $Y = (Y_1, \dots, Y_n)'$ , thus yielding a more compact in-line representation. Each possible *realization*,

$$(1.1.4) \quad y = (y_1, \dots, y_n)' = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

of the random vector,  $Y$ , then denotes a possible set of specific observations (such as the temperatures at each location  $i = 1, \dots, n$ ).

Most of our analysis will focus on the means and variances of these random variables, as well as the covariances between them. Again, following standard notation we shall usually denote the *mean* of each random variable,  $Y(s_i)$ , by

$$(1.1.5) \quad E[Y(s_i)] = \mu(s_i) = \mu_i, \quad i = 1, \dots, n$$

so that the corresponding *mean vector* for  $Y$  is given by

$$(1.1.6) \quad E(Y) = [E(Y_1), \dots, E(Y_n)]' = (\mu_1, \dots, \mu_n)' = \mu$$

Similarly, the *variance* of random variable,  $Y(s_i)$ , can be denoted in a number of alternative ways as:

$$(1.1.7) \quad \text{var}(Y_i) = E[(Y_i - \mu_i)^2] = \sigma^2(s_i) = \sigma_i^2 = \sigma_{ii}$$

The last representation facilitates comparison with the *covariance* of two random variables,  $Y(s_i)$  and  $Y(s_j)$ , as defined by

$$(1.1.8) \quad \text{cov}[Y(s_i), Y(s_j)] = E[(Y_i - \mu_i)(Y_j - \mu_j)] = \sigma_{ij}$$

The full matrix of variances and covariances for the components of  $Y$  is then designated as the *covariance matrix* for  $Y$ , and is written alternatively as

$$(1.1.9) \quad \text{cov}(Y) = \begin{bmatrix} \text{cov}(Y_1, Y_1) & \cdots & \text{cov}(Y_1, Y_n) \\ \vdots & \ddots & \vdots \\ \text{cov}(Y_n, Y_1) & \cdots & \text{cov}(Y_n, Y_n) \end{bmatrix} = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1n} \\ \vdots & \ddots & \vdots \\ \sigma_{n1} & \cdots & \sigma_{nn} \end{pmatrix} = \Sigma$$

where by definition,  $\text{cov}(Y_i, Y_i) = \text{var}(Y_i)$ .

As we shall see below, spatial stochastic processes can be often be usefully studied in terms of these first and second moments (means and covariances). This is especially true for the important case of *multivariate normally distributed* random vectors that will be discussed in some detail below. For the present, it suffices to say that much of our effort to model spatial stochastic processes will focus on the structure of these means and covariances for finite samples. To do so, it is convenient to start with the following overall conceptual framework.

## 1.2 Basic Modeling Framework

Essentially all spatial statistical models that we shall consider start by decomposing the statistical variation of random variables,  $Y(s)$ , into a deterministic *trend term*,  $\mu(s)$ , and a stochastic *residual term*,  $\varepsilon(s)$ , as follows

$$(1.2.1) \quad Y(s) = \mu(s) + \varepsilon(s) \quad , \quad s \in R$$

Here  $\mu(s)$  is almost always take to be the mean of  $Y(s)$ , so that by definition,

$$(1.2.2) \quad \varepsilon(s) = Y(s) - \mu(s) \Rightarrow E[\varepsilon(s)] = E[Y(s)] - \mu(s) \\ \Rightarrow E[\varepsilon(s)] = 0 \quad , \quad s \in R$$

Expressions (1.2.1) and (1.2.2) together constitute the basic modeling framework to be used throughout the analyses to follow. It should be emphasized that this framework is simply a convenient representation of  $Y(s)$ , and involves no substantive assumptions. But it is nonetheless very suggestive. In particular, since  $\mu(\cdot)$  defines a deterministic function on  $R$ , it is useful to think of  $\mu(\cdot)$  as a *spatial trend function* representing the typical values of the given spatial stochastic process over all  $R$ , i.e., the *global structure* of the  $Y$ -process. Similarly, since  $\varepsilon(\cdot)$  is by definition a spatial stochastic process on  $R$  with mean identically zero, it is useful to think of  $\varepsilon(\cdot)$  as a *spatial residual process* representing local variations about  $\mu(\cdot)$ , i.e., the *local structure* of the  $Y$ -process.

Within this framework, our *basic modeling strategy* will be to identify a spatial trend function,  $\mu(\cdot)$ , that fits the  $Y$ -process so well that the resulting residual process,  $\varepsilon(\cdot)$ , is not statistically distinguishable from “random noise”. To make this strategy precise, we must of course develop appropriate models of “random noise” (in a manner paralleling the *CSR* hypothesis above). But first, we consider a range of motivating examples.