

https://dspace.org/

Analysing spatial data via geostatistical methods

Morgan, Craig John

2006-11-16T08:23:15Z
http://hdl.handle.net/10539/1768

Downloaded from a DSpace repository. https://dspace.org/

CHAPTER 4
THE THEORY OF REGIONALIZED VARIABLES

4.1 Introduction

It is pointed out by Armstrong (1998 : 16) that Matheron (1963b), realizing the significance of the spatial aspect of geostatistical data, coined the term regionalized variable to represent variables distributed in space. Such variables possess both

- "…a random aspect, which accounts for local irregularities…", and
- "…a structured aspect, which reflects large scale tendencies"

(Armstrong, 1998 : 16)

and are ideally suited to describe "…almost all variables encountered in the earth sciences…" (Journel & Huijbregts, 1978 : 27).

Consider, for example, a gold-ore deposit. The grade of the gold-ore can be thought of as a regionalized variable, since values of the grade are distributed in space and display both a structured aspect (areas of rich and poor grade) and a random aspect (local variability in grade measurements).

"From a mathematical point of view, a … [regionalized variable] is simply a function $f(x)$ which takes a value at every point x [within a studied spatial region]…" (Journel & Huijbregts, 1978 : 27). So, a random variable can be seen as varying randomly according to some probability distribution, while a regionalized variable can be seen as a variable varying regionally (i.e. according to its spatial location).

4.2 The Probabilistic Model of Geostatistics

Geostatistics utilizes a probabilistic model which is based on random functions and regionalized variables (Journel & Huijbregts, 1978 : 29 and Wackernagel,

WitsETD

1995 : 25). A random function is merely a set of random variables and is also commonly referred to as a stochastic process or random field (Cressie, 1991 : 30 and Armstrong, 1998 : 17). Herewith follows a description of the probabilistic model of geostatistics.

Consider a spatial attribute (such as gold-ore grade) defined over some spatial region Ω, and suppose further that n values of this attribute have been measured at various locations within Ω. In order to perform any geostatistical calculations on the attribute, the probabilistic model of geostatistics assumes the existence of the random function (i.e. the set of random variables)

$$\{Z(x), x \in \Omega\} \tag{4.1}$$

where $Z(x)$ is the random variable defining the value of the spatial attribute at spatial location x, and Ω is a fixed region whose spatial index x varies continuously throughout Ω (Cressie, 1991 : 8, 30).

Note 4.1 In geostatistics, the random function in Equation 4.1 is often simply referred to as the random function $Z(x)$. Thus, $Z(x)$ can refer to either a random function as in Equation 4.1, or it can refer to a random variable situated at spatial location x. It is generally always understood which meaning of $Z(x)$ is intended, and throughout this dissertation it will be made clear which is intended.

Note 4.2 It is generally accepted in geostatistics that $\Omega \subseteq \mathbb{R}^2$ or $\Omega \subseteq \mathbb{R}^3$ (i.e. the studied spatial region is generally of either two- or three-dimensions). Thus, although spatial locations in this dissertation will generally be referred to simply by a single letter, say x (as has been done thus far in Chapter 4), this notation actually implies a spatial location vector in two- or three-dimensions. In other words, x will either imply the spatial location (x_1, x_2) in two-dimensions, or the spatial location (x_1, x_2, x_3) in three-dimensions.

WiSETD

A realization of the random function $Z(x)$ in Equation 4.1 is classified as the regionalized variable $z(x)$, and therefore the available spatial data set

$$\{z(x_i) : i = 1, ..., n\} \tag{4.2}$$

is simply a number of observations selected from the regionalized variable $z(x)$ (Journel & Huijbregts, 1978 : 30 and Wackernagel, 1995 : 26-27). Cressie (1991: 8-11, 29-30) provides an excellent explanation of this probabilistic model.

Note 4.3 As with Note 4.1, the notation $z(x)$ can refer to both a realization of the random function $Z(x)$ (i.e. a regionalized variable), or a realization of the random variable $Z(x)$ defining an attribute value at spatial location x. This dissertation will again insure that it is clear as to which meaning of $z(x)$ is intended.

Note 4.4 Journel & Huijbregts (1978 : 29-30) point out that the random function construct in Equation 4.1 is ideally suited to the definition of regionalized variables, since the value at each location within the spatial area is theoretically the outcome of a random variable, which need not be independent of the random variables at the other locations.

Since geostatistical data is viewed as a partial realization of a random function, it is not possible to deduce the probabilistic laws central to this random function solely from the available spatial data. Journel & Huijbregts (1978 : 30) state:

> Obviously, it is not rigorously possible to infer the probability law of a … [random function] $Z(x)$ from a single realization $z(x)$ which is, in addition, limited to a finite number of sample points x_i.

Therefore, in order to proceed with the probabilistic model of geostatistics, certain assumptions need to be made. Some of the most important of these assumptions are discussed in Section 4.3.

WitsETD

4.3 Assumptions

The following assumptions permit estimation of the probabilistic laws governing the random function $Z(x)$ from a single partial realization.

4.3.1 Second-order stationarity

The random function defined in Equation 4.1 is deemed second-order stationary if:

$$\mathrm{E}\{Z(x)\} = m, \qquad \forall x \in \Omega \tag{4.3}$$

and

$$\mathrm{cov}(Z(x+h), Z(x)) = C(h), \qquad \forall x, x+h \in \Omega \tag{4.4}$$

This assumption implies that the first two moments of all the random variables contained in the random function exist and are constant, and that the covariance between any two random variables depends only on the distance vector between them and not on their actual positions in Ω (Journel & Huijbregts, 1978 : 32 and Armstrong, 1998 : 18).

In fact, according to Journel & Huijbregts (1978 : 33), in many situations in geostatistics the assumption of second-order stationarity can be weakened such that a covariance between the random variables as in Equation 4.4 need not even exist at all. This weaker assumption is called intrinsic stationarity.

4.3.2 Intrinsic stationarity

The random function defined in Equation 4.1 is said to be intrinsically stationary if:

$$\mathrm{E}(Z(x+h) - Z(x)) = 0 \qquad \forall x, x+h \in \Omega \tag{4.5}$$

and

$$\mathrm{var}(Z(x+h) - Z(x)) = 2\gamma(h), \qquad \forall x, x+h \in \Omega \tag{4.6}$$

According to Armstrong (1998 : 19) the intrinsic assumption simply assumes that the increments of the random variables from the random function $Z(x)$ are

WitsETD

second-order stationary. Intrinsic stationarity assumes that the variability of the difference between two random variables from $Z(x)$ depends only on the distance vector between them, and not on their actual positions within Ω.

Note 4.5 A second-order stationary random function is also intrinsically stationary, but the converse is not necessarily true (Journel & Huijbregts, 1978 : 33).

Note 4.6 The previous two assumptions are commonly associated with linear stationary geostatistics (see Section 1.4). Some authors, such as Rendu (1981), express Equation 4.5 more generally as

$$\mathrm{E}\left(Z\left(x+h\right)-Z\left(x\right)\right)=m\left(h\right) \tag{4.7}$$

where $m(h)$ is a linear drift function,

which "…implies the absence of drift, or the presence of a linear drift" (Rendu, 1981 : 17). Although in such cases the definition of intrinsic stationarity is slightly different to that given in this dissertation, these authors all concede that when applying the assumption of intrinsic stationarity to linear stationary geostatistics, $m(h)$ in Equation 4.7 must be equal to zero.

Note 4.7 The functions $C(h)$ in Equation 4.4 and $2\gamma(h)$ in Equation 4.6 are commonly referred to as the covariance function and variogram respectively. The function $\gamma(h)$ (i.e. half of the variogram) is called the semi-variogram[†].

4.3.3 Ergodicity: An assumption that is rarely stated is that of ergodicity. Ergodic random functions are second-order stationary (Cressie, 1991 : 53), and the assumption of ergodicity in geostatistics is discussed in Matheron (1989 : 81-

[†] Certain texts carelessly define $\gamma(h)$ as the variogram, while other inappropriately use the term variogram and semi-variogram interchangeably.

WitsETD

83), Cressie (1991 : 53-58) and Chilès & Delfiner (1999 : 19-22). The discussion in Chilès & Delfiner (1999) is particularly good.

According to Chilès & Delfiner (1999 : 19-20), a spatial random function $Z(x)$ is said to be ergodic in the mean if, for all realizations $z(x)$,

$$\lim_{V \to \infty} \frac{1}{|V|} \int_V z(x)dx = m$$

where V denotes the spatial domain over which $Z(x)$ is defined, $|V|$ denotes the volume of V, V is assumed to tend towards infinity in all directions, and m is a constant.

In other words, ergodicity in the mean "…allows the determination of the mean from a single realization of the stationary random function…" (Chilès & Delfiner, 1999 : 20).

Ergodicity in the covariance can also be defined (see Chilès & Delfiner, 1999 : 21), as can ergodicity in any other moment of the random function. Thus the assumption of ergodicity in spatial statistics implies that the moments of all realizations $z(x)$ of a random function $Z(x)$ converge to the moments of $Z(x)$ as the spatial area in which $z(x)$ occurs tends towards infinity (Chilès & Delfiner, 1999: 19-22 and Chilès, J.-P. 2004, pers. comm., 21 April). However, in geostatistics, the spatial region studied does not tend towards infinity[†], and only one realization of $Z(x)$ is considered – thus , "…there is no way of knowing if the … [moments] would have converged to a different value on another realization" (Chilès & Delfiner, 1999 : 21).

Therefore, for the probabilistic model of geostatistics, ergodicity is simply assumed as a property of the random function in Equation 4.1. In fact, according to Cressie (1991 : 58), the assumption of ergodicity in geostatistics can be

[†] According to Chilès & Delfiner (1999 : 22), Matheron (1978) introduced the concept of micro-ergodicity to deal with cases of spatial domains that do not tend towards infinity.

WitsETD

weakened to simply assume that the mean and covariance (or semi-variogram) of the single realization $z(x)$ converges to the true mean and covariance (or semi-variogram) of the random function $Z(x)$.

4.4 Alternative Approaches to Spatial Estimation

Up to now the probabilistic model of geostatistics and the assumptions which allow for geostatistical estimation have been discussed. Various other approaches to spatial estimation also exist, each with their own model and assumptions. Besides geostatistics, three of the more common spatial estimation techniques include the inverse distance weighting method, the method of least squares regression and the polygonal method.

4.4.1 Inverse distance weighting

This is an uncomplicated and appealing technique that assumes that "…the relationship between values depends [only] on the distance between their locations [and sometimes, also the direction]" (Clark and Harper, 2000 : 190). Suppose that the value of a spatial attribute has been measured at n spatial locations, resulting in the spatial data set

$$\{z(x_i) : i = 1, ..., n\}.$$

If the unknown value of the attribute at spatial location x_0 is estimated by Q, the inverse distance weighting technique assumes that

$$Q = \frac{\sum_{i=1}^{n} w_i z(x_i)}{\sum_{i=1}^{n} w_i} \tag{4.8}$$

where the weights w_i are based on some inverse function of the distance d_i between the locations x_i and x_0 (Isaaks & Srivastava, 1989 : 257-258).

According to Clark & Harper (2000 : 190), examples of weight functions regularly used are $1/d_i$, $1/d_i^2$ and e^{-d_i}.

WitsETD

To illustrate the inverse distance weighting method, consider the following example taken from Isaaks & Srivastava (1989 : 257):

Example 4.1

Suppose that the value of seven point support samples has been measured and that an estimate of the value $z(x_0)$ at the location (65E, 137N) is desired (see Figure 4.1). Suppose further that it is assumed that the relationship between sample values depends only on the inverse of their Euclidean distance d_i apart. Table 4.1 displays the required calculations for estimating the value $z(x_0)$ using the method of inverse distance weighting.

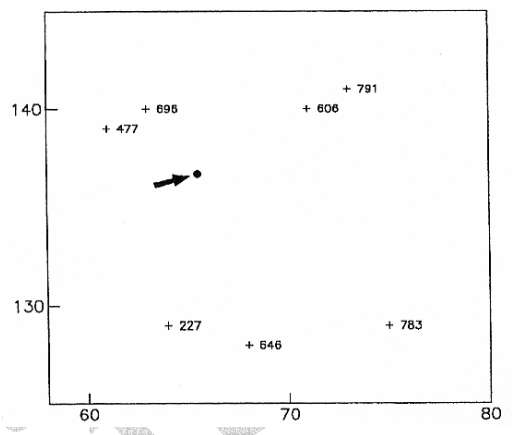

Figure 4.1 The location of the seven measured point support samples, as well as the location of the point to be estimated (which is indicated by the arrow). Figure 4.1 has been taken directly from Isaaks & Srivastava (1989 : 251).

Using the calculations displayed in Table 4.1 together with Equation 4.8, the inverse distance weighting estimate of the value $z(x_0)$ is $Q = 594$.

Regrettably, this straightforward method possesses many shortcomings which are highlighted by a large amount of geostatistical literature (e.g. Clark and Harper 2000, Journel & Huijbrechts 1978, Wackernagel 1995, and Wellmer 1998). According to these authors, inverse distance weighting methods:

Table 4.1 Inverse distance weighting calculations for estimating the point $z(x_0)$, using the seven available sample values in Figure 4.1

	***X* co-ordinate** (measured in units in easterly direction)	***Y* co-ordinate** (measured in units in northerly direction)	**Value of sample**	**Euclidean distance d_i between sample and $z(x_0)$**	**Inverse distance weighting w_i associated with sample $(w_i = 1/d_i)$**
$z(x_0)$	65	137	---	---	---
Sample 1	61	139	477	4.5	0.22
Sample 2	63	140	696	3.6	0.28
Sample 3	64	129	227	8.1	0.12
Sample 4	68	128	646	9.5	0.11
Sample 5	71	140	606	6.7	0.15
Sample 6	73	141	791	8.9	0.11
Sample 7	75	128	783	13.5	0.07
					$\Sigma w_i = 1.06$

- Fail to define measures of dealing with samples that are not all of the same support (consider the implications by reviewing Section 2.4.1 and Note 3.1).

- Often break down in trying to provide estimates at the location of samples used in the calculation. This problem is highlighted by the question of Clark & Harper (2000 : 192): "What happens when d becomes zero? None of the $1/d$ type functions can be calculated if d is zero."

- May supply biased solutions.

These are but a few of the many problems associated with the method of inverse distance weighting – most of which are not problems for geostatistical estimation.

4.4.2 Least squares regression

Suppose the value of a spatial attribute has been measured at n locations. The theory of least squares regression can be used to produce a prediction surface over the study area. A key assumption of this technique is the independence of residuals, which is unfortunately not generally the case for geological variables (Armstrong, 1998 : 16). Fitting a surface which does result in uncorrelated residuals is both tedious and results in complex regression equations (see Note 4.8). Unlike kriging, least squares regression is not an exact interpolator (Koch & Link, 1971 : 43-44) and, according to Journel & Huijbregts (1978 : 343), least squares regression estimates are not as good (in terms of minimum estimation variance) as kriging estimates.

Note 4.8 According to Armstrong (1998), Gomez & Hazen (1970) fitted a regression equation to describe the proportion of pyritic sulphur in a certain type of coal, which resulted in uncorrelated residuals. This regression equation is copied straight from Armstrong (1998 : 17) and displayed in Table 4.2.

Armstrong (1998 : 17) writes: "This …[equation in Table 4.2] assumes uncorrelated errors, which forces the … [prediction] surface to twist and turn

WiSETD

rapidly, hence the trigonometric and exponential terms." The difficulties associated with using such a complicated equation (e.g., the estimation of the equation parameters, the concept of parsimony, etc.) are obvious.

Table 4.2 The regression equation fitted to describe the proportion of pyritic sulphur in a certain type of coal (taken from Armstrong, 1998 : 17).

Equation A−1 (Pyritic sulfur, coarse coal) =

$$14.0548 - 5.97910 \times AS \div 1.35753 \times SU + 1.34232 \times SU \times (AS - SU) - 0.419448 \times AS \times SU \times (AS - SU) + 4.95307 \times 10^{-3} [AS \times SU \times (AS - SU)]^2 - 2.69728 \times 10^{-4} [AS \times SU \times (AS - SU)]^3 + 5.88963 \times 10^{-11} \times e^{(AS \times SU)} - 1.38995 \times 10^{-6} [e^{(AS \times (AS - SU))}] + 0.0103637 [e^{[SU \times (AS - SU)]}] - 16.70984 [AS \times SU / e^{(AS \times SU)}] + 5.67080 \times 10^{-3} [AS \times (AS - SU) / e^{[AS \times (AS - SU)]}] - 1.96079 \times 10^{-5} [SU \times (AS - SU) / e^{[SU \times (AS - SU)]}] \div 0.104688 \times \sin(AS \times SU)^3 \times \cos(AS \times SU)^2 - 9.53418 \times 10^{-3} \times sin[AS \times (AS - SU)]^3 \times cos[AS \times (AS - SU)]^2 - 0.0848224 \times sin[SU \times (AS - SU)]^3 \times cos[SU \times (AS - SU)]^2 - 2.08676 \times sin(AS) - 2.19124 \times sin(SU) + 1.49662 \times sin(AS - SU) - 0.107983 \times e^{AS} - 0.0504733 \times e^{SU} + 0.312656 \times {}^{(AS - SU)} - 0.0491919 \times \sin(AS^3) \times \cos(AS^2) + 0.0434771 \times sin(SU^3) \times \cos(SU^2) - 1.81229 \times 10^{-4} \times sin(AS - SU)^3 \times cos(AS - SU)^2 - 2.01703 \times 10^{-4} \times e^{(AS + SU)} \times e^{(AS - SU)}$$

Variables used in equation.

AS	= Ash in coal, percent
SU	= Sulfur in coal, percent

4.4.3 Polygonal estimation

Polygonal estimation is nothing more than "…choos[ing] as an estimate the sample value that is closest to the point [or block being estimated]..." (Isaaks & Srivastava, 1989 : 250). In the case where two or more samples are equally close to the point or block being estimated, the arithmetic average of these sample values is used as the estimate.

Example 4.2

Consider the estimation of $z(x_0)$, the point located at (65E, 137N) in Example 4.1. In this example, $z(x_0)$ was estimated to be equal to 594. If polygonal estimation was used instead to estimate the value of $z(x_0)$, it is seen in Table 4.1 that sample

two is located the closest to (65E, 137N), and hence $z(x_0)$ is estimated as 696 (see Isaaks & Srivastava, 1989 : 251).

The benefit of polygonal estimation is that it is straight-forward to apply. Unfortunately, this estimation technique also possesses numerous shortcomings – some of which are mentioned in Isaaks & Srivastava (1989 : 251-252), Goovaerts (1997 : 356) and Armstrong (1998 : 9). These shortcomings include the fact that

- polygonal estimation does not take into account the spatial structure and continuity of the attribute being studied and estimated.

- polygonal estimation does not take into account the support effect (see Section 2.4.1 and Note 3.1). Suppose, for instance, that point support samples are used to estimate a block support area/volume via the polygonal estimation method. Armstrong (1998 : 9), describing this scenario in a mining context, rightly warns that,

  > ...the polygonal method equates the grades of the samples (i.e. a small support) with those of the blocks … [and hence does not] … take account of the difference between the supports of the samples and the blocks to be estimated; that is, of the support effect.

- the prediction surface of the polygonal estimation method is discontinuous.

Note 4.9 Polygonal estimation is also known by the names Thiessen polygons and Voronoi polygons (Olea, 1990).

Note 4.10 Polygonal estimation can be seen as an extreme case of inverse distance weighting, where the relationship between samples is assumed to be adequately explained by $1/d^p$ where p is assumed to be extremely large (in fact, tending towards infinity).

WITSETD

4.4.4 Additional comments

Although geostatistics would appear superior to inverse distance weighting, least squares regression and the polygonal method (since geostatistics is not affected by the various shortcoming listed for each of the methods discussed in Section 4.4), these three methods still retain important roles in spatial estimation. Not only are the mathematical calculations of these methods undemanding, but regression, for instance, also plays a significant role in identifying and eliminating trend from non-stationary data (see Section 7.12.2).

Note 4.11 An experiment using a simulated spatial data set is presented in Section 8.3, and in this experiment the estimation results of kriging, inverse distance weighting and polygonal estimation are compared.

WitsETD