# SPATIAL STATISTICS

**Dale L. Zimmerman**
*Department of Statistics and Actuarial Science, University of Iowa, USA*

**Keywords:** Geostatistics, Isotropy, Kriging, Lattice Data, Spatial point patterns, Stationarity

## Contents





## Summary

This article gives a brief overview of classical spatial statistics. Three main types of spatial data are considered: geostatistical data, lattice data, and spatial point patterns. Modeling themes common to all three data types are emphasized. In addition, some basic structural identification procedures for each data type are described. Finally, maximum likelihood estimation and spatial prediction (kriging) are briefly reviewed.

## 1. Introduction

Many observational studies and scientific investigations involve making observations of one or more variables at multiple, identifiable sites within some region in two-dimensional or three-dimensional space. If the locations of these sites (in some coordinate system) are observed and attached, as labels, to the observations, the resulting data are called *spatial data*. A *spatial data analysis* is an analysis of spatial data in which the set of spatial locations are taken into account. *Spatial statistics* is a particular kind of spatial data analysis in which the observations or locations (or both) are modeled as random variables, and inferences are made about these models and/or about additional unobserved quantities. Spatial statistics also includes methods, based on the same stochastic models, for determining where and how the observations are to be taken (spatial design).

Spatial statistics is a vast subject, in large part because the observations, the data locations, and the mechanisms that tie the two together can be of so many different types. The observations, for example, may be univariate or multivariate, categorical or continuous, real-valued or not real-valued (e.g. set-valued); and they may arise from either an observational study or a well-designed experiment or sample survey. The data locations may be points, regions, line segments, or curves, they may be regularly or irregularly spaced; they may be regularly or irregularly shaped; and they may belong to a Euclidean or non-Euclidean space (e.g. the surface of a sphere). The mechanism that generates the data locations may be random or non-random, and may be related or unrelated to the processes that govern the observations.

Notwithstanding the immensity of the subject, three specific subfields of spatial statistics are most important, as measured by the amount of attention they have received and the amount of methodology that has been developed. These three subfields, known as geostatistics, lattice data analysis, and spatial point pattern analysis, are distinguished by data type. Roughly, geostatistical data are point observations of a continuously varying quantity over a region in space; lattice data are counts or spatial averages of a quantity over sub-regions of a larger region; and a spatial point pattern is an arrangement of a countable number of points within a region. Examples of geostatistical data abound in geology and mining, from which the name was originally derived, but they also occur in hydrology, atmospheric science, and other fields. Specific examples would include the richness of iron ore within an ore body, annual acid rain deposition at point sites in the eastern U.S., and the level of electrical activity at point sites in the human brain in response to a specific stimulus. Lattice data include, for example, pixel values from remote sensing of natural resources. Further examples would include the presence or absence of a plant species in square blocks laid out over a prairie remnant, and the number of deaths due to lung cancer in the counties (or other administrative districts) of a nation. The third type, spatial point pattern data, arise in many diverse fields including forestry (e.g. locations of trees in a forest), astronomy (e.g. locations of craters on the moon), and epidemiology (e.g. locations of lung cancer cases in relation to the location of an incinerator).

The distinctions between these three main types are not always clear-cut. In particular, lattice data have some similarities with the other two types. Indeed, some lattice data may be the result of integrating a geostatistical process, and other lattice data may result from aggregating a spatial point pattern. Thus, statistical methodology for lattice data borrows substantially from that for the other two types.

Why has spatial spatisitcs emerged, over the past 20 years, as a distinct and important area within statistics? There are many reasons. One factor has been a growing awareness that data collected in space, like data collected over time, tend to exhibit statistical dependence. One commonly exhibited form of dependence is *spatial continuity*, which says merely that observations taken at two sites tend to be more alike if the sites are close together than if the sites are far apart. The existence of this or other kinds of dependence would cast the results of a classical statistical analysis, based on an assumption of independent observations, into doubt. Two examples will serve to make this point.

Consider a situation in which 16 observations are taken at point sites forming a 4 by 4 square grid. Suppose that these observations have common mean $\mu$ and common variance 1, but are spatially correlated. Specifically, suppose that the correlation between an observation taken in row *i* and column *j* and an observation taken in row *k* and column *l* is equal to $0.5^{|i-k|+|j-l|}$. Suppose we wish to estimate $\mu$ by the sample mean, $\overline{Y}$, of the observations. If there were no spatial correlation, then the variance of $\overline{Y}$ would of course be equal to 1/16. But with the spatial correlation properly accounted for, it can be shown that the variance of $\overline{Y}$ is equal to ¼. Thus, failing to account for spatial correlation would lead to understating the sampling variance by a factor of four.

As a second example of the importance of accounting for spatial structure, consider a situation in which one wants to estimate the number of plants, *N*, of a certain species that live within a region of interest. One method for estimating *N* is based on measuring the distance, say $X_i$, to the nearest plant from each of several, say *m*, fixed points interspersed throughout the region. If the plant locations are *completely spatially random*, i.e. if they are a random sample from a uniform distribution over the study area, then the maximum likelihood estimator of *N* is $\widehat{N} = mA/\sum_{i=1}^{m} \pi X_i^2$, where *A* is the area of the region of interest. If the plant locations are not completely spatially random, however, then $\hat{N}$ can be badly biased. For example, if the plants tend to occur in clusters separated by large areas void of plants, then $\hat{N}$ can badly underestimate the total number of plants.

Another factor that has contributed to the rise of spatial statistics is an increased interest in estimation or otherwise characterizing spatial dependence for its own sake. For instance, knowledge of the extent to which the spatial distribution of two plant species co-vary can shed light on whether the two species have some kind of mutualistic association, no association, or a competitive association. Knowledge of the nature of spatial dependence of crop yields in a field can lead to a selection of plot size and shape that maximize the information content of subsequent variety trials or other experiments carried out on that field.

Finally, the tremendous increase in computational capability has played an important role in the development of spatial statistics. Many inference procedures for spatial data are much more computationally intensive than classical procedures. Faster computers and better algorithms have rendered practical computations that would otherwise be impractical.

This article gives a brief overview of statistical models for univariate geostatistical, lattice, and spatial point pattern data. Modeling themes common to all three data types are emphasized. In addition, some basic structural identification procedures for each data type are described. Finally, maximum likelihood estimation and spatial prediction are briefly reviewed. For concreteness we assume throughout that the region of interest lies in two-dimensional Euclidean space.

## 2. Models

### 2.1. Geostatistical Models

Let $Z_1, Z_2, ...., Z_n$ denote $n$ observations of a quantity of interest at $n$ point sites $\mathbf{s}_1,\ldots,\mathbf{s}_n$ in a region of interest $D$. The point sites are assumed fixed, and if this is not so then all inferences must be regarded as conditional on the observed point sites. This is a situation similar to that of classical linear regression, in which inferences must be interpreted as conditional on the realized values of the regressor variables.

The framework for modeling geostatistical data is the assumption that the observations represent a finite sample from a realization of a random process $\{Z(\mathbf{s}) : \mathbf{s} \in D\}$. Within this framework it is further assumed that

$$Z(\mathbf{s}) = m(\mathbf{s}) + \epsilon(\mathbf{s})$$

where $m(\mathbf{s}) \equiv E[Z(\mathbf{s})]$ is known as the *mean function* of the process and $\{\epsilon(\mathbf{s}) : \mathbf{s} \in D\}$ is a zero-mean random process. This model decomposes the total variation of the quantity of interest over $D$ into large-scale variation (the mean function) and small-scale variation (the residual process). The residual process has associated with it a *covariance function*, which expresses the covariance between two values of $\epsilon(\cdot)$ as a function of the coordinates of the two corresponding sites, i.e.

$$C(\mathbf{s},\mathbf{t}) \equiv \text{cov}\{\epsilon(\mathbf{s}), \epsilon(\mathbf{t})\}$$
$$= \text{cov}\{Z(\mathbf{s}), Z(\mathbf{t})\}.$$

The mean function is generally not constrained, but the principle of spatial continuity suggests that it may often be sensible to take it to be continuous and relatively smooth. The covariance function, however, must satisfy two properties. First, it must be symmetric, i.e., $C(\mathbf{s},\mathbf{t}) = C(\mathbf{t},\mathbf{s})$ for all $\mathbf{s},\mathbf{t} \in D$. Second, it must be nonnegative definite, i.e.,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j C(\mathbf{s}_i, \mathbf{s}_j) \geq 0$$

for all $n$, all sequences $\{a_i : \ i = 1,...,n\}$, and all sequences of spatial locations $\{\mathbf{s}_i \in D : \ i = 1,...,n\}$.

Thus, the modeling of geostatistical data involves supposing that $Z_i = Z(\mathbf{s}_i)$ and then making choices of a mean function and a valid covariance function. Generally, it is assumed that these functions belong to certain parametric families with unknown parameter values. Accordingly, the mean function is rewritten as $m(\mathbf{s};\boldsymbol{\beta})$ and the covariance function as $C(\mathbf{s},\mathbf{t};\boldsymbol{\theta})$ where $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are unknown parameters to be estimated. Since the data generally originate from only one realization, some further assumptions about the residual process must be made for parameter estimation and other types of inference to be possible. One useful assumption is *stationarity*, which asserts that even a single realization of the process has a certain kind of replication built into it. Two well-known types of stationarity are strict stationarity and second-order (or covariance) stationarity. The former stipulates that the joint probability distribution of the residuals depends only on the relative positions of the sites at which the data were taken. The latter is weaker than the former and stipulates only that the covariance between residuals at two sites depends on the sites' relative positions. In practice, an assumption of second-order stationarity is sufficient for point estimation but a distributional assumption (e.g. normality) may be needed for other inference purposes. Furthermore, the principle of spatial continuity suggests that emphasis be given to covariance functions that decrease monotonically to zero as the inter-site distance, $[(\mathbf{s}-\mathbf{t})'(\mathbf{s}-\mathbf{t})]^{1/2}$ increases (in any direction). Additionally, the property of *isotropy*, whereby the covariance function depends on sites only through the Euclidean distance between them, may be assumed for convenience.

What are some common choices for the mean and covariance functions? Ideally, a parametric mean function should be flexible enough to approximate closely surfaces of various shapes yet parsimonious enough to be estimated well. A family of mean functions that has proven to be useful is the polynomial family of order *q*, given by

$$m(\mathbf{s};\boldsymbol{\beta}) = \beta_0 + \beta_1 f_1(\mathbf{s}) + \ldots + \beta_p f_p(\mathbf{s}),$$

where $\{f_1(\mathbf{s}),\ldots,f_p(\mathbf{s})\}$ are pure and mixed monomials of degree $\leq q$ in the coordinates of **s**. For example, in two dimensions, with a site's coordinates denoted as **s** = *(x, y)*, the full first-order and second-order polynomials are as follows:

$$m(x,y;\boldsymbol{\beta}) = \beta_0 + \beta_1 x + \beta_2 y,$$

$$m(x,y;\boldsymbol{\beta}) = \beta_0 + \beta_1 x + \beta_2 y + \beta_{11} x^2 + \beta_{12} xy + \beta_{22} y^2$$

Other parametric families, including trigonometric and other nonlinear functions, are possible but are used rather less often.

Ideally a covariance function, like a mean function, should be flexible yet parsimonious. Three popular parsimonious covariance models are the spherical model

$$C(r;\boldsymbol{\theta}) = \begin{cases} \theta_0 I_{\{r=0\}} + \\ \theta_1(1 - \frac{3r}{2\theta_2} + \frac{r^3}{2\theta_2^3}) & \text{for } 0 \le r \le \theta_2 \\ 0 & \text{for } r > \theta_2 \end{cases}$$

the exponential model

$$C(r;\boldsymbol{\theta}) = \theta_0 I_{\{r=0\}} + \theta_1 \exp(-\theta_2 r),$$

and the Gaussian model

$$C(r;\boldsymbol{\theta}) = \theta_0 I_{\{r=0\}} + \theta_1 \exp(-\theta_2 r^2).$$

Here, $r$ is the Euclidean distance between sites, $I_{\{r=0\}}$ is an indicator function equal to 1 if $r = 0$ and 0 otherwise, and the allowable parameter space is

$$\{(\theta_0, \theta_1, \theta_2): \\ \theta_0 > 0, \theta_1 \ge 0, \theta_2 \ge 0\}.$$

In each of these models, the covariance is a monotone decreasing function of inter-site distance, vanishing to zero as the inter-site decreases. However, the parabolic behavior of the Gaussian model near $r = 0$ contrasts with the linear behavior there of the other two models, and confers a much smoother behavior upon the corresponding random process.

The isotropic models just given may be generalized to *geometrically anisotropic* models by replacing $r$ with $[(\mathbf{s}-\mathbf{t})'\mathbf{A}(\mathbf{s}-\mathbf{t})]^{1/2}$ where $\mathbf{A}$ is any positive definite matrix. Geometrically anisotropic models allow for the modeling of stronger dependence in some directions than in others.

Historically, practitioners of geostatistics have adopted a slightly more general kind of stationarity assumption than second-order stationarity, and they have modeled the small-scale spatial dependence through a function somewhat different from the covariance function. The more general stationarity assumption is called *intrinsic stationarity*, and it specifies that the residuals have mean zero and that $\frac{1}{2}\text{var}[\epsilon(\mathbf{s}) - \epsilon(\mathbf{t})]$ depends only on the lag $\mathbf{s} - \mathbf{t}$, i.e.,

$$\frac{1}{2}\text{var}[\epsilon(\mathbf{s}) - \epsilon(\mathbf{t})] = \gamma(\mathbf{s}-\mathbf{t}), \text{ for all } \mathbf{s},\mathbf{t} \in D.$$

The function $\gamma(\cdot)$ just defined is called the *semivariogram*. A second-order stationary random process with covariance function $C(\cdot)$ is intrinsically stationary, with semivariogram

$$\gamma(\mathbf{h}) = C(\mathbf{0}) - C(\mathbf{h})$$

but the converse is not true in general. That is, there exist processes that are intrinsically stationary but not second-order stationary. For second-order stationary processes, however, the spatial dependence can be described by either the covariance function or the semivariogram.

**2.2 Lattice Models**

Let $Z_1, Z_2, ..., Z_n$ denote lattice data at $n$ sites. As is the case for geostatistical data, it is useful to regard lattice data as derived from a single realization of a random process. In contrast to geostatistical data, however, a lattice process is often observed at every site at which it occurs. Therefore, it suffices to adopt a model for the quantity of interest at the sites where it is actually observed rather than at all points within a region

Nevertheless, models similar to geostatistical models can be used for lattice data. That is, a mean function and a covariance function can be defined that are entirely equivalent to functions used for geostatistical data save that the index set on which they are defined is finite. When the sites are regions, this commonly involves assuming that the observation for each region occurs at the centroid of the region. This assumption is somewhat arbitrary, so alternative models analogous to time series models have received considerable attention. These analogues require that a set of "neighbors" be defined for each site. For example, if sites are contiguous regions (e.g. counties or other administrative units), then commonly a site's neighbors are defined to be those other sites with which it shares a border.

Two popular models that incorporate this discrete neighbor information are the simultaneous autoregressive (SAR) and conditional autoregressive (CAR) models. The Gaussian SAR model is given by

$$Z_i - \mu_i = \sum_j S_{ij}\left(Z_j - \mu_j\right) + \epsilon_i,$$

$$\epsilon_i \sim N\left(0, \sigma^2\right), \quad i = 1, ..., n \quad ,$$

where $\mu_i = E(Z_i)$ and $\mathbf{S} \equiv \{S_{ij}\}$ is a matrix of fixed parameters such that $S_{ii} = 0$ and **I – S** is nonsingular. The Gaussian CAR is specified by

$$Z_i \big| Z_j, j \neq i \sim N(\mu_i + \sum_j C_{ij}(Z_j - \mu_j), \sigma^2), \quad i = 1,...,n$$

where $\mathbf{C} \equiv \{C_{ij}\}$ is such that $C_{ii} = 0$ and $\mathbf{I} - \mathbf{C}$ is symmetric and positive definite.

Though similar in form, the SAR and CAR models are different, i.e., if we take $C_{ij} = S_{ij}$ the CAR yields responses whose joint distribution is different from for the SAR. The joint distribution of $Z_1,...,Z_n$ is $N(\boldsymbol{\mu}, \sigma^2(\mathbf{I} - \mathbf{S})^{-1}(\mathbf{I} - \mathbf{S}')^{-1})$ under the SAR and $N(\boldsymbol{\mu}, \sigma^2(\mathbf{I} - \mathbf{C})^{-1})$ under the CAR, when $\boldsymbol{\mu} = (\mu_i)$

The SAR and CAR models as given above have too many unknown parameters to be useful in practice. Typically, these models are parameterized in terms of a single parameter and a given neighborhood structure. Thus, for the SAR model $\mathbf{S} = \rho_s \mathbf{W}$ and for the CAR model $\mathbf{C} = \rho_c \mathbf{W}$ where $\mathbf{W}$ is a *neighborhood matrix* whose (*i, j*)th element is equal to 1 if region *i* and region $j (i \neq j)$ share a common boundary, and is equal to zero otherwise, and $\rho_s$ or $\rho_c$ is a spatial dependence parameter to be estimated.

For completeness we mention spatial moving average models. These models can be represented, in general, as follows:

$$Z_i - \mu_i = \sum_j M_{ij} \epsilon_j, \quad i = 1,...,n$$

where the $M_{ij}$'s are unknown parameters satisfying $M_{ii} = 1$ and the $\epsilon_i$'s are independent and identically distributed as $N(0, \sigma^2)$. The joint distribution of $Z_1,...,Z_n$ is $N(\boldsymbol{\mu}, \sigma^2 \mathbf{M}\mathbf{M}')$ where $\mathbf{M} = (M_{ij})$. Similar to the case for SAR and CAR models, $\mathbf{M}$ is usually modeled very parsimoniously, e.g. as $\mathbf{M} = \rho_m \mathbf{W}$ where $\rho_m$ is a spatial dependence parameter and $\mathbf{W}$ is a neighborhood matrix.

-
-
-

**Bibliography**

Chilés, J.-P. and Delfiner, P. (1999). *Geostatistics: Modeling Spatial Uncertainty*, New York: Wiley. [A thorough exposition of the theory underlying geostatistics.]

Cliff, A.D. and Ord, J.K. (1981). *Spatial Processes: Models and Applications*, London: Pion. [One of the earliest books on spatial statistics. Emphasizes measures of autocorrelation, models, and inference for lattice data.]

Cressie, N.A.C. (1993). *Statistics for Spatial Data*, New York: Wiley. [A comprehensive reference on spatial statistics, reviewing most of the relevant literature through the early 1990's. Covers geostatistics, lattice data, spatial point patterns, and random sets.]

Diggle, P.J. (1983). *Statistical Analysis of Spatial Point Patterns*, London: Academic Press. [A classic on the theory and analysis of spatial point patterns. Treats univariate and multivariate patterns at a moderate level of difficulty.]

Haining, R. (1990). *Spatial Data Analysis in the Social and Environmental Sciences*, Cambridge: Cambridge University Press. [A readable account of geostatistics and lattice data analysis, with emphasis on the latter. Does not include spatial point pattern analysis.]

Ripley, B.D. (1981). *Spatial Statistics*, New York: Wiley. [One of the first books on spatial statistics. Treats spatial regression, lattice data analysis, spatial point patterns, and image analysis.]

Stoyan, D. and Stoyan, H. (1994). *Fractals, Random Shapes and Point Fields: Methods of Geometrical Statistics*, Chichester: Wiley. [A geometrical approach to the modeling and analysis of spatial point patterns, random shapes, and fractals.]

Upton, G.J. and Fingleton, B. (1985). *Spatial Data Analysis By Example*, New York: Wiley. [An elementary treatment of spatial point pattern and lattice data analysis, replete with numerous interesting examples. Does not include geostatistics.]

**Biographical Sketch**

**Dale Zimmerman** is Professor of Statistics at the University of Iowa. His research interests include ecological statistics, spatial statistics, and environmetrics. One active area of his research is statistical methodology for combining pollutant concentration data from several monitoring networks to produce better predictions of pollutant levels at unsampled sites. He is also currently developing stochastic models for animal movement in heterogeneous environments that combine temporal correlation in movements with resource selection.