DURÃO, F., CORTEZ, L., and MAGALHÃES, Á. A comparative study of two approaches to modelling and prediction in spatial statistics. *Application of Computers and Operations Research in the Minerals Industries,* South African Institute of Mining and Metallurgy, 2003.

A comparative study of two approaches to modelling and prediction in spatial statistics

F. DURÃO, L. CORTEZ, and Á. MAGALHÃES
C.V.R.M., I.S.T., Technical University of Lisbon, Portugal

Methods based on regionalized variables theory, commonly known as geostatistics, have been widely used since more than three decades to predict/estimate spatial distributed data, namely mineral resources, with quite good results. However, some restrictive theoretical assumptions, such as the need of some kind of stationarity of data and the linearity of the predictor/estimator, are often wrong options in many applications. An alternative approach based on a finite mixture modelling, or cluster weighted modelling (CWM) technique, is presented and compared, in its theoretical and practical aspects, with the geostatistical methodology. CWM is a general non-linear framework based on the estimation of joint density functions allowing the calculation of the conditional probability function of the target values given their locations. Validation tests are performed in a 2D case study, where testing data proceeds from a soil contaminated by a heavy metal. The results of ordinary kriging and CWM estimations show a very good agreement with actual data.

Keywords: Non-linear regression, Geostatistics, Ordinary kriging, Finite mixture modelling, Cluster weighted modelling.

Introduction

The problem of predicting/inferring the value of an unknown function $z(x)$ where $x \in D \subseteq \mathrm{R}^d$ (d=1,2,3), given their measured/observed values at N spatial locations $z(x_i)$, $x_i \in D$, i=1,2...N, is usually made in two steps: a modelling step based on the available data $z(x_i)$ and a estimation and prediction step based on this inferred model (and occasionally also on data).

These estimations/predictions are normally made in the scope of a branch of spatial statistics known as Geostatistics, or regionalized variables theory[1,2]. Its theoretical framework has good grounds on the random functions model and has been widely and successfully used since more than three decades, in spite of several attempts to present alternative approaches[3]. These attempts have failed mainly because their theoretical foundations and practical performances were not convincing.

In order to overtake the main limitations appointed to geostatistics (the stationarity assumptions, the linearity of the estimator/predictor and some subjectivity in the choice of the model parameters), in former papers the authors have tried the application of artificial neural networks (multilayer perceptrons, MLP, and generalized regression neural networks, GRNN) to the estimation of mineral resources [4–6], with quite good results. However, MLP presents also some limitations: their architecture is usually chosen in a trial and error basis, they take long to converge and model parameters are seldom meaningful. On the other hand, GRNN show excellent results and have a solid background[7], but its study is much more interesting when embodied in a more general framework.

The alternative approach presented here is a finite mixture (or cluster weighted) modelling—which includes GRNN as a particular case—with a sound theoretical basis avoiding the aforesaid drawbacks. It is plainly suitable for 3D mineral resources estimation, but by reason of easiness in showing the results, we chose for validation a simple 2D case study.

Geostatistical approach

Modelling step

The assumed function $z(x)$ is defined as a sum of two components:

$$z(x) = m(x) + \varepsilon(x)$$

where $m(x)$ is a deterministic function of the spatial coordinates x, called mean, trend or drift, and $\varepsilon(x)$ is a zero-mean spatial random (stochastic) function.

The function $m(x)$ can be modelled as a linear combination of basis functions. The basis functions may be polynomial functions of spatial coordinates x. More often, $m(x)$ is taken as a constant function, m, independent of the spatial coordinates. In this case, one has:

$$z(x) = m + \varepsilon(x)$$

The function $\varepsilon(x)$, or $z(x)$, is assumed to be a *second order intrinsically stationary random function*, that is,

$$\mathrm{E}\left[z(x) - z(x')\right] = 0$$

$$\mathrm{E}\left[\left(z(x) - z(x')\right)^2\right] = 2\gamma(x - x') = 2\gamma(h)$$

where $\gamma(h)$, called the *variogram*, is a function of the incremental/separation vector h. If the variogram, $\gamma(h)$, depends only on the distance between the two spatial locations x e x', then $z(x)$ is said to be an *intrinsic isotropic function*, with:

$$E\left[z(x)-z(x')\right]=0$$

$$E\left[\left(z(x)-z(x')\right)^2\right]=2\gamma\left(\|x-x'\|_2\right)=2\gamma\left(\|h\|_2\right)$$

If $z(x)$ is a *second order stationary random function*, then:

$$E\left[z(x)\right]=m$$

$$E\left[\left(z(x)-m\right)\left(z(x')-m\right)\right]=R(x-x')=R(h)$$

where $R(h)$ is the covariance function. The relation between the two second moments is given by:

$$\gamma(h)=\tfrac{1}{2}E\left[\left(z(x)-z(x')\right)^2\right]=-R(h)+R(0)$$

Estimation step

The estimation step amounts to model the function $\gamma(h)$ or $R(h)$.

The usual procedure first starts by building the so-called experimental variograms for a set of predefined directions through their rotation angles about the coordinate axes. In a second stage, the experimental variograms for the principal directions are modelled, manually, through the use of interactive software tools.

The output of the estimation step can be summarized as follows:

$$\gamma(h)=\sum_{k=1}^{n_e}\left(\sigma^2\right)^{(k)}\gamma^{(k)}\left(\left\|h^*\right\|_2^{(k)}\right)$$

where n_e is the number of elementary variogram models, $(\sigma^2)^{(k)}$ is the sill of k-th elementary variogram, $\gamma^{(k)}(\|h^*\|_2^{(k)})$ is the k-th elementary variogram model and $\|h^*\|_2^{(k)}$ is defined by:

$$\left\|h^*\right\|_2^{(k)}=\sqrt{\left(h_x^*\right)^2+\left(h_y^*\right)^2+\left(h_z^*\right)^2}$$

with h^* given by:

$$h^*=S\left(a_x^k,a_y^k a_z^k\right)Rot\left(\varphi_x,\varphi_y,\varphi_z\right)h$$

where $S(a_x^{(k)}, a_y^{(k)}, a_z^{(k)})$ is a diagonal scaling matrix and $\mathrm{Rot}(\varphi_x, \varphi_y, \varphi_z)$ is the rotation matrix. The elements $a_x^{(k)}$, $a_y^{(k)}$ and $a_z^{(k)}$ are the ranges of the k-th elementary structure along the principal directions. The rotation angles (φ_x, φ_y, and φ_z) are positive for counterclockwise rotations about the axes.

Some of the more common valid covariance/variogram models are presented in Table I, where $h^* = \| h^* \|_2$.

Prediction Step (Ordinary Kriging)

The prediction step, known as *Ordinary Kriging* (OK), in the scope of the geostatistical approach, consists of predicting the z value, $\hat{z}(x_0)$, at a new spatial location x_0. It assumes a linear predictor, defined as linear combination of the N observed z values, $\{z(x_i)\}_{i=1}^N$, and given as:

$$\hat{z}\left(x_o\right)=\hat{z}_o=\sum_{i=1}^{N}\lambda_i z\left(x_i\right)$$

The unknown coefficients, λ_i, $i = 1,2,...,N$, are determined such that the predictor is the *Best Linear Unbiased Predictor (BLUP)*. Thus, the BLUP enjoys the two most important properties: unbiasedness and minimum mean-squared prediction error.

It can be shown[8,9] that the unbiasedness condition amounts to:

$$\sum_{i=1}^{N}\lambda_i=1$$

and that the *mean-squared prediction error* at x_0,σ_o^2, can be expressed as:

$$\sigma_o^2=E\left[\left(\hat{z}_o-z\left(x_o\right)\right)^2\right]=-\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j\gamma\left(\left\|x_i-x_j\right\|_2\right)+$$

$$2\sum_{i=1}^{N}\lambda_i\gamma\left(\left\|x_i-x_o\right\|_2\right)$$

or as:

$$\sigma_o^2=\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j R\left(\left\|x_i-x_j\right\|_2\right)+$$

$$R(0)-2\sum_{i=1}^{N}\lambda_i R\left(\left\|x_i-x_o\right\|_2\right)$$

Adopting the matrix-vector notation, the two expressions above can be written as:

$$\sigma_o^2=\lambda^T c_\gamma+\tfrac{1}{2}\lambda^T Q_\gamma\lambda$$

or as:

$$\sigma_o^2=R(0)+\lambda^T c_R+\tfrac{1}{2}\lambda^T Q_R\lambda$$

with: $c_\gamma = 2\gamma$, $Q_\gamma = -2\Gamma$, $c_R = -2r$ and $Q_R = 2R$. $\Gamma \in \Re^{N\times N}$ is a symmetric indefinite matrix with no 0-eigenvalues, $R \in \Re^{N\times N}$ is a symmetric definite matrix and γ, r, $\lambda \in \Re^N$ are (column) vectors.

According to the two conditions mentioned above, the unknown coefficients, λ_i,i=1,2,..,N, are found by solving one of the following Quadratic Optimization problems:

$$\min_{\lambda_i,i=1,2,...,N}\sigma_o^2=\lambda^T c_\gamma+\tfrac{1}{2}\lambda^T Q_\gamma\lambda$$

subject to

$$1^T\lambda=1$$

or:

$$\min_{\lambda_i,i=1,2,...,N}\sigma_o^2=R(0)+\lambda^T c_R+\tfrac{1}{2}\lambda^T Q_R\lambda$$

subject to

$$1^T\lambda=1$$

where 1 is the vector with all N entries equal to 1.

The first problem is an Indefinite Quadratic Optimization problem, while the second one is a Positive Definite (Convex) Quadratic Optimization problem.

The optimality conditions for the solution of both problems state that if λ^* is the optimal solution vector, then there is a pair (λ^*, μ^*) that satisfy the following first and

Table I
Common covariance/variogram models

	$R(h^*)$	$\gamma(h^*)$
Nugget effect	$\begin{cases}0, & h^*>0\\ 1, & h^*=0\end{cases}$	$\begin{cases}1, & h^*>0\\ 0, & h^*=0\end{cases}$
Exponential	$\exp(-h^*)$	$1-\exp(-h^*)$
Spherical	$\begin{cases}(1-\frac{3}{2}h^*+\frac{1}{2}h^{*3}), & 0\le h^*\le 1\\ 0, & h^*>1\end{cases}$	$\begin{cases}(\frac{3}{2}h^*-\frac{1}{2}h^{*3}), & 0\le h^*\le 1\\ 1, & h^*>1\end{cases}$
Gaussian	$\exp(-h^{*2})$	$1-\exp(-h^{*2})$

second order conditions:

First order conditions

$$\begin{cases}\nabla_\lambda L(\lambda^*,\mu^*)=0 \Rightarrow c+Q\lambda^*+\mu^*1=0 \Leftrightarrow Q\lambda^*+\mu^*1=-c\\ \nabla_\mu L(\lambda^*,\mu^*)=0 \Rightarrow \quad 1^T\lambda^*-1=0 \Leftrightarrow \quad 1^T\lambda^*=1\end{cases}$$

Second order conditions

$$\delta^T\nabla_{\lambda\lambda}L(\lambda^*,\mu^*)\delta\geq 0,\ \forall\delta\in\Re^N,\delta\neq 0 \mid 1^T\delta=0$$

where: $c = c_\gamma$ or $c = c_R$, $Q = Q_\gamma$ or $Q = Q_R$, and $L(\lambda,\mu)$ is the Lagrangian associated to the problem with Lagrange multiplier μ, defined as:

$$L(\lambda,\mu)=\lambda^T c+\tfrac{1}{2}\lambda^T Q\lambda+\mu(1^T\lambda-1)$$

The operators $\nabla_\lambda L(\lambda,\mu)$ and $\nabla_\mu L(\lambda,\mu)$ denote the gradients of the Lagrangian with respect to λ and μ, respectively. The operator $\nabla_{\lambda\lambda} L(\lambda,\mu)$ denotes the Hessian of Lagrangian with respect to vector λ, defined as:

$$\nabla_{\lambda\lambda}L(\lambda,\mu)=Q$$

It can be shown that $\delta^T Q\delta>0$, $\delta\in\Re^N, \delta\neq 0 \mid 1^T\delta=0$ for both optimization problems. That means that any λ satisfying the first order conditions (stationary point) is a global minimizer in the feasible solution set of the problem. Within this setting, the solution of the optimization problem becomes a problem of solving the following system of $(N+1)$ linear equations with $(N+1)$ unknowns, (λ^*, μ^*):

$$\left[\begin{array}{c|c}Q & 1\\ \hline 1^T & 0\end{array}\right]\left[\begin{array}{c}\lambda^*\\ \hline \mu^*\end{array}\right]=\left[\begin{array}{c}-c\\ \hline 1\end{array}\right]$$

Taking into account the fact that the matrix Q is non-singular, for both problems, it can be proved that the solution of the above system is:

$$\begin{cases}\mu^*=\dfrac{1+1^T z}{-1^T y}\\ \lambda^*=-z-\mu^* y\end{cases}$$

where $z = Q^{-1}c$ and $y = Q^{-1}1$ with $1^T y\neq 0$.

The corresponding optimal value of σ_0^2 is:

$$\sigma_o^{2^*}=\lambda^{*T}\gamma-\tfrac{1}{2}\mu^*$$

or:

$$\sigma_o^{2^*}=R(0)-\lambda^{*T}r-\tfrac{1}{2}\mu^*$$

Finite mixture modelling (cluster weighted modelling)

Introduction

The following presentation is a short summary of the main steps fully presented in Schoner[10]. A short presentation is found in Gershenfeld[11].

Modelling step

According to the estimation theory[12], the *Best Least Squares predictor*, $z(x)_{BLS}$, of $z(x)$, which minimizes the expected value (*Risk function*) of the squared prediction error (*Loss or Cost function*), is the conditional expectation (mean of the conditional distribution of $z(x)$ given x), i.e.,

$$\hat{z}(x)_{BLS}=\int\hat{z}(x)\,p(z\mid x)dz=\mathrm{E}\left[\hat{z}(x)\mid x\right]$$

where $p(z|x)$ is the conditional probability density function of z given x.

The conditional probability density function $p(z|x)$ is given by:

$$p(z\mid x)=\frac{p(z,x)}{p(x)}=\frac{p(z,x)}{\int p(z,x)dz}$$

The joint probability density function, $p(z,x)$, can be approximated by an identifiable finite mixture of known parametric probability density functions as follows:

$$p(z,x)=\sum_{j=1}^{K}\pi_j p_j(z,x)=\sum_{j=1}^{K}\pi_j p_j(z\mid x)p_j(x)$$
$$\pi_j>0, j=1,2,...,K,\ \sum_{j=1}^{K}\pi_j=1$$

where π_j, $j=1,2,...,K$, are the mixing proportions of the K components in the mixture. It is also the probability, π_j, of a given observation of z and x coming from the j-th component distribution. $p_j(x)$ and $p_j(z|x)$ are the parametric probability density functions of component j.

The most used and well-known parametric probability density function is the Gaussian density function. Thus, the component density functions can be defined as follows:

$$p_k(x)=(2\pi)^{-d/2}|C_k|^{-1/2}\exp\left[-\tfrac{1}{2}(x-\mu_k)^T C_j^{-1}(x-\mu_k)\right],$$
$$k\quad =1,2,...,K$$

where μ_j is the mean vector of the j-th multivariate Gaussian distribution of x and C_j^{-1} the inverse of the covariance matrix. The $|\ .\ |^1/_2$ is the square root of the determinant.

$$p_k(z\mid x)=\frac{1}{\sqrt{2\pi\sigma_{k,z}^2}}\exp\left[-\tfrac{1}{2}\left(\frac{z-\mu_{k,z}}{\sigma_{k,z}}\right)^2\right]=$$
$$\frac{1}{\sqrt{2\pi\sigma_{k,z}^2}}\exp\left[-\tfrac{1}{2}\left(\frac{z-f(x,\beta_k)}{\sigma_{k,z}}\right)^2\right],\ k=1,2,...,K$$

where the mean, $\mu_{k,z}$, of the Gaussian density is a function k of x and a set of adjustable parameters β_k and $\sigma_{k,z}^2$ is the variance of z conditionally on j-th component.

It is assumed that $f(x,\beta_k)=\sum_{1=0}^{M}\beta_{k,1}f_1(x)$ is a linear combination of $M+1$ basis functions f_i of x. The $f(x,\beta_k)$ are local trend models. The basis functions can be polynomial functions of order up to p, commonly p = 0, 1 or 2. The total number of terms, M+1, is a function of p.

Estimation step

Estimation of the parameter set θ defining the mixture model

Let θ be an element of the parameter space Ω, defined as follows:

$$\theta\in\Omega\overset{\Delta}{=}\left\{\begin{array}{l}\pi_k>0,\sum_{k=1}^{K}\pi_k=1,\mu_k\in\Re^d,C_j\in\Re^{dxd}pds,\sigma_{k,z}^2\\ >0,\beta_k\in\Re^{M+1},k=1,2,..,K\end{array}\right\}$$

The parameter estimation step is done using the *maximum likelihood principle*. Given a set of N independent and identically distributed (*i.i.d.*) observations $\{x_i,z_i\}_{i=1}^N$ where x_i are the known spatial locations and the z_i the corresponding observed or measured values of the z function (grade, concentration,...), the maximum likelihood estimate of θ, θ_{MLE}, is the argument $\theta\in\Omega$ that maximizes the likelihood

function of the observed data or the logarithm of the likelihood function.

The likelihood function of the observed data is defined as follows:

$$L\left(\theta;\{x_i,z_i\}_{i=1}^{N}\right)=\prod_{i=1}^{N}p(z_i,x_i;\theta)=\prod_{i=1}^{N}\sum_{j=1}^{K}\pi_j p_j\left(z_i,x_i:\theta_j\right)$$

and the corresponding observed data log-likelihood:

$$l\left(\theta;\{x_i,z_i\}_{i=1}^{N}\right)=\log\left[L\left(\theta;\{x_i,z_i\}_{i=1}^{N}\right)\right]=\log\left(\prod_{i=1}^{N}p(z_i,x_i;\theta)\right)=\sum_{i=1}^{N}\log\left(\sum_{j=1}^{K}\pi_j p_j\left(z_i,x_i;\theta_j\right)\right)$$

Summarizing, the estimation step amounts to solve the following problem:

$$\theta_{MLE}=\arg\min_{\theta\in\Omega}\; l\left(\theta;\{x_i,z_i\}_{i=1}^{N}\right)=\arg\min_{\theta\in\Omega}\sum_{i=1}^{N}\log\left(\sum_{j=1}^{K}\pi_j p_j\left(z_i,x_i;\theta_j\right)\right)$$

This difficult non-linear constrained optimization problem is currently solved by using the Expectation-Maximization (EM) algorithm, widely disseminated after the work published by Dempster, Laird and Rubin[13]. In Titterington *et al.*[14] a full presentation of the EM algorithm can also be found. A good introduction to the EM algorithm, with numerical examples is given in Flury[15].

The EM algorithm maximizes the observed data log-likelihood function with respect to θ by maximizing, iteractively, the expectation of the complete data log-likelihood with respect to the conditional distribution of the missing data W, given the observed data and the current estimate $\theta^{(s)}$.

If one knew the component codes (missing data) of the observed data, the maximum likelihood estimation of the complete data log-likelihood would consist of separated maximum likelihood estimation of the parametric Gaussian density functions, which is quite easy and has a well-known closed form.

The EM algorithm first computes the expectation of the complete data log-likelihood, by replacing the unknown component codes variables by their expectations given the observations and the current approximation $\theta^{(s)}$ (E-step), followed by its maximization (M-step).

The basic EM algorithm for Gaussian (Normal) Mixtures is summarized in Figure 1.

Assuming: $f(x,\beta_k)=\sum_{l=0}^{M}\beta_{k,l}f_l(x)$ as a linear combination of basis functions of x, the generic elements of matrix A_k and vector b_k are given by:

$$a_{r,s}^{k}=\frac{1}{N\pi_k^{(s+1)}}\sum_{i=1}^{N}\pi_{k,i}^{(s+1)}f_r(x_1)f_s(x_i),\quad r=0,1,2,...,M;\ s=0.12,...,M$$

$$b_s^k=\frac{1}{N\pi_k^{(s+1)}}\sum_{i=1}^{N}\pi_{k,i}^{(s+1)}z_i f_s(x_i),\qquad s=0,1,2,...,M$$

Like many other known unconstrained/constrained optimization methods, the EM algorithm only guarantees the convergence to a local maximizer of the non-convex observed data log-likelihood function. This means that the local maximizer is critically dependent on the initialization of the parameter set θ. The other drawback is the convergence to a point on the boundary of parameter space, specifically when one π_k becomes almost 0.

As suggested in several published papers[16], the initial values of the most model parameters is done using the *K-means clustering algorithm*[17]. The initial values of the parameter vectors β_k are arbitrary. For instance, 1 for the first element, constant term, and 0 for the remaining elements.

The convergence to the parameter space boundary can be prevented by adding very small positive values to the main diagonal as long as the covariance matrices are not positive definite. Other approaches can be used such as those based on the estimation or computation of the minimum eigenvalue of the covariance matrices.

In the Figure 1, SVD denotes the Singular Value Decomposition of the matrix A_k.

Estimation of the hyper parameter K (Number of components of the mixture)

The number of components, K, is taken as the value that minimizes the mean squared prediction error corrected for the bias. The approach closely follows that described in Efron and Tibshirani[18] regarding refined *bootstrap estimates* of the prediction error. An algorithmic description is given below in Figure 2.

This approach can be computer demanding. Other approaches based on information criteria (AIC, BIC, MDL) are currently used in the closely related area of unsupervised classification. A very interesting algorithm is the Agglomerative EM (AEM), based on Mixture Minimum Descriptive Length (MMDL), a modification of MDL criterion proposed by[16] combined with a merging step of the two less 'significant' and 'close' components. However, this approach still requires a validation in the scope of regression.

Prediction step

The conditional prediction/forecast is given by:

$$\hat{z}(x_o)_{BLS}=E\left[z\mid x_o\right]=\int zp(z\mid x_o)dz=\frac{\sum_{k=1}^{K}f(x_o,\beta_k)p_k(x_o)\pi_k}{\sum_{k=1}^{K}p_k(x_o)\pi_k}=\sum_{k=1}^{K}\lambda_k f(x_o,\beta_k),\ \text{with}\ \lambda_k=\frac{p_k(x_o)\pi_k}{\sum_{k=1}^{K}p_k(x_o)\pi_k}$$

and the conditional variance of prediction error is:

$$\sigma_o^2=E\left[\left(z-E\left[z\mid x_o\right]\right)^2\mid x_o\right]=\int\left(z-E\left[z\mid x_o\right]\right)^2p(z\mid x_o)dz=\frac{\sum_{k=1}^{K}\left[\sigma_{k,z}^2+f(x_o,\beta_k)^2\right]p_k(x_o)\pi_k}{\sum_{k=1}^{K}p_k(x_o)\pi_k}-E\left[z\mid x_o\right]^2$$

$$=\sum_{k=1}^{K}\lambda_k\left[\sigma_{k,z}^2+f(x_o,\beta_k)^2\right]-E\left[z\mid x_o\right]^2,\ \text{with}\ \lambda_k=\frac{p_k(x_o)\pi_k}{\sum_{k=1}^{K}p_k(x_o)\pi_k}$$

Example

Brief description

The example considers a sample of 100 observations of lead concentration (ppm) from a contaminated soil with an area approximately 4500 x 3000 m. The basic statistics are summarized in Table II.

Basic EM Algorithm

Input:

$\{\mathbf{x}_i, z_i\}_{i=1}^N$	(Observed data)
K	(Number of components in the mixture)
$\theta^{(0)}$	(Initial values of model's parameters)
ε	(Stopping criteria tolerances)
s_{max}	(Maximum number of iterations)

$s = 0$! Iteration Counter

While ($s \leq s_{max}$)

1. Compute current value of observed data log-likelihood

2. **Expectation (E) Step**

Compute the expected values of the missing data conditionally on the observed data and current $\theta^{(s)}$ (The expected values of component codes variables)

$$\pi_{k,i} = \frac{\pi_k^{(s)} p_k\left(z_i|\mathbf{x}_i;\theta^{(s)}\right) p_k\left(\mathbf{x}_i;\theta^{(s)}\right)}{\sum_{j=1}^{K} \pi_j^{(s)} p_j\left(z_i|\mathbf{x}_i;\theta^{(s)}\right) p_j\left(\mathbf{x}_i;\theta^{(s)}\right)}, \quad k = 1,2,\ldots,K;\ i = 1,2,\ldots,N$$

3. **Maximization (M) Step**

Maximize w.r.t. θ the complete data log-likelihood with their component codes variables replaced by their expectations (Compute $\theta^{(s+1)}$)

For $k = 1,2,\ldots,K$

3.1 $\pi_k^{(s+1)} = \frac{1}{N}\sum_{i=1}^{N} \pi_{k,i}$

3.2 $\mu_k^{(s+1)} = \frac{1}{N\pi_k^{(s+1)}}\sum_{i=1}^{N} \pi_{k,i}\mathbf{x}_i$

3.3 $\mathbf{C}_k^{(s+1)} = \frac{1}{N\pi_k^{(s+1)}}\sum_{i=1}^{N} \pi_{k,i}\left(\mathbf{x}_i - \mu_k^{(s+1)}\right)\left(\mathbf{x}_i - \mu_k^{(s+1)}\right)^T$

3.4 Solve the linear system of M+1 equations

$\mathbf{A}_k\beta_k^{(s+1)} = \mathbf{b}_k$

with respect to $\beta_k^{(s+1)}$, of size $(M+1)\times 1$, using SVD

3.5 $\mu_{k,z_i}^{(s+1)} = f\left(\mathbf{x}_i, \beta_k^{(s+1)}\right), \qquad i = 1,2,\ldots,N$

3.6 $\sigma_k^{2(s+1)} = \frac{1}{N\pi_k^{(s+1)}}\sum_{i=1}^{N} \pi_{k,i}\left(z_i - \mu_{k,z_i}^{(s+1)}\right)^2$

End

4. **If Stopping Criteria satisfied**

Output: $\theta_{MLE} = \theta^{(s)}$

STOP

End

5. $s = s+1$

End

Figure 1. The Basic EM algorithm for normal (Gaussian) mixtures

Bootstrap estimate of prediction error

Input: K_{min}, K_{max} (Minimum and maximum numbers of components)
B (Number of bootstrap samples – resampling stages)
N_B (Number of observations in each bootstrap sample)
N_V (Number of observations in validation/testing sample: $N_V=N-N_B$)

for K = K_{min} to K_{max}
for i=1to B (Number of bootstrap samples)

1. **Generate bootstrap sample**
   (Randomly select N_B observations from original sample of N observations)
2. **Initialize parameter set**: $\theta^{(0)}$
3. **Compute** θ_{MLE} **using EM algorithm**
4. **Predict** z(x) values for observations in bootstrap sample and compute (cumulative) **mean squared apparent errors**
5. **Predict** z(x) values for observations in validation/testing sample and compute (cumulative) **mean squared prediction errors**
6. **Compute bias** (optimism): Difference between mean squared prediction errors and mean squared apparent errors
7. **Compute corrected mean squared prediction errors**: Add **bias** to the average of the mean squared apparent and prediction errors.

end
Save mean squared errors (apparent, prediction and corrected prediction)
end
Output: $\hat{K}$ as the argument that minimizes the corrected prediction error as a function of **K**

Figure 2. Bootstrap estimate of the prediction error with determination of *K*

Table II
Basic statistics of sample data (Pb ppm)

Mean: 53.05 Variance: 736.72 St. Dev.: 27.14	Minimum: 25.0 Median: 45.85 Maximum: 189.9

Software tools

The modelling and estimation steps in the scope of the geostatistical approach were made by using the *GeoMS* software tool[19]. All the other steps were carried out by writing and using software to be run under MATLAB environment.

Results

Results of modelling and estimation steps (geostatistical approach)

The results of the modelling and estimation steps are summarized in Table III. The basic results of the modelling step are the parameters of the variogram (or covariance) functions in the different directions. A geometric anisotropy was identified with two main directions.

Results of modelling and estimation steps (CWM approach)

Computation of optimal number of components in the finite mixture, considering linear and quadratic local models, was carried out with the following setup: Bootstrap samples: 20; Working samples: N_B=0.8N; Validation samples (random): N_V=0.2N.

The results are shown in Figure 3. The optimal numbers of components are, approximately, 10–11 for linear local models and 5–6 for quadratic local models.

Figure 4 presents the contour levels of the quadratic local models, considering K=6, and the corresponding elliptical 'influence' regions.

Cross-validation

The geostatistical approach was cross-validated using a *leave one out* scheme. In Figures 5 and 6 the histograms of deviations and the normal probability plot of deviations are presented, respectively, for the geostatistical and finite mixture modelling (CWM) approaches. Figure 7 presents the respective scatter diagrams for both approaches.

In Table IV are summarized the main statistics of both estimators.

Maps

Figures 8 and 10 present the predicted values printed in grey tones, using a 100 x 100 grid, respectively for the geostatistical and CWM approaches. The values and location of measured data are also shown. Figures 9 and 11 present the respective maps of standard deviations of the prediction errors.

Table III

Direction	Angle (°)	Model	Sill (ppm²)	Range (m)
1	90	Spherical	736.71	1479
2	0	Spherical	736.71	1212

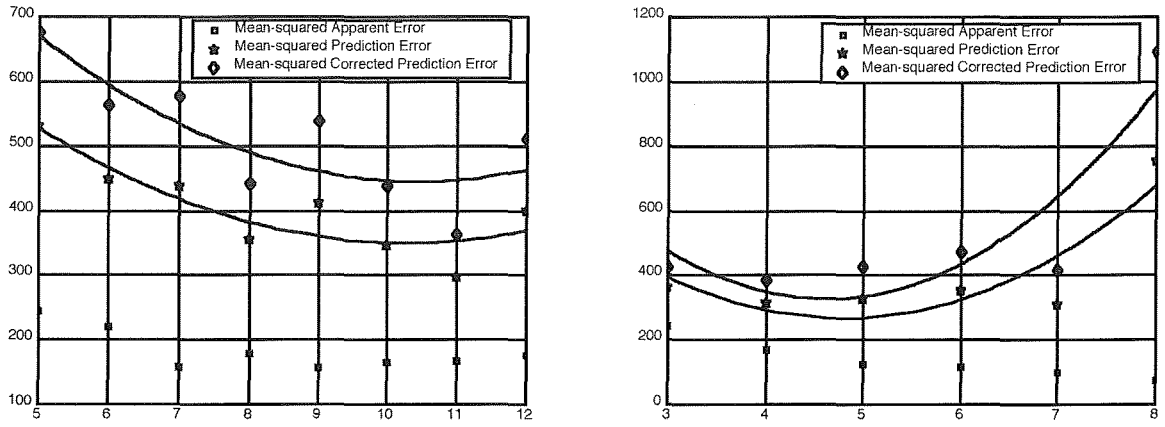

Figure 3. Variation of the mean-squared apparent and prediction errors as a function of the number of components, *K*, for linear local models (left) and for quadratic local models (right)

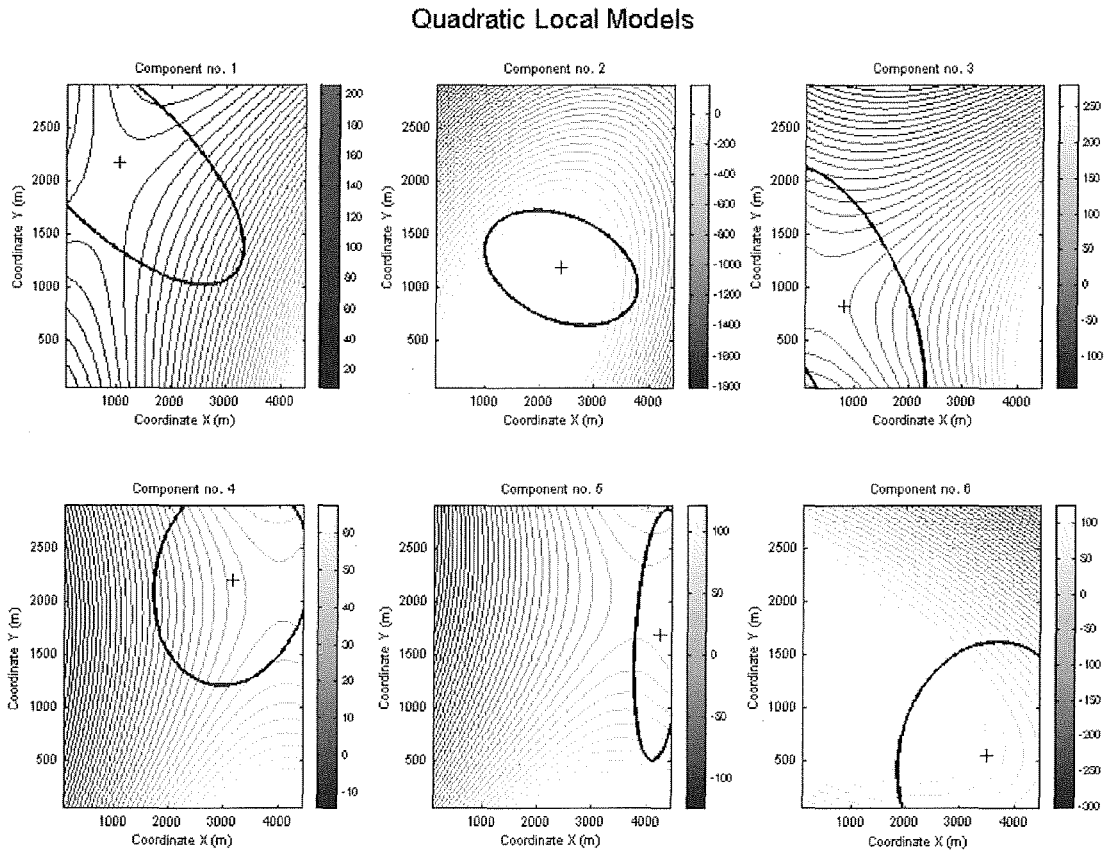

Figure 4. Quadratic local models, considering *K*=6, and the corresponding elliptical 'influence' regions

Conclusions

The geostatistical approach can be regarded as a non-parametric regression approach, while finite mixture modelling can be viewed as a semi-parametric approach.

As referred to in[11], the global model provided by the finite mixture modelling (cluster weighted modelling) approach as an assembling of local models can handle **non-linearity, non-Gaussianity**, by overlap of several components/clusters, and **non-stationarity**. The kernel based estimation methods, such as the GRNN, can be seen as special cases of this more general approach, where each point (sample) is the centre of a radial basis function with one global adjustable parameter.

It must be said that geostatistics can handle also with non stationary data (universal kriging, kriging with external drift, IRF-K, etc), but it seems that these techniques are much more complex than the alternative model presented here.

The prediction step of finite mixture modelling is much faster than the geostatistical approach, because there is no need to solve any numerical problem.

References

1. MATHERON, G. *Les Variables Regionalisées et leur Estimation*, Masson, Paris, 1965, 305 pp.
2. JOURNEL, A. Geostatistics Models and Tools for the

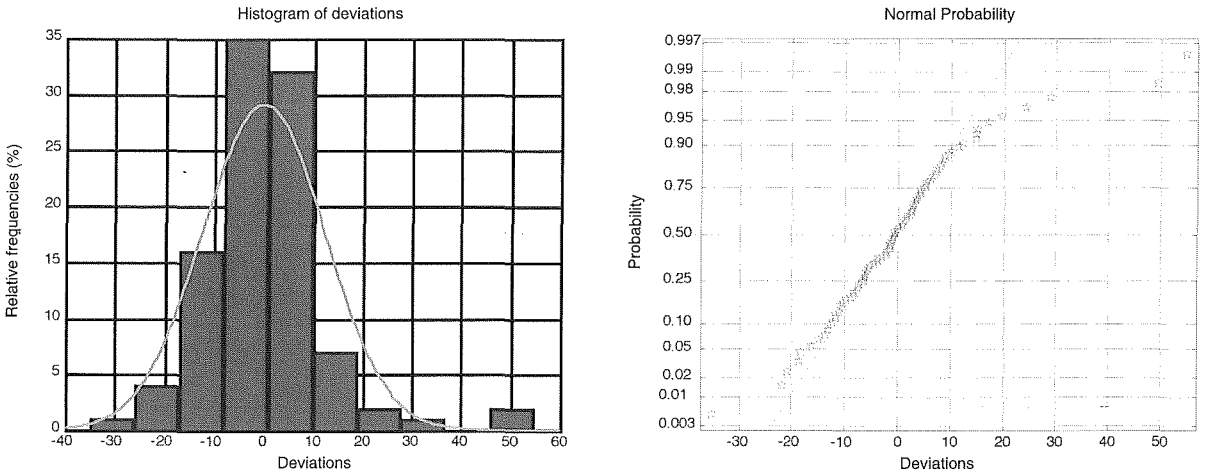

Figure 5. Cross-validation (Geostatistical approach). Histogram of deviations (left) and normal probability plot of deviations (right)

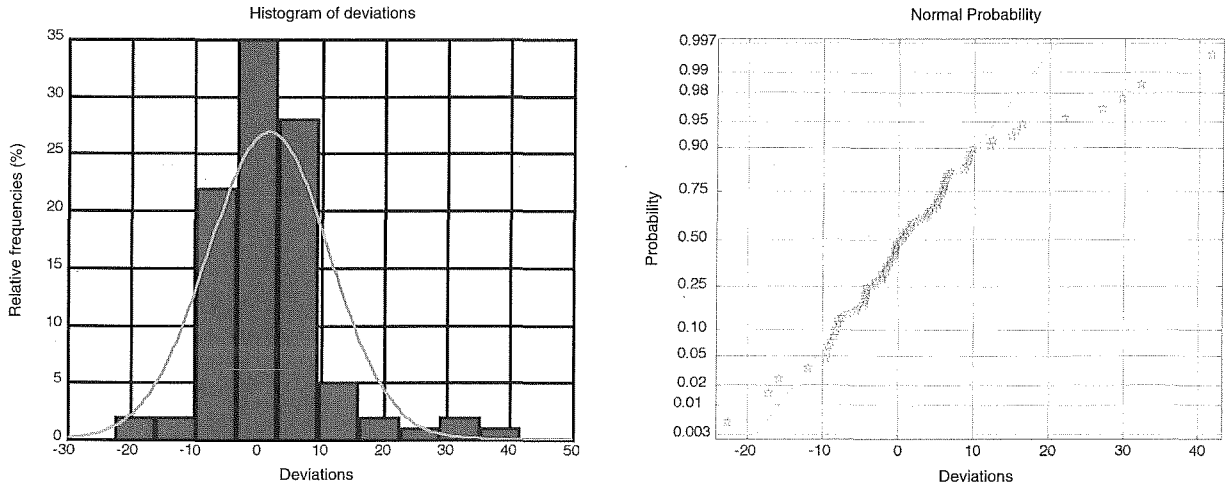

Figure 6. Model-validation (CWM). Histogram of deviations (left) and normal probability plot of deviations (right)

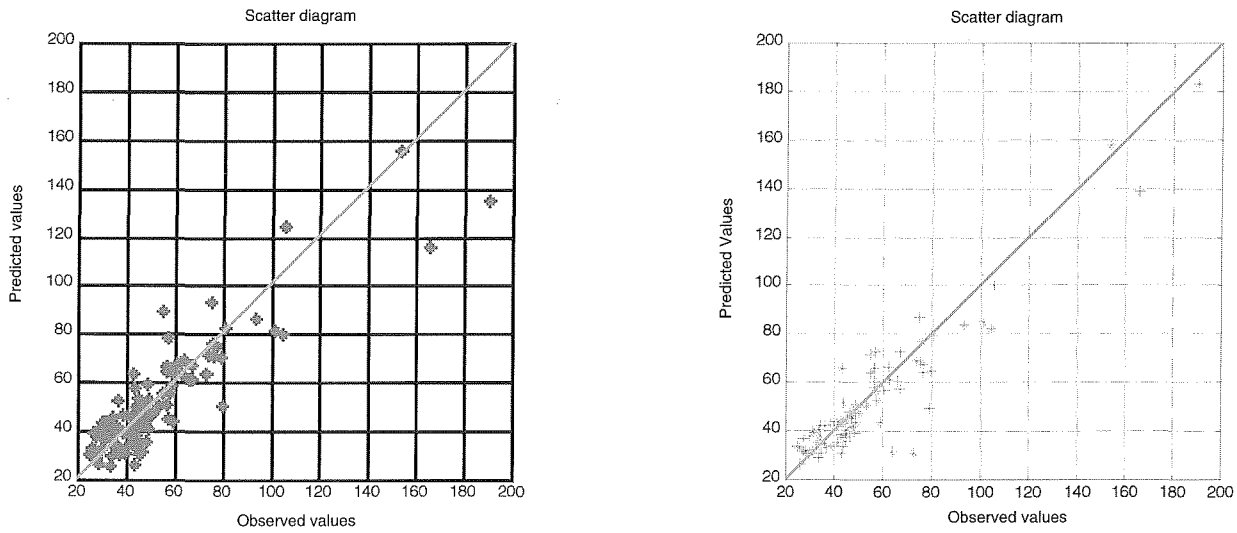

Figure 7. Scatter Diagrams (Cross-validation). Geostatistical approach (left) and finite mixture model (right)

Table IV

	Geostatistical approach	CWM approach
Correlation Coef.	0.894	0.937
Mean	52.97	51.20
St. Deviation	22.98	24.93

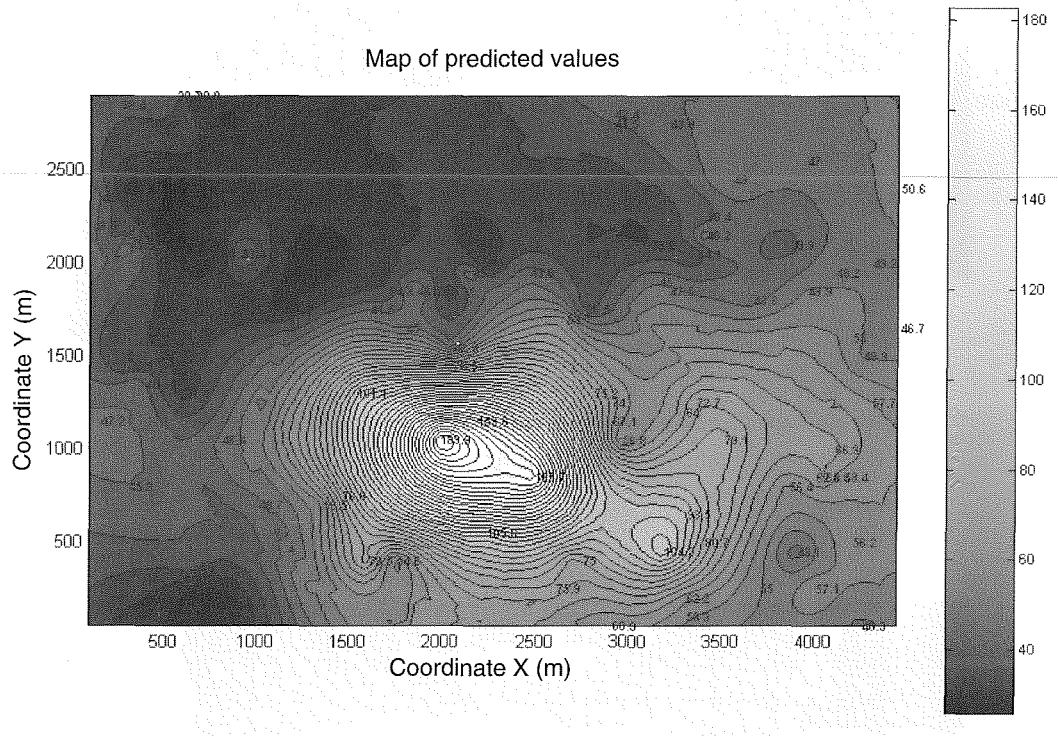

Figure 8. Map of predicted values for geostatistical approach (grid 100×100)

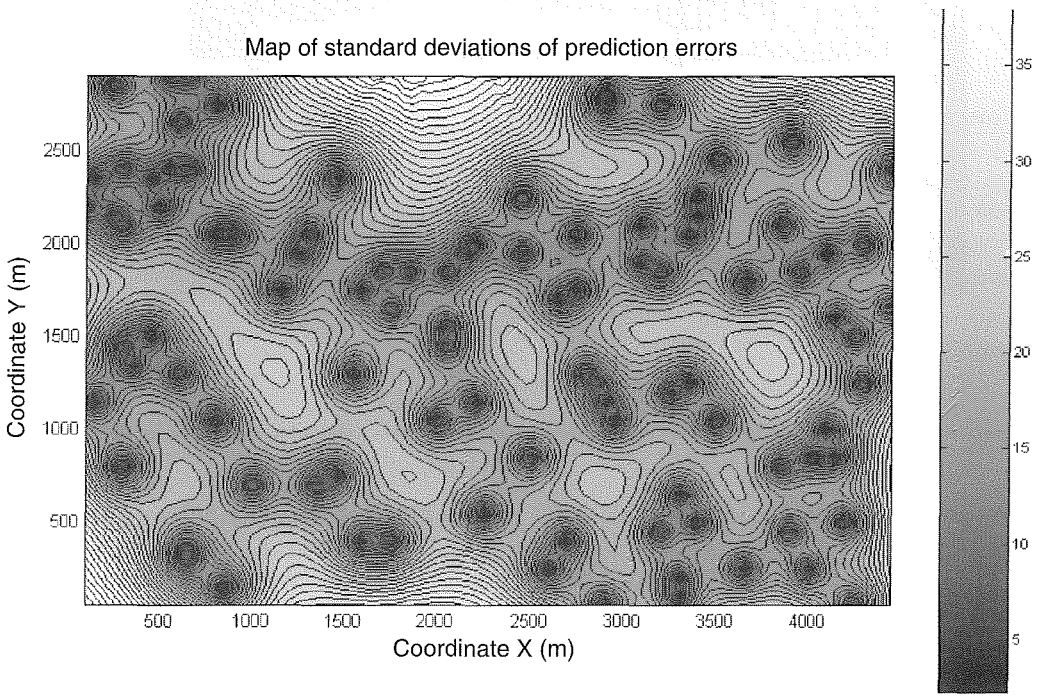

Figure 9. Map of standard deviations of prediction errors for geostatistical approach (grid 100×100)

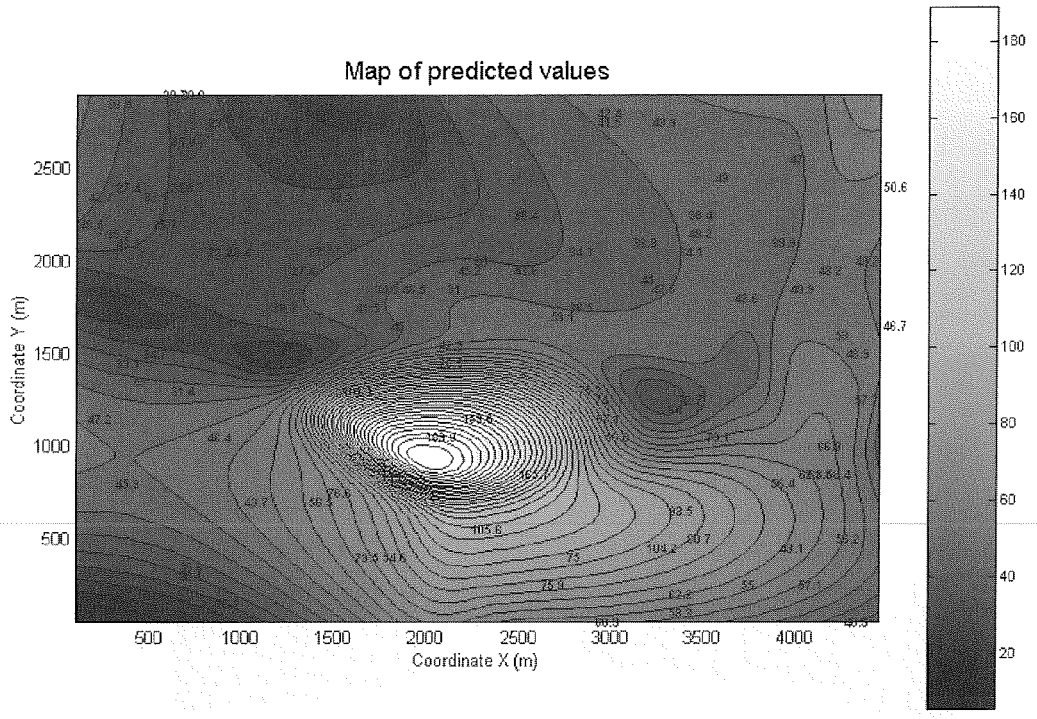

Figure 10. Map of predicted values for cluster weighted modelling (grid 100×100)

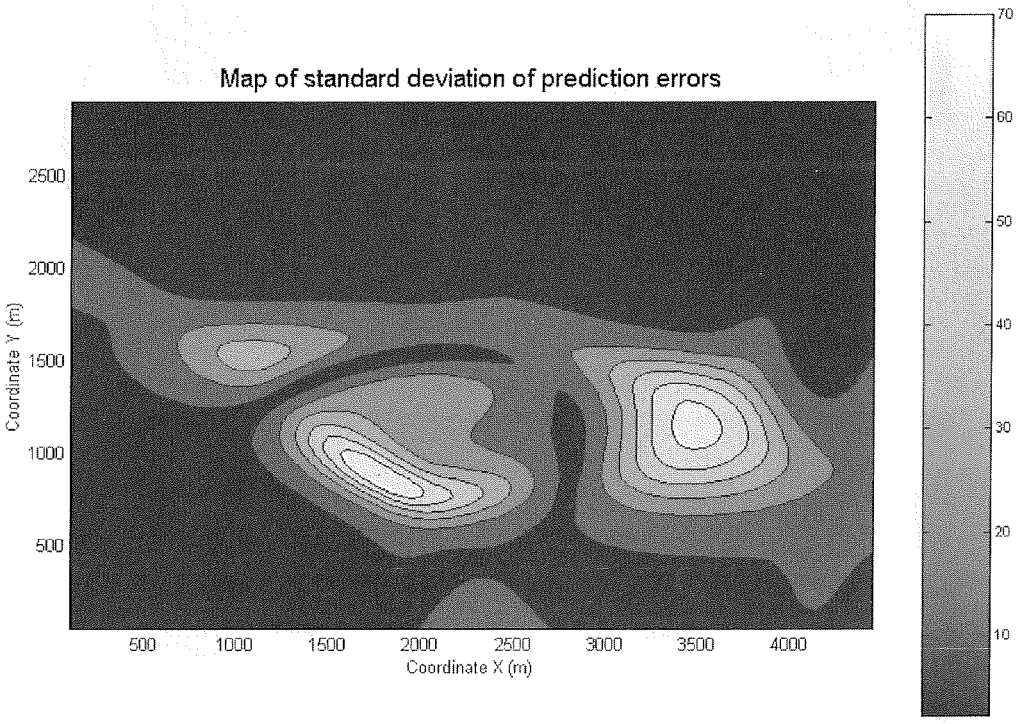

Figure 11. Map of standard deviations of prediction errors for cluster weighted modelling (grid 100×100)

Earth Sciences, *Mathematical Geology*, 1986, vol. 18, no. 1, pp. 119–140.

3. HENLEY, S. and WATSON, D.F. Possible Alternatives to Geostatistics, *Proceedings XXVII APCOM Conference*, London, IMM ed., 1998, pp. 337–353.
4. CORTEZ, L., SOUSA, A., DURÃO, F., ROGADO, J., and SIMÕES, J. A Neural Network Approach for Natural Resources Estimation, *Geostatistics'96*, Wollongong, E. Baafi and N. Schofield (Editors), Kluwer Acad. Publ., 1996, vol. 2, pp. 1149–1162.
5. CORTEZ, L. DURÃO, F., and SOUSA, A. Mineral Resources Estimation using Neural Networks and Geostatistical Techniques, *Proceedings XXVII APCOM Conference*, London, IMM ed., 1998, pp. 305–314.

6. CORTEZ, L., DURÃO, F., and SOUSA, A. Mineral Resources Estimation Methods: a Comparative Study, *Proceedings XXVIII APCOM Conference*, Golden, K. Dagdelen (Editor), 1999, pp. 425–434.
7. SPECHT, D. A General Regression Neural Network, *IEEE Trans. on Neural Networks*, 1991, vol. 2, no. 6, pp. 568–576.
8. CRESSIE, N. *Statistics for Spatial Data*, John Wiley & Sons, 1991, 900 pp.
9. KITANIDIS, P.K. *Introduction to Geostatistics*. Applications in Hydrogeology, Cambridge University Press, 1997, 249 pp.
10. SCHONER, B. Probabilistic Characterization and Synthesis of Complex Driven Systems, Ph. D. thesis, MIT ed., 2000, 204 pp.
11. GERSHENFELD, N. *The Nature of Mathematical Modeling*, Cambridge University Press, 1999, 344 pp.
12. DEUTSCH, R. *Estimation Theory*, Prentice-Hall Inc, 1965, 269 pp.
13. DEMPSTER, LAIRD and RUBIN, Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society*, Series B, vol.39, 1977, pp. 1–38.
14. TITTERINGTON, D.M., SMITH, A.F., and MARKOV, U.E. *Statistical Analysis of Finite Mixture Distributions*, John Wiley & Sons, 1985, 243 pp.
15. FLURY, B. *A First Course in Multivariate Statistics*, Springer-Verlag, 1997, 713 pp.
16. FIGUEIREDO, M.A.T., LEITÃO, M.N., and JAIN, A.K. On Fitting Mixture Models, in: *Energy Minimization Methods in Computer Vision and Pattern Recognition*, Hancock, E. and Pellilon, M. (Editors), 1999, pp. 54–69.
17. GRIFFITHS, P. and HILL I.D. (Editors), *Applied Statistics Algorithms*, Ellis Horwood Limited, 1985.
18. EFRON, B. and TIBSHIRANI, R.J. *An Introduction to the Bootstrap*, 1993, 436 pp.
19. RODRIGUES, J.A. Sistema de Modelização Geostatistica Aplicada aos Recursos Naturais, M.Sc. thesis, I.S.T., Lisbon, 2000.