Advances in Water Resources 32 (2009) 225–230

Contents lists available at ScienceDirect

Advances in Water Resources

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

Delineation of geological facies from poorly differentiated data

Brendt Wohlberg [a,1], Daniel M. Tartakovsky [b,*,2]

[a] Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA
[b] Department of Mechanical and Aerospace Engineering, University of California, San Diego, 9500 Gilman Dr., MC 0411, La Jolla, CA 92093, USA

ARTICLE INFO

Article history:
Received 8 July 2008
Received in revised form 23 October 2008
Accepted 29 October 2008
Available online 6 November 2008

Keywords:
Geostatistics
Nearest neighbor
Undifferentiated
Classification
Measurement error

ABSTRACT

The ability to delineate geologic facies and to estimate their properties from sparse data is essential for modeling physical and biochemical processes occurring in the subsurface. If such data are poorly differentiated, this challenging task is complicated further by the absence of a clear distinction between different hydrofacies at locations where data *are* available. We consider three alternative approaches for analysis of poorly differentiated data: a *k*-means clustering algorithm, an expectation–maximization algorithm, and a minimum-variance algorithm. Two distinct synthetically generated geological settings are used to analyze the ability of these algorithms to assign accurately the membership of such data in a given geologic facies. On average, the minimum-variance algorithm provides a more robust performance than its two counterparts, and when combined with a nearest neighbor algorithm, it also yields the most accurate reconstruction of the boundaries between the facies.

1. Introduction

Our knowledge of the spatial distribution of the physical properties of geologic formations is often uncertain because of ubiquitous heterogeneity and the scarcity and sparsity of information. Yet capturing the complexity of natural hydrogeological systems and quantifying the associated uncertainty is of paramount importance for reliable groundwater flow and transport assessments. While many studies combine several types of information (including hydraulic conductivity, electrical resistivity, hydraulic heads and/or solute travel times) to predict the salient features of flow and transport in heterogeneous subsurface environments, the uncertainty associated with the delineation of lithofacies and their hydraulic properties (e.g., hydraulic conductivity and porosity) from limited geological and geophysical data is only marginally analyzed. Such data, which include grain size distribution curves, are typically derived from core samples and are often poorly differentiated, further compounding predictive uncertainty.

Geostatistics has become an invaluable tool for estimating facies distributions and attributes of facies at points in a computational domain where data are not available, as well as for quantifying the corresponding uncertainty [3]. In the presence of poorly differentiated data, or data with low signal-to-noise ratios, identification of heterogeneous aquifer structure is often performed in two steps. First, a multivariate facies-based parameterization approach relying on multivariate cluster analysis [5] is applied to classify aquifer materials and to estimate their spatial arrangement [7]. Second, Kriging is used to estimate hydraulic and other properties within each cluster (a sedimentological facies).

Geostatistical frameworks treat the properties of a formation, such as hydraulic conductivity K, as a random process that is characterized by multivariate probability density functions or, equivalently, by ensemble moments. Whereas spatial moments of K are obtained by sampling K in physical space, its ensemble moments are defined in terms of samples collected in probability space. In reality only a single realization of a geologic site exists. Therefore, it is necessary to invoke the ergodicity hypothesis in order to substitute the sample spatial statistics, which can be calculated, for the ensemble statistics, which are actually required as input to a stochastic model of flow or contaminant transport. Ergodicity cannot be proved, and requires a number of modeling assumptions. Alternatives to geostatistics include neural networks [6], support vector machines [9,11], and nearest neighbor classifications [10].

These and other similar approaches to facies delineation rely on one's ability to classify available data, i.e., to establish their membership in a given geological facies. The task of assigning the values of an indicator function to hydraulic and soil properties data is nontrivial if properties in question are either poorly differentiated or characterized by low signal-to-noise ratios, a situation often encountered in geophysical site characterization. Section 2 contains a mathematical formulation of this problem. We present three alternative approaches to classify poorly differentiated data

* Corresponding author. Fax: +1 505 665 5757.
E-mail addresses: brendt@t7.lanl.gov (B. Wohlberg), dmt@ucsd.edu (D.M. Tartakovsky).

[1] This research was supported by the NNSA's Laboratory Directed Research and Development Program.
[2] This research was supported in part by the DOE's Office of Advanced Scientific Computing Research.

0309-1708/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.advwatres.2008.10.014

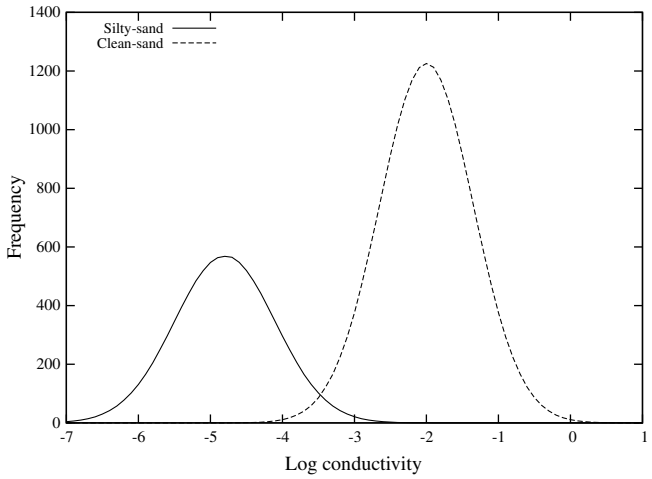

Fig. 1. A typical sample frequency distribution of the log hydraulic conductivity $Y = \ln K$ of a subsurface environment composed of silty sand and clean sand (reference fields). The log hydraulic conductivity of the silty sand and clean sand facies ranges between −7.00 and −2.70 and −4.15 and 0.60, respectively.

in Section 3, and use this classification to reconstruct the boundaries between geological facies by means of a nearest neighbor classification in Section 4. The proposed approaches are analyzed by considering two synthetic porous media in Section 5.

2. Facies delineation from poorly differentiated data

We consider a problem of reconstructing a boundary between two heterogeneous materials M_1 and M_2 from spatially distributed parameter data. The latter can consist of hydraulic data (e.g., hydraulic conductivity), geophysical data (e.g., electric resistivity), and/or sedimentological data, $\{K_i \equiv K(\mathbf{x}_i)\}_{i=1}^N$ collected at N locations $\mathbf{x}_i = (x_i, y_i)^T$, where $i \in \{1, \ldots, N\}$ and the superscript T denotes the transpose. (While all three methods presented below are applicable to three-dimensional settings, we present our results in two dimensions to simplify the presentation.) The first step in our facies delineation procedure is to analyze the distributions of samples with the goal of assigning an indicator function

$$I(\mathbf{x}) = \begin{cases} 1, & \mathbf{x} \in M_1, \\ 0, & \mathbf{x} \in M_2 \end{cases} \tag{1}$$

to each point where data are available. This is precisely the step that is affected most by the poor differentiation of data. Consider, for example, a subsurface environment consisting of two heterogeneous facies that are formed by clean sand and silty sand. A typical histogram of hydraulic conductivity data for such an environment is shown in Fig. 1. The measurements falling in the overlapping region between the two distributions do not render themselves to a straightforward classification by (1). We refer to such measurements as poorly differentiated data.

To assign values of the indicator function (1) to such data, we consider three alternative statistical approaches, which are described in some detail in Section 3.

3. Classification of poorly differentiated data

By their very definition, poorly differentiated data do not lend themselves to an unambiguous classification. Instead, such a classification has to be estimated. We compare the relative performance of three alternative strategies: a k-means clustering algorithm, an expectation–maximization algorithm, and a minimum-variance algorithm.

3.1. k-Means clustering algorithm

The k-means clustering algorithm [4, p. 412], one of the first and still most popular classification algorithms, consists of the following steps:

(1) Identify the number of clusters – in our example, one cluster for each of the two geologic facies.
(2) Treat the minimum and maximum value of hydraulic conductivity as initial values for the means (centroid positions) of the respective populations.
(3) Assign each of the conductivity measurements to the cluster with the closest centroid.
(4) Recalculate the centroids based on the current cluster assignments.
(5) Repeat steps 3 and 4 until the centroid positions stabilize.

In our experiments, we used the `kmeans` function from a Matlab clustering toolbox [2].

3.2. Expectation–maximization algorithm

The expectation–maximization (EM) algorithm [4, p. 236] takes advantage of the fact that material properties of individual geological units can often be characterized by classical unimodal distributions, while their counterparts sampled across various geological units comprising the subsurface cannot. For example, many geological facies are routinely characterized by log-normally distributed hydraulic conductivity [8] and grain sizes [1]. While the EM algorithm is equally applicable to any number of geological facies and distributions, our presentation below is limited to two hydrofacies whose log conductivities are Gaussian.

The EM algorithm treats the data $\{Y_i \equiv \ln K(\mathbf{x}_i)\}_{i=1}^N$ as samples from a population $\mathscr{Y}$ that represents a mixture of two Gaussian populations $\mathscr{Y}_1$ and $\mathscr{Y}_2$

$$\mathscr{Y} = (1-\lambda)\mathscr{Y}_1 + \lambda\mathscr{Y}_2, \quad \mathscr{Y}_k = N(\overline{\mathscr{Y}}_k, \sigma_k^2), \quad k \in \{1,2\}. \tag{2}$$

The random variable λ takes the value of 1 with the probability $Pr[\lambda = 1] = p$ and of 0 with the probability $Pr[\lambda = 0] = 1 - p$. The mean $\overline{\mathscr{Y}}_k$ and variance σ_k^2 of the kth population (geological facies) and the value of the probability p are determined by maximizing the likelihood function $\mathscr{L}$

$$\max_{p, \overline{Y}_k, \sigma_k^2} \mathscr{L}, \quad \mathscr{L} \equiv \sum_{i=1}^N \ln f_{\mathscr{Y}}(Y_i), \quad f_{\mathscr{Y}}(y) = (1-p) f_{Y_1}(y) + p f_{Y_2}(y), \tag{3}$$

where $f_{\mathscr{Y}}$ is the probability density function (PDF) of the random field $\mathscr{Y}$ in (2), and f_{Y_i} is the Gaussian PDF of the random variable $\mathscr{Y}_i$ ($i \in \{1,2\}$).

The EM algorithm for solving (3) consists of the following steps [4, p. 238]:

(1) Make an initial guess for p, $\overline{\mathscr{Y}}_k$ and σ_k^2 ($k \in \{1,2\}$).
(2) Compute the so-called responsibilities γ_i (the expectation step)

$$\gamma_i = \frac{p f_{Y_2}(Y_i)}{f_{\mathscr{Y}}(Y_i)}, \quad i = 1, \ldots, N, \quad \Gamma_1 = \sum_{i=1}^N (1-\gamma_i), \quad \Gamma_2 = \sum_{i=1}^N \gamma_i.$$

(3) Modify the initial guess (the maximization step) by computing the means

$$\overline{Y}_1 = \frac{1}{\Gamma_1} \sum_{i=1}^N (1-\gamma_i) Y_i, \quad \overline{Y}_2 = \frac{1}{\Gamma_2} \sum_{i=1}^N \gamma_i Y_i,$$

variances

$$\sigma_1^2 = \frac{1}{\Gamma_1}\sum_{i=1}^{N}(1-\gamma_i)(Y_i-\overline{Y}_1)^2, \quad \sigma_2^2 = \frac{1}{\Gamma_2}\sum_{i=1}^{N}\gamma_i(Y_i-\overline{Y}_2)^2$$

and the membership probability

$$p = \frac{1}{N}\sum_{i=1}^{N}\gamma_i.$$

(4) Repeat steps 2 and 3 until convergence with a prescribed tolerance is achieved.

In our experiments, we employed the `mixtureEM` function from a Matlab clustering toolbox [2]. The convergence of the EM depends on the choice of an initial guess. To facilitate the convergence, we used the k-means clustering results to provide initial mean values, instead of the default random initialization.

3.3. Minimum-variance algorithm

We compare these two algorithms with an algorithm that partitions data, i.e., assigns the values of the indicator function, in a way that minimizes the variability within each geologic facies. To the best of our knowledge, this approach is new, at least in the present context. We accomplish this goal with the following algorithm:

(1) Sort the values in the data set $\{K_i\}_{i=1}^N$ from the smallest to the largest.
(2) Let N_1 be a cutoff point separating this set into two, $\{K_i\}_{i=1}^{N_1}$ and $\{K_i\}_{i=N_1+1}^{N}$.
(3) Consider the sum of the variances in both sets:

$$\Sigma = \frac{1}{N_1}\sum_{i=1}^{N_1}(K_i-\mu_1)^2 + \frac{1}{N-N_1}\sum_{i=N_1+1}^{N}(K_i-\mu_2)^2,$$

where μ_1 and μ_2 denote the corresponding means.
(4) The partition is defined by $N_1^\star$ that minimizes Σ.

4. Delineation of geological facies

A variety of conceptual frameworks and computational approaches have been proposed to estimate boundaries between geological facies from sparse data (see Section 1). The starting point of such approaches is to assign the values of the indicator function (1), a task that is prone to interpretive errors if available data are poorly differentiated and/or the signal-to-noise ratio is small (Section 2).

This task can be achieved with the three alternative approaches described in Section 3. We now proceed by describing nearest neighbor classification (NNC) as a means for reconstructing geological facies from an estimated indicator function data set $\{I_i\}_{i=1}^N$ where $I_i \equiv I(\mathbf{x}_i)$. We use NNC because it outperforms both a geostatistical approach and support vector machines when applied to well-differentiated data [10].

Given a set of data points $\{\mathbf{x}_i\}_{i=1}^N$ with corresponding indicator function values I_i, NNC uses the following algorithm to assign the

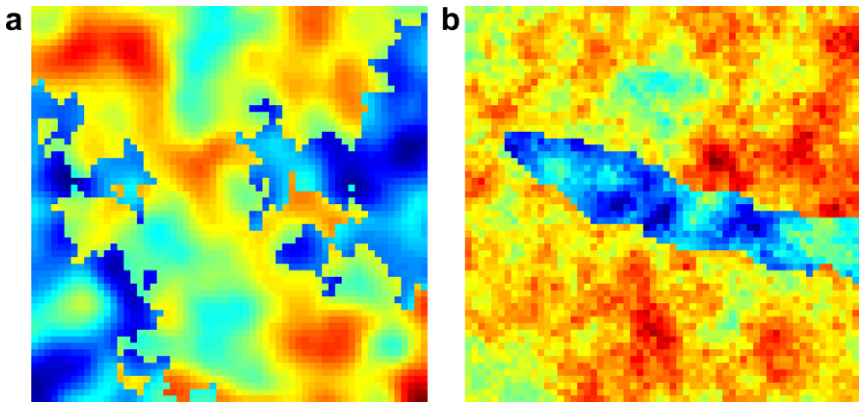

Fig. 2. Synthetic porous media, whose log hydraulic conductivity takes values between −6.9 (dark blue) and 0.6 (red). These computational examples pose different reconstruction challenges: the porous medium (a) exhibits highly irregular internal boundary, while the porous medium (b) contains a preferentially directed small inclusion. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

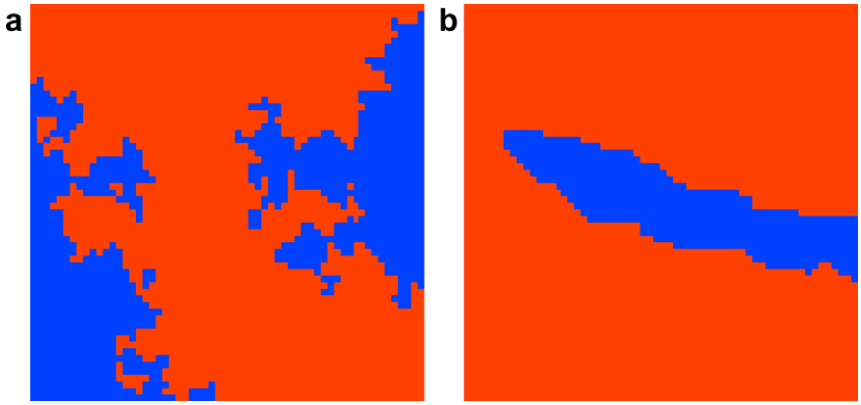

Fig. 3. Facies maps (indicator functions) used to generate the two synthetic porous media in Fig. 2.

value of the indicator function I at a point $\mathbf{x}$ where measurements are not available:

(1) Define j as the index of the training data point, from the set $\{\mathbf{x}_i\}_{i=1}^N$, which is closest to the point $\mathbf{x}$; i.e., $j = \text{argmin}_i \|\mathbf{x} - \mathbf{x}_i\|_2$.
(2) Assign the indicator function value I_j at the data point $\mathbf{x}_j$ to the indicator function value at the point $\mathbf{x}$.

It is worthwhile noting that in addition to better performance, NNC makes no operational assumptions and has no free (fitting) parameters.

5. Computational examples

To test our approach for facies delineation from poorly differentiated data, we consider the two synthetic porous media shown in Fig. 2. The following two-step procedure was used to generate both examples, i.e., to assign a value of log hydraulic conductivity to each point (pixel). First, we generated two autocorrelated, weakly stationary Gaussian fields with ensemble means of -4.96 and -2.30, respectively. (The mutually uncorrelated random fields had unit variance and Gaussian autocorrelation with unit correlation scale.) Second, these fields were superimposed onto the facies map in Fig. 3.

The goal of our numerical experiments is to reconstruct the boundaries between the two materials in Fig. 3 from a few (randomly selected) measurements of log conductivity (Fig. 2). We considered data sets consisting of 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, and 500 measurements. Figs. 4 and 5 present typical histograms of data sets used to reconstruct the two media in Fig. 2.

The reconstruction quality clearly depends on the locations of the sampling points. To minimize this effect, we averaged simulation results over 10,000 randomly generated realizations of the locations of data points for each sample size. Sample locations were selected from a uniform distribution, and each random realization was assigned an equal weight.

Our facies delineation approach consists of an initial step to estimate the classification of the poorly differentiated data, followed by a facies delineation step using the estimated classifications. We provide results for the initial data classification step in Section 5.1, and results for the full facies delineation problem in Section 5.2.

5.1. Data classification

Fig. 6 presents the classification errors (for classification of the poorly differentiated data) corresponding to the three alternative classification approaches described in Section 3. The error for each realization is defined as the number of misclassified data points relative to the total number of sample points in that realization, and the overall error reported for each sample size is the average over all realizations for that sample size. The classification errors for the porous medium in Fig. 2a are larger than those for its counterpart in Fig. 2b. This is to be expected since the boundary in the former is much more extensive and irregular than in the latter. When averaged over the two examples presented in Fig. 2, the minimum-variance approach performs slightly better than both k-means and expectation–maximization algorithms. The k-means

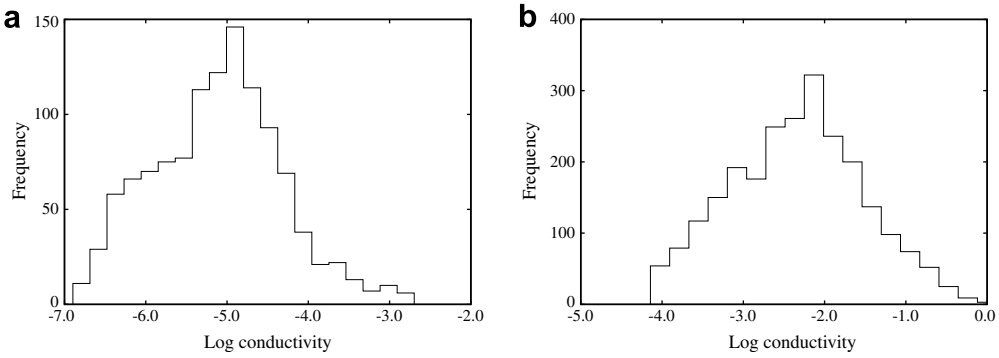

Fig. 4. Histograms of the log hydraulic conductivity values in the bluish (a) and reddish (b) regions of Fig. 2a.

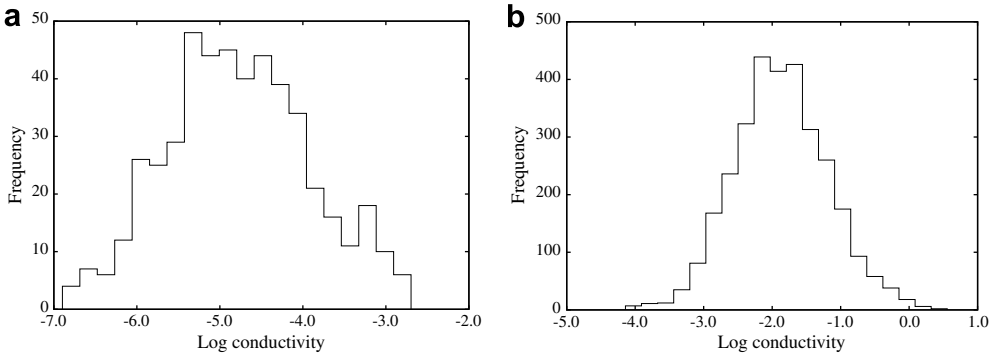

Fig. 5. Histograms of the log hydraulic conductivity values in the bluish (a) and reddish (b) regions of Fig. 2b.

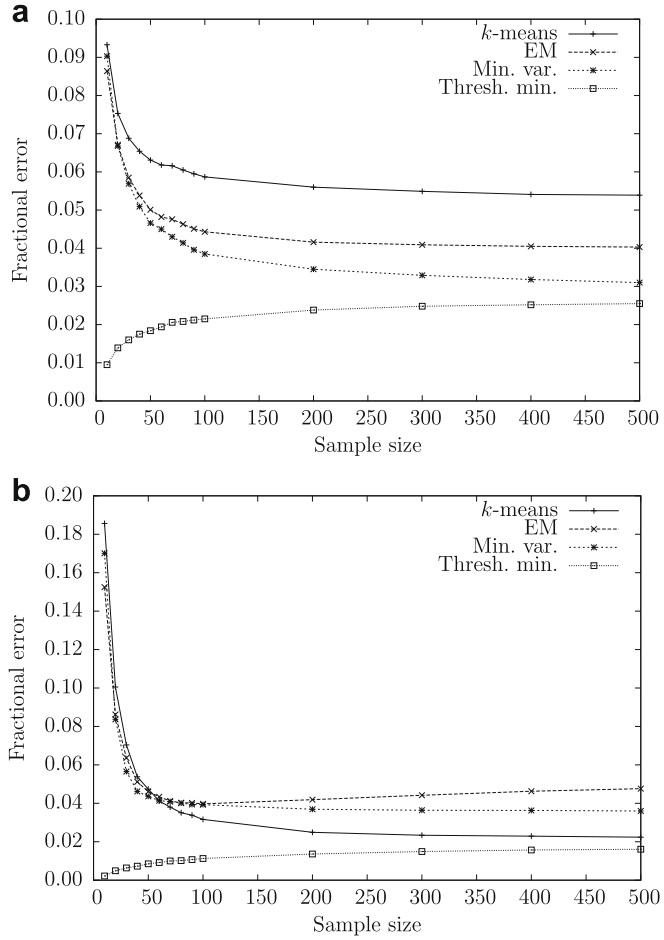

Fig. 6. Errors (the number of misclassified data points relative to the total number of sample points) in classification of conductivity samples from the porous media in Fig. 2a and b, respectively. The figure compares the performance of three alternative approaches to data classification – a k-means clustering algorithm, an expectation–maximization algorithm (EM), and a minimum-variance algorithm (Min. var.) – and the smallest possible threshold-based classification error (Thresh. min.).

provides the best performance on the computational example in Fig. 2b, but the worst performance on the computational example in Fig. 2a.

With a notable exception of the EM algorithm's performance in Fig. 6b, all three approaches to classification of poorly differentiated data exhibit uniformly convergent behavior in that the classification error decreases as the number of available measurements increases (Fig. 6). After a certain limit, the addition of more data provides little discernible gain in reducing the classification error. However, it is worthwhile recalling that in our analysis the measurement locations are selected at random. In actual field applications, one would expect a sampling strategy that relies on geophysical site characterization, expert opinion and other soft data to guide the selection of new sample locations.

Also shown in Fig. 6 is minimum threshold-based classification error, measured as the fraction of points misclassified with respect to the ground-truth classification (i.e., the true reference classification, determined by the facies maps in Fig. 3, from which the synthetic fields were generated). This error increases with sampling density, which may seem counterintuitive. To understand this behavior, it is important to recognize that the ground-truth classification may be such that it cannot be obtained from a threshold on the corresponding scalar values. This can be demonstrated by the two sample sets (sorted conductivity values and corresponding ground-truth classification) shown in Table 1. While the data in Sample 1 can be perfectly classified by a threshold of $(-4.2$

Table 1
Examples of threshold-based minimum classification errors.

Sample 1						
Measurement	−5.1	−4.5	−4.2	−3.4	−2.0	−1.9
Indicator	−1	−1	−1	+1	+1	+1
Sample 2						
Measurement	−5.4	−5.2	−4.3	−4.1	−3.6	−2.1
Indicator	−1	−1	+1	−1	+1	+1

$-3.4)/2$, the data in Sample 2 has a minimum threshold-based classification error of 2/6. We observe the increase of the minimum possible threshold-based estimation error as the number, N, of samples grows, the total number of ways of classifying the points (2^N) grows much faster than the number of ways of partitioning the points based on a threshold ($N+1$).

5.2. Facies reconstruction

After identifying the facies membership of the data points, i.e., after assigning the values of the indicator function to each data point, we use the NNC described in Section 4 to estimate the boundaries between the two facies in the two distinct geological settings shown in Fig. 2. Fig. 7 exhibits the boundary reconstruction errors introduced by this procedure when applied to the indicator function data estimated with the k-means clustering, expectation–maximization, and minimum-variance algorithms. The errors are reported as a number of misclassified pixels relative to the total number of pixels. Also presented in this figure are the

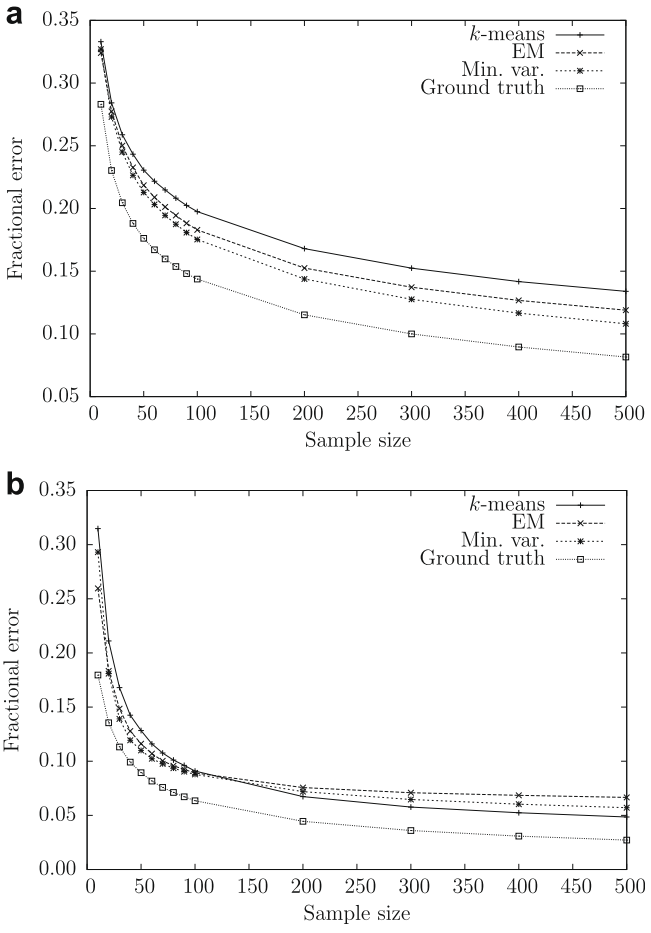

Fig. 7. Boundary reconstruction errors for the two synthetic porous media shown in Fig. 2. The errors are defined as a fraction of misclassified pixels relative to the total number of pixels.

reconstruction errors one would obtain with NCC if none of the data points were misclassified. This ground-truth reconstruction relies on the true indicator values shown in Fig. 3.

As can be expected, the misclassification errors decrease with the number of samples (data points) increases. The introduction of more data leads to the reduction in the reconstruction error, with a rate that is much faster at low sampling densities in the example in Fig. 2b than in the example in Fig. 2a. At higher sampling densities, adding more poorly differentiated data reduces the reconstruction errors in the example in Fig. 2a, while the effect on the reconstruction errors in the example in Fig. 2b is minimal. This effect is a result of the much greater complexity of the region boundary in Fig. 2a than in Fig. 2b, and is also observed for well-differentiated data. At low sampling densities, each additional point more accurately constrains the simple boundary in Fig. 2b than the complex boundary in Fig. 2a. At higher sampling densities, additional points are more likely to lie along the boundary, and therefore play a more prominent role in reducing the error in Fig. 2a than in Fig. 2b. In both cases the asymptotic error rate is non-zero, as a result of the non-zero asymptotic error rate for the underlying poorly differentiated data classification, on which the reconstruction is based. (In contrast, the reconstruction error rate at 100% sampling density is zero for well-differentiated data, for which the true classification is known a priori.) Finally, while the three alternative approaches to classification of poorly differentiated data result in considerably different outcomes in terms of classification error (Fig. 6), their impact on the reconstruction error is less pronounced (Fig. 7).

6. Conclusions

We analyzed the value (information content) of poorly differentiated data or data with low signal-to-noise ratios for the task of facies delineation. To classify such data, we considered two existing approaches, *k*-means clustering and expectation–maximization algorithms, and proposed a new one, the minimum-variance algorithm. Once classified, the data were used in conjunction with nearest neighbor classification to reconstruct two synthetic randomly generated porous media consisting of two heterogeneous materials. Our analysis leads to the following major conclusions:

(1) The selection of a proper classification algorithm has a significant impact on the data classification, with the minimum-variance algorithm being the most robust.
(2) The impact of this selection on errors in reconstruction of geological facies is significantly smaller.
(3) At low sampling densities, the addition of new data leads to a nearly exponential decrease in both classification and reconstruction errors.
(4) The value of additional data at high sampling densities is limited, with both errors reaching their asymptotic values.

It is worthwhile recalling that our results and conclusions hold on average, so that the impact of a fortuitous selection of measurement locations is either minimized or eliminated all together.

References

[1] Clausnitzer V, Hopmans JW. Determination of phase-volume fractions from tomographic measurements in two-phase systems. Adv Water Resour 1999;22(6):577–84.
[2] F. Dellaert, Matlab clustering package, Available from the author's website at College of Computing, Georgia Tech, USA, 2001. <http://www.cc.gatech.edu/dellaert/clusters.tgz>.
[3] Guadagnini L, Guadagnini A, Tartakovsky DM. Probabilistic reconstruction of geologic facies. J Hydrol 2004;294:57–67.
[4] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. New York, NY, USA: Springer; 2001.
[5] J. McQueen, Some methods for classification and analysis of multivariate observations, in: Fifth Berkeley Symposium on Mathematics, Statistics and Probability, vol. 1, 1967, pp. 281–98.
[6] Moysey S, Caers J, Knight R, Allen-King RM. Stochastic estimation of facies using ground penetrating radar data. Stoch Environ Res Risk Assess 2003;17:306–18.
[7] M. Riva, L. Guadagnini, A. Guadagnini, E. Martac, T. Ptak, A composite medium approach for probabilistic modelling of contaminant travel time distribution to a pumping well in a heterogeneous aquifer, in: Proceedings of the Fifth International Conference on Calibration and Reliability in Groundwater Modelling (ModelCARE 2005), 2005.
[8] Rubin Y. Applied stochastic hydrogeology. New York: Oxford University Press; 2003.
[9] Tartakovsky DM, Wohlberg B. Delineation of geologic facies with statistical learning theory. Geophys Res Lett 2004;31(18):L18502.
[10] Tartakovsky DM, Wohlberg B, Guadagnini A. Nearest neighbor classification for facies delineation. Water Resour Res 2007;34:L05404. doi:10.1029/2007GL029245.
[11] Wohlberg B, Tartakovsky DM, Guadagnini A. Subsurface characterization with support vector machines. IEEE Trans Geosci Remote Sens 2006;44(1):47–57.