

# Geostatistics without Stationarity Assumptions within GIS

Alexander Brenning\*

K. Gerald van den Boogaart<sup>†</sup>

## Abstract

The present work deals with two challenging problems of applied geostatistics: (i) Stationarity assumptions often do not hold under real-world conditions. (ii) Geostatistical methods have to be linked with spatial databases in order to be applicable in non-stationary situations. Solutions for both problems are proposed and implemented.

(i) A central assumption in geostatistics is the stationarity of the process. However the spatial variability of many natural phenomena heavily depends on the local geology, which is non-stationary in most cases. To deal with this, the concept of process stationarity is replaced by a stationarity of the governing influence relating the local semivariogram and the local geology as stored in a Geographical Information System (GIS). A construction method is used, which can meaningfully incorporate additional spatial information from GIS, e.g. smoothly varying geology in the investigated area or geological faults interrupting continuity. Least-squares parameter estimation is used for fitting instationary semivariogram models in typical example situations, leading to non-linear optimization problems.

(ii) Geostatistical tools that make use of the local geology need direct access to the data stored in the GIS. A link between the presented geostatistical tools and the GIS software ArcView was established. Thus, spatial data such as soil properties and morphology can be incorporated in geostatistical analyses.

R code that fits instationary semivariogram models and performs kriging is provided. It is applied to simulated data.

## Resumen

Este trabajo trata dos problemas desafiantes de la geoestadística aplicada: (i) Bajo condiciones reales, muchas veces no son válidas las suposiciones de estacionaridad. (ii) Los métodos geoestadísticos deben ser vinculados con bases de datos espaciales para ser aplicables en situaciones no estacionarias. Se proponen e implementan soluciones para ambos problemas.

(i) En geoestadística, una suposición central es la estacionaridad del proceso. Sin embargo, la variabilidad espacial de muchos procesos depende estrechamente de la geología local, la cual es inestacionaria en la mayoría de los casos. Para tratar esto, el concepto de estacionaridad del proceso es reemplazado por una estacionaridad de una ley de influencia; ésta vincula el semivariograma con la geología local tal como está guardada dentro de un Sistema de Información Geográfico. Se usa un método de construcción que permite incorporar información espacial de un SIG, por ejemplo una geología que varía de manera continua dentro del área de estudio o fallas geológicas que interrumpen la continuidad. Se aplica una estimación de mínimos cuadrados para ajustar modelos paramétricos de semivariogramas inestacionarios en situaciones ejemplares típicas, resultando en problemas de optimización no lineales.

(ii) Herramientas geoestadísticas que se apoyan en la geología local precisan de un acceso directo a los datos guardados en el SIG. Se estableció un vínculo entre la herramienta geoestadística presentada y el SIG ArcView. De esta manera pueden incorporarse a los análisis geoestadísticos datos espaciales tales como propiedades del suelo y la morfología.

Se facilita un código de R que ajusta modelos de semivariogramas inestacionarios y realiza kriging. El código es aplicado a datos simulados.

---

\*Friedrich-Alexander-Universität Erlangen-Nürnberg, Institut für Geographie, Kochstr. 4/4, 91054 Erlangen, Germany, e-mail [ali@proforma.de](mailto:ali@proforma.de).

<sup>†</sup>TU Freiberg, Institut für Geologie, B.-v.-Cotta-Str. 2, 09596 Freiberg, Germany, e-mail [boogaart@iname.com](mailto:boogaart@iname.com).

# 1 Introduction

The amount of spatial information that is available within GIS is increasing rapidly due to the collection of thematic data using remote sensing techniques. These data are often useful for modeling local (non-geometrical) anisotropies following morphology or tectonic structures, e. g. This development creates a need for instationary geostatistical methods that meaningfully incorporate this information, enabling us to model complex geologic situations such as tectonic faults interrupting continuity or drainage systems inducing nested patterns of anisotropy. These methods must be linked with a GIS.

## 2 Theoretical Background

We consider a real-valued stochastic processes  $Z = (Z_s)_{s \in D}$  of second order with parameter set  $D \subset \mathbb{R}^d$ ,  $d \geq 2$ .

### 2.1 A Method of Instationary Covariogram Construction

We study processes with covariograms that are induced by a weight function. We will see that these induced covariograms approximate stationary covariograms arbitrarily well. Of particular interest will be weight functions that incorporate covariables describing local conditions such as anisotropy. Let  $E$  be a non-empty measurable Borel set in  $\mathbb{R}^d$ .

**Definition 2.1** A *weight function* on  $D \times E$  is an arbitrary function  $w : D \times E \rightarrow \mathbb{R}$  such that for all  $s \in D$

$$\nu_w(s) := \int_E w(s, p)^2 dp < \infty.$$

A weight function on  $D \times E$  is called *isotropic*, if there exists a function  $w_i : \mathbb{R} \rightarrow \mathbb{R}$  such that  $w(s, p) = w_i(\|p - s\|)$  for all  $s \in D$ ,  $p \in E$ .

**Theorem 2.2** For an arbitrary weight function  $w$  on  $D \times E$  and all  $s, t \in D$  there exists the integral

$$C_w(s, t) := \int_E w(s, p)w(t, p) dp, \tag{1}$$

and the function  $C_w$  is positive semidefinite. Furthermore,  $C_w$  is the covariogram of a second-order process on  $D$ ; it is called the *covariogram* induced by  $w$ .

*Proof:* The integral in (1) exists and is finite. Furthermore, if for  $n \in \mathbb{N}$ , we choose arbitrary  $t_1, \dots, t_n \in D$  and  $a_1, \dots, a_n \in \mathbb{R}$ , then we obtain

$$\begin{aligned} \sum_{i=1}^n \sum_{j=1}^n a_i a_j C_w(t_i, t_j) &= \sum_{i=1}^n \sum_{j=1}^n a_i a_j \int_E w(t_i, p)w(t_j, p) dp \\ &= \int_E \sum_{i=1}^n a_i w(t_i, p) \sum_{j=1}^n w(t_j, p) dp = \int_E \left( \sum_{i=1}^n a_i w(t_i, p) \right)^2 dp \geq 0. \end{aligned}$$

Hence  $C_w$  is positive semidefinite.  $C_w$  is also symmetric, so there exists a second-order process on  $D$  with covariogram  $C_w$ .  $\square$

**Remark 2.3** The semivariogram  $\gamma$  corresponding to a covariogram  $C$  that is induced by an arbitrary weight function  $w$  is of the form

$$\begin{aligned} \gamma(s, t) &= \frac{1}{2}C(s, s) + \frac{1}{2}C(t, t) - C(s, t) \\ &= \frac{1}{2} \int_{\mathbb{R}^d} (w(s, p)^2 + w(t, p)^2 - 2w(s, p)w(t, p)) dp \\ &= \frac{1}{2} \int_{\mathbb{R}^d} (w(s, p) - w(t, p))^2 dp. \end{aligned} \tag{2}$$

**Theorem 2.4 (Approximation of stationary covariograms)** *Every stationary covariogram  $C$  on  $\mathbb{R}^d$  that has a spectral density  $g = dG/d\lambda$  can be approximated arbitrarily well with respect to the essential supremum  $\|\cdot\|_\infty$  by an induced covariogram, i. e.: For every  $\varepsilon > 0$  there exists a translation invariant weight function  $w(s, t) = w_s(t - s)$  that induces a covariogram  $C_w$  with*

$$\|C - C_w\|_\infty < \varepsilon.$$

*Proof:* Cf. van den Boogaart (1999).

**Remark 2.5** Assuming the existence of a spectral density in Theorem 2.4 implies that neither a nugget effect nor a covariogram not vanishing as  $\|h\| \rightarrow \infty$  can be approximated arbitrarily well by weight functions. However a nugget effect can be added a posteriori to an induced covariogram.

For motivation we now present a well-known example of an induced covariogram that is isotropic (cf. Wackernagel (1998)). Later it will be used as a point of departure for modeling anisotropies.

**Example 2.6 (Spherical covariograms)** For fixed  $a = 2R > 0$  and  $E = \mathbb{R}^d \supset D$ , consider the isotropic weight function

$$w : D \times \mathbb{R}^d \rightarrow \mathbb{R}, \quad w(s, p) = \mathbf{1}_{[0, R]}(\|p - s\|), \quad s \in D, p \in \mathbb{R}^d.$$

and its normalization scaled by  $\sigma > 0$ ,

$$\tilde{w}(s, p) = \sigma \cdot w(s, p) / \sqrt{\nu_w(s)} = \sigma \cdot \mathbf{1}_{[0, R]}(\|p - s\|) / \lambda^d(B^d(0, R)).$$

We study the covariograms induced by  $\tilde{w}$  for  $d = 3, 2$ , which writes

$$C_{\tilde{w}}(s, t) = \sigma^2 \cdot \lambda^d(B^d(s, R) \cap B^d(t, R)) / \lambda^d(0, R).$$

i)  $d = 3$ : We get the spherical covariogram in three dimensions. The usual two-dimensional version is obtained by taking  $D' = D \times \{0\}$ ,  $D \subset \mathbb{R}^2$ .

ii)  $d = 2$ : We have to calculate the area of dissection of two discs of radius  $R$  in  $\mathbb{R}^2$ . Geometrical considerations yield

$$C_{\tilde{w}}(h) = \frac{\sigma^2}{\pi} \left( 2 \arccos(h/a) - \frac{h}{4a^2} \sqrt{a^2 - h^2} \right), \quad \text{if } h < a,$$

0 otherwise. The isotropic covariogram  $C_{\tilde{w}}$  is continuous, but its first derivative has a singularity at  $h = 2R$ , unlike the spherical covariogram.

## 2.2 Modeling Local Anisotropy

Theorem 2.4 showed that the class of covariograms induced by a weight function is sufficiently large as to be useful instruments for covariogram modeling. The construction method will be used for creating a class of semivariograms and covariograms that can adapt to local anisotropies that may be quite irregular, but follow a known pattern.

From now on we choose  $d = 2$  and  $E \subset \mathbb{R}^2$ .

**Definition 2.7 (Elliptical covariograms)** Let  $w_\tau^\circ : \mathbb{R} \rightarrow \mathbb{R}$ ,  $\tau \in T \neq \emptyset$ , be an isotropic weight function with support  $\subset B^2(0, 1)$ . For  $r \geq 0$ ,  $q \in (0, 1]$  and  $\phi \in [0, \pi)$  we define

$$R(r, q, \phi) = \frac{1}{r} \begin{pmatrix} \cos \phi & \sin \phi \\ -q^{-1} \sin \phi & q^{-1} \cos \phi \end{pmatrix}$$

to be a combined contraction and rotation by  $-\phi$  satisfying

$$R(r, q, \phi) \text{Ell}(0; r, q, \phi) = B^2(0, 1),$$

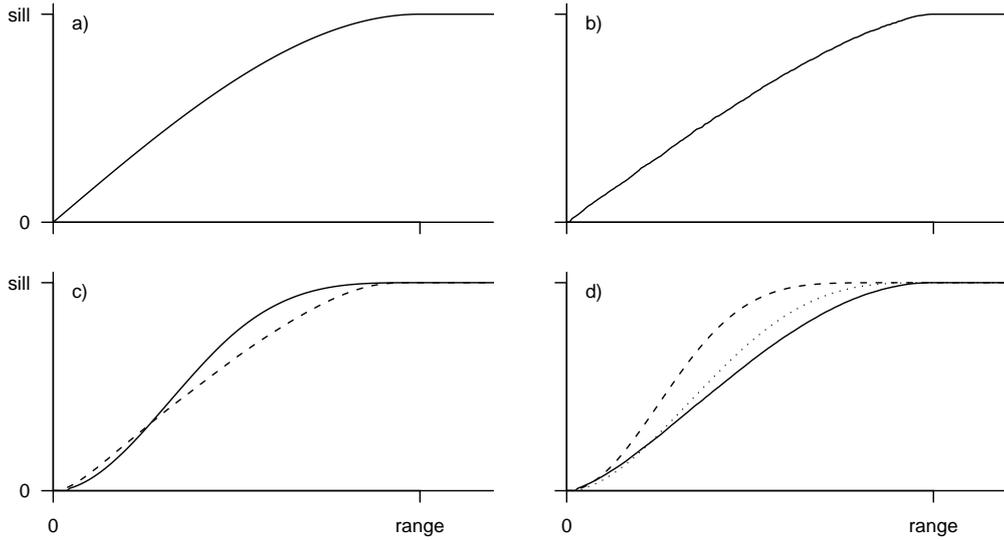


Figure 1: Semivariogram plots: a) Spherical semivariogram (Example 2.6 i)). b) Simple elliptical semivariogram (Remark 2.9, Example 2.6 ii)). c) Elliptical semivariograms corresponding to the piecewise linear kernel function with break points at  $b = 0$  (solid line) and  $b = 0.8$  (dashed). d) Elliptical semivariograms corresponding to the Bezier kernel function with exponents  $\nu = 0.3$  (solid line), 1 (dotted) and 3 (dashed).

where  $\text{Ell}(0; r, q, \phi)$  denotes the open two-dimensional ellipse around 0 with longer radius  $r$  in an angle of  $\phi$  with the  $x_1$ -axis and axis ratio  $q$ .

Now consider a function  $\theta = (\sigma^2, a, q, \phi, \tau) : D \rightarrow \mathbb{R}_0^+ \times \mathbb{R}^+ \times (0, 1] \times [0, \pi) \times T$ . Then we define

$$\begin{aligned} w_\theta^* : D \times \mathbb{R}^d &\rightarrow \mathbb{R}, & w_\theta^*(s, p) &= \mathbf{1}_E(p) w_{\tau(s)}^o(\|R(\frac{a(s)}{2}, q(s), \phi(s)) \cdot (p - s)\|), \\ w_\theta^{\text{ell}} : D \times E &\rightarrow \mathbb{R}, & w_\theta^{\text{ell}}(s, p) &= \sigma w_\theta^*(s, p) \cdot \nu_\theta^*(s)^{(-1/2)}. \end{aligned}$$

Then the weight function  $w_\theta^{\text{ell}}$  is called an *elliptical weight function* on  $D$ . Covariograms and semivariograms induced by elliptical weight functions are also said to be *elliptical*.  $w_\tau^o$  will be referred to as a *kernel function*. A component of  $\theta$  is called a *parameter*, if it is a constant function, otherwise a *covariable*.

**Remark 2.8** In this work, the elliptical semivariograms considered have parameters  $\sigma^2$ ,  $a$ ,  $q$  and  $\tau$ , and one single covariable  $\phi$ , unless specified otherwise.

**Remark 2.9** Kernel functions for elliptical semivariograms are for example

$$\begin{aligned} w^{\text{ind}}(h) &= \mathbf{1}_{[-1,1]}(h), && \text{("simple kernel function")} \\ w_b^{\text{pwl}}(h) &= \begin{cases} 1 & \text{if } |h| \leq b, \\ 1 - (|h| - b)/(1 - b) & \text{if } b < |h| < 1, \\ 0 & \text{otherwise,} \end{cases} && \text{("piecewise linear kernel fn.")} \\ w_\nu^{\text{bez}}(h) &= \begin{cases} (h + 1)^\nu (h - 1)^\nu & \text{if } |h| < 1, \\ 0 & \text{otherwise,} \end{cases} && \text{("Bezier kernel fn.")} \end{aligned}$$

Simple elliptical covariograms (induced by the simple kernel function) with  $q = 1$  are isotropic and coincide with the covariogram  $C_{\tilde{w}}$  in Example 2.6 ii).

Even the simple area of dissection of these ellipses is difficult to determine analytically. In this work quasi-Monte Carlo integration methods will be applied for approximating elliptical semivariograms. The covariograms and semivariograms induced by the Bezier kernel function (see figure 1) are  $\lfloor 2\nu \rfloor$  times continuously differentiable because  $w_0^{\text{bez}}$  is (exactly)  $\lfloor \nu \rfloor$  times continuously differentiable. The other kernel functions presented are not differentiable on  $\partial B^2(0, 1)$ .

The following example shows how the class of elliptical semivariograms can be extended, and how parameters and covariables can be chosen in order to represent qualitative knowledge of the geologic processes to model.

**Example 2.10 (Modeling soil loss by water run-off)** Soil erosion by water run-off is a geomorphological process that basically depends on morphology, vegetation, land use, soil properties and precipitation regime. At a small scale, however, the size of the catchment area and the slope's inclination are the most important factors that influence soil erosion. The soil loss at one point depends on the soil loss uphill in the same catchment area, and little correlation with soil loss on the other side of a ridge or on the opposite side of a valley can be assumed.

For a point  $s \in D$ , let  $A(s) \subset \mathbb{R}^2$  denote its catchment area. Then a weight function  $w$  on  $D \times \mathbb{R}^d$  with  $\text{supp } w(s, \cdot) = A(s)$  incorporates our knowledge of the relation between soil erosion and morphology.

i) As an example, we define a weight function by (the normalization of)

$$w_{\tilde{\theta}}(s, p) = w_{\theta}^{\text{ell}}(s, p) \mathbf{1}_{A(s)}(p), \quad \tilde{\theta} = (\theta, \mathbf{1}_{A(\cdot)}),$$

where  $w_{\theta}^{\text{ell}}$  is an arbitrary elliptical weight function with covariable  $a : s \mapsto 2 \sup_{p \in A(s)} \|p - s\|$  and parameters  $\sigma^2$  and  $\tau$ , the axis ratio  $q = 1$  being constant and hence  $\phi$  without any effect.

The catchment area of a point can be determined by digital terrain modeling, and a large number of expensive GIS calls to evaluate  $\mathbf{1}_{A(s)}(p)$  is necessary to approximate the induced semivariogram by numerical integration.

ii) A computationally less demanding method is the following, which approximates  $w_{\tilde{\theta}}$  quite well in sufficiently regular relief. Let  $R(s)$  be the shortest distance to the ridge that lies uphill from  $s$ , and let  $\phi(s) \in [0, 2\pi)$  denote the gradient of topography at  $s$  expressed as an angle, and  $\delta(s) \in (0, \pi]$  an opening angle. Then for an arbitrary elliptical weight function  $w_{\theta}^{\text{ell}}$  with  $q = 1$  and covariables  $a = 2R$  and  $\phi$ , we take (the normalization of)

$$w'_{\tilde{\theta}}(s, p) = w_{\theta}^{\text{ell}}(s, p) \mathbf{1}_{\text{Sec}(s; \phi(s), \delta(s), a(s)/2)}(p),$$

where  $\text{Sec}(s; a(s)/2, \phi(s), \delta(s)) \subset B^2(s, a(s)/2)$  denotes the disc sector with opening angle  $\delta(s)$  and radius  $a(s)/2$  oriented according to  $\phi(s)$ . The opening angle  $\delta(s)$  could for example be constant or a function of local curvature at  $s$ .

## 2.3 Modeling Boundaries between Subprocesses

In many geostatistical applications we find the following situation: The parameter set  $D$  of the process  $Z$  of interest decomposes into  $\eta$  disjoint subsets  $D_1, \dots, D_{\eta} \subset D$  such that the subprocesses  $Z_1 := Z|_{D_1}, \dots, Z_{\eta} := Z|_{D_{\eta}}$  have little or no correlation between each other. Instead of studying each subprocess separately, one might wish to study the process  $Z$  as a whole.

Let  $w_1^o, \dots, w_{\eta}^o$  be weight functions on  $D \times \mathbb{R}^2$ . We consider the following situations ( $i = 1, \dots, \eta$ ):

$$\begin{aligned} \text{"strong boundaries":} & \quad w_i^*(s, p) := w_i^o(s, p) \mathbf{1}_{D_i}(s) \mathbf{1}_{D_i}(p), & C^* &= \sum_k C_{w_k^*}, \\ \text{"ordinary boundaries":} & \quad w_i(s, p) := w_i^o(s, p) \mathbf{1}_{D_i}(s), & C &= \sum_k C_{w_k}, \\ \text{"soft boundaries":} & \quad w_i^{\#}(s, p) := \sum_k w_i^o(s, p) \mathbf{1}_{D_i}(s), & C^{\#} &= C_{w^{\#}}. \end{aligned}$$

**Remark 2.11** i) The subprocesses of a process with ordinary or strong boundaries are uncorrelated; those of a process with soft boundaries are generally correlated.

ii) Near the boundaries between  $D_1, \dots, D_{\eta}$ , covariograms with strong boundaries are greater than their analogues with ordinary boundaries (figure 2). This is due to the normalizing factor that decreases rapidly as the integration domain is being cut by  $\mathbf{1}_{D_i}$ .

iii) When approximating covariograms with strong boundaries using numerical integration, the functions  $\mathbf{1}_{D_i}$ ,  $i = 1, \dots, \eta$ , have to be evaluated at each node  $p \in E$ . In practice,  $D_1, \dots, D_{\eta}$  generally are polygons stored within a Geographical Information System, and evaluations of  $\mathbf{1}_{D_i}$  have to be considered as expensive GIS queries.

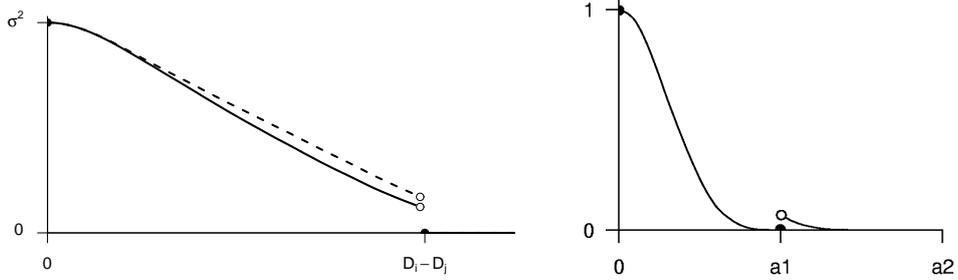


Figure 2: Left: Covariograms with ordinary (solid line) and strong boundaries (dashed) near the boundaries. Right: The covariogram from Example 2.12 with soft boundaries.

**Example 2.12 (Soft boundaries)** Suppose  $D_1 \cup D_2 = D$ ,  $D_1 \cap D_2 = \emptyset$ , and let  $C$  denote a covariogram with soft boundaries between  $D_1$  and  $D_2$  induced by a weight function

$$w_{(1,a_1,1,0)}^{\text{pw1}}(s,p)\mathbf{1}_{D_1}(s) + w_{(1,a_2,1,0)}^{\text{pw1}}(s,p)\mathbf{1}_{D_2}(s)$$

(cf. Definition 2.7, Remark 2.9), where  $a_1 < a_2$  are different range parameters. This may produce a covariogram as shown in figure 2. It will generally not be desirable to introduce a positive correlation  $C(s,t)$  for  $s \in A_1$  and  $\|t-s\| > a_1$ , as seen in the figure.

**Remark 2.13 (Smooth transitions)** The problem encountered in Example 2.12 with soft boundaries may be avoided by allowing the weight function's parameter  $\theta \in \Theta$  to vary smoothly in space, i. e. modeling it as a covariable.

## 2.4 Generic Stationarity: towards Stationarizing Instationarity

By construction the (generally non-geometrical) anisotropy of elliptical covariograms can be considered as rather regular because it is determined by the direction  $\phi(s)$ ,  $s \in D$ , of local anisotropy, which is assumed to be known. Therefore we want to introduce a concept that is more general than that of stationarity and includes the elliptical case and other situations with “understandable” instationarities.

**Remark 2.14** In a more general setting than in the preceding example, we consider processes that are embedded within a geological environment. The covariogram of such a process should be a function of local geology, we write

$$C(s,t) = C_g(s,t;g(\cdot)).$$

We wish to model the function  $C_g$  that maps local geology to the proces' covariogram. The resulting covariogram is no longer stationary, but depends on local geology. However, some degree of stationarity remains, since the physical laws reflected in a model for  $C_g$  are the same at any place. In particular, we assume that we would get the same covariogram in another place, if we had the same geology there:

$$C(s+h,t+h) = C_g(s,t;\Phi(\cdot+h)).$$

**Example 2.15** Let  $Z$  be a process on  $D = \mathbb{R}^2$  with elliptical covariogram  $C$  with parameter  $\theta \in \Theta$  and direction of local anisotropy  $\phi : \mathbb{R}^2 \rightarrow [0, \pi)$ . Then for all  $s, t, h \in \mathbb{R}^2$  with  $\phi(s) = \phi(s+h)$  and  $\phi(t) = \phi(t+h)$  it holds

$$C(s,t) = C(s+h,t+h). \tag{3}$$

That is, it holds “something like” a stationarity property conditional on the direction of local anisotropy  $\phi$ .

However, there will probably exist points  $s+h', t+h' \in \mathbb{R}^2$  with  $\phi(s) \neq \phi(s+h')$  or  $\phi(t) \neq \phi(t+h')$ . If we go beyond the covariogram itself and study the way how its generation depends on the covariable  $\phi$ , we will find out that for all  $s, t \in \mathbb{R}^2$ , we have

$$\forall h \in \mathbb{R}^2 : \quad C(s+h, t+h) = C_g(s, t; \phi(\cdot + h)), \quad (4)$$

where the function  $C_g : \mathbb{R}^2 \times \mathbb{R}^2 \times \{\phi(\cdot + h') : h' \in \mathbb{R}^2\} \rightarrow \mathbb{R}$ , is defined by

$$C_g(s, t; \phi(\cdot + h)) = \frac{\sigma^2}{\nu} \int_{\mathbb{R}^2} w_\theta^g(\|R(\phi(s+h); \theta)(p-s)\|) w_\theta^g(\|R(\phi(t+h); \theta)(p-t)\|) dp \quad (5)$$

with an appropriate normalizer  $\nu$ . Thus,  $C(s+h, t+h)$  only depends on  $s, t$  and the moved geology represented by  $\phi(\cdot + h)$ .

The last two equations suggest that “if we had  $\phi(s) = \phi(s+h')$  and  $\phi(t) = \phi(t+h')$ , then we would get  $C(s, t) = C(s+h', t+h')$ ” —a hypothetical version of (3), based on our belief in the validity of the law expressed in (4) and (5).

The following definition reflects this concept in a precise way.

**Definition 2.16 (Generic stationarity)** Consider a process  $Z$  on  $D = \mathbb{R}^d$  with covariogram  $C$ , semivariogram  $\gamma$  and mean  $m$ , and let  $g : \mathbb{R}^d \rightarrow T$  be a mapping onto an arbitrary set  $T$ . Then we define:

- i) The process  $Z$  is *strongly generically stationary with respect to  $g$* , if there exists a function  $P_g$  such that

$$\forall B \forall h \in \mathbb{R}^d : \quad P(Z_{\cdot+h} \in B) = P_g(B, g(\cdot + h)).$$

Or, more precisely,

$$\forall B \in \mathcal{A} \forall h \in \mathbb{R}^d : \quad P(B_h) = P_g(B; g(\cdot + h))$$

where  $B_h = \{\omega \in \Omega : \exists \tilde{\omega} \in B : Z(\omega) = Z_{\cdot+h}(\tilde{\omega})\}$ ,  $(\Omega, \mathcal{A}, P \in \mathbb{P})$  is a statistical model and  $P_g$  a mapping  $\{g : \mathbb{R}^d \rightarrow T\} \rightarrow \mathbb{P}$ .

- ii)  $\Gamma \in \{C, \gamma\}$  is *generically stationary with respect to  $g$* , if there exists a function  $\Gamma_g$  such that

$$\forall s, t, h \in \mathbb{R}^d : \quad \Gamma(s+h, t+h) = \Gamma_g(s, t; g(\cdot + h)).$$

- iii) If there exists a function  $m_g$  such that

$$\forall s, h \in \mathbb{R}^d : \quad m(s+h) = m_g(s, g(\cdot + h)),$$

then  $m$  is *generically stationary with respect to  $g$* , and  $Z$  is *first-order generically stationary with respect to  $g$* .

- iv)  $Z$  is (*second-order or weakly*) *generically stationary*, if  $m$  and  $C$  are, and the process is *intrinsically generically stationary*, if  $m$  and  $\gamma$  are generically stationary.

$P_g, \Gamma_g$  and  $m_g$  are called *influence laws of generic stationarity*, and  $g$  an *influence function*.

**Remark 2.17** Generic stationarity says that the distribution laws of  $Z_s$  and  $Z_t$ ,  $s, t \in D$ , are or become the same, if local geology around  $s$  and  $t$  are the same or are “forced” to be the same. That is, generic stationarity assumes that there is something like a law of nature that determines the distribution of a random variable given the local geology  $g$ . Depending on how we choose the function  $g$ , generic stationarity becomes a triviality or an instrument that describes how a distribution law is determined by the environment.

**Example 2.18** i) Elliptical semivariograms are stationary with respect to their covariables.  
ii) If  $Z$  is a process with a deterministic trend,  $Z_s = Y_s + \beta^T f(s)$  for all  $s \in D$ , where  $(Y_s)_{s \in D}$  is a zero-mean stationary process, then  $Z$  is generically stationary with respect to  $g = \beta^T f$ .

**Remark 2.19** Consider a generically stationary process  $Z$  with influence function  $g$ .  $g$  can take two “extreme” cases: i)  $g$  is constant. Then  $Z$  is stationary in the usual sense. ii)  $g$  is bijective. Then there is no condition on the distribution of  $Z$ . One could say that a constant influence function gives no information on local geology, while a bijection contains complete knowledge of local geology and hence explains arbitrary spatial variation of the process’ distribution. Between these two extremes, there is a broad variety of meaningful influence laws (see the examples above).

### 3 Geostatistics and GIS

Our goal is to implement instationary geostatistical methods fully within a general-purpose data analysis environment, and to provide direct access to this functionality from an exemplarily chosen GIS. The applicability of the geostatistical routines will however not be restricted to a specific GIS software.

For the implementation of geostatistical methods, the data analysis language and environment R was chosen. R offers a simple and effective programming language that includes conditionals, loops and user defined functions, and there exists a kind of object-oriented design. A great variety of statistical models and methods is also available, among them linear models, clustering and time-series analysis.

The R environment is open source. Open-source software is not only cheaper than commercial alternatives; the main advantage of using it is related to the efficiency of problem-solving and debugging within a community of users. Furthermore, full access to source code makes it possible to find out details of the implementation that are not documented in the manuals (Bivand 1999). R differs very little from the language S and its derivative S-PLUS.

As GIS platform, the commercial software ArcView 3.1 was chosen, because of its wide use in the public and private sector. However, a relatively small amount of code has to be generated in order to create an interface for a different GIS environment.

In total, about 100 KB of R code, 30 KB of C code and 30 KB of AVENUE source code were produced. The code was developed for R 1.2.2, Microsoft Visual C++ 5.0 and ArcView 3.1 and used within a Microsoft Windows 2000 5.0 / NT 4.0 client–server environment on Pentium II family processors at Freiberg University.

Source code and further documentation can be obtained from the authors.

#### 3.1 Approximating Semivariograms with Quasi-Monte Carlo Integration

When we fit a semivariogram model or do kriging, a great number of semivariogram evaluations is needed, which are generally computed by approximating the integral in (1), if a model of the elliptical class is used. Quasi-Monte Carlo techniques possess good asymptotic properties and are often preferred to numerical methods, especially in higher-dimensional integration. For details of quasi-Monte Carlo methods, we refer to Niederreiter (1992), Evans and Swartz (2000) and Press et al. (1992).

Consider a covariogram  $C$  induced by a weight function  $w$  with bounded support. We want to approximate (non-zero values of)  $C(s, t)$ ,  $s, t \in D$ ,  $E = \mathbb{R}^d$ , i. e. the integral of  $f_{s,t}(p) = w(s, p)w(t, p)$  over a suitable interval  $A \supset \text{supp } f_{s,t} = \text{supp } w(s, \cdot) \cap \text{supp } w(t, \cdot)$ . We use the quasi-Monte Carlo approximation

$$\hat{C}_K(s, t) := \frac{\lambda^d(A)}{K} \sum_{i=1}^K f_{s,t}(p_i) \approx C(s, t)$$

with  $p_1, \dots, p_K \in A$ . Using *low-discrepancy sequences* of nodes, the error of the quasi-Monte Carlo approximation is of the order of  $(\log K)^d/K$ . Such a sequence is the Sobol sequence, which is used

in this work in a modified version proposed by Antonov and Saleev. The implementation of the generator is taken from Press et al. (1992).

We use an algorithm that computes the quasi-Monte Carlo approximation of covariograms induced by elliptical weight functions. Both algorithms support ordinary and strong boundaries. The algorithm uses nodes that are created a priori and recycled in every evaluation of  $\hat{C}(s_i, s_j)$ . Thus, the number of oracle calls  $\mathbf{1}_{D_i}(p)$  can be reduced significantly. Furthermore, all  $K$  oracle calls can be performed at once, e. g. by just one call to a GIS.

### 3.2 Minimizing the Mean Squared Error

**Remark 3.1 (Minimization in R)** In R, non-linear minimization can be carried out using a Newton-type algorithm available through the function `nlm`. In this work, `nlm` is used for minimizing mean squared error functions. It computes numerical derivatives of the target function and consequently needs a high number of (expensive) function evaluations. Derivative-free techniques may therefore be more efficient.

**Observation 3.2** The mean squared error function corresponding to the simple elliptical semivariogram (Example 2.6, Remark 2.9) usually possesses local minima. This is related to the corresponding semivariogram approximation being discontinuous with respect to the range parameter. In contrast, the piecewise linear elliptical kernel function (Remark 2.9) with breaking point parameter  $b$  close to 1 is almost identical to the simple one, but it is continuous in its arguments and hence the corresponding semivariogram approximation is continuous with respect to its range parameter. The minimization of mean squared error performed quite well with this semivariogram approximation using  $b = 0.95$ , for example, since small local minima are smoothed out now.

### 3.3 Implementation of Geostatistical Methods

Because of our emphasis on generically stationary processes, the generated R and C code will be called MoGeS, which stands for “**M**odeling **G**eneric **S**tationarity”.

The current implementation deals with the following geostatistical tasks<sup>1</sup>:

**Semivariogram modeling:** Select from a variety of stationary and instationary models, determine covariables, assign fixed values to parameters, and add semivariograms. Furthermore, visualize anisotropic semivariograms along a given path.

**Semivariogram fitting:** Estimate semivariogram parameters by minimizing the mean squared error function.

**Kriging:** Perform universal and ordinary kriging.

**Interaction with a GIS:** Read and write geostatistical datasets in an interchangeable file format.

**Simulation of datasets:** Create random data to a given semivariogram.

The specific tasks result in a series of routines, most of them being represented as methods of object classes.

**Semivariogram** (and covariogram) **models** are represented as objects of class `sv`, which is an abstract parent of `svfn`, `svc` and `csv` representing different levels of specification and aggregation. Semivariogram models “know” which parameters they need, if they are stationary etc.

**Parameters** are modeled as an independent object class `param` in order to allow meta-data such as semantics or the range of valid parameter values to be handled consistently together with the parameter vector itself.

**Geostatistical data** consists of locations, measurements and covariables stored in a `svm.data` object.

---

<sup>1</sup>Furthermore, fitting of semivariograms in the presence of trend was implemented, which is discussed in the presentation by K. G. van den Boogaart and A. Brenning: Why is Universal Kriging Better than IRFk-Kriging: Estimation of Variograms in the Presence of Trend.

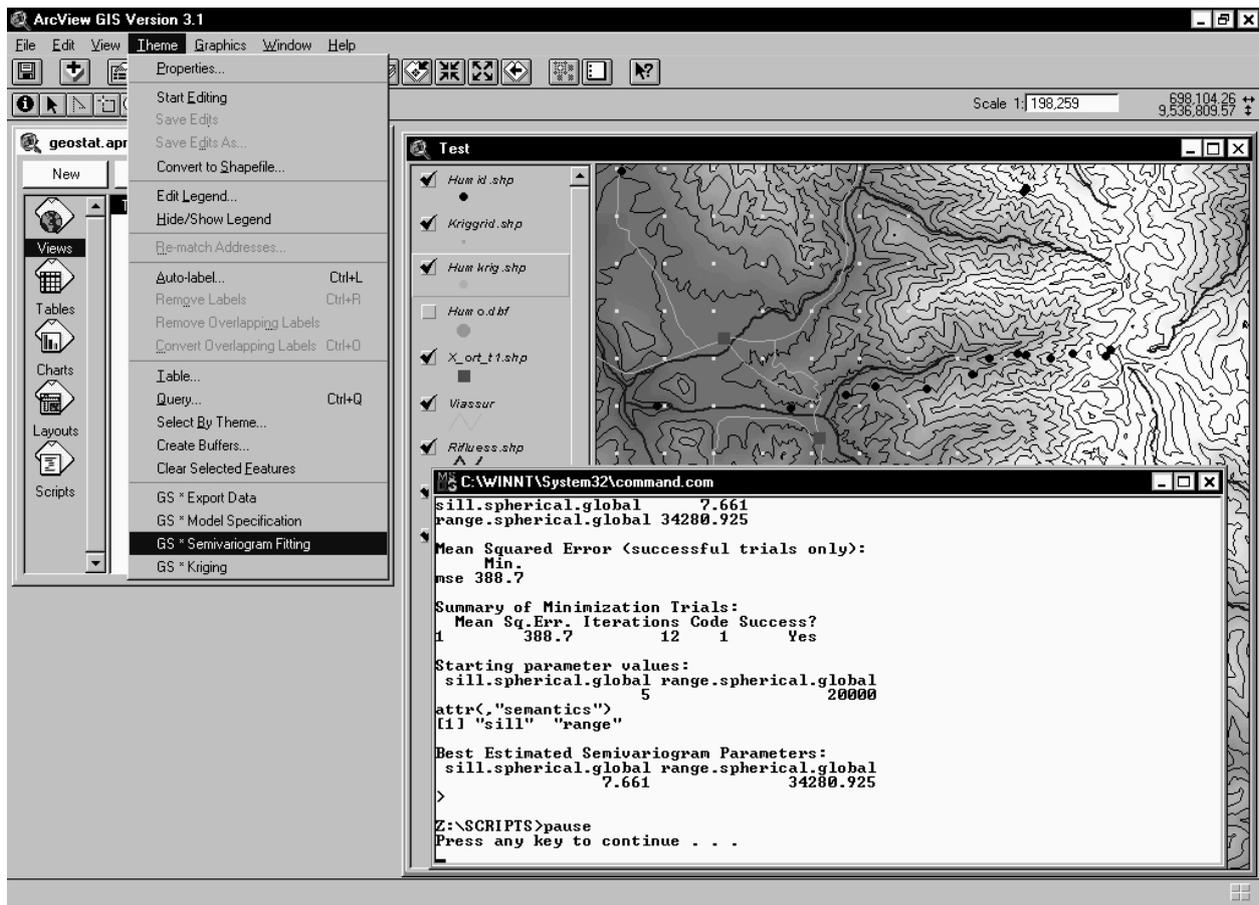


Figure 3: Semivariogram fitting within ArcView using the ArcView/MoGeS interface.

Figure 4: R code produced by the ArcView/MoGeS interface in an example situation.

```

[...] # initialization
##### read geostatistical data
d <- read.svm.data( file="z:/scripts/humid.csv",
  xnames="x", ynames="y", znames="Humo", gnames=c("orientation") )
n <- nrow(d$xy)

##### specify a semivariogram model
fpa <- param( c(0.9), # we have one fixed parameter
  nm = c("break.elliptical.pwlinear.global"), sem = c("break") )
pa.al <- setnames( c("sill.elliptical.pwlinear.global", # parameter aliases
  "range.elliptical.pwlinear.global", "q.elliptical.pwlinear.global",
  "break.elliptical.pwlinear.global"),
  c("sill", "range", "q", "break") )
g.al <- setnames( c("orientation.elliptical.pwlinear.global", # covariable aliases
  c("orientation") )
svc1 <- svc( svfn.elliptical.pwlinear.global, # a 'svc' semivariogram object
  fix.param=fpa, param.alias=pa.al, g.alias=g.al )
sv <- csv( list( svc1 = svc1 ) ) # this is our composed 'csv' semivariogram object

##### fit the semivariogram model:
start <- param( c(5,10000,0.7), # starting 'param'eter object
  nm = c("sill.elliptical.pwlinear.global", "range.elliptical.pwlinear.global",
  "q.elliptical.pwlinear.global"),
  sem = c("sill", "range", "q") ) # parameter semantics
smp <- NULL
if (needs.smp.data(sv)) # generate a priori nodes for quasi-monte carlo integration
  # the 'Rmax' argument must be sufficiently large!
  smp <- smp.data(d, Rmax=1.7*max(getrange(sv,start))/2, N=5000)
svmfit <- svm(sv,d=d,param=start,smp=smp,trend=FALSE) # fit the model
print(summary(svmfit))

```

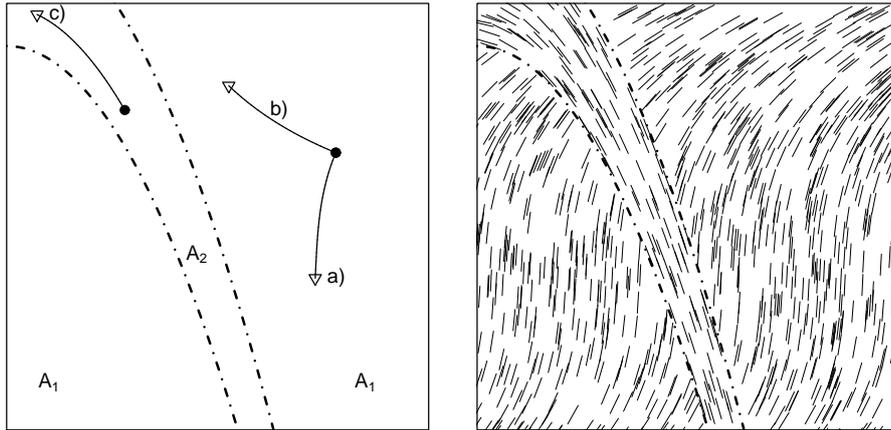


Figure 5: The situation of the simulated dataset with local anisotropy: Two independent subprocesses on  $A_1$  and  $A_2$  are considered (left). The curves represent the paths for which semivariograms are shown in figure 6. The sketch at the right visualizes the directions of local anisotropy.

**Nodes** are generated a priori for quasi-Monte Carlo integration and stored in a `smp.data` object. **Fitted semivariogram models** are represented by a `svm` object, which contains information on parameter estimates and other results of mean squared error minimization trials. (They are handled in a similar way as fitted linear models in R.)

### 3.4 Implementation of a GIS Interface

The ArcView/MoGeS interface performs geostatistical modeling in four steps:

**Export data:** Select a set of points and corresponding data from a *point theme* and its database or *table* (in ArcView terminology) and convert it to a text file format that can be read by MoGeS.

**Specify a semivariogram model:** Select semivariogram models, and link covariables with fields in the theme's database. If desired, assign fixed values to parameters or identify parameters.

**Fit the semivariogram model:** Choose starting parameter values, and perform the fitting through a call to the R environment. See figure 3 for a screen shot and figure 4 for an example of R code generated by the ArcView/MoGeS interface.

**Perform kriging:** Select measurement and prediction locations, export the corresponding data and perform kriging by calling the R environment.

The ArcView/MoGeS interface is a collection of AVENUE scripts that perform these tasks. It is accessible through commands added to the **Theme** menu (see figure 3). The scripts of course do not cover the complete MoGeS functionality available within R. Nevertheless, the problems mentioned above can be solved more easily than by hand, since R code is generated and executed automatically, and a user who is familiar with R will be able to add flexibility by modifying this code or doing additional analyses using the whole spectrum of R and MoGeS functions.

The AVENUE scripts model parameter vectors, names, aliases and semantics just as MoGeS does, however as separated lists rather than object classes or names vectors.

In its current implementation, the ArcView/MoGeS interface strongly depends on the MoGeS implementation (i.e. its function identifiers, argument names etc.), which makes it very sensible to small changes in MoGeS. This could be overcome by using a meta-language for geostatistical modeling that makes for example model specifications independent of the actual implementation that executes it.

## 4 Application: A Simulated Dataset in Complex Geology

We consider a Gaussian process  $Z$  on  $D = [0, 1]^2$  made up of two independent subprocesses  $Z_1$  on  $A_1$  and  $Z_2$  on  $A_2$  with location-dependent directions of anisotropy  $\phi(s)$  (see figure 5 for illustration).

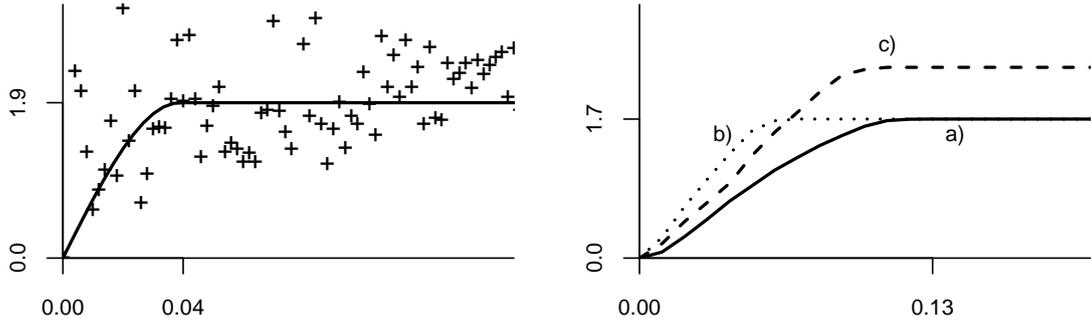


Figure 6: Left: Empirical semivariogram and the fitted spherical semivariogram with sill 1.90 and range 0.04.

Right: The fitted “true” semivariogram model, evaluated along the three paths shown in Figure 5: a) following the local direction of anisotropy in  $A_1$ ; b) orthogonal to anisotropy in  $A_1$ ; c) following anisotropy in  $A_2$ .

We select a constant mean  $m = 5$  and elliptical semivariograms with a piecewise linear kernel function and parameters  $\theta_1^T = (\sigma_1^2, a_1, q_1, b) = (1, 0.2, 0.6, 0.6)$  on  $A_1$  and  $\theta_2^T = (0.8, 0.1, 0.4, 0.95)$  on  $A_2$ , and on  $A_2$ , we also add a nugget effect with parameter  $\sigma_{\text{nug}(2)}^2 = 0.4$ . The directions of anisotropy on  $A_1$  and  $A_2$  are defined by polynomials and visualized in figure 5.

A geologic setting that hosts such a process could for example be the following: Suppose that  $Z$  represents some (lithogenic) soil property. On  $A_2$ , geologically young river sediments with strong fabric orientation down-stream and high local irregularity host the subprocess  $Z_2$  with analogous properties (anisotropy with  $q_2 = 0.4$ , small range, nugget effect). On  $A_1$ , things are smoother ( $a_1 = 2a_2$ , no nugget effect), but oblique sediment layers with folding structures originate an anisotropy ( $q_1 = 0.6$ ).

A total of  $n = 259$  locations was generated, 170 of which are uniformly distributed over  $A_1$ , and the remaining 89 are uniformly distributed over  $A_2$ . Then realizations of  $Z$  at these points were simulated using a Cholesky decomposition of the covariance matrix.

Kriging was performed on a  $60 \times 60$  grid using 10 000 a priori nodes. The computation of all the kriging predictions presented below took about a quarter of an hour in total, and each fitting trial a few minutes, depending on the number of iterations needed.

For comparison with more sophisticated models, we fitted stationary models with and without anisotropy, first of all the spherical semivariogram model using four different starting values. All trials succeeded and yielded practically the same estimates, namely a sill of 1.90 and a range of 0.039 (mse: 32.08). (See figure 6 for a comparison with the empirical semivariogram.) However, fitting an elliptical semivariogram with a piecewise linear kernel function (with  $b = 0.7$ ) and fixed direction of (geometrical) anisotropy did not yield consistent results for various choices of starting values and (fixed) orientation parameters.

Finally we fitted the “true” semivariogram model. Due to the high number of parameters, fitting took several steps, fixing some parameters at each step.

Comparing fitted and true parameters,

	$\sigma_1^2$	$a_1$	$q_1$	$b_1$	$\sigma_2^2$	$a_2$	$q_2$	$b_2$	$\sigma_{\text{nug}(2)}^2$
fitted	1.70	0.13	0.55	(0.70)	2.27	0.12	0.34	(0.70)	0.06
true	1.00	0.20	0.60	0.95	0.8	0.10	0.40	0.95	0.40

we observe that the fitted nugget effect almost vanishes, the sill parameters were overestimated, axis ratios were estimated quite well, and the range in the smoother area  $A_1$  was underestimated. Overestimation of the sill may be caused by an additional randomness due to integration errors during simulation.

See figure 6 for some sample plots of the fitted semivariogram along the paths shown in figure 5.

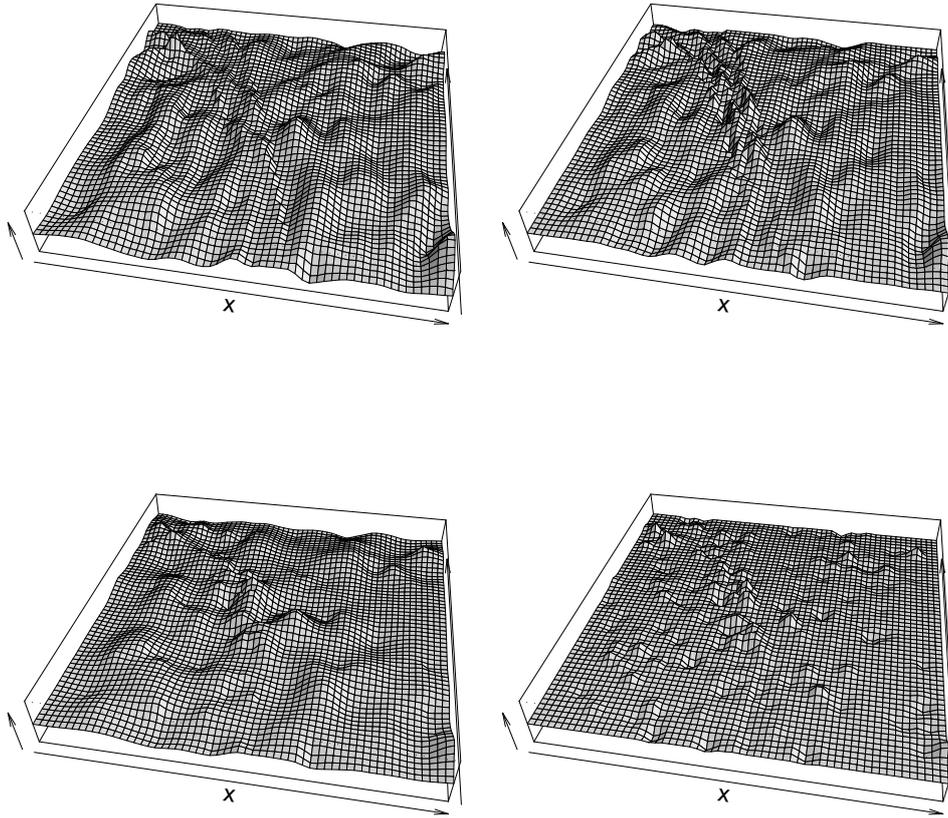


Figure 7: Kriging surfaces using the true semivariogram (top left), using the fitted generically stationary model (top right), using a fitted spherical semivariogram (sill 1.90, range 0.039; bottom right), and using a spherical semivariogram with sill 1.90 and range 0.13 (bottom left).

Note that these plots are representative in the sense that any semivariogram evaluation along anisotropy direction within  $A_1$  will look like graph a), etc. This is due to the generic stationarity property of elliptical semivariograms.

Our next aim is to compare kriging predictions obtained with different fitted models. We use the following semivariograms:

- the true semivariogram,
- the fitted semivariogram consisting of piecewise linear elliptical semivariograms on  $A_1$  and  $A_2$  plus a nugget effect on  $A_2$ ,
- the fitted spherical semivariogram with sill 1.90 and range 0.39, and
- a spherical semivariogram with sill 1.90 and the (more reasonable) range 0.13 taken from the fitted semivariogram with local anisotropy.

Kriging results are shown in Figure 7. It can clearly be seen that both predictions based on spherical semivariograms do not reflect the strong anisotropies present in our dataset, whereas the fitted generically stationary model with anisotropies leads to predictions that are very close to those obtained with the true semivariogram.

## 5 Conclusions

The construction method presented in this work has shown to be a useful instrument for incorporating knowledge of local geology as stored in a GIS into semivariogram models. Many situations of local anisotropy can be modeled using the class of elliptical semivariograms, which was studied in detail. Using the code developed for the present work, such models were successfully fitted, and

in an example situation it could be seen that the corresponding kriging results are also consistent with our knowledge of local anisotropy of the process, in contrast to isotropic or geometrically anisotropic semivariograms.

The study of the class of elliptical covariogram models motivated the introduction of the concept of generic stationarity. This concept reflects our belief in the existence of natural laws that determine a process' distribution law depending on local geology. The less knowledge of local geology is necessary to determine the distribution law, "the more stationary" is the process. Thus, generic stationarity becomes a means for stationarizing instationarity conditional on local geology.

First practical results were obtained by modeling the influence of local geology by means of generically stationary semivariograms in some cases of particular interest. However, it remains for future work to investigate and apply other generically stationary models to geological problems involving data from GIS.

## References

- Bivand, R. S. (1999): Integrating GRASS 5.0 and R: GIS and modern statistics for data analysis. In: Proc. 7th Scandinavian Research Conf. on Geogr. Information Science, pp. 111-127. Aalborg, Denmark.
- Cressie, N. A. C. (1993): Statistics for spatial data. Wiley, New York.
- Evans, M., and T. Swartz (2000): Approximating integrals via Monte Carlo and deterministic methods. Oxford Univ. Press, Oxford et al.
- Goovaerts, P. (1997): Geostatistics for natural resources evaluation. Oxford University Press, New York et al.
- Niederreiter, H. (1992): Random number generation and quasi-Monte Carlo methods. CBMS-NSF regional conference series in applied mathematics. SIAM, Philadelphia.
- Press, W. H., S. A. Teukolsky, W. T. Vetterling and B. P. Flannery (1992): Numerical recipes in C. 2nd ed., Cambridge Univ. Press, Cambridge et al.
- Stein, M. L. (1999): Interpolation of spatial data: some theory for kriging. Springer, New York.
- van den Boogaart, K. G. (1999): A new possibility for modelling variograms in complex geology. Proc. of StatGIS Klagenfurt 1999, to appear.
- Wackernagel, H. (1998): Multivariate geostatistics: an introduction with applications. Springer, Berlin et al.