Chapter 2
Regionalized Variable Theory

Abstract Many physical, chemical and biological processes have acted to create the current environment with the result that the variation appears to be random. Practical geostatistics treats the results as if they were the outcomes of correlated random processes and is underpinned by assumptions of stationarity. Variation may be treated as second-order stationary and represented by covariance functions. The somewhat weaker assumption of intrinsic stationarity leads to a more general analysis based on the variogram as a description of the variation. Quasi-stationarity limits stationarity to local areas, and with sufficient data the assumptions can be applied locally. If there is trend then more complex assumptions are needed; these usually comprise a combination of deterministic spatially smooth trend plus random residuals that are spatially correlated and stationary to some degree.

Keywords Random function · Random variable · Regionalized variable · Stationarity · Intrinsic hypothesis · Quasi-stationarity · Trend

2.1 Random Variables and Regionalized Variable Theory

The processes that act in the environment obey the laws of physics, and are in that sense deterministic: the variation we observe has its physical causes. Nevertheless, numerous processes have combined and interacted to produce the current environment, and the results are so complex that the variation appears as though it were generated randomly (Webster 2000). In geostatistical terms, we regard the grades of ores, properties of the soil, or the rainfall of a region, of almost any size, as the realizations of random processes.

Based on this view the value of, for example, a soil property such as its pH, at any place, $\mathbf{x}$ denoting its coordinates in two dimensions, is just one of the infinitely many that are possible there. We associate with each place $\mathbf{x}$ not just one value but a whole suite of values with a mean, a variance and higher-order moments of a distribution. The actual value at $\mathbf{x}$ is regarded as just one value from that

distribution, allocated at random. Thus the value of the variable at **x** is treated as a random variable, which we denote with the capital Z. The set of random variables for all **x** in $\mathfrak{R}$ constitutes a random function, random process or stochastic process. Random variables in the real space, which may be one-, two- or three-dimensional, are also called 'regionalized variables', and hence we have the theory of regionalized variables mentioned above.

A random function has no mathematical description in the way that a deterministic one has, i.e. it cannot be written as an equation. Nevertheless, it may have 'structure' in that there is correlation in space, or in time (for signals). This means that values at different places may be related to one another in a statistical sense. Intuitively, we expect the features of the environment at places near to one another to be similar, whereas those at widely separated places are less likely to be. This intuition is formalized in the theory of random functions. We must realise that the randomness is a mental model of the world and not a property of the environment.

2.1.1 Stationarity

Stationarity underpins the practicality of geostatistics; it is an assumption that enables us to treat data as though they have the same degree of variation over a region of interest. We can represent the random process by the model

$$Z(\mathbf{x}) = \mu + \varepsilon(\mathbf{x}), \tag{2.1}$$

where μ is the mean of the process and $\varepsilon(\mathbf{x})$ is a random quantity with a mean of zero and a covariance, $C(\mathbf{h})$, where **h** is the separation in space and known as the lag. The covariance is

$$C(\mathbf{h}) = \mathrm{E}[\varepsilon(\mathbf{x})\varepsilon(\mathbf{x}+\mathbf{h})], \tag{2.2}$$

which is equivalent to

$$C(\mathbf{h}) = \mathrm{E}[\{Z(\mathrm{x}) - \mu\}\{Z(\mathbf{x}+\mathbf{h})\} - \mu\}] = \mathrm{E}[\{Z(\mathbf{x})\}\{Z(\mathbf{x}+\mathbf{h})\} - \mu^2]. \tag{2.3}$$

Here $Z(\mathbf{x})$ and $Z(\mathbf{x}+\mathbf{h})$ are the values of the random variable Z at places **x** and **x** + **h** and E denotes the expectation. This covariance depends on **h** and only on **h**, the separation between samples in both distance and direction; it is a function of **h**. The assumption on which this is based is that of second-order stationarity. In the real world, we often encounter situations in which we cannot assume that the mean is constant, and if so the covariance cannot exist. Such a situation need not be a stumbling block; we can simply weaken the assumption of stationarity to that of what Matheron (1963) called *intrinsic stationarity* in which the expected differences are zero,

$$E[Z(\mathbf{x}) - Z(\mathbf{x} + \mathbf{h})] = 0, \tag{2.4}$$

and the covariance of the residuals is replaced by the variance of the differences to measure the spatial relations:

$$\text{var}[Z(\mathbf{x}) - Z(\mathbf{x} + \mathbf{h})] = E\left[\{Z(\mathbf{x}) - Z(\mathbf{x} + \mathbf{h})\}^2\right] = 2\gamma(\mathbf{h}). \tag{2.5}$$

Here $\gamma(\mathbf{h})$ is the semivariance at lag $\mathbf{h}$, and as a function of $\mathbf{h}$ it is the variogram. The variogram is based on differences, and provided Eq. (2.4) holds locally it is valid. This property makes the variogram more generally useful than the covariance function. In Chap. 3 we describe how to compute the covariance and variogram functions. We focus on the variogram because of its generality and go on in Chap. 3 to describe variogram modelling.

For second-order stationary processes the covariance function and variogram are equivalent:

$$\gamma(\mathbf{h}) = C(\mathbf{0}) - C(\mathbf{h}), \tag{2.6}$$

where $C(\mathbf{0}) = \sigma^2$ is the variance of the process.

A process that appears stationary at one scale might at another scale appear to embody trend, that is, a systematic component. At this scale we might have to elaborate the simple model represented in Eq. (2.1) by

$$Z(\mathbf{x}) = u(\mathbf{x}) + \varepsilon(\mathbf{x}), \tag{2.7}$$

in which $u(\mathbf{x})$ is a deterministic trend term that replaces the constant mean, μ. Its variogram,

$$\gamma(\mathbf{h}) = \frac{1}{2} E\left[\{\varepsilon(\mathbf{x}) - \varepsilon(\mathbf{x} + \mathbf{h})\}^2\right], \tag{2.8}$$

is no longer the same as

$$\gamma(\mathbf{h}) = \frac{1}{2} E\left[\{Z(\mathbf{x}) - Z(\mathbf{x} + \mathbf{h})\}^2\right], \tag{2.9}$$

of Eq. (1.5). It is the variogram of the residuals from the trend. We explain what to do to estimate the variogram in the presence of trend in Chap. 6.