Mathematical Geology, Vol. 36, No. 4, May 2004 (© 2004)

Comparison of Kriging and Neural Networks With Application to the Exploitation of a Slate Mine[1]

J. M. Matías,[2] A. Vaamonde,[3] J. Taboada,[4] and W. González-Manteiga[5]

To carry out an efficient and effective exploitation of a slate mine, it is necessary to have detailed information about the production potential of the site. To assist us in estimating the quality of slate from a small set of drilling data within an unexploited portion of the mine, the following estimation techniques were applied: kriging, regularization networks (RN), multilayer perceptron (MLP) networks, and radial basis function (RBF) networks. Our numerical results for the test holes show that the best results were obtained using an RN (kriging) which takes into account the known anisotropy. Differing deposit configurations were obtained, depending on the method applied. Variations in the form of pockets were obtained when using a radial pattern with RBF, RN, and kriging models while a stratified pattern was obtained with the MLP model. Pockets are more suitable for a slate mine, which indicates that the selection of a technique should take account of the specific configuration of the deposit according to mineral type.

KEY WORDS: kernels, kriging, neural networks, regularization, splines, slate.

INTRODUCTION

The spatial prediction of random functions from a realization of the process is extremely useful in disciplines such as geology, hydrogeology, meteorology, mining, etc. This problem has traditionally been tackled using methods such as kriging, inverse distance weighting, interpolating polynomials, splines, etc. The range of procedures has given rise, moreover, to comparative studies as to the suitability of the different methods in different contexts. Thus, Yakowitz and Szidarovszky

[1] Received 13 August 2001; accepted 21 January 2004.

[2] Department of Statistics, University of Vigo, 36200 Vigo, Spain; e-mail: jmmatias@uvigo.es

[3] Department of Statistics, University of Vigo, 36200 Vigo, Spain; e-mail: vaamonde@uvigo.es

[4] Department of Mining, University of Vigo, 36200 Vigo, Spain; e-mail: jtaboada@uvigo.es

[5] Department of Statistics, University of Santiago de Compostela, 15782 Santiago de Compostela, Spain; e-mail: wencesalo@usc.es

0882-8121/04/0500-0463/1 © 2004 International Association for Mathematical Geology

(1985), Weber and Englund (1992, 1994), and Zimmerman and others (1999), among others, describe theoretical and computational comparisons for a range of predictors.

Another method of prediction is the neural network, capable of recognizing and reproducing the internal structure of processes from observations. Neural networks have proven to be extremely useful in the resolution of a wide range of statistical problems and engineering applications, such as pattern recognition and the classification of data with multiple attributes. A general overview of these techniques, their computational problems, and statistical interpretation can be found in the literature (Bishop, 2000; Haykin, 1999; Ripley, 1996).

Of particular interest is the application of neural networks to the treatment of dependent data, given the difficulties involved in recognizing and modeling the underlying dependency structure. Chakraborty and others (1992) and Koike, Matsuda, and Gu (2001) are just some of the works in which different techniques of network design and training are described with a view to obtaining predictions of spatial or temporal processes that are as precise as possible.

This article describes the statistical bases for and the relationships between kriging and a variety of neural networks and highlights points in common and differences. This will facilitate a comprehension of the procedures as well as of the alternatives available when faced with our specific problem, namely the reconstruction of a slate deposit model using a sample set of borehole data.

The structure of the document is as follows.

1. We commence showing the formal reslationship between kriging and regularization networks (Girosi, Jones, and Poggio, 1995), as a generalization of the well-known relationship between kriging and splines (Cressie, 1993; Laslett, 1994; Wahba, 1990a). Both techniques result in the same formal solution even they start from different statistical hypothesis.
2. Originating from interpolation techniques and closely related to kernel smothers, a radial basis function (RBF) network is a linear smoother with fewer centers than data. We will describe an efficient method for RBF model selection based on the fundamental role that the width of the basic functions plays in the complexity of the network.
3. We introduce multilayer perceptron networks, a nonradial technique for multivariate regression, estimated using nonlinear optimization algorithms. The main problem with this method is the possibility of obtaining local optimums in the optimization problem. Bayesian model selection methods are a better choice than are other more heuristic methods.
4. We apply all these techniques to the problem of predicting the exploitability of a slate mine and compare the performance and complexity of each. Finally, we summarize our conclusions.

UNIVERSAL KRIGING

The stochastic hypotheses of universal kriging (Cressie, 1993; Journel, 1977; Matheron, 1973; Ripley, 1981; Ribeiro and others, 1997) can be formulated as described immediately below.

Let the random function of interest be

$$\mathsf{Z}(\mathrm{x}) = \sum_{i=1}^{m} \alpha_i g_i(\mathrm{x}) + \mathsf{U}(\mathrm{x}) \tag{1}$$

where $\mathrm{x} \in C \subset \mathbb{R}^d$, $\alpha = (\alpha_1, \ldots, \alpha_m) \in \mathbb{R}^m$ is an unknown parameter vector; $g_i : \mathbb{R}^d \to \mathbb{R}$, $i = 1, \ldots, m$ are known functions; and $\mathsf{U}(\mathrm{x})$ is a zero-mean stochastic process, second-order stationary with a covariance function:

$$\kappa(\mathrm{x}, \mathrm{x}') = \mathrm{Cov}(\mathsf{U}(\mathrm{x}), \mathsf{U}(\mathrm{x}')) = \begin{cases} k(\mathrm{x}, \mathrm{x}) + \sigma^2 & \text{if } \mathrm{x} = \mathrm{x}' \\ k(\mathrm{x}, \mathrm{x}') & \text{if } \mathrm{x} \neq \mathrm{x}' \end{cases}$$

where k is a conditionally positive definite function and where σ^2 (the nugget effect) may be zero.

The objective is to obtain an estimator $\hat{\mathsf{Z}}(\mathrm{x})$ for the process $\mathsf{Z}(\mathrm{x})$ from n observations $\{\mathsf{Z}(\mathrm{x}_i) = z_i\}_{i=1}^n$ (henceforth we will assume a fixed design for x).

The *universal kriging estimator* $\hat{\mathsf{Z}}(\mathrm{x}_0)$ of Z for a new point $\mathrm{x}_0 \in C$ is the linear estimator, $\hat{\mathsf{Z}}(\mathrm{x}_0) = \lambda^t \mathsf{Z}$, unbiased with minimum variance, where $\lambda \in \mathbb{R}^n$, $\mathsf{Z} = (\mathsf{Z}_1, \ldots, \mathsf{Z}_n)^t$ with $\mathsf{Z}_i = \mathsf{Z}(\mathrm{x}_i)$. The optimality conditions of the corresponding Lagrange problem are

$$\begin{cases} (K + \sigma^2 I)\lambda + Q\mu = \mathrm{k}_0 \\ Q^t\lambda - \mathrm{g}_0 = 0 \end{cases} \equiv \begin{bmatrix} K + \sigma^2 I & Q \\ Q^t & 0 \end{bmatrix} \begin{bmatrix} \lambda \\ \mu \end{bmatrix} = \begin{bmatrix} \mathrm{k}_0 \\ \mathrm{g}_0 \end{bmatrix} \tag{2}$$

where $(K)_{ij} = k(\mathrm{x}_i, \mathrm{x}_j)$, $(Q)_{ij} = g_j(\mathrm{x}_i)$, $(\mathrm{k}_0)_i = k(\mathrm{x}_0, \mathrm{x}_i)$, $(g_0)_j = \mathrm{g}_j(\mathrm{x}_0)$, and where $\mu \in \mathbb{R}^m$ is the Lagrange multiplier vector. We will denote these conditions more briefly as $\bar{K}\bar{\lambda} = \bar{\mathrm{k}}_0$. The solution to the problem is $\bar{\lambda} = \bar{K}^{-1}\bar{\mathrm{k}}_0$ and so the universal kriging estimator $\hat{\mathsf{Z}}_0 = \hat{\mathsf{Z}}(\mathrm{x}_0)$ of $\mathsf{Z}_0 = \mathsf{Z}(\mathrm{x}_0)$ can be alternatively written as

$$\hat{\mathsf{Z}}_0 = \lambda^t \mathsf{Z} = \bar{\lambda}^t \bar{\mathsf{Z}} = \sum_{i=1}^{n} \lambda_i \mathsf{Z}_i \tag{3}$$

$$= \bar{\mathrm{k}}_0^t \bar{\mathrm{c}} = \sum_{i=1}^{n} c_i k(\mathrm{x}_0, \mathrm{x}_i) + \sum_{j=1}^{m} d_j g_j(\mathrm{x}_0) \tag{4}$$

where $\overline{\mathrm{Z}} = (\mathrm{Z}^t, 0_{1\times m})^t$ with $0_{1\times m}$ the vector of m zeros, and $\overline{\mathrm{c}^t} = (\mathrm{c}^t, \mathrm{d}^t)$ with $\mathrm{c} \in \mathbb{R}^n$, $\mathrm{d} \in \mathbb{R}^m$. Using Equation (2), we immediately obtain the error variance:

$$\mathrm{Var}(\mathrm{Z}_0 - \hat{\mathrm{Z}}_0) = k(\mathrm{x}_0, \mathrm{x}_0) - \overline{\lambda}^t \overline{\mathrm{k}_0} \tag{5}$$

A particular case to be examined further below is when the functions g_j are polynomials of lesser degree than m.

The expression of the estimator in Equation (3) in terms of $\overline{\lambda}$ with the optimality conditions described above is known as the *primal* kriging formulation, whereas its expression in Equation (4) in terms of $\overline{\mathrm{c}} = \overline{K}^{-1}\overline{\mathrm{z}}$ where $\overline{\mathrm{z}} = (z^t, 0_{1\times m})^t$ with $\mathrm{z} = (z_1, \ldots, z_n)^t$ is known as the *dual* kriging formulation, where $\overline{\mathrm{c}}$ is the solution to the system $\overline{K}\overline{\mathrm{c}} = \overline{z}$:

$$\begin{cases} (K + \sigma^2 I)\mathrm{c} + Q\mathrm{d} = \mathrm{z} \\ Q^t\mathrm{c} = 0 \end{cases} \equiv \begin{bmatrix} K + \sigma^2 I & Q \\ Q^t & 0 \end{bmatrix} \begin{bmatrix} \mathrm{c} \\ \mathrm{d} \end{bmatrix} = \begin{bmatrix} \mathrm{z} \\ 0 \end{bmatrix} \tag{6}$$

In the above expressions, the term $\sigma^2 I$ guarantees the nonsingularity of the matrix $\overline{K}$ in a way analogous to ridge regression as a result of applying the regularization approach to the regression problem.

REGULARIZATION NETWORKS (SPLINES)

In the framework of the regression problem

$$\mathsf{Y} \equiv \mathsf{Y}(\mathrm{x}) = E(\mathsf{Y}/\mathrm{x}) + \varepsilon$$

with ε random noise, Girosi, Jones, and Poggio (1995) generalized the results of Wahba (1990a) with splines and defined the regularization networks (RNs) as those resulting from resolving the following regularization problem:

$$\min_{f\in\mathcal{H}} L(f) = \min_{f\in\mathcal{H}}\{\|\mathrm{y} - \mathrm{f}\|^2 + \lambda R(f)\} \tag{7}$$

where $\mathrm{y} = (y_1, \ldots, y_n)^t$ is a vector of independent and identically distributed observations for a set of points $\{\mathrm{x}_i\}_{i=1}^n$; $\mathrm{f} = (f(\mathrm{x}_1), \ldots, f(\mathrm{x}_n))^t$ is a vector of estimations for these points; λ is a regularizing constant; and R is a stabilizer (Tikhonov and Arsenin, 1977) defined in the family $\mathcal{H}$ of the functions under consideration.

Regularization converts an ill-posed problem (without a solution, without a unique solution, or nonstable) into a well-posed problem. In this respect, the first term in Equation (7) reflects the degree of realiability with which the estimator f reproduces the data whereas the second term penalizes its degree of complexity, thus endeavoring to stabilize the problem and equip it with a unique solution. The greater the regularizer λ the more importance is given to the smoothness of the function and the less importance is attached to the degree of fit. In the opposite case, $\lambda = 0$, the priority is the fit and if the family $\mathcal{H}$ is sufficiently rich then interpolation occurs.

It can be demonstrated (Girosi, Jones, and Poggio, 1995; Haykin, 1999, p. 273; Wahba, 1990a) that the space of possible functions takes the form:

$$\mathcal{H} = \{ f = \sum_i a_i k(\cdot, \mathrm{x}_i) : a_i \in \mathbb{R}, \mathrm{x}_i \in C\}$$

where $k(\cdot, \mathrm{x}_i)(\mathrm{x}) = k(\mathrm{x}, \mathrm{x}_i)$ and k is the Green function of an operator associated with the stabilizer R. If k is positive definite, the solution to the problem (7) is

$$\hat{f}(\mathrm{x}) = \sum_{i=1}^{n} c_i k(\mathrm{x}, \mathrm{x}_i)$$

where

$$\mathrm{c} = (K + \lambda I)^{-1} \mathrm{y} \tag{8}$$

and if, in general, the function k is conditionally positive definite of order m, then the solution takes the form:

$$\hat{f}(\mathrm{x}) = \sum_{i=1}^{n} c_i k(\mathrm{x}, \mathrm{x}_i) + \sum_{j=1}^{m} d_j g_j(\mathrm{x}) \tag{9}$$

where $\{g_j : g_j \in \prod_m(\mathbb{R}^d)\}_{j=1}^m$ is a basis of the space of polynomials of degree less than m. (If $m = 0$, we have the previous case.) The coefficients $\mathrm{c} = (c_1, \ldots, c_n)^t$ and $\mathrm{d} = (d_1, \ldots, d_m)$ result from the equations:

$$\begin{cases} (K + \lambda I)\mathrm{c} + Q\mathrm{d} = \mathrm{y} \\ Q^t \mathrm{c} = 0 \end{cases} \tag{10}$$

where the parameter λ contributes to the good conditioning of the matrix $(K + \lambda I)$.

The Smoothness of a Function

A link between the above formulation and the information that may be available a priori on the smoothness of the regression function to be estimated is obtained by means of a stabilizer proposed by Girosi, Jones, and Poggio (1995):

$$R(f) = \int_{\mathbb{R}^d} \frac{|\tilde{f}(\omega)|^2}{\overline{k}(\omega)}\, d\omega$$

where $\tilde{f}$ is the Fourier transform of f and $\overline{k}$ is a positive function, symmetrical and integrable with $\lim_{\|\omega\|\to\infty} \overline{k}(\omega) = 0$. The expression $R(f)$ is a measure of the smoothness of the function f in the frequency domain, given that from the Parseval identity

$$\int_{\mathbb{R}^d} |\hat{f}(\omega)|^2 d\omega = \int_{\mathbb{R}^d} |f(\mathrm{x})|^2\, d\mathrm{x}$$

is the *quantity of energy* of the function f that gives a measurement of its oscillatory behavior. Since the quotient $1/\overline{k}(\omega)$ is a filter of the low frequencies of f (*high pass filter*), $R(f)$ gives a measure of the energy of f at its higher frequencies.

Special cases of the above (Girosi, Jones, and Poggio, 1995; Wahba, 1990a) are smoothing spline, thin-plate splines and the RNs with Gaussian radial basic functions:

$$\overline{\mathrm{k}}(\omega) = e^{-\|\omega\|^2\sigma_k^2/2} \to k(\mathrm{x}-\mathrm{x}') = e^{-\|\mathrm{x}-\mathrm{x}'\|^2/2\sigma_k^2}$$

$$f(\mathrm{x}) = \sum_{i=1}^{n} c_i \exp\left(\frac{-1}{2\sigma_k^2}\|\mathrm{x}-\mathrm{x}_i\|^2\right)$$

A COMPARISON OF KRIGING AND REGULARIZATION NETWORKS

The comparison between kriging and RNs inherits many of the characteristics of the comparison between kriging and splines, a fact which has caused certain controversy (Cressie, 1989, 1990; Wahba, 1990b).

From a comparison of the dual equations (Eq. (6)) for kriging and for RNs (Eq. (10)) it can be concluded that universal kriging and RNs (splines) will have the same solution when the covariance function k of the process coincides with the kernel k of the latter and when the quantity of regularization applied to RN coincides with the nugget effect ($\lambda = \sigma^2$).

More specifically, the fact that the kernel k selected for RN is a conditionally positive definite function of order m is equivalent to considering, for kriging, an intrinsic random function of order m (equivalence class of processes with mean $\sum_{j=1}^{m} \alpha_j g(\mathrm{x}) \in \prod_m$; Matheron, 1973), with a generalized covariance function k of the same order. Consequently, the solution provided by RNs corresponds to the dual formulation for universal kriging:

$$\hat{Z}(\mathrm{x}) = \sum_{i=1}^{n} c_i k(\mathrm{x}, \mathrm{x}_i) + \sum_{j=1}^{m} d_j g_j(\mathrm{x}) = \hat{f}(\mathrm{x}) \tag{11}$$

In particular, RN with a kernel of the order $m = 1$ is equivalent to ordinary kriging, if the kernel is positive definite ($m = 0$) then the result is simple kriging.

Finally, for all the above cases, the absence of a nugget effect for kriging is equivalent to considering a null regularizing parameter $\lambda = 0$ for RN, in other words, giving precedence to the fit of the data in the regularization problem in Equation (7).

This formal equivalence between the two techniques permits the solution (Eq. (11)) to be interpreted in terms of two mutually dual approaches, which can be distinguished, fundamentally, by the proportion of variability that they assign to a (deterministic) trend component and to a stochastic component.

1. The first approach (regression using independent observations) assigns all the variability to the trend function and postulates a model that will be flexible enough to adapt to the data and smooth enough to reproduce the degree of regularity of the phenomenon. Moreover, the attempt to avoid the learning of possible noise in the data may impose additional smoothing conditions, via regularization techniques, whose intensity (the regularizing parameter) is estimated a posteriori using model selection techniques (cross-validation, bootstrapping, etc.)
2. The second approach (prediction of a random function) distributes the variability of the phenomenon between the trend component and the stochastic component, firstly modeling the latter from the data (variogram or covariogram) and assuming a parametric model for the former.

In general, for the latter case, the trend model is simpler than in the former since a significant part of the variability in the phenomenon is assigned to the stochastic component. Nonetheless, the final trade-off between the two components depends on prior information being available about the problem under consideration.

Under the above-mentioned duality, the term $\sum_{i=1}^{n} c_i k(x, x_i)$ in the expression (11) can be seen as a term expressing trend (RN) or as a term expressing dependency (kriging), and in both cases, the analytical properties of k determine the degree of regularity in the estimation.

The term $\sum_{j=1}^{m} d_j g_j(x)$ from the expression (11), on the other hand, constitutes, for both approaches, a trend term (even if its use in RNs is infrequent, given that these networks do not need it to approximate any continuous function; Poggio and Girosi, 1989).

A synthesis between the two methods is given by the Bayesian approach in the regression problem $\mathsf{Y}(\mathrm{x}) = E[\mathsf{Y}(\mathrm{x})/\mathrm{x}] + \varepsilon$ (e.g., MacKay, 1998), where ε is random independent noise with a zero mean and where the regression function is subject to uncertainty and is modeled as a random function $\mathsf{F}(\mathrm{x}) = E[\mathsf{Y}(\mathrm{x})/\mathrm{x}]$. In this context it makes sense to consider the dependency measure $\mathrm{Cov}(\mathsf{F}(\mathrm{x}), \mathsf{F}(\mathrm{x}'))$, which can be estimated using an estimator of

$$\mathrm{Cov}(\mathsf{Y}(\mathrm{x}), \mathsf{Y}(\mathrm{x}')) = \mathrm{Cov}(\mathsf{F}(\mathrm{x}) + \varepsilon, \mathsf{F}(\mathrm{x}') + \varepsilon') = \mathrm{Cov}(\mathsf{F}(\mathrm{x}), \mathsf{F}(\mathrm{x}'))$$

This fact would justify using, in RNs and RBFs, a kernel estimated from data as in spatial statistics, the benefits of which we will evaluate below in our application problem.

Three major differences exist between kriging and RNs in their normal application.

1. The first difference resides in the selection method for the k function. In kriging this function (or the variogram) is estimated from the data using proven methods. For RNs, however, the kernel is selected assuming a smoothness hypothesis for the regression function. Nonetheless, in practice it is often difficult to codify the prior information in these terms. For this reason, standard kernels, such as the Gaussian, multiquadric, or polynomial functions (splines) are often used, despite the lack of a strong empirical basis, in the belief that the choice of kernel has less influence than the estimation of the parameters for the corresponding family.

   That said, as we shall see in our applications this practice is somewhat risky, and it does not take advantage of the information contained in the data to formulate a hypothesis in regard to the kernel.

   Probably one of the circumstances giving rise to this habit is the frequent high dimensionality of the application problems typical of the neural networks, in which the above task is extremely difficult to perform.
2. The second major difference lies in the fact that because of the statistical hypotheses governing kriging, this—unlike RN—provides the error variance (Eq. (5)) (even though splines can be formally seen as a particular case of kriging; Laslett, 1994).

   RNs (and RBFs) are linear smoothers, and this permits the variance of the estimator of the regression function to be estimated. However, the variance of the predictor cannot be estimated, unless the noise variance is estimated by another method.

3. Finally, we should point out kriging's capacity to predict specific functions of the random process in the framework of different supports (Cressie, 1990), a facility not shared with RNs.

The characteristics of our application problem do not permit an evaluation of the impact of the latter two aspects, but they do permit the effects of the first to be tested experimentally.

A priori, these circumstances tip the balance in favor of the spatial statistical techniques when the dimensionality of the input space is no greater than 3 (which occurs in the application problem described below). For problems of greater dimensionality, the estimation techniques for the dependency structure, although desirable, are somewhat more problematic.

RBF NETWORKS

RBF networks essentially arose in the context of interpolation technique, (Broomhead and Lowe, 1988; Moody and Darken, 1989), although they were also inspired by kernel-type smoothing techniques (Schiöler and Hartmann, 1992). They were intended as a means of applying the said techniques to the problem of regression with observations subject to noise for which data interpolation was not appropriate. The RBF model is very similar to kriging and the regularized networks, but its training is different, giving rise to estimators different from those described above.

The radial basis network is an expansion in basic functions, in general, non orthogonal:

$$f(\mathrm{x}) = \sum_{j=1}^{m} c_j k_j(\|\mathrm{x} - \mu_j\|) + b$$

where $b, c_j \in \mathbb{R}$ and each $k_j : \mathbb{R}^+ \to \mathbb{R}$ is a function of the distance between x and its *center* μ_j. In general, $m \ll n$, and so this model tends to be more parsimonious than the ones described above. Moreover, if $k_j = k\, \forall j$ and k is a density function, if $k_j(\|\mathrm{x} - \mu_j\|)/\sum_{j=1}^{n} k_j(\|\mathrm{x} - \mu_j\|)$ are considered as normalized distance, if the centers coincide with the data $\mu_j = \mathrm{x}_j$, $j = 1, \ldots, n$, and if the coefficients are prefixed: $c_j = y_j$, $j = 1, \ldots, n$, then we have the traditional kernel estimator, which has the advantage of being capable of immediate training and the disadvantage of having as many basic functions as data.

For its good smoothing properties the basic functions k_j are frequently chosen as Gaussian:

$$k_j(h) \propto \exp\left(-\frac{1}{2\sigma_j^2} h^2\right)$$

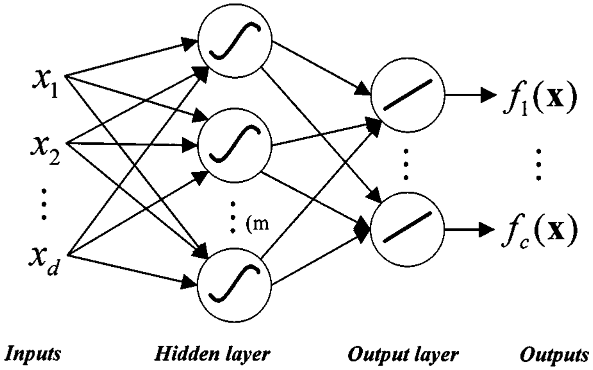

Figure 1. Feed-forward neural network with d input units, m neurons in the hidden layer, and c neurons in the output layer.

but other classical basic functions (multiquadric, etc.) from the interpolation field are also used. Below we will assume basic Gaussian functions all with the same radius σ_r.

In neural network terms RBF is a feed-forward network with a hidden layer (Fig. 1), the units of which possess as an activation function the basic functions k_j, in addition to an output layer with a linear activation function with coefficients c_j and b.

The different approaches for training the RBF networks (see, for an overall view, for example, Bishop, 2000; Haykin, 1999) are basically distinguished on the basis of being performed in a single phase by estimating all the parameters at once, or in two phases, by estimating first the parameters in the hidden layer and then the parameters in the output layer. The latter approach has the disadvantage that the estimation of the hidden layer is performed without taking into account the values of the response variable. Of the first approach, of note are the algorithms which perform nonlinear optimization of all the parameters simultaneously and those in which the output linear regression is performed by means of a forward selection of variables via orthogonal lest squares (Chen, Cowan, and Grant, 1991) introducing centers gradually from the sample data. The first method has the problem of local minimums; the second, although also suboptimal, is often preferable because of being less costly in computational terms.

The main difficulty with the RBF networks is model selection, in other words, the determination of the number of basic functions (or centers). To facilitate the task of model selection, we use the training algorithm described as follows: if, for example, basic Gaussian functions are used, their width σ_r determines the degree of redundancy of the hidden variables and so the final selection of this parameter determines the number of basic functions required. For this reason the algorithm starts off from a prefixed set A of possible radii σ_r and for each $\sigma_r \in A$ the schematic development is as follows.

1. *Determination of the number of basic functions*. Calculate K with $K_{ij} = k(\|\mathrm{x}_i - \mathrm{x}_j\|), i, j = 1, \ldots, n,$ obtain the eigenvalues $\{\alpha_j\}$ and, once ordered, calculate

$$l = \min\left\{ j \in \{1, \ldots, n\} : \mathrm{cond}(K) = \frac{\alpha_n}{\alpha_j} < \mathrm{cond}_0 \right\}$$

where cond_0 is a condition number fixed a priori that ensures the nonsingularity of the matrix $K^t K$ at the software precision level. Finally, select $m_r = n - l$ as the number of basic functions associated with σ_r.
2. *Training*. Apply orthogonal least squares forward selection to the linear regression of the output level. The solution is the classical expression of linear regression in terms of the hidden variables:

$$\hat{\mathrm{y}} = K^t (K^t K)^{-1} K^t \mathrm{y} = S\mathrm{y} \tag{12}$$

where $(K)_{ij} = k(\|\mathrm{x}_i - \mathrm{x}_j\|), i = 1, \ldots, n, j = 1, \ldots, m_r$.

Once trained the RBF network is a linear smoother, the equivalent kernels of which are the rows of the *hat matrix S*. The trace of the matrix provides a measure of its complexity (*effective number of parameters*). Thus, model selection methods can be applied that are available in analytical form, such as crossed validation, Bayesian information criterion (BIC), etc., or else more recent model combination methods.

MULTILAYER PERCEPTRON (MLP) NETWORKS

It would be impossible to summarize adequately here the mathematical, statistical, and computational aspects of the MLP neural networks, and we therefore refer the reader to the excellent monographs existing in the literature (e.g., Bishop, 2000; Haykin, 1999).

Focussing on the networks with one hidden layer and a single output, the MLP model—a special case of the feed-forward multilayer network—has the following formulation,

$$f(\mathrm{x}) = \psi\left(\sum_{j=1}^{m} c_j \phi\left(\mathrm{u}_j^t \mathrm{x} + u_{j0}\right) + c_0\right)$$

where ϕ is the activation function for the units in the hidden layer—generally sigmoid (logistic, hyperbolic tangent, etc.)—and ψ is the activation function for the output level, which may be of the Heaviside type for a classification problem, or sigmoid or linear for a regression problem.

The model selection problem for the MLP networks consists of the selection of an optimum number of basic functions that provide a maximum generalization capacity. Among the various techniques proposed, we highlight the Bayesian method implemented by Foresee and Hagan (1997), following on from MacKay (1992), which depends to a lesser extent on expert criterion. This method is really a regularization technique, the objective function of which—under a hypothesis of normality in the data and in the prior distribution of the parameters—can be viewed from a Bayesian perspective as the logarithm of the posterior distribution to be maximized (apart from a constant):

$$\ln p(\mathrm{w}/D) \equiv L(\mathrm{w}) = \frac{\beta}{2}\|\mathrm{y} - \mathrm{f}\|^2 + \frac{\alpha}{2}\sum_{j=1}^{m} w_i^2$$

where D represents the data, w is the vector of all the parameters of the network and α, β are the hyperparameters that are selected by means of maximization of the *evidence* $p(D/\alpha, \beta)$ (MacKay, 1992),

$$p(\alpha, \beta/D) \propto p(D/\alpha, \beta)p(\alpha, \beta)$$

where $p(\alpha, \beta)$ is the prior distribution of the hyperparameters which is assumed to be noninformative.

APPLICATION OF THE DIFFERENT TECHNIQUES TO A SLATE MINE

Objectives and Methodology

The techniques described above were applied to the estimation of the exploitability of a slate mine on the basis of a set of 10 continuous-core sampling boreholes (Fig. 2), which uncovered the exploitable levels of slate (Taboada, Saavedra, and Vaamonde, 2001). The cores had a diameter of 63 mm and were perfomed for each 0.5 m, to evaluate the quality variables of the rock bed.

To determine exploitability, the geotechnical quality of the rock bed was analyzed (García and others, 1998), in other words, the possibility of obtaining blocks of slate sufficiently large to make the process of producing roofing slate economically viable, a conditioning factor common to all ornamental

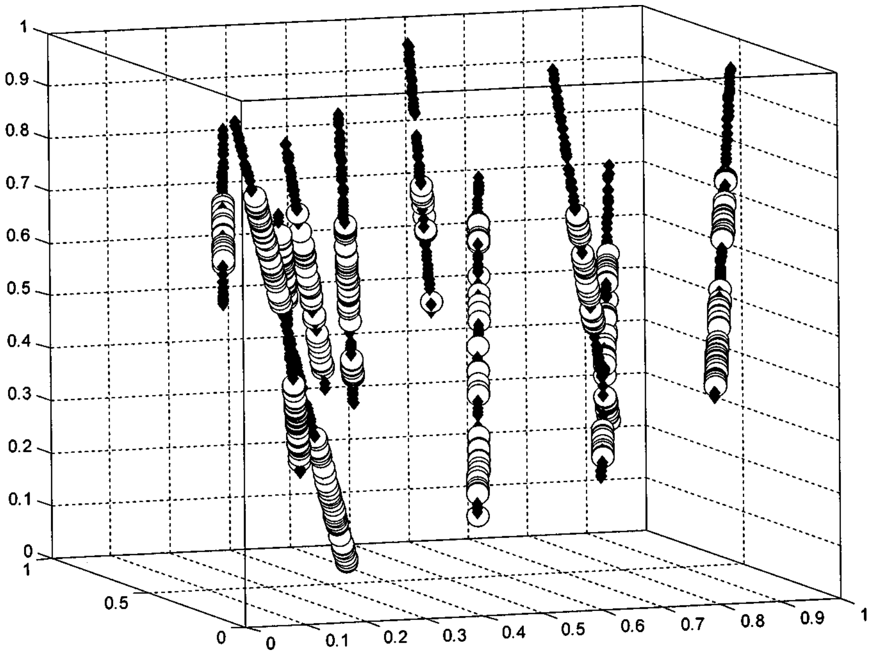

Figure 2. Exploitability training sample. The circles indicate exploitability (Y(x)= 1). The axes (scaled to [0,1]) represent: X east, Y north, and Z, deposit depth.

rock-quarrying activities (Taboada and others, 1997). The variables considered in this case (Taboada and others, 1998) were rock quality designation (RQD), fractures, sandy intercalations, surface alterations, crenulation schistosity, kink bands, and quartz veins.

As a result of the drilling, available was a sample $\{(\mathrm{x}_i, y_i)\}_{i=1}^{N}$ with $N = 1932$ observations from the 10 boreholes (Fig. 2), where $\mathrm{Y(x)} \in \{0, 1\}$ indicates if the slate is exploitable ($\mathrm{Y(x)} = 1$) or nonexploitable, and where $\mathrm{x} \in \mathbb{R}^3$ identifies the geographic location of the observation.

The data for each borehole were assigned aleatorily to two sets, the first a training sample of size $n = 1000$ and the second a test sample of size $n' = 932$ for evaluating the quality of the estimation provided by each model.

In this context, the main objectives of the study were as follows:

1. to compare the different estimations of the deposit produced by kriging and the neural networks, in terms of both performance and morphology;
2. to compare the estimations of the nugget effect σ^2 in kriging and of the regularizer λ in RNs, given their equivalence, and bearing in mind the different methods of estimation used in each case: parametric estimation

of the dependency structure for the kriging and cross-validation for RNs; and

3. to evaluate the benefits of using kernels estimated using spatial statistical methods in the RN and RBF neural networks.

Given the discrete character of the variable $\mathsf{Y}(x) \in \{0, 1\}$, the estimators $\hat{f}(\mathrm{x})$ obtained using neural networks were used to estimate the a posteriori probabilities $\mathsf{P}(\mathsf{Y}(\mathrm{x}) = 1/\mathrm{x})$ (Bishop, 2000; Ripley, 1996) that each observation was exploitable. The following decision rule was subsequently adopted:

$$\hat{\mathsf{Y}}(\mathrm{x}) = \begin{cases} 1 & \text{if } \hat{f}(\mathrm{x}) > \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

where $\hat{f}(\mathrm{x}) = \hat{\mathsf{P}}(\mathsf{Y}(\mathrm{x}) = 1/\mathrm{x})$ is the estimation of the said probabilities by each model.

As a measure of performance the classification error for the test sample was used:

$$\text{Test Error} = \frac{1}{n'} \sum_{i=1}^{n'} 1_{\{\hat{\mathsf{Y}}(\mathrm{x}_i) \neq \mathsf{Y}(\mathrm{x}_i)\}}$$

For kriging, the same methodology can be based as follows in the context of indicator kriging (Cressie, 1993, pp. 281–283): the exploitability of the slate is determined by evaluating 12 variables $\{\xi_i, (x)\}_{i=1}^{12}$ that reflect the characteristics of the slate that negatively affect its exploitability (slate is not characterized by grades or continuous variables typical of other mining scenarios). Of these, 11 are discrete variables (many of them indicators) that reflect the existence or otherwise of specific macroscopic properties, and the 12th variable is a continuous variable (RQD) that reflects the degree of fragility of the slate. These 12 variables are evaluated together by an expert with a view to determining the exploitability of the slate, either visually or using physical tests.

Therefore, the existence of an underlying process $\mathsf{Z}(\mathrm{x})$ can be postulated, continuous but unknown, a function of the previous set of variables:

$$\mathsf{Z}(\mathrm{x}) = g(\xi(\mathrm{x}))$$

with $\xi(\mathrm{x}) = (\xi_i(\mathrm{x}), \ldots, \xi_{12}(\mathrm{x}))^t$. This function g will reflect the expert decision process, depending on the values for the vector $\xi(\mathrm{x})$. We will not go into the estimation of the g function here, given that the above variables ξ_i were not defined specifically for this particular purpose. This classification problem will be tackled in future research, once the variables ξ_i have been carefully redefined.

In the above context, the observed process $\mathsf{Y}(\mathrm{x})$ is defined as

$$\mathsf{Y}(\mathrm{x}) = \begin{cases} 1 & \text{if } \mathsf{Z}(\mathrm{x}) \leq z_0 \\ 0 & \text{if } \mathsf{Z}(\mathrm{x}) > z_0 \end{cases}$$

in such a way that $E[\mathsf{Y}(\mathrm{x})/\mathrm{x}] = \mathsf{P}[\mathsf{Y}(\mathrm{x}) = 1/\mathrm{x}]$ is unknown. In this way the application of indicator kriging to the process $\mathsf{Y}(\mathrm{x})$ produces (Cressie 1993, p. 282) estimators $\hat{\mathsf{Y}}(\mathrm{x})$ of the probabilities:

$$\mathsf{P}[\mathsf{Z}(\mathrm{x}) \leq z_0/\mathsf{Y}(\mathrm{x}_1), \ldots, \mathsf{Y}(\mathrm{x}_n)] = \mathsf{P}[\mathsf{Y}(\mathrm{x}) = 1/\mathsf{Y}(\mathrm{x}_1), \ldots, \mathsf{Y}(\mathrm{x}_n)]$$

in which we have applied the definition of the variable $\mathsf{Y}(\mathrm{x})$. Thus, a criterion available for decision making as to exploitability is the estimator:

$$\hat{\mathsf{Y}}(\mathrm{x}) = \hat{\mathsf{P}}[\mathsf{Y}(\mathrm{x}) = 1/\mathsf{Y}(\mathrm{x}_1), \ldots, \mathsf{Y}(\mathrm{x}_n)]$$

through the decision *plug-in* rule:

$$\tilde{\mathsf{Y}}(\mathrm{x}) = \begin{cases} 1 & \text{if } \hat{\mathsf{Y}}(\mathrm{x}) > \frac{1}{2} \\ 0 & \text{if } \hat{\mathsf{Y}}(\mathrm{x}) \leq \frac{1}{2} \end{cases}$$

In this framework, if the estimated covariogram for $\mathsf{Y}(\mathrm{x})$ does not possess discontinuity at the origin, the kriging predictor $\hat{\mathsf{Y}}(\mathrm{x})$ becomes an interpolator for the points of the sample. If, on the other hand, the estimated covariogram is discontinuous at the origin, $\hat{\mathsf{Y}}(\mathrm{x})$ could lose this interpolating property, although the resulting plug-in estimator $\tilde{\mathsf{Y}}(\mathrm{x})$ may well keep it.

Modeling Spatial Variability

To be able to use the estimated covariogram as the kernel for the radial neural networks, we first estimated the following isotropic Gaussian-exponential covariogram model:

$$C(h) = \begin{cases} 0.23 - 0.104(1 - \exp(-h/0.0083)) - 0.126(1 - \exp(-(h/0.043)^2 \text{ if } h>0 \\ 0.26 \text{ if } h = 0 \end{cases} \tag{13}$$

which, as can be observed, has discontinuity at the origin of a magnitude of 0.03.

Nonetheless, in view of the observations of the referees, a subsequent anisotropic study was made so as to compare kriging with the neural networks using all the predictor potential of the former.

Slate is a particularly anisotropic material, given that it originates from a process of regional metamorphism. Slate is used in roofing because it has a schistosity plane, which, for the deposit studied, coincides with the azimuth N120°E and has a slope of 60°S. Figure 3 illustrates the experimental variograms obtained for the three anisotropic directions of the slate: the direction of maximum weakness (L1) (azimuth N120°E, dip 0°), the schistosity dip (N210°, dip 60°) and the direction prependicular to the schistosity plane (N210°E, dip −30°). The variogram range is observed to be greatest in the direction N120°E, i.e., the direction of maximum weakness (L1).

On the basis of this data, the following nested model was considered,

$$\gamma(\mathrm{h}) = \gamma_0(|\mathrm{h}|) + \gamma_1(|\mathrm{h}_1|) \qquad (14)$$

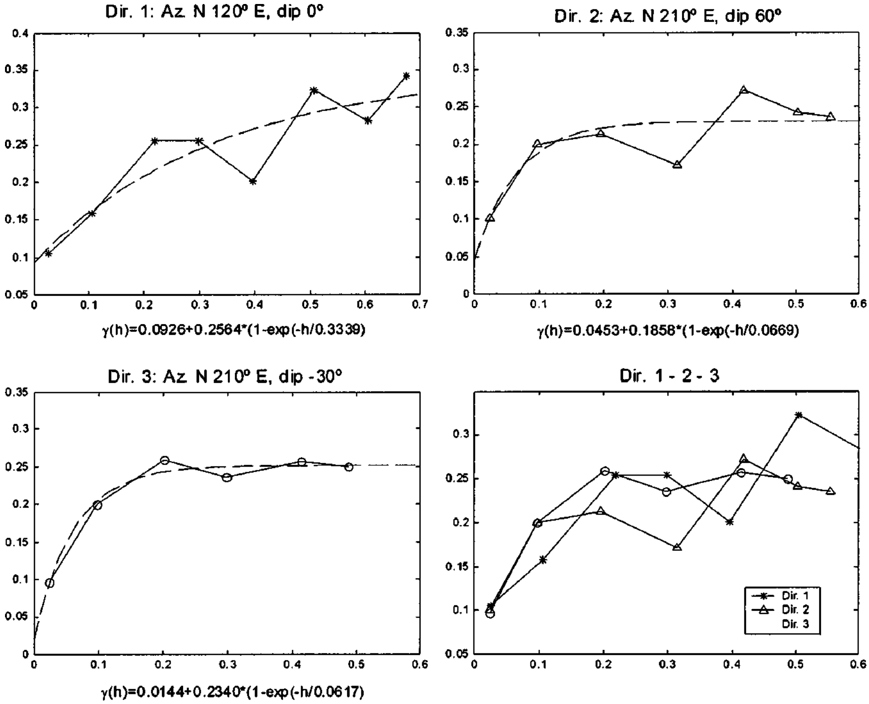

Figure 3. Experimental variograms in the main anisotropic directions of the slate: [N120° E, dip 0°], [N210°, dip 60°] and [N210° E, dip −30°], and overall representation.

where $\gamma_0(|\mathrm{h}|) = 0.07$ is the nugget effect and $\gamma_1(|\mathrm{h}_1|)$ is a model with geometric anisotropy, exponential in nature, with a sill of 0.25 and ranges of 0.3339, 0.0669, and 0.0617 in the directions [N120°E, dip 0°], [N210°E, dip 60°], and [N210°E, dip −30°], respectively.

RESULTS

The models used for the comparison were as follows.

1. Ordinary kriging (OK) using the estimated anisotropic variogram (Eq. (14)).
2. Ordinary kriging using the isotropic covariogram with nuggest effect (Eq. (13)).
3. Regularized neural network using as kernel the function:

$$\begin{aligned} C'(h) = 0.23 - 0.104(1 - \exp(-h/0.0083)) \\ - 0.126(1 - \exp(-h/0.043)^2) \end{aligned} \tag{15}$$

   in other words, the isotropic covariogram of Equation (13) without the nugget effect $\sigma^2 = 0.03$. The nugget effect $\sigma^2 = 0.03$ was not included here with a view to comparing the value of this parameter with that of regularizer λ estimated using 10-fold cross-validation (see below) which finally resulted in $\lambda = 0.04$. This RN is equivalent to an OK with the same covariogram.
4. Regularized neural network with a Gaussian kernel of radius $\sigma_r = 0.022$ and with a regularizer value $\lambda = 0.9$. Both the radius and the regularizer were selected by means of 10-fold cross-validation. This model is equivalent to ordinary kriging with the same Gaussian covariogram plus a nugget effect of $\sigma^2 = \lambda = 0.9$.
5. Radial basis neural network with Gaussian basic functions having the same width $\sigma_r = 0.09$, with the model (width) selection also made by means of 10-fold cross-validation.
6. Radial basis neural network using the isotropic covariogram in Equation (13) as the basic function.
7. Multilayer perceptron network containing 14 sigmoid hidden units, with Bayesian training. Although our problem is essentially a classification problem, we preferred to use a regression focus for comparing the MLP network to the other techniques in equal conditions. Because of its simplicity, we selected a linear activation function at the output level and least squares loss.

The s-fold cross-validation method consists of randomly allocating the training data to s groups of more or less equal size and calculating, for the data in each group, the performance of the estimator trained using the data of the remaining $s-1$ groups. if $s = n$, this is the leave-one-out method of cross-validation. More specifically, if $g : \{1, \ldots, n\} \to \{1, \ldots, s\}$ is the allocation of the points to the s groups, and $\tilde{y}^{-\gamma}(\mathrm{x})$ is the estimation obtained without the data of the γth group, the selection criterion is

$$s\text{-fold-CV} = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \tilde{y}^{-g(i)}(\mathrm{x}_i))$$

where ℓ is the loss function. In our case, $\ell(y_i, \tilde{y}^{-g(i)}(\mathrm{x}_i)) = 1_{\{y_i \neq \tilde{y}^{-g(i)}(\mathrm{x}_i)\}}$ is the loss $0-1$.

As can be observed, models 4, 5, and 7 above do not use prior information nor the information that may be contained in the data on the spatial variability of the deposit (as it happens, models 4 and 5 use erroneous information). Models 2 and 3 only use this information partially.

The results obtained are reproduced in Table 1, where for each of the models the following data is shown: the *hyperparameters* used (*Hyperparam*), the error in the test sample (*Test Error*), the number of basic functions (*NBF*), the effective number of parameters (*ENP*), and the interpolating character of the plug-in estimator $\tilde{\mathrm{Y}}$ (x) (*Int*). In view of the aims of this study, the following comments are in order.

1. First of all, regarding the test error, OK with an anisotropic variogram produced interesting results, which would indicate the benefits of an appropriate estimation of the spatial dependency structure. The improvements are evident with respect to kriging without anisotropy and above all, in

Table 1. Performance of the Estimators

Model	Hyperparam.	Test error	NBF	ENP	Int.
Ordinary kriging – anisotropic variogram	Eq. (14)	0.108	1000	1000.0	Yes
Ordinary kriging – isotropic covariogram	Eq. (13); $\sigma^2 = 0.03$	0.117	1000	697.1	Yes
RN – isotropic covariogram	Eq. (15); $\lambda = 0.04$	0.114	1000	643.3	Yes
RN – gaussian kernel	$\sigma_r = 0.022, \lambda = 0.9$	0.121	1000	311.3	No
RBF – gaussian kernel	$\sigma_r = 0.09$	0.127	163	164.0	No
RBF – isotropic covariogram	Eq. (15)	0.119	1000	1000.0	Yes
MLP (14 hidden units)	—	0.118	14	64.5	No

Note. The columns represent, in this order, the hyperparameters for the model, the error in the test sample, the number of resulting basic functions, the effective number of parameters (see text), and the interpolating character of the resulting estimator.

relation to the radial neural networks, whose kernels are selected a priori, with no particular justification for this decision.

For the case of the RN and the RBF networks, it is interesting to observe the positive effect of using an estimated isotropic covariogram.

In the specific case of RBF, the good results produced by the use of the estimated omnidirectional covariogram were surprising, given that all the sample data were selected as centers in the feed-forward linear regression of the output level, which is normally accompanied by a strong over-fit effect. In fact, RBF with covariogram was an interpolator of the data, unlike what usually happens with this kind of networks. Thus, with as many centers as data, this RBF with the estimated isotropic covariogram produces exactly the same estimator as simple kriging without noise in the data:

$$\hat{\mathrm{y}} = K\mathrm{c} \quad \text{when} \quad \mathrm{c} = (K'K)^{-1}K'y = K^{-1}y$$

with $K_{ij} = k(\mathrm{x}_i, \mathrm{x}_j) = C(\|\mathrm{x}_i - \mathrm{x}_j\|)$.

2. The MLP network produced satisfactory results in terms of test error, speed of training, and parsimony (see below). Nonetheless, if initial random conditions are used, each new training session can produce a distinct solution, and the estimation of the exploitable areas may thus result somewhat arbitrary. An interesting line of research could be the incorporation of prior information in these networks, for example, establishing the main direction beforehand in training.
3. It is interesting to observe how RNs with isotropic covariogram and $\lambda = 0.04$ (equivalent to isotropic kriging with a nugget effect of 0.04) improve the results of the kriging with a nugget effect of 0.03, suggesting that this parameter had initially been undervalued. This would indicate that it might be a good idea to occasionally contrast the original estimation of this crucial parameter using an a posteriori model selection technique, such as, cross-validation.
4. Analyzing the exploitability maps for the deposit (Figs. 4 and 5 for the Gaussian RBF and the MLP), we can divide the techniques used into two groups: those, such as the MLP networks, which use projection-type basic functions to produce a reconstructed deposit in stratified form; and those, such as the RBF networks and kriging, based on a distance function (whether kernel or covariogram), which produce reconstructions in pocket form.

   The case of anisotropic kriging (Fig. 6) is particularly interesting since it incorporates anisotropy in its representation of the deposit. This kind of estimator could be considered as an intermediate case between the radial configuration and projection configuration, the latter occurring at the limit.

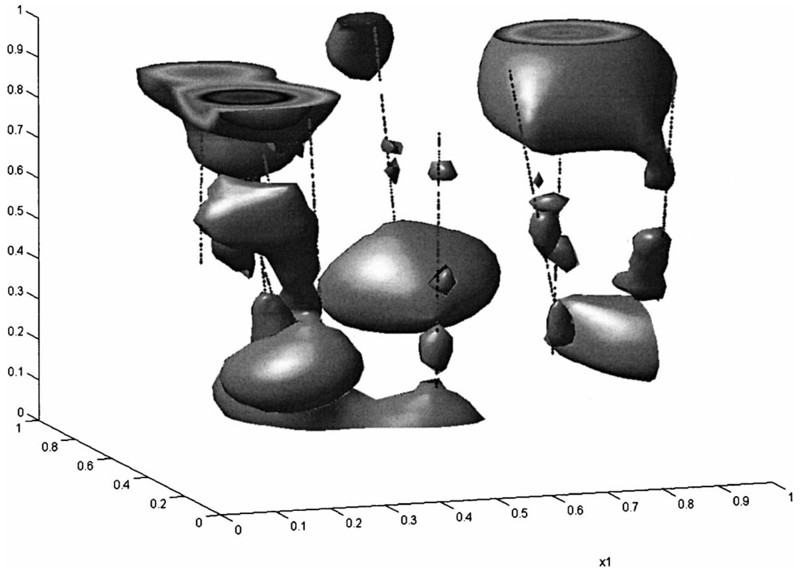

Figure 4. Estimation of the exploitable zones using Gaussian RBF. The image shows the regions in which $\hat{f}(X) \leq 1/2$, where $\hat{f}$ is the estimator of $\mathsf{P}[\mathsf{Y}(\mathrm{x}) = 1/\mathrm{x}]$ provided by RBF. Test error: 0.12661.

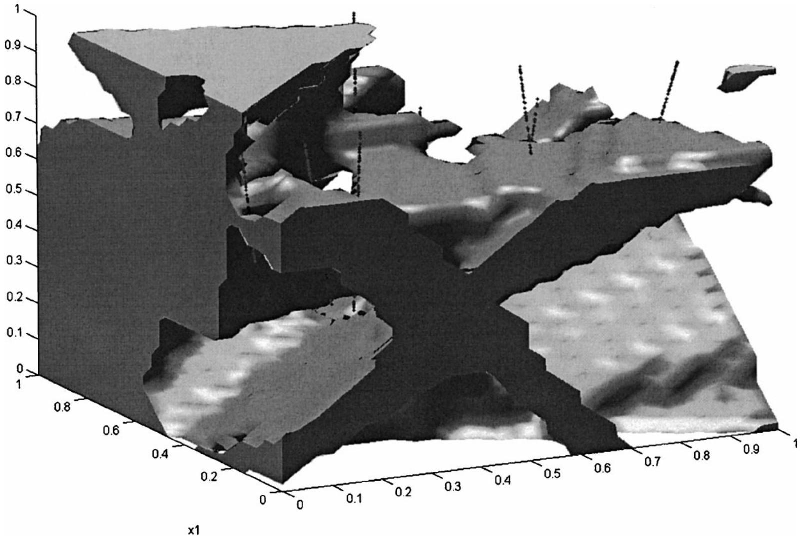

Figure 5. Estimation of the exploitable zones using an MLP neural network. The image shows the regions in which $\hat{f}(\mathrm{x}) \geq 1/2$, where $\hat{f}$ is the estimator of $\mathsf{P}[\mathsf{Y}(\mathrm{x}) = 1/\mathrm{x}]$ provided by the MLP. Test error: 0.11803.

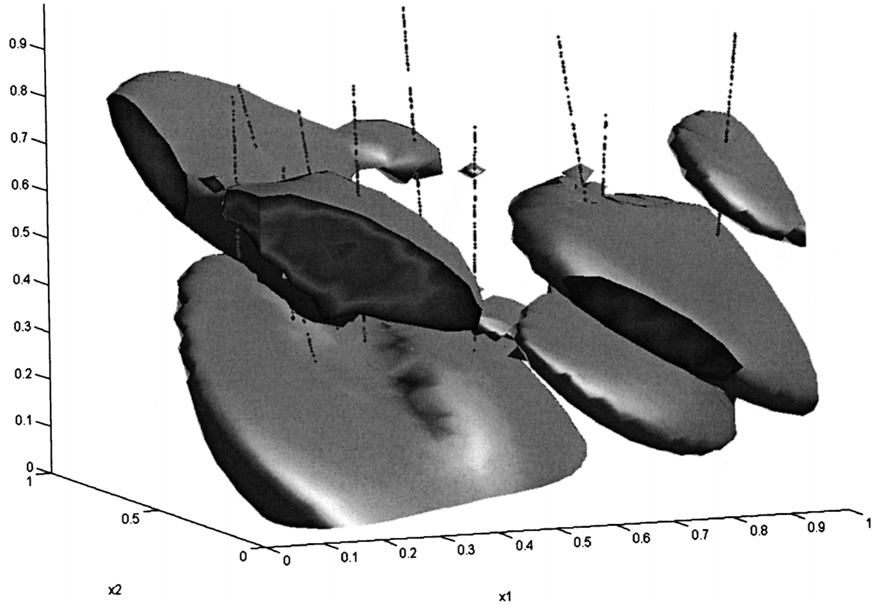

Figure 6. Estimation of the exploitable zones using kriging with an anisotropic variogram. The image shows the regions in which $\hat{Y}(x) \leq 1/2$, where $\hat{Y}(x)$ is the estimator of $P[Y(x) = 1/x]$ provided by the indicator kriging. Test error: 0.10837.

The *ENP* column reflects the *effective* number of parameters for each model, giving an idea of its complexity. In the case of the MLP networks, this quantity is obtained form the Bayesian training algorithm, giving also an idea of the degrees of freedom in the model (Foresee and Hagan, 1997). In the case of the RBF network this quantity is the trace of the *hat matrix S* of the equation in (12).

Given that the kriging estimator and, analogously, the RN estimator are also linear estimators, we applied the said concept to these estimators. Thus, the kriging estimator in the data is written as (Eq. (3)):

$$\hat{y} = [\bar{k}_{x_1}, \ldots, \bar{k}_{x_n}]^t \bar{K}^{-1}\bar{y} = [K^t Q^t]\bar{K}^{-1}\bar{y}$$

where $\bar{y} = (y^t, 0_{1\times m})^t$, therefore the squared $n \times n$ left-superior submatrix of $[K^t Q^t]\bar{K}^{-1}$ contains the coefficients of the observations $y_1, \ldots, y_n$ that make up each estimation $\hat{y}_i$. Denominating the submatrix as S, we define

$$\text{ENP (Kriging)} = \text{trace}(S)$$

a definition which we also apply to the estimator of the regularization networks.

The most parsimonious of all the techniques used was the MLP network. With regard to the radial techniques, it can be observed from Table 1 how the models with a noise/nugget hypothesis in the data possess a more reduced effective number of parameters as a result of the smoothing they perform. Despite this, in some cases (isotropic kriging and RN with covariogram and $\lambda = 0.04$) the interpolating property of the plug-in estimator based on indicator kriging is preserved.

CONCLUSIONS

This article has compared estimation and prediction techniques from fields as different as Spatial Statistics and Neural Networks and applied them to the estimation of the exploitability of a slate quarry. Our conclusions are as follows.

Firstly, whenever possible (problems of dimensionality less than or equal to 3 as in the case of the spatial problems), the dependency structure of the process under consideration should be carefully modeled. Spatial statistical techniques have important advantages over the neural networks whose kernels are selected arbitrarily.

Secondly, of note are the different morphologies of the estimations of the deposit produced by the compared estimators. The reason for these differences lie with the specific architecture of the estimators whose basic functions possess different contour surfaces: ridges for the MLP networks and hyperspheres (hyperellipses for the anisotropic case) for kriging, RN, and RBF. This kind of distinction is not a novel one; see for example Donoho and Johnstone (1989) for a comparison of kernel estimators with the projection pursuit regression (Friedman and Stuetzle, 1981) another technique based on projections such as MLP.

The third aspect of relevance in the tests was the degree of smoothing/interpolation of the estimators. In this respect, the MLP network was not capable of interpolating while maintaining an adequate generalization capacity. In various tests carried out without controlling the complexity of the model (increasing gradually the number of basic functions/epochs), the test error increased as the training error was reduced. In contrast, the other models permit the interpolation capacity to be controlled without restricting the generalization capacity. The key to the interpolation of RBF lay in the use of the estimated isotropic covariogram as a basic function. This opens a wide range of possibilities when selecting the basic functions for these networks.

Aspects that require further exploration in the future are, for example, the use of prefixed directions for the MLP networks as a means of incorporating prior information into the model, and the use of norms of the type $\|x - x'\|_{\sum} = (x - x')^t \sum (x - x')$ in the RBF networks with a view to including anisotropy.

ACKNOWLEDGMENTS

We extend our thanks to the anonymous referees whose comments have led to important improvements in this paper.

Dr J. Taboada's research was supported by Project 1FD97-0091 of the European Union ERDF Programme. Dr W. González-Manteiga's research was supported by Project BFM 2002-03213 of the Spanish Ministry of Science and Technology and the European Union ERDF Programme.

REFERENCES

Bishop, C. M., 2000, Neural networks for pattern recognition: Cambridge University Press, Cambridge, UK, 482 p.

Broomhead, D. S., and Lowe, D., 1988, Multivariable functional interpolation and adaptive networks: complex syst. v. 2, p. 321–355.

Chakraborty, K., Mehrotra, K., Mohan, C. K., and Ranka, S., 1992. Forecasting the behavior of multivariate time series using neural networks: Neural Netw. v. 5, p. 961–970.

Chen, S., Cowan, C. F. N., and Grant, P. M., 1991. Orthogonal least squares learning algorithm for radial basis function networks: IEEE Trans. on Neural Netw. v. 2, no. 2, p. 302–309.

Cressie, N., 1989: Geostatistics: Am. Stat. v. 43, p. 197–202.

Cressie, N., 1990: Reply to Wahba's letter. Am. Stat. v. 44, p. 256–258.

Cressie, N., 1993: Statistics for spatial data: Wiley, New York, 900 p.

Donoho, D. L., and Johnstone, I. M., 1989. Projection-based approximation and a duality with Kernel methods: Ann. Stat. v. 17, p. 58–106.

Foresee, F. D., and Hagan, T., 1997, Gauss–Newton approximation to Bayesian regularization: Proceedings of the 1997 International Joint Conference on Neural Networks, Houston, Texas, p. 1930–1935.

Friedman, J. H., and Stuetzle, W., 1981, Projection pursuit regression: J. Am. Stat. Assoc. v. 76, p. 817–823.

García-Guinea, J., Lombardero, M., Roberts, B., Taboada, J., and Peto, A., 1998, Mineralogía y microestructura de la pizarra de techar: comportamiento termoóptico y fisilidad: Materiales de Construcció n, v. 48, n. 251, p. 37–48.

Girosi, F., Jones, M., and Poggio, T., 1995, Regularization theory and neural networks architectures: Neural Comput. vol. 7, no. 2, p. 219–269.

Haykin, S., 1999, Neural networks. A comprehensive foundation: Prentice-Hall, Upper Saddle River, NJ, 824 p.

Journal, A. G., 1977, Kriging in terms of projections: J. Int. Assoc. Math. Geol. v. 9, p. 563–586.

Koike, K., Matsuda, S., and Gu, B., 2001, Evaluation of interpolation accuracy of neural kriging with application to temperature-distribution analysis: Math. Geol. v. 33, no. 4, p.421–448.

Laslett, G. M., 1994, Kriging and splines: An empirical comparison of their predictive performance in some applications: J. Am. Stat. Assoc., v. 89, no. 426, p. 391–409.

MacKay, D. J. C., 1998, Introduction to Gaussian processes, *in* Bishop, C. M., ed., Neural networks and machine learning: NATO Asi Series F, computer and systems sciences, Vol. 168, Morgan Kaufmann, San Mateo. CA, p. 133–165.

MacKay, D. J. C., 1992, A practical Bayesian framework for backpropagation networks: Neural Comput. v. 4, p. 448–472.

Matheron, G., 1973, The intrinsic random functions and their applications: Ad. Appl. Probability, v.5, p. 439–468.

Moody, J. E., and Darken, C.J., 1989, Fast learning in networks of locally-tuned processing units: Neural Comput. v. 1, 281–294.

Poggio, T., and Girosi, F., 1989, Networks for approximation and learning: Proc. IEEE, v.78, p. 1481-1497.

Ribeiro, J., Pereira, H. G., Sousa, A. J., Albuquerque, T., Tirabasso, F., Taboada, J., and Rico, J. G., 1997, Geostatistical characterization of natural stone quarries, *in* Baafi, E. Y., and Schofield N. A., eds., Geostatistics wollongong '96: Kluwer, Dordrecht, the Netherlands, p. 905–915.

Ripley, B. D., 1981, Spatial statistics: Wiley, New York.

Ripley, B. D., 1996, Pattern recognition and neural networks: Cambridge University Press, Cambridge, UK, 403 p.

Schioler, H., and Hartmann, U., 1992, Mapping neural network derived from the Parzen Window Estimator: Neural Netw., vol. 5, no. 6, p. 903–909.

Taboada, J., Vaamonde, A., Saavedra, A., and Alejano, L., 1997, Application of geostatistical techniques to exploitation planning in slate quarries: Eng. Geol. v. 47, p. 269–277.

Taboada, J., Vaamonde, A., Saavedra, A., and Argüelles, A., 1998: Quality index for ornamental slate deposits: Eng. Geol. v. 50, p. 203–210.

Taboada, J., Saavedra, A., and Vaamonde, A., 2001, Evaluation of a slate extraction bank: Trans Inst. Min. Metallurgy, A. v. 110, p. 40–46.

Tikhonov, A. N., and Arsenin, V. Y., 1977, Solutions of ill-posed problems: Wiley, New York, 258 p.

Wahba, G., 1990a, Spline models for observational data: SIAM, Philadelphia, 169 p.

Wahba, G., 1990b, Reply to Cressie: Am. Stat. v. 44, p. 255–256.

Weber, D. D., and Englund, E. J., 1992, Evaluation and comparison of spatial interpolators: Math. Geol., v. 24, no. 4, p. 381–391.

Weber, D. D., and Englund, E. J., 1994, Evaluation and comparison of spatial interpolators, II: Math. Geology, v. 26 no. 5, p. 589–603.

Yakowitz, S. J., and Szidarovszky, F., 1985, A comparison of kriging with nonparametric regression methods: J. Multivariate Anal., v.16, p. 21–53.

Zimmerman, D., Pavlik, C., Ruggles, A., and Armstrong, M.P., 1999, An experimental comparison of ordinary and universal kriging and inverse distance weighting: Math. Geol., v. 31, no. 4, p. 375–390.