A locally adaptive kernel regression method for facies delineation

D. Fernàndez-Garcia, M. Barahona, C. Henri, X. Sanchez-Vila

Hydrogeology Group (UPC-CSIC), Department of Geotechnical Engineering and Geosciences, Universitat Politècnica de Catalunya (UPC-BarcelonaTech), 08034 Barcelona, Spain

Abstract

Facies delineation is defined as the separation of geological units with distinct intrinsic characteristics (grain size, hydraulic conductivity, mineralogical composition). A major challenge in this area stems from the fact that only a few scattered pieces of hydrogeological information are available to delineate geological facies. Several methods to delineate facies are available in the literature, ranging from those based only on existing hard data, to those including secondary data or external knowledge about sedimentological patterns. This paper describes a methodology to use kernel regression methods as an effective tool for facies delineation. The method uses both the spatial and the actual sampled values to produce, for each individual hard data point, a locally adaptive steering kernel function, self-adjusting the principal directions of the local anisotropic kernels to the direction of highest local spatial correlation. The method is shown to outperform the nearest neighbor classification method in a number of synthetic aquifers whenever the available number of hard data is small and randomly distributed in space. In the case of exhaustive sampling, the steering kernel regression method converges to the true solution. Simulations ran in a suite of synthetic examples are used to explore the selection of kernel parameters in typical field settings. It is shown that, in practice, a rule of thumb can be used to obtain suboptimal results. The performance of the method is demonstrated to significantly improve when external information regarding facies proportions is incorporated. Remarkably, the method allows for a reasonable reconstruction of the facies connectivity patterns, shown in terms of breakthrough curves performance.

1. Introduction

Image reconstruction has a long history in a number of disciplines such as satellite image mapping, shape recognition in robotics, face recognition, and license plate reading, among other uses [*Bughin et al.* 2008, *Daoudi et al.* 1999, *Yang & Huang* 1994, *Lin & Chen* 2007]. The topic can be loosely subdivided into two main groups: (a) The reconstruction of incomplete images where some of the pixels have no information; and (b) The reconstruction of noisy images, where some of the pixels display wrong information and the main problem is detecting and reclassifying the misclassified pixels.

A good reconstruction work relies heavily on the presence of data and on an efficient reconstruction algorithm that can either complete information gaps, or else filter noisy signals. A particular case of reconstruction appears in subsurface hydrology, where the information relies on very few points (well logs), so that the initial available picture for reconstruction is mostly a black signal (meaning no information) with some sparse data scattered throughout the medium. Reconstruction is, thus, a really difficult and error prone task.

Many methods for the interpolation of scattered data exist [*Franke,* 1982] and some of them have been used for geologic facies reconstruction [i.e., *Ritzi et al.,* 1994, *Guadagnini et al*., 2004, *Tartakovsky and Wohlberg,* 2004, *Wohlberg et al*., 2006, *Tartakovsky et al.,* 2007]. In particular, *Tartakovsky et al*. [2007] compared the fractional error obtained in two synthetic examples using three approaches: indicator kriging (IK) [*Isaaks & Srivastava,* 1990, *Ritzi et al.,* 1994, *Guadagnini et al*., 2004], support vector machines (SVM) [*Tartakovsky and Wohlberg* 2004, *Wohlberg et al*., 2006] and nearest-neighbor classification (NNC) [*Dixon,* 2002]. Different sampling densities, ranging from 0.28% to 3.06%, and random sampling data generated following a 2D Poisson random process were used for comparison. Here sampling density refers to the proportion of pixels where hard data is available (pixels that are univocally classified). Their analysis indicated that NNC outperformed IK, in terms of proportion of correctly classified pixels, in both examples, and that SVM slightly outperformed NNC in one of the examples.

There exist a number of reconstruction methods available in different disciplines that to our knowledge have never been used in geological facies reconstruction. A potential reason for this is that these methods were devised for the presence of massive data sets

that are never available in hydrogeology. One family of methods is based on kernel regression functions, widely used in signal theory for solving different problems such as image denoising, upscaling, interpolation, fusion, etc. Such methods have proved to be efficient for problems such as restoration and enhancement of noisy and/or incomplete sampled images. Even though regression methods have been used for reconstruction of images from extensive data sets, in principle, there is no reason not to use them when information is sparse. As an example, *Takeda et al.* [2007] tested a kernel regression method on an image reconstruction case in which only 15% of the pixels were informed, obtaining a very good reconstruction of a 2D image.

Making an analogy between image reconstruction (from irregularly sampled data) and facies delineation (from scattered sampling points), we investigate the performance of a Steering Kernel Regression (SKR) method for the latter problem. The aim is to describe a methodology to use kernel regression as an effective tool for facies delineation, an application involving far less information available for image delineation from that for what it was originally developed (reconstruction). In doing this, we investigate the optimal tuning parameters to be used in the reconstruction of geological facies and their connectivity patterns.

This paper is structured as follows; Section 2 briefly describes the fundamental concepts of facies reconstruction. Section 3 presents the details of the data-adapted kernel regression method. We test this method with respect to the NNC method in Section 4 by means of four synthetic images, here including the two figures profusely investigated by *Tartakovsky et al.* [2007] to allow for performance comparisons.

2. The concept of facies reconstruction

The term facies is used in geology to differentiate among geological units on the basis of interpretive or descriptive characteristics, such as sedimentological conditions of formation, mineralogical composition, presence of fossils (biofacies), structures, grain size, etc. [*Tarbuck et al.*, 2002]. In this work, we consider that each facies is a clear distinctive geology unit, understood in a descriptive sense. Keeping this in mind, facies reconstruction is defined as the process of assigning each unsampled point (eventually also the sampled ones if misclassification errors are admitted) to one facies. Formally,

for any given facies F_k, the reconstruction problem can be addressed using an indicator function defined as

$$I(\mathbf{x}, F_k) = \begin{cases} 1 & \mathbf{x} \in F_k \\ 0 & otherwise \end{cases} \qquad (1)$$

where the indicator variable $I(\mathbf{x},F_k)$ is equal to 1 when a particular point in the domain, **x**, can be classified as belonging to facies F_k and zero otherwise. In this work we assume that the available data from the sampling points are clearly distinctive in order to be unmistakably classified as indicated in (1) without interpretation errors. From now on, we consider that only two facies are used for geological mapping. However, the method can be easily extended to any finite number of facies by direct superposition.

Several methods have been proposed in the literature to estimate the spatial distribution of the indicator variable $I(\mathbf{x},F_1)$. Here we compile only three of such methods. The first one is indicator kriging (IK) [*Journel*, 1983], a method that provides a least-squares estimate of the probability that **x** belongs to F_1 conditioned to nearby data. Once a threshold value is given, a distinction between categories (facies) can be done. The method relies on the theory of random functions to model the uncertainty of not having data at unknown locations. It accounts for the inherent spatial correlation of data but typically fails to properly estimate curvilinear geological bodies. Multiple point geostatistics [e.g., *Strebelle*, 2000] can overcome most of these problems by largely relying on an empirical multivariate distribution inferred from training images, i.e., under the assumption that significant information about the spatial distribution of facies is known from external sources (outcrops, modeling of sedimentological processes,…); these information is directly transferred to the final images.

Alternatively, Support Vector Machine (SVM) methods are a set of popular tools for data mining tasks such as classification, regression, and novelty detection [*Vapnik*, 1963; *Bennett and Campbell*, 2000]. SVM takes a training data, i.e., a set of *n* data points $J_i = J(\mathbf{x}_i,F_1) \in \{-1,1\}$, $i=1,..,n$, and separates them into two classes by delineating the hyperplane that has the largest distance to the nearest training data point of any class.

Last, the nearest-neighbor classification (NNC) simply classifies each point in the domain by finding the nearest (not necessarily in the Euclidean sense) training point, assigning to the unsampled location the class corresponding to that training point.

A comparison of the three methods presented is provided in a series of papers by *Tartakovsky and Wholberg* [2004], *Wholberg et al.* [2006], and *Tartakovsky et al.* [2007]. Surprisingly, the NNC method outperformed the more sophisticated ones, i.e., SVM and IK, indicating the validity of the parsimony principle for this problem. Yet, the comparison between methods in such works was done only in terms of the number of misclassified points without considering other performance metrics, such as connectivity features inherent in geological facies that can strongly impact contaminant transport simulations (e.g., Fernàndez-Garcia et al., 2010). We consider this issue as non-ideal and in the next section we seek for a method that can actually represent the presence of connected geological bodies with elongated and curvilinear shapes.

3. Kernel regression approaches for facies classification

Kernel regression methods have been developed in statistics to estimate the conditional expectation of a random variable without assumptions about its probability distribution function. These methods are well documented and summarized in the literature [e.g., Hardle, 1990; Simonoff, 1996; Li et al., 2007]. Suppose that we ignore the fact that the target classification output is a binary function $I(\mathbf{x},F_1)$. Instead, we consider that it is a continuous function that depends on the location $\mathbf{x}$ and a number of (yet unknown) parameters $\mathbf{b}=[b_0,b_1,\ldots,b_N]^T$. The regression model proposed here for facies classification assumes that the measured data $I_i=I(\mathbf{x}_i,F_1)$, $i=1,\ldots,n$, can be expressed as

$$I_i = m(\mathbf{x}_i;\mathbf{b}) + \varepsilon_i, \qquad i = 1,..,n, \tag{2}$$

where $m(\mathbf{x}_i,\mathbf{b})$ is the regression function to be determined, and ε_i are independent and identically distributed zero mean noise values. Kernel regression is a form of regression analysis in which the function m is exclusively dictated by the data, and not prespecified a priori (no model assumed). At each point $\mathbf{x}$ the conditional expected value of the dependent (indicator) variable can be estimated, i.e., $m(\mathbf{x},\mathbf{b})=\mathrm{E}[I(\mathbf{x},F_1)]$. The interest of kernel regression to facies reconstruction resides on the fact that the conditional expected value of the indicator variable is exactly the probability that the given facies F_1 prevails at that location, since

$$E\{I(\mathbf{x},F_1)\} = 1\cdot \mathrm{Prob}\{\mathbf{x}\in F_1\} + 0\cdot \mathrm{Prob}\{\mathbf{x}\notin F_1\} = \mathrm{Prob}\{\mathbf{x}\in F_1\} \tag{3}$$

By definition, the probability of occurrence of a given facies is a continuous variable ranging between 0 and 1. In order to separate the data into classes or facies we must then establish a cut-off in the estimate of the indicator variable. This is similar to the facies reconstruction problem posed by the geostatistical indicator kriging approach. In this case, *Ritzi et al.* [1994] has suggested to define the boundary between facies by the isoline Prob{$\mathbf{x}\in F_k$}=p_k, where p_k is estimated as either the global mean of the indicator values or the empirical relative volumetric fraction of the facies F_k. We propose here to use the same approach for classifying facies with regression methods. The benefits of such approach will be explored in section 4.

Two kernel regression methods, namely the classical (CKR) and the adaptive steering (SKR) are presented next, and later their performance is compared in a number of synthetic cases.

3.1. Classical kernel regression (CKR)

Let us consider a local Taylor expansion of the mean response $m(\mathbf{x},\mathbf{b})$ of the indicator values around the estimation location $\mathbf{x}_0$,

$$m(\mathbf{x};\mathbf{b}) \approx m(\mathbf{x};\mathbf{b},\mathbf{x}_0) = b_0 + b_1 x' + b_2 y' + b_3 z' + b_4 x'^2 + b_5 x' y' + b_6 y'^2 + b_7 x' z' ... \quad (4)$$

where $\mathbf{x}'=\mathbf{x}-\mathbf{x}_0$ is the distance between any point and that being estimated, b_0 is the mean response at $\mathbf{x}_0$, $[b_1,b_2,b_3]^T$ is the gradient of the mean response at $\mathbf{x}_0$, and so on. The order of the polynomial is in principle arbitrary. Nonparametric regression generalizes the standard regression approach by locally estimating **b** at a given location $\mathbf{x}_0$ using only nearby data. This is done by weighting data located far away from the estimation location with a kernel function K_H defined as

$$K_H(\mathbf{x}) = \frac{1}{\det(\mathbf{H})} K\left(\mathbf{H}^{-1}\mathbf{x}\right) \quad (5)$$

where **H** is a matrix that controls the degree of smoothing and is user dependent. The kernel associates a very low weight to points located far from the estimation point. Section 4 will explore the choice of kernel parameters for optimal facies reconstruction.

The kernel function K is a continuous, bounded, and symmetric real function centered at zero that integrates to one and typically decays with distance. The choice of the kernel is known to not affect significantly the final solution and therefore a standard Gaussian

distribution is typically used for mathematical convenience. In *n* dimensions this is written as

$$K(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}} \exp\left[\frac{1}{2}\mathbf{x}^T\mathbf{x}\right] \qquad (6)$$

For any given estimation location $\mathbf{x}_0$, the principle of least squares expresses that one should choose as estimates of **b** those values that minimize the weighted sum of squared residuals, S(**b**), the residual being the difference between data values and model predictions,

$$S(\mathbf{b}) = \sum_{i=1}^{n} [I_i - m(\mathbf{x}_i, \mathbf{b}, \mathbf{x}_0)]^2 K_H(\mathbf{x}_i - \mathbf{x}_0) \qquad (7)$$

Let us express equation (2) in matrix form,

$$\mathbf{I} = \mathbf{X}\mathbf{b} + \mathbf{e} \qquad (8)$$

where $\mathbf{I}=[I_1,..,I_n]^T$, $\mathbf{e}=[\varepsilon_1, \ldots,\varepsilon_n]^T$, and $\mathbf{X}$ is a matrix composed of n rows and a number of columns that is associated with the degree of the polynomial chosen for **b** (i.e., in 3-D would be 4 for order 1, 10 for order 2,…)

$$\mathbf{X} = \begin{bmatrix} 1 & x_1' & y_1' & z_1' & x_1'^2 & x_1'y_1' & y_1'^2 & x_1'z_1' & \ldots \\ \ldots & & \ldots & & \ldots & & \ldots & & \ldots \\ 1 & x_n' & y_n' & z_n' & x_n'^2 & x_n'y_n' & y_n'^2 & x_n'z_n' & \ldots \end{bmatrix} \qquad (9)$$

Then, the optimization problem is written as

$$\min_{\mathbf{b}} S(\mathbf{b}) = \min_{\mathbf{b}} (\mathbf{I} - \mathbf{X}\mathbf{b})^{\mathbf{T}} \mathbf{W} (\mathbf{I} - \mathbf{X}\mathbf{b}) \qquad (10)$$

where **W** is a diagonal weight matrix given by

$$\mathbf{W} = \mathrm{diag}\{\mathrm{K_H}(\mathbf{x}_1 - \mathbf{x}_0), ..., \mathrm{K_H}(\mathbf{x}_n - \mathbf{x}_0)\} \qquad (11)$$

Setting $\partial S(\mathbf{b})/\partial b_j = 0$ to each parameter b_j we obtain the following solution

$$\hat{\mathbf{b}} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{I} \qquad (12)$$

This solution is formally the same to that of standard regression but the matrices **W** and **X** depend now on the estimation location $\mathbf{x}_0$. Knowing the optimal estimate of **b**, the probability that **x** belongs to F_1 can be estimated by

$$\text{Prob}\{\mathbf{x} \in F_1\} = E\{I(\mathbf{x}, F_1)|\mathbf{x}\} = m(\mathbf{x}_0, F_1, \mathbf{x}_0) = \hat{b}_0 \qquad (13)$$

Let us define $\mathbf{W}_{\text{eq}}$ by

$$\mathbf{W}_{eq} = \mathbf{e}_1^T \left(\mathbf{X}^T \mathbf{W} \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{W} \qquad (14)$$

where $\mathbf{e}_1$ is a column vector with first element equal to one, and the rest equal to zero. Then, the Classical Kernel Regression (CKR) algorithm can be seen as a local weighted averaging of the data in which the probability that **x** belongs to F_1 is determined by the following linear interpolation of indicator values

$$\hat{b}_0 = \mathbf{W}_{eq}^T \cdot \mathbf{I} \qquad (15)$$

Hence, $\mathbf{W}_{\text{eq}}$ is a vector containing the equivalent weights of the indicator data values. The forms of these equivalent weights are exclusively dictated by the polynomial order chosen in (4).

3.2. Steering kernel regression (SKR)

The SKR method comes as a direct extension of the CKR algorithm. Since the latter is nothing but a weighted average of indicator data values, the final regression estimate of Prob{$\mathbf{x} \in F_1$} only depends on the geometric configuration of the data, and therefore ignores the inherent correlations between data positions and their values. *Takeda et al.* [2007] developed a SKR algorithm to include key structural features into the estimated fields.

The key idea behind the SKR algorithm is to modify the size and orientation of the regression kernel to assign more weight along the direction of highest local spatial correlation. The advantage of doing this to classify facies is the following: consider a point $\mathbf{x}_0 \in F_1$ located close to a facies boundary; the conventional CKR algorithm (symmetric spherical kernel) will estimate the probability that $\mathbf{x}_0$ belongs to F_1 by equally considering both nearby samples of the same facies F_1 and samples of other facies located beyond the boundaries. The SKR method is designed to adapt the regression kernel to the boundary isosurface so as to assign more weight to those samples belonging to the same facies. This way, the denoising is affected most strongly along the boundaries, rather than across them, resulting in a strong preservation of details in the final output.

The algorithm works by reorienting the smoothing matrix in the direction of the gradients of the mean response m(**x**,**b**) through a redefinition of the kernel matrix

$$\mathbf{H}_i^{steer} = h\mathbf{C}_i^{-1/2} \tag{16}$$

$$\mathbf{C}_i \approx \left(\overline{\nabla m(\mathbf{x}_j, \hat{\mathbf{b}}) \cdot \nabla^T m(\mathbf{x}_j, \hat{\mathbf{b}})}\right), \qquad \mathbf{x}_j \in w_i \tag{17}$$

where the overbar stands for averaging over the mean response adjacent to $\mathbf{x}_i$, w_i is the window search around $\mathbf{x}_i$, and h is a global smoothing parameter.

In contrast to the CKR algorithm, the smoothing matrix $\mathbf{H}^{steer}$ at each individual point $\mathbf{x}_i$ depends now on the solution of the regression function m(**x**,F_1). This makes the SKR method to be nonlinear in nature. Its application must be therefore iterative, starting with a first initial estimate of m(**x**,F_1) computed, for instance, with the CKR method. This estimate is used to measure the dominant orientation of the local gradients, then used to sequentially steer the local kernel function through (17), resulting in elongated, ellipsoidal contours spread along the indicator isosurface (or isocurve in 2D).

We must state that while the method is applicable to 3D reconstruction problems, here we present the details only for the 2D problems. The main reason is to be able to use the same synthetic examples available in the literature for geologic facies reconstruction using IK, SVM or NNC methods. Under these conditions, and from (16), the new form of the regression kernel is

$$K_{H_i^{steer}}(\mathbf{x}_i - \mathbf{x}_0) = \frac{\sqrt{\det(\mathbf{C}_i)}}{2\pi h} \exp\left[\frac{(\mathbf{x}_i - \mathbf{x}_0)^T \mathbf{C}_i (\mathbf{x}_i - \mathbf{x}_0)}{2h^2}\right] \tag{18}$$

When estimating the covariance matrix $\mathbf{C}_i$ through (17), the resulting matrix can be rank deficient and unstable. To overcome this problem, a multiscale technique for estimating local gradients [*Takeda et al.*, 2007] can be adopted. Let us consider the following matrix $\mathbf{G}_i$ formed by a collection of p estimated gradient vectors at the neighborhood of the sampled location $\mathbf{x}_i$

$$\mathbf{G}_i = \begin{bmatrix} \nabla m(\mathbf{x}_1, \mathbf{b}) \\ \ldots \\ \nabla m(\mathbf{x}_p, \mathbf{b}) \end{bmatrix}, \qquad \mathbf{x}_j \in w_i, \quad j = 1,\ldots,p \tag{19}$$

The singular value decomposition of $\mathbf{G}_i$ factorizes this matrix in the following form

$$\mathbf{G}_i = \mathbf{U}_i \mathbf{S}_i \mathbf{V}_i \tag{20}$$

where the diagonal entries S_{jj} of $\mathbf{S}_i$ (singular values) represent the energy in the dominant directions (singular vectors) of the local gradient field. These dominant directions are given by the column vectors of the matrix $\mathbf{V}_i$. In particular, the second column of $\mathbf{V}_i$, $[V_{12}, V_{22}]^T$, determines the direction of smallest energy and represents the dominant orientation angle of $\mathbf{C}_i$ (direction with highest local spatial correlation) by

$$\theta_i = \arctan\left(\frac{V_{12}}{V_{22}}\right) \tag{21}$$

The actual shape of the regression kernel is then calculated from the energy associated with the dominant gradient directions,

$$E_{\max} = \frac{S_{11} + \lambda_1}{S_{22} + \lambda_1}, \qquad E_{\min} = \frac{S_{22} + \lambda_1}{S_{11} + \lambda_1} \tag{22}$$

where λ_1 is a regularization parameter that dampens the effect of noise and restricts the ratio from becoming degenerate. Knowing these parameters, the covariance matrix can be calculated by the combination of a scaling parameter γ_i, a rotation matrix $\mathbf{R}_i$, and an elongation matrix $\mathbf{E}_i$ by means of

$$\mathbf{C}_i = \gamma_i \mathbf{R}_i \mathbf{E}_i \mathbf{R}_i^T \tag{23}$$

The different terms in (23) are defined as

$$\mathbf{R}_i = \begin{bmatrix} \cos\theta_i & \sin\theta_i \\ -\sin\theta_i & \cos\theta_i \end{bmatrix} \tag{24}$$

$$\mathbf{E}_i = \begin{bmatrix} E_{\max} & 0 \\ 0 & E_{\min} \end{bmatrix} \tag{25}$$

$$\gamma_i = \left(\frac{S_{11} S_{22} + \lambda_2}{M}\right)^{\alpha}, \tag{26}$$

where λ_2 is another regularization parameter aimed at dampening the effect of noise and keeping the scaling parameter from becoming zero, α is a structure sensitive parameter satisfying that $0<\alpha<1$, and M is the number of samples in the local analysis window w_i.

3.3. Uncertainty in facies classification

At this stage, it is important to highlight the following advantage of the SKR method compared to deterministic algorithms, e.g., the nearest neighbor classification (NNC). Statistical approaches not only provide a map of the spatial distribution of the estimates of indicator values (i.e., the probability that a given point belongs to a facies), but also the error variance of the estimates of **b**. If the error terms ε_i are uncorrelated, and all have the same variance σ^2, then it can be shown that the estimator (12) is an unbiased estimate of **b**, and that the variance-covariance of the estimation matrix is

$$\mathbf{C}_b = \sigma^2 \left(\mathbf{X}^T \mathbf{W} \mathbf{X} \right)^{-1} \qquad (27)$$

Thus, the variance of the estimate of Prob{$\mathbf{x} \in F_1$} can be determined by

$$\sigma^2_{SKR} = \sigma^2 \mathbf{e}_1 \left(\mathbf{X}^T \mathbf{W} \mathbf{X} \right)^{-1} \mathbf{e}_1^T \qquad (28)$$

and the error variance can be estimated as

$$\sigma^2 \approx s^2 = \frac{S(\hat{\mathbf{b}})}{n - N} \qquad (29)$$

where the integer N is the number of estimated parameters. Knowing this, one can define an approximate confidence region in which the border between facies is most expected to be found. This will be illustrated in the synthetic example presented next.

4. Synthetic Examples

4.1. Methodology

Since the NNC has been already demonstrated to outperform SVM and IK approaches [*Tartakovsky et al.*, 2007], in this section we only compare the performance of the SKR method with that of NNC. The NNC algorithm is provided in the Appendix for completeness. Four synthetic geological field settings formed with two distinct facies (see Figure 1) were used to test the performance of SKR. Two of these fields, Figures 1a and 1b, are identical to the ones presented by *Tartakovsky et al.* [2007]; the remaining two were specifically generated for this work. Figure 1c is a curvilinear shape, obtained from an abandoned meander in the Ebro river (Spain), potentially

indicating the shapes of paleochannels that could be found in the subsurface. Figure 1d is just a circle, in order to test the performance of the algorithm for a very simple shape that is easily reproduced with any reconstruction algorithm. Each of the figures consists of an image data discretized in 60×60 (=3600) pixels. Red and blue pixels correspond to facies F_1 and F_2, respectively. In accordance to previous sections, the following indicator function is used for facies reconstruction purposes,

$$I(\mathbf{x}, F_1) = \begin{cases} 1 & \mathbf{x} \in F_1 \\ 0 & \mathbf{x} \in F_2 \end{cases} \qquad (30)$$

The objective of the numerical simulations is to reconstruct the facies depicted in each individual image from a few measurements. We consider a random data set consisting of 10, 20, 30, 50, 80, and 110 measurements, corresponding to a range of 0.28% to 3.06% of the total pixels investigated. Emphasis is given to the lowest sample densities (below 1%), which illustrate the most typical problem encountered in subsurface hydrology, i.e., those with scarce information covering a very low portion of the simulation domain. The SKR is used with a quadratic polynomial approximation of the mean response $m(\mathbf{x},F_1)$ in (4) and a Gaussian kernel.

An analysis of the fractional error of the reconstructed images is used to compare the performance of the SKR and NNC methods. For each realization, the fractional error was obtained as the ratio of misclassified pixels to the total number of pixels in the images. One hundred realizations were created for each sample density, and the fractional error reported is the average over the ensemble of realizations. For comparison purposes, selected points associated with each sample density were the same for the SKR and the NNC methods. It is important to notice that *Tartakovsky et al.* [2007] used only 20 (rather than 100) randomly generated realizations for each sample density; for this reason, our calculated fractional error for NNC, although similar, is slightly different to theirs.

The SKR method provides the probability of occurrence of facies F_1 at a given location. Therefore, the output data is a continuous variable (i.e., $m(\mathbf{x},\mathbf{b})\in[0,1]$). A cut-off in the estimated values is then necessary to classify the data into facies. We explore two different strategies to introduce this cut-off. The first strategy considers that no *prior information* on the relative volumetric proportion of facies is known. In this case, the boundary between facies is determined by the isoline $\text{Prob}\{\mathbf{x}\in F_1\}=\text{Prob}\{\mathbf{x}\in F_2\}=0.5$,

expressing that both facies have the same probability of existence at the facies boundaries. We denote this method as SKR(0). When prior information on the relative volumetric proportion of facies is known, then one can define the boundary between facies by the isoline Prob{$\mathbf{x}\in F_1$}=p_1 (i.e., Prob{$\mathbf{x}\in F_2$}=1- p_1), where p_1 is estimated either by the global mean of the indicator values or the empirical relative volumetric fraction of facies F_1. The latter method is similar to the facies reconstruction problem posed by the geostatistical indicator kriging approach proposed by *Ritzi et al.* [1994]. We will denote this strategy as SKR(%).

4.2. Choosing the kernel parameters

Five different parameters control the solution of the SKR method: (1) the global smoothing parameter *h*, equation (16); (2) the size of the local orientation analysis window *w*, equation (17); (3) the regularization parameter λ_1, equation (22); (4) the structure sensitive parameter α, equation (26); and (5) a second regularization parameter λ_2, equation (26). This last one was directly fixed to 10^{-7}. A sensitivity analysis of the lowest fractional error was carried for the remaining four parameters.

Figure 2 provides a series of contour plots of the lowest fractional error associated with the image shown in Figure 1a and only for the case of lowest sample density (10 data points). Each contour plot displays the lowest fractional error as a function of two parameters. Blue dots correspond to the estimated values used to generate the contour plots. In general, the lowest fractional error is mainly controlled by *h* and α, being the output solution quite insensitive to *w* and λ_1. A good quality of facies classification reconstruction is typically obtained with *h*=1 (pixel), *w*=5 (pixel), λ_1=500 (-) and α=0.01 (-). This optimum combination of parameter values is explained as follows:

- The structure sensitivity parameter, α, which must satisfy the condition $0<\alpha<1$, is devised to increase the steering kernel area in regions where large fluctuations exist (high-frequency data); so, large α values produce smooth estimates in high-frequency data regions. The reconstruction problem in hydrogeology typically involves small densities and low-frequency data (scarce data) and thereby this correction is somehow uncalled for. Accordingly, the sensitivity analysis yields α=0.01, which basically expresses that the scaling factor γ_i is always close to 1.

- The window size, *w*, defines the search area over which the gradients ∇m used to determine the local covariance function $\mathbf{C}_i$ at a data point location are estimated. Results show that a relatively small region (*w*=5 pixels) is sufficient to properly capture the patterns of $\mathbf{C}_i$, which is most likely due to the use of a small sample population and the lack of noise in the data values.
- The parameter λ_1 is a regularization parameter used to avoid numerical singularities during the estimation of the principal components of the elongation matrix **E**. Results show that the lowest fractional error decays with increasing λ_1. Large values are needed here because S_{11} and S_{22} in equation (21) are relatively large for the field conditions considered.
- Given that scaling is not required ($\gamma_i \approx 1$) and that the solution is not much sensitive to both *w* and λ_1, the global smoothing parameter *h* appears as the main controlling factor. This parameter determines the area underneath the steering kernel so that large *h* values will increase the influence of distant data points to the final estimation. Results show that a small *h* value close to 1 pixel is required in this synthetic example, which implies small steering kernel areas.

An illustration of the shape of the steering kernel ellipses obtained during the iterative solution of the SKR method is shown in Figure 3 for a sample density of 0.83% (corresponding to 30 data points over 3600 pixels). Figure 3a shows the reference image, whereas the series of Figures (b)-(e) display the reconstruction solution at different iterations. Initially, there is no information on the local correlation of data values (gradients) and therefore the ellipses are circles of radius close to 1 pixel. Notice that circles in this method are uninformative. In subsequent iterations, a better gradient estimation is increasingly achieved and circles are reshaped to ellipses elongated in the direction of the highest local correlation (smallest gradient). As a result, large weights are given to the data values located in the direction of the local highest correlation while other data points are practically ignored. Based on this observation, the application of the SKR method to facies reconstruction can be seen as a specialized nearest neighbor procedure in which the distance metric is not measured by an Euclidian distance but in terms of the highest local correlation, changing for each data location.

The parameter sensitivity analysis presented here considers a given sample density and a particular image. To complete the analysis, Figure 4 presents the global smoothing

parameter *h* as a function of sample density and for each reference image. The remaining parameters were set to *w*=5 pixels, λ_1=500 and α=0.01. For a given sample density, the *h* value provided is the best estimate obtained manually by trial-and-error to minimize the fractional error. Figure 4 shows that, in the lowest sampling density, which is the typical scenario in subsurface hydrology, the lowest fractional error is always achieved when *h*=1 for both methods, i.e., SKR(0) and SKR(%). It was also observed that, in most cases, the SKR(0) method with no prior information on the volumetric proportion of facies yielded larger fractional errors as compared to the SKR(%) method. This effect was more significant for the smaller sample densities, the typical scenario in real applications.

4.3. Simulation results

Figure 5 shows the fractional error as a function of sample density for the different methods employed. In all cases, the fractional errors associated with both the SKR(0) and the SKR(%) methods were smaller than that of the NNC. Interestingly, while the performance of the SKR(0) method is only slightly better than that of the NNC method, with a relative error difference no larger than 1% in most cases, the introduction of prior information into the analysis via the SKR(%) method was capable to significantly outperform the other two approaches. This impact was most noticeable in Figure 1c. It is important to highlight here that for all the evaluated images, the benefit (in relative terms) given by the SKR(%) method was higher for the smaller sampling densities. This is an important finding in itself. Under real circumstances, in typical hydrogeology problems it is likely that the number of data points will be rather limited, rendering the SKR(%) method a valuable instrument to interpret facies delineation with the lowest estimation error.

Let us emphasize the real benefit of using SKR(%) compared to the NNC algorithm. Consider the problem of reconstructing the image shown in Figure 1a from only 30 data points randomly located. Figure 6 compares the true image (cross symbols represent the sampling points) with the output of NNC and 4 iterations of the SKR(%) method. In this case, the fractional error associated with the SKR(%) method is only slightly better than that of the NNC but still important reconstruction features can be distinguished. NNC only depends on data configuration and not on the actual values or their spatial

correlation. As such, its reconstructed image (Figure 6b) fails to represent the central spatial continuity observed in facies F_1, clearly extending from the northern to the southern boundaries. Instead, with only four iterations, the SKR(%) is able to correctly identify this spatial continuity of data values and properly represent the true connection north-south. Figure 3 illustrates the evolution of the local kernel functions associated to each data point in the same problem. In these images, the variable represented is the direct output data given by the SKR method without applying a classification strategy, and the progressive increase in the ratio of the two axes of the ellipse can be observed.

In addition to the recognition of spatial continuity, the SKR(%) mehod is also capable of providing a measure of uncertainty in the delineation of the facies boundary. In principle this is not possible for any deterministic approach, such as that of the NNC algorithm. Figure 7 presents different maps to evaluate the uncertainty in the estimation corresponding to the same example already used previously. Interestingly, there is a very good correlation between low variance and high sampling density areas and viceversa. From this map, one can also delineate a safe zone for drawing the border between facies, plotted as gray areas in Figure 7e, those corresponding to values above 0.3 times the standard deviation. By visual inspection, a very good agreement between the results from the SKR(%) method (Figure 7e) and the original facies boundaries visible in Figure 7a, can be appreciated.

4.4. Impact on transport predictions

In this paper, we contend that a key aspect to consider during the reconstruction of geological facies is the representation of connectivity. Even though the SKR method is shown to only slightly outperform the NNC in terms of volumetric fractional errors (see Figure 5), results demonstrated that the NNC is often not capable to properly describe the spatial continuity of the facies body. Solute transport simulations in a Monte Carlo framework were further performed to illustrate the impact that this effect can have on contaminant transport predictions. To do this, we considered the synthetic field presented in Figure 1a as a reference geological setting. The hydraulic conductivity is assumed to vary in space. A hydraulic conductivity of K=100 m/day and K=1 m/day was respectively assigned to the blue and red facies. Figure 8 shows the setup of the simulations. A non-reactive contaminant source was assumed to be originally located in a southern block region of size 5×5 m^2 (it is assumed that each pixel has a length of 1 m). Groundwater is assumed at steady-state and moves from south to north along the

main facies direction driven by a hydraulic gradient of 0.001 in the x-direction and 0.002 in the y-direction. Prescribed heads are fixed at all boundaries according to this hydraulic gradient.

Solute transport was simulated with a random walk code that solves the advection-dispersion equation [Fernàndez-Garcia et al., 2005; Henri and Fernàndez-Garcia, 2014]. Transport parameters were considered constant with a porosity of 0.3, a longitudinal dispersivity of 0.1 m, and a transverse dispersivity of 0.01 m. The effect of heterogeneity inside each facies was not considered to only focus on the reconstruction problem. Contaminant concentrations were observed at a control plane located at y=5 m. We then compare the transport simulations obtained with the reference hydraulic conductivity field with those resulting from the one hundred SKR(%) and NNC realizations generated using a sample density of 30 data points.

The cumulative breakthrough curves are shown in Figure 9 normalized by the total mass injected. The ensemble of solutions provided by the SKR(%) and NNC methods is represented by the median and the 95% confidence interval (yellow region in this figure). Individual realizations are also depicted. Results clearly show that the SKR is more robust than the NNC method in terms of transport predictions. Even though the median solution provided by both methods is close to the true solution, the confidence interval of the SKR method is strikingly smaller than that obtained by the NNC, an effect that is more pronounced at late times. This indicates that the probability that reality is not properly represented by the SKR method is substantially smaller. Remarkably, this also reflects that, in many of the NNC realizations, the contaminant is forced to move through inexistent small permeability areas, resulting in artificial tailing and an artificial retardation. We also note that in some realizations, the poor south-north connection described by the NNC is such that the contaminant is partially exiting the system from the east and west boundaries without reaching at the control plane (note that some breakthrough curves do not contain all the mass injected).

5. Conclusions

A non-parametric method, SKR, originally designed for image processing [*Takeda et al.* 2007], has been presented and tested for its application as a facies delineation algorithm. The performance of the method was compared with the nearest neighbor classification,

a method that has proven to be more efficient than others discussed in the literature [*Tartakovsky et al.*, 2007]. Four synthetic scenarios were used for the comparison: two of them identical to the figures presented by *Tartakovsky et al.* [2007], and the other two figures are new for this work, one inspired on a cartographied river meander, and the other being a representation of a simple geometry. For each example different tests were studied ranging from very sparse to sparse number of data points available.

Two variations of the SKR method were tested depending on whether additional information about the exact proportion of facies was introduced in the algorithm (SKR(%)) or not (SKR(0)). Our results indicate that the SKR(0) method had similar or lower fractional errors than those obtained with NNC, except for two cases (Figure 1(c) and (d), with a sampling density of 0.28%). The SKR(%) outperformed all methods, with improvements up to 5% in terms of reduction in misclassified points. The improvement is better in relative terms for the lowest sampling densities. This finding leads us to believe that the SKR(%) method would be an useful tool on real cases, when scattered and few sampling data points are expected.

One of the major advantages of the SKR method is the quantification of the uncertainty in the delineation of the facies boundaries. In this context, we presented a method to stochastically generate variance maps that allows one to identify potential areas where a boundary between facies is more likely to exist. An example of application for one of the study cases is provided, leading to the delineation of an area over which there is most probably a boundary between facies.

Acknowledgements

This work has been supported by the Spanish Ministry of Science and Innovation through projects Consolider-Ingenio 2010 CSD2009-00065 and FEAR CGL2012-38120. XS acknowledges support of Program ICREA Acadèmia.

Appendix: The nearest-neighbor classification (NNC)

The nearest-neighbor classification (NNC) employed by *Tartakovsky et al.* [2007] is a k-nearest-neighbor classification [*Hastie et al.*, 2001] in which the classification of a test point is determined by majority vote amongst the k nearest-neighbor points in the

training set, *Tartakovsky et al.* [2007] considered the case in which k=1, for which the classification of each point in the domain is determined by finding the nearest training point, and assigning the known class of that point. Given a set of training data points $I_i=I(\mathbf{x}_i, F_k)$, i=1,…,n, the NNC classification for an arbitrary point **x** in the domain is computed as follows: (1) Define j as the index of the training data point, from the set $\{x_i\}_{i=1}^{N}$, which is closest to query point x; that is, $j = \mathbf{argmin}_i \|x - x_i\|_2$. Usually an Euclidean measure is prefer as distance metric, for simplicity, however, other metric can be used; (2) Assign the indicator function value of training data point x_j (i.e., $I(x_j)$) as the indicator function value of query point x. This classification is simple to compute, and has no free parameters to estimate (no optimization of the method is possible).

References

Bennett, K.P., Campbell, C., 2000. Support vector machines: Hype or hallelujah? SIGKDD Explorations, 2(2).

Bughin, E., Blanc-Feraud, L., Zerubia, J., 2008. Satellite image reconstruction from an irregular sampling. Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 849-852. doi: 10.1109/ICASSP.2008.4517743

Daoudi, M., Ghorbel, F., Mokadem, A., Avaro, O., Sanson, H., 1999. Shape distances for contour tracking and motion estimation. Pattern Recognition, 1297-1306.

Dixon, P.M., 2002. Nearest neighbour methods, in Encyclopedia of Environmetrics, vol. 3, edited by A. H. El-Shaarawi andW.W. Piegorsch, pp. 1370– 1383, John Wiley, New York.

Guadagnini, L., Guadagnini, A., Tartakovsky, D.M., 2004. Probabilistic reconstruction of geologic facies. Journal of Hydrology, 294, 57-67.

Feng, X., Milanfar, P., 2002. Multiscale principal components analysis for image local orientation estimation. 36th Asilomar Conf. Signals, Systems and Computers.

Fernàndez-Garcia, D., Illangasekare, T. H., Rajaram, H., 2005. Differences in the scale-dependence of dispersivity estimated from temporal and spatial moments in physically and chemically heterogeneous porous media, Adv. Water Res., 28, 745-759, 2005.

Fernàndez-Garcia, D., Trinchero, P., Sanchez-Vila, X., 2010. Conditional stochastic mapping of transport connectivity, Water Resour. Res., 46, Art. No. W10515.

Franke, R., 1982. Scattered Data Interpolation: Tests of Some Methods. Mathematics of Computation, 38(157), 181-200.

Hardle, W., 1990. Applied nonparametric regression, Econometric Society Monographs No. 19, Cambridge University Press, 333 p.,

Henri, C. V., Fernàndez-Garcia D., 2014. Toward efficiency in heterogeneous multispecies reactive transport modeling: A particle-tracking solution for first-order network reactions, Water Resour. Res., 50, doi:10.1002/2013WR014956.

Isaaks, E. H., Srivastava, R. M., 1990. An Introduction to Applied Geostatistics, Oxford Univ. Press, New York.

Journel, A.G., 1983. Non-parametric estimation of spatial distribution. Mathematical Geology, 15(3).

Li, Qi; Racine, Jeffrey S., 2007. Nonparametric Econometrics: Theory and Practice. Princeton University Press.

Lin, S.C., Chen, C.T., 2008. Reconstructing vehicle license plate Image from low resolution images using nonuniform interpolation method. International Journal of Image Processing, 1(2).

Ollero Ojeda, A., 1996. Dinámica de meandros y riesgos hidrogeomorfológicos en Alcalá de Ebro y Cabañas de Ebro (Zaragoza). In: IV Reunión de

Geomorfología, Grandal d'Anglade, A. and Pagés Valcarlos, J. Eds. Sociedad Española de Geomorfología.

Ritzi, R.W. Jr., Jayne, D.F., Zahradnik, A.J.Jr., Field, A.A., Fogg, G.E., 1994. Geostatistical Modeling of Heterogeneity in Glaciofluvial, Buried-Valley Aquifers. Ground Water 32(4), 666-674.

Simonoff, J.S., 1996. Smoothing Methods in Statistics. Springer.

Strebelle, S., 2000. Sequential simulation drawing structures from training images: Unpublished doctoral dissertation, Stanford University, 200 p.

Takeda, H., Farsiu, S., Milanfar, P., 2007. Kernel Regression for Image Processing and Reconstruction. IEEE Transactions on Image Processing, Vol. 16, No. 2, February 2007.

Tarbuck, E.J., Lutgens, F.K., 2002. Earth: An Introduction to Physical Geology. Seventh Edition. Prentice Hall

Tartakovsky, D.M., Wohlberg, B.E., 2004. Delineation of geologic facies with statistical learning theory. Geophysical Research Letters, Vol. 31, L18502, doi:10.1029/2004GL020864

Tartakovsky, D.M., Wohlberg, B., Guadagnini, A., 2007. Nearest-neighbor classification for facies delineation. Water Resour. Res., 43, W07201, doi: 10.1029/2007WR005968

Vapnik, V., Lerner, A., 1963. Pattern recognition using generalized portrait method. Automation and Remote Control, 24, 774–780.

Wohlberg, B., Tartakovsky D.M., Guadagnini A., 2006. Subsurface characterization with support vector machines. IEEE Transactions on Geoscience and Remote Sensing, Vol. 44, No. 1, January 2006.

Yang, G., Huang, T.S., 1994. Human face detection in a complex background. Pattern Recognition. Volume 27, Issue 1, January 1994, Pages 53–63.

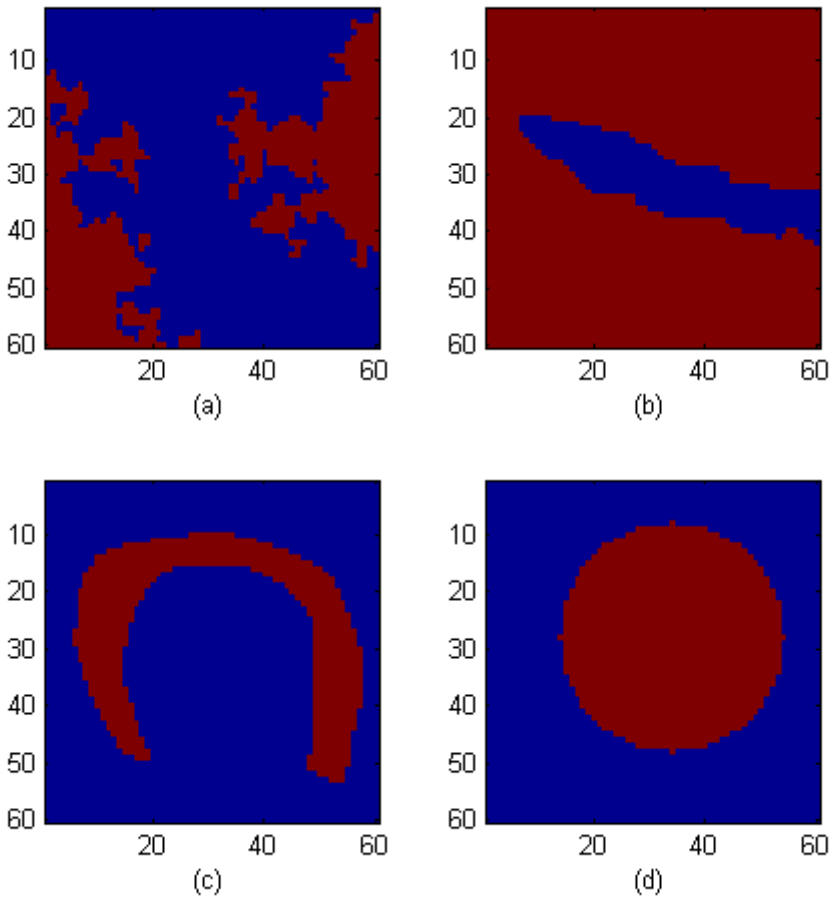

Figure 1. Synthetic fields used for facies delineation: a and b are the same figures presented by Tartakovsky et al. [2007]. We generated Figure 1 (c) and (d) considering a real case scenario (a meander from the Ebro river, Spain), and a simple geometric figure (circle). Blue and red colors indicate the two distinct facies.

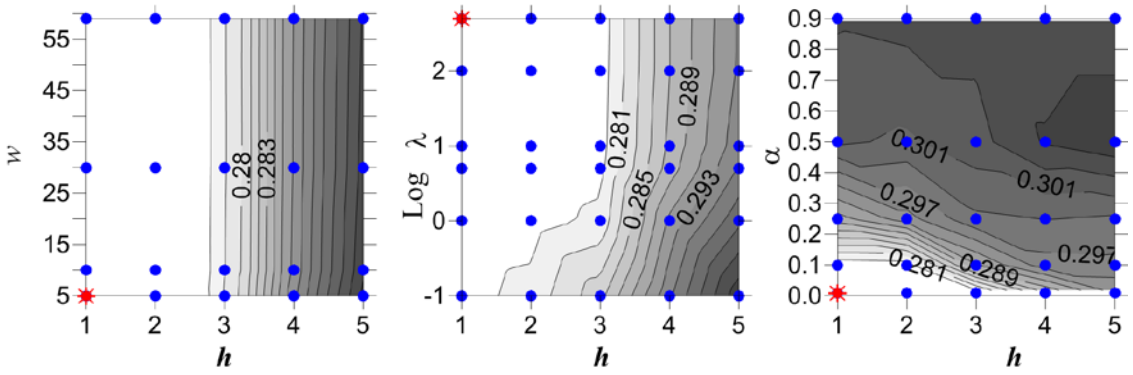

Figure 2. Sensitivity analysis of the parameters needed to reconstruct geological facies with the SKR method: the local orientation analysis window (w), the regularization for the elongation parameter (λ), the structure sensitive parameter (α) and the global smoothing parameter (h). Blue dots indicate the different value choices for the calculation of the fractional errors and the red star indicates the value used for our calculations, coincidently with the lowest fractional error.

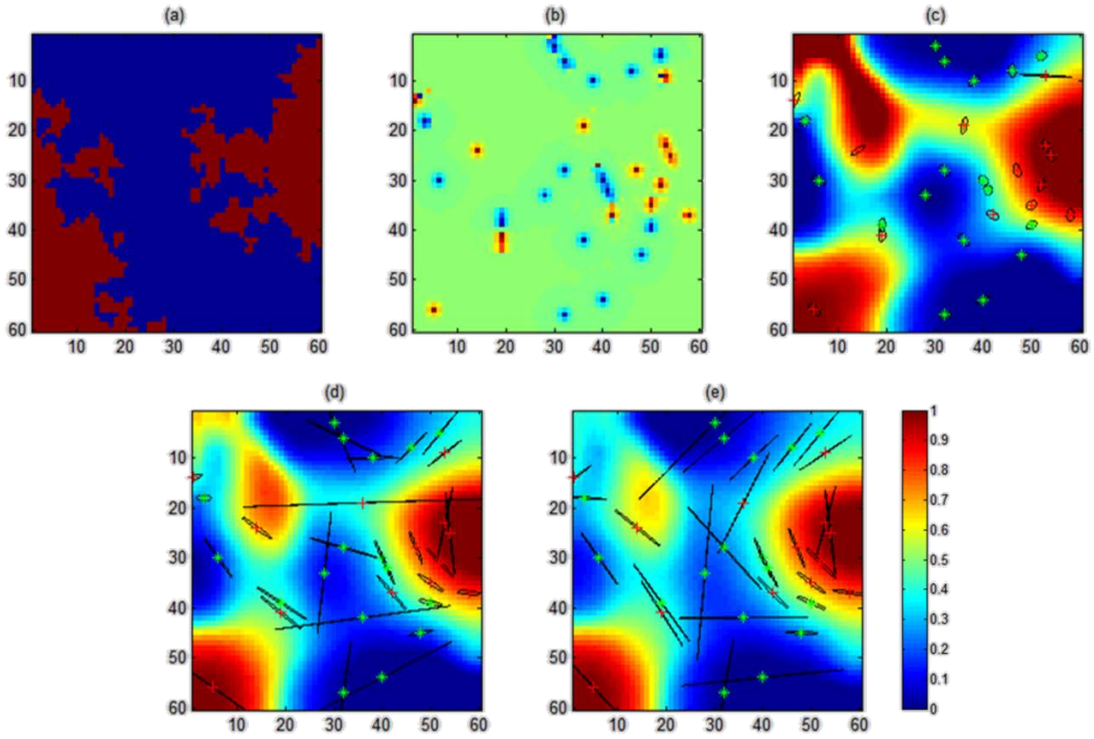

Figure 3. Iteration comparison: a) Original figure corresponding to Figure 1a. Random sampling points are shown as blue and red squares (example with a sample density of 30); b) Classical Kernel Regression results. The first, second and third iteration of the Steering Kernel is shown in c), d) and e), respectively.

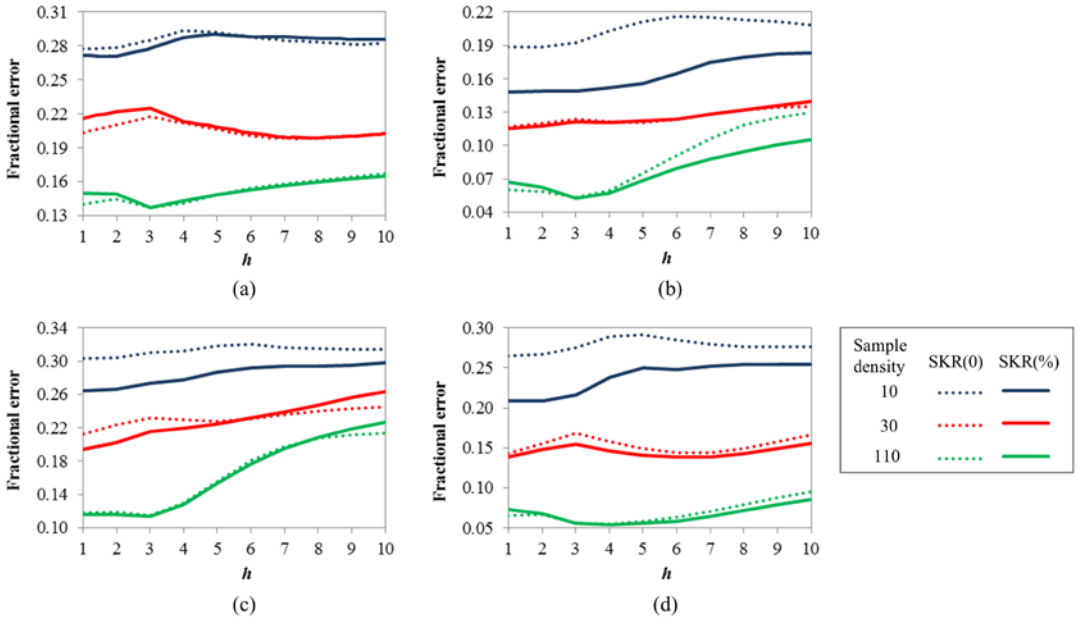

Figure 4. Fractional error variation as a function of the smoothing parameter *h* for the four figures analyzed. Figures presented in the same order as shown in Figure 1: a) Figure A, b) Figure B, c) Meander, d) Ball. Discontinuous and continuous lines represent respectively the fractional error when SKR(0) and SKR(%) are considered.

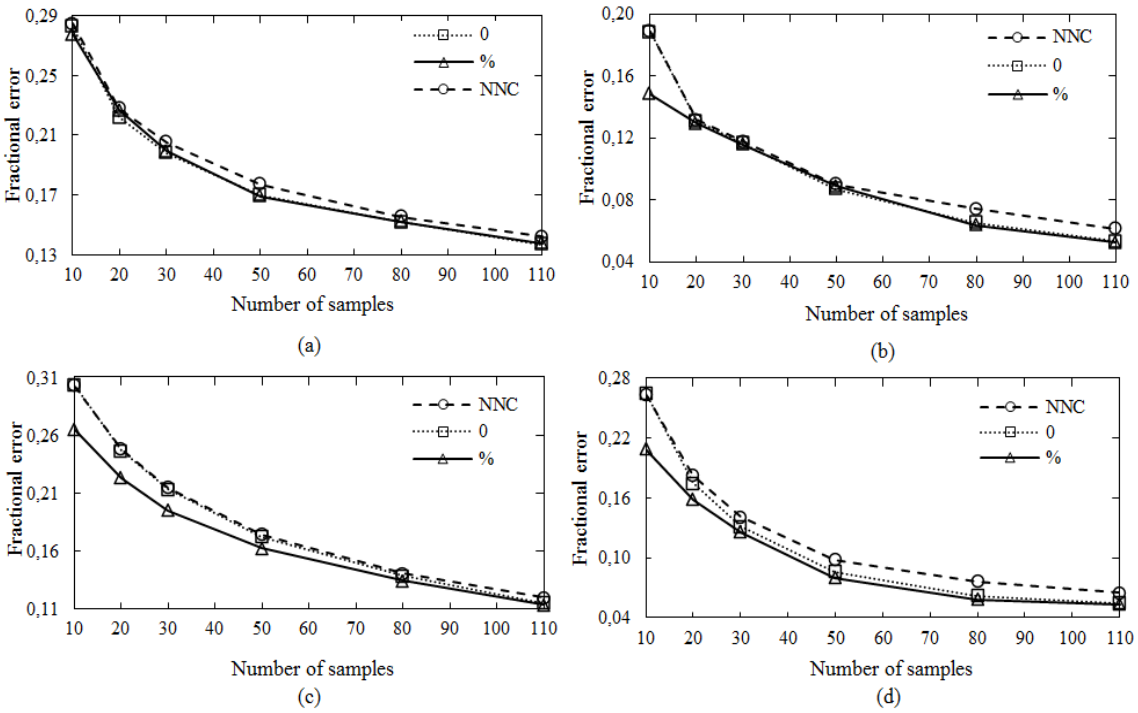

Figure 5. Fractional error comparison: From top to bottom, synthetic fields (a), (b), (c) and (d) ordered according to Figure 1. NNC stands for nearest neighbour classification, 0 for SKR(0) and % for SKR(%).

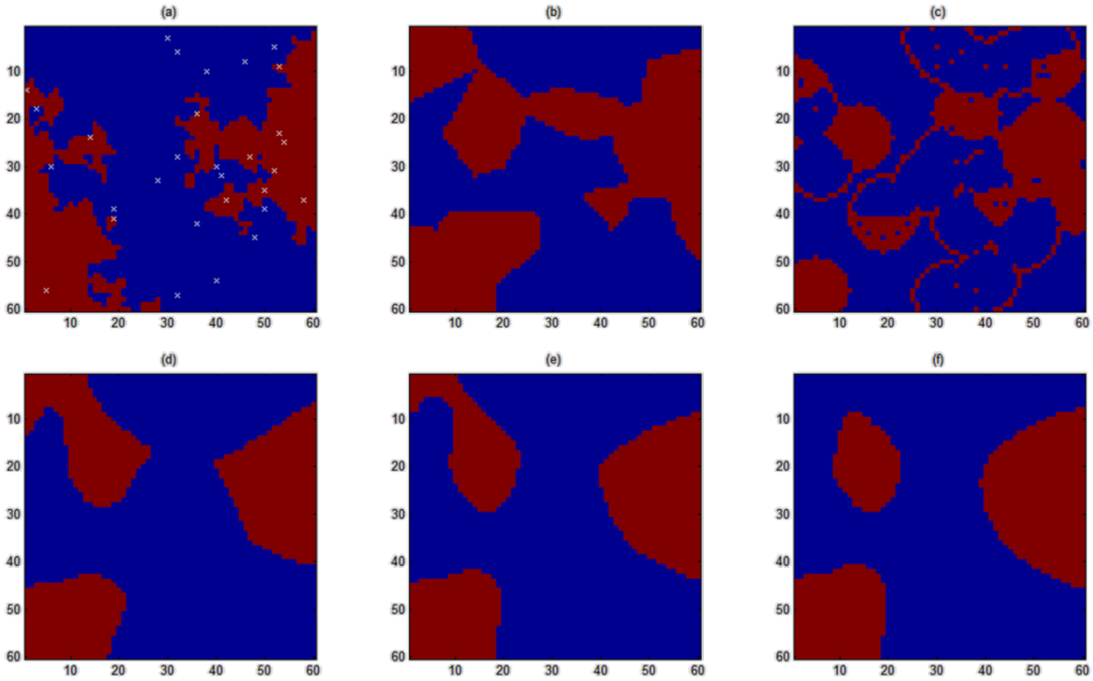

Figure 6. (a) Original figure showing the location of the random samples considered; (b) Nearest-neighbor classification; (c) Classic kernel regression h=1. Steering kernel regression: (d) iteration 1, (e) iteration 2, (f) iteration 3. Figures d, e and f are the result of equation (15) with (18).

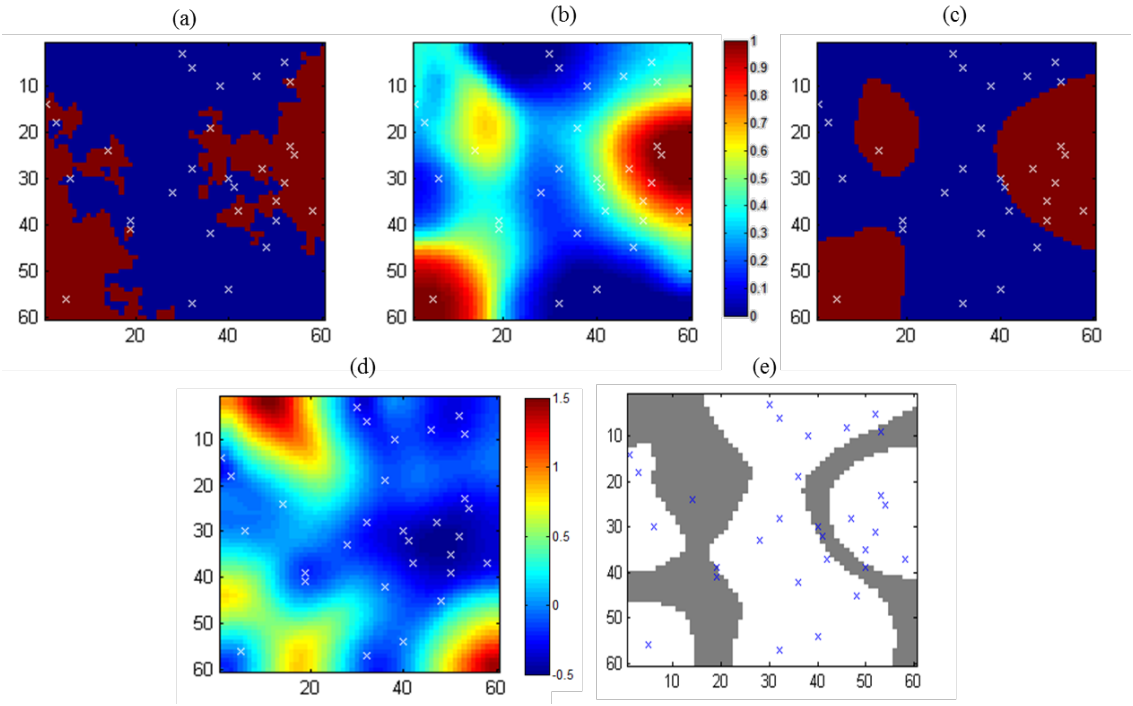

Figure 7. (a) Original figure, (b) steering kernel iteration 3, (c) steering kernel iteration 3 after equation (15) with (18), (d) Variance map showing the areas with the highest and lowest uncertainty (red and blue zones), (e) standard deviation map, showing in gray the area where the border between facies is more likely located.

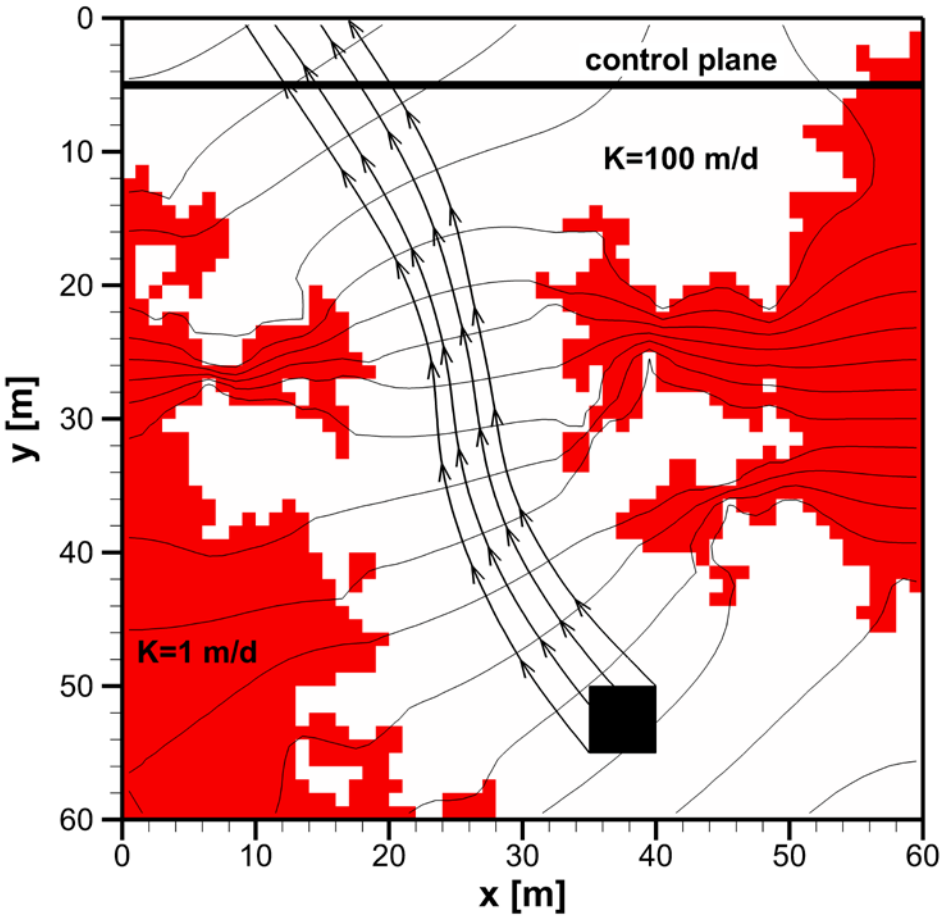

Figure 8. Setup of transport simulations.

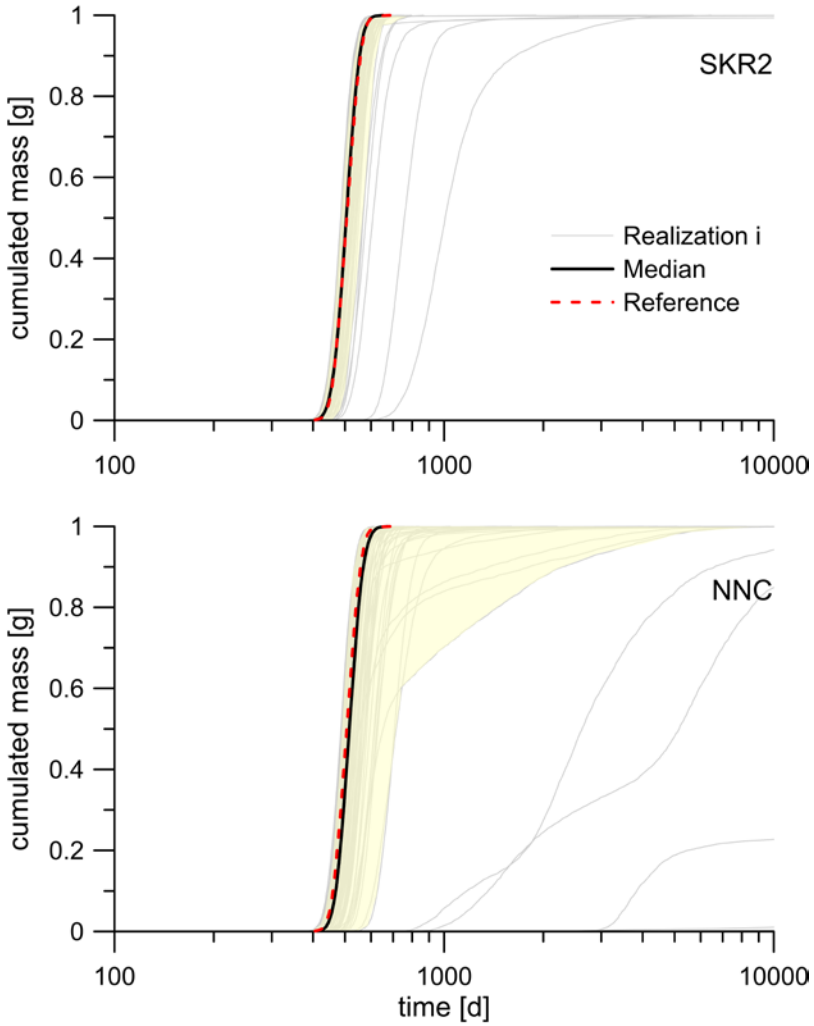

Figure 9. Cumulative breakthrough curves normalized by the total mass injected associated with the geological setting presented in Figure 1a. Comparison of the reference solution with the SKR and the NNC reconstructed fields with 30 sample locations. Yellow region depicts the 95% confidence interval over 100 realizations.

.