Journal of Hydrology 531 (2015) 62–72

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

A locally adaptive kernel regression method for facies delineation

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

Department of Geotechnical Engineering and Geosciences, Universitat Politècnica de Catalunya (UPC), Jordi Girona 1-3, 08034 Barcelona, Spain
Associated Unit: Hydrogeology Group (UPC-CSIC), Spain

ARTICLE INFO

Article history:
Available online 30 September 2015

Keywords:
Stochastic groundwater hydrology
Geological facies reconstruction
Geostatistics
Solute transport

SUMMARY

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 and Huang, 1994; Lin and Chen, 2008). 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 and Srivastava, 1990; Ritzi et al., 1994; Guadagnini et al., 2004), support Vector Machines (SVMs) (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.

* Corresponding author at: Department of Geotechnical Engineering and Geosciences, Universitat Politècnica de Catalunya (UPC), Jordi Girona 1-3, 08034 Barcelona, Spain.

E-mail address: daniel.fernandez.g@upc.edu (D. Fernàndez-Garcia).

http://dx.doi.org/10.1016/j.jhydrol.2015.09.066
0022-1694/

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, and fusion. 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. Since the NNC was already shown to outperform the IK and the SVM methods in these examples, we limited the comparison to the NNC method.

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, and grain size (Tarbuck and Lutgens, 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

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

where the indicator variable $I(\mathbf{x}, F_k)$ is equal to 1 when a particular point in the domain, $\mathbf{x}$, 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 $\mathbf{x}$ 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 and Lerner, 1963; Bennett and Campbell, 2000). SVM takes a training data, i.e., a set of n data points $J_i = J(\mathbf{x}_i, F_1) \in \{-1, 1\}$, $i = 1, \ldots, n$, 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 Wohlberg (2004), Wohlberg 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 and Racine, 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

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

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., $m(\mathbf{x}, \mathbf{b}) = E[I(\mathbf{x}, F_1)]$. 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