SUPPORT VECTOR MACHINES FOR DELINEATION OF GEOLOGIC FACIES FROM POORLY DIFFERENTIATED DATA

ALBERTO GUADAGNINI[1], BRENDT E. WOHLBERG[2], DANIEL M. TARTAKOVSKY[2,3], MICHELA DE SIMONI[1]

[1] *Politecnico di Milano, Dipartimento di Ingegneria Idraulica, Ambientale, Infrastrutture viarie, Rilevamento, Piazza L. Da Vinci, 32; 20133 Milano (Italy)*

[2] *Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545 (USA)*

[3] *Department of Mechanical & Aerospace Engineering, University of California, San Diego, La Jolla, CA 92093 (USA)*

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 preventing a clear distinction between different hydrofacies even at locations where data are available. We study the problem of facies delineation in geologic formations by means of the Support Vector Machine (SVM). To show the potential of the SVM, we randomly generate a two-dimensional porous medium composed of two heterogeneous materials, and then reconstruct boundaries between these materials from a few data points. We assess the performance and accuracy of the SVM-based facies delineation technique and assess in the presence of either well differentiated or poorly differentiated information about hydraulic parameters, such as hydraulic conductivity.

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 consider the effects of incorporating various types of information (including hydraulic conductivity, electrical resistivity, hydraulic heads and/or solute travel times) on predicting the salient features of flow and transport in heterogeneous underground reservoirs, the uncertainty associated with the delineation of lithofacies and associated hydraulic conductivity (and eventually porosity) from limited geological and geophysical data are only marginally analyzed. Such data, which include grain size distribution curves, are typically derived from core samples and are often poorly differentiated thus further compounding predictive uncertainty.

Since the classification of soils and other natural porous media is usually performed upon integrating information of percent sand, silt, and clay, and some measured value of hydraulic

parameters (*e.g.*, hydraulic conductivity), it is often subjective and therefore somewhat arbitrary. This forces the introduction of modeling approximations and can give rise to cross-correlations between material attributes [*Neuman*, 2003]. While such correlations between blocks might occur in systems where lenses of one material are laid down simultaneously with a base material for a while so that uncertain boundaries result [*Rajaram and McLaughlin*, 1990], hydrofacies delineation in the presence of poor data classification can yield apparent cross-correlations between materials even if the hydrogeologic attributes of geologic materials are themselves independent [*Winter and Tartakovsky*, 2002].

Geostatistics has become an invaluable tool for estimating facies distributions and their attributes at points in a computational domain where data are not available, as well as for quantifying the corresponding uncertainty [*Guadagnini et al*., 2004]. In the presence of poorly differentiated information, characterization of the heterogeneous aquifer structure is often performed in two steps. First, a multivariate facies-based parameterization approach relying on multivariate cluster analysis [*McQueen*, 1967] is applied to classify aquifer materials and to describe the heterogeneity of the aquifer lithology [*Ptak and Liedl*, 2002; *Riva et al*., 2005]. The resulting clusters are representative of sedimentological facies. Then, parameters distributions within each identified material block are estimated. Geostatistical frameworks treat a formation's property 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 realisation 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.

Machine learning provides an alternative to the geostatistical framework, allowing one to make predictions in the absence of sufficient data parameterization, without treating geologic parameters as random and, hence, without the need for the ergodicity assumptions. Intimately connected to the field of pattern recognition, machine learning refers to a family of computational algorithms for data analysis that are designed to automatically tune themselves in response to data. *Tartakovsky and Wohlberg* [2004] used a subset of the machine learning techniques - the Support Vector Machine (SVM) and its mathematical underpinning, the Statistical Learning Theory (SLT) of *Vapnik* [1998], which is ideally suited for the problem of facies delineation in geologic formations. While similar to neural networks in its goals, the SVM is firmly grounded in rigorous mathematical analysis, which allows one not only to assess its performance, but to bound the corresponding errors as well. Like other machine learning techniques, SVMs enable one to treat the subsurface environment and its parameters as deterministic. Uncertainty associated with insufficient data parameterization is then represented by treating sampling locations as a random subset of all possible measurement locations.

Recently we [*Wohlberg et al*., 2006] used a synthetic example to demonstrate that SVMs provide a viable alternative to geostatistical frameworks by allowing one to delineate lithofacies from well-differentiated hydraulic conductivity data. We found that: (i) SVMs slightly outperfom geostatistical approaches in reconstructing the boundary separating disjoint blocks of geologic facies, while being significantly less labor intensive; and (2) when data

sampling density is low (e.g., 0.25% or ten data points) the geostatistical inference becomes meaningless, while SVMs are capable of providing reasonable estimates of internal blocks boundaries. The use of SVMs for facies delineation depends critically on the ability to obtain a reliable classification (i.e., determine the corresponding material identity) of a data point by thresholding the measured hydraulic conductivity, since the classification for each measurement point is required for the SVM training process. The main goal of this study is to extend the analysis of *Wohlberg et al*. [2006] to account for poorly differentiated hydraulic conductivity data.

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 hydrodynamic data (e.g., hydraulic conductivity), geophysical data (e.g., electric resistivity), and/or sedimentological data, collected at N selected locations $\mathbf{x}_i = (x_i, y_i)^T$, where $i = 1, ..., N$ and ^{T} is transpose. The first step in our facies delineation procedure is to analyze the distributions of samples with the goal of assigning an indicator function

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

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 Figure 1. The data falling in the overlapping region between the two distributions do not render themselves to a straightforward classification by (1).

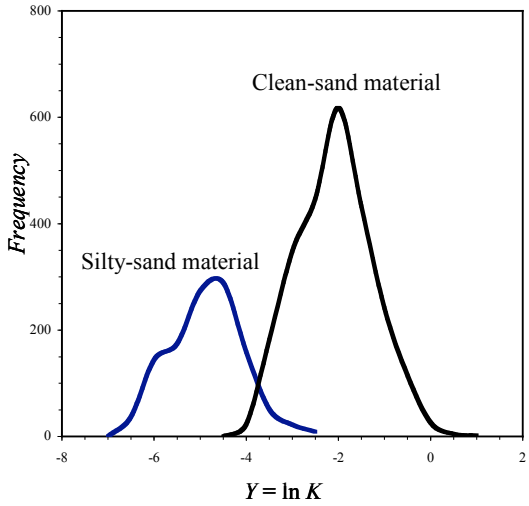

FIGURE 1. A typical sample frequency distribution of the log hydraulic conductivity $Y = \ln K$ of a subsurface environment composed of silty-sand and clean-sand (reference fields). The log hydraulic conductivity of the silty-sand and clean-sand facies ranges between -7.00 and -2.70 and -4.15 and 0.60,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

To assigning the indicators (1) to such data, we employ the so-called *k*-means clustering algorithm, which consists of (i) identifying a number of clusters – two geologic facies in our example, (ii) treating the minimum and maximum value of hydraulic conductivity as initial values for the means (centroid positions) of the respective populations; (iii) assign each conductivity measurements to the clusters with the closest centroid; (iv) recalculating the centroids based on the current cluster assignments, and (v) repeating (iii) and (iv) until the centroid positions stabilized.

Let $\bar{I}(\mathbf{x}, \alpha)$ be an estimate of a "true" indicator field $I(\mathbf{x})$, whose adjustable parameters α are consistent with, and determined from, the available information. Our objective is to construct an estimate that is as close to the true field as possible, i.e., to minimize the difference between the two, $\|I-\bar{I}\|$.

3. SUPPORT VECTOR MACHINE

The theoretical foundation of SVM techniques relies on the definition of a bound for the expected risk, which is provided by the maximum margin SVM [e.g., *Cristianini and Shawe-Taylor*, 2000]. The simplest maximum margin SVM deals with linearly separable data collected from perfectly stratified geologic media, where different geologic facies are separated by planes (in three dimensions) or straight lines (in two dimensions), as analyzed by *Tartakovsky and Wohlberg* [2004]. In most practical problems, boundaries between geologic facies are significantly more complex than a straight line or a plane. To account for this geometric complexity, one can generalize the linear maximum margin SVM by noting that data which cannot be separated by a straight line or plane in the two- or three-dimensional space of observation often become linearly separable (by a hyperplane) when projected onto another, usually higher-dimensional space. This is made computationally feasible by the use of kernels (see *Wohlberg et al.* [2006] for details).

4. SYNTHETIC EXAMPLE

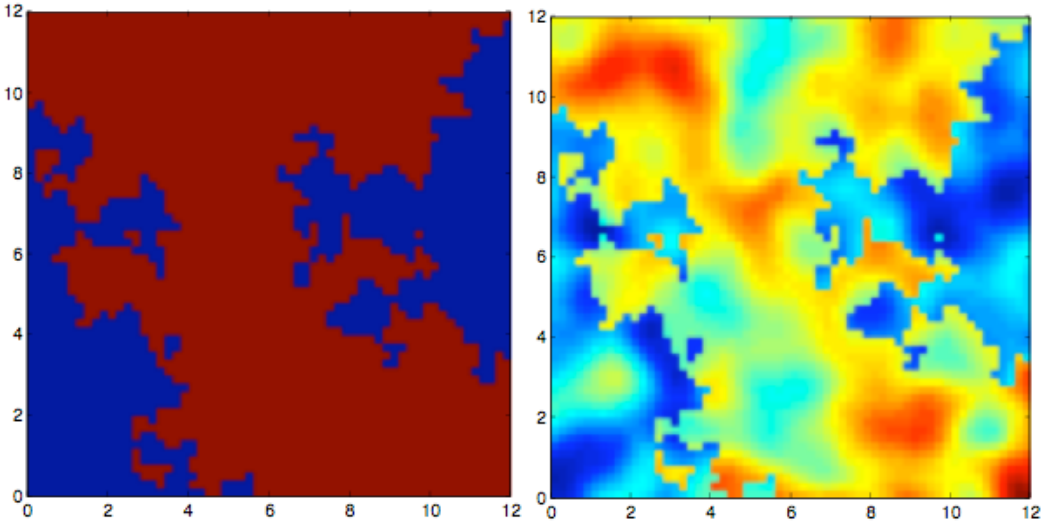

FIGURE 2. (a) Material classification of *Wohlberg et al*. [2006]; and (b) Distribution of log hydraulic conductivity, whose values range from -7.00 (dark blue) to 0.60 (red).

We use SVMs to reconstruct, from a few data points selected at random from a uniform distribution, the boundaries between two heterogeneous geologic facies in a synthetic porous medium shown in Figure 2. The boundaries between the silty-sand (an ambient formation) and clean-sand (an inclusion) facies are the same as those used in the analysis of *Wohlberg et al*. [2006]. The following procedure was used to assign a value of hydraulic conductivity to each point (pixel). First, we generated two autocorrelated, weakly stationary Gaussian fields with the ensemble means of −4.96 and −2.30, respectively. (The mutually uncorrelated random fields had unit variance and Gaussian autocorrelation with unit correlation scale.) These values correspond to the geometric means of hydraulic conductivity equal to 7×10^{-3} and 1×10^{-1}, respectively, when conductivities are measured in [cm/s], which represent silty-sand and clean-sand materials [*Freeze and Cherry*, 1979]. Then, we superimposed these fields onto the facies map (Figure 2a) to obtain the conductivity field in Figure 2b. The resulting distributions of log hydraulic conductivities within the silty-sand and clean-sand facies are shown in Figures 3a and 3b, respectively.

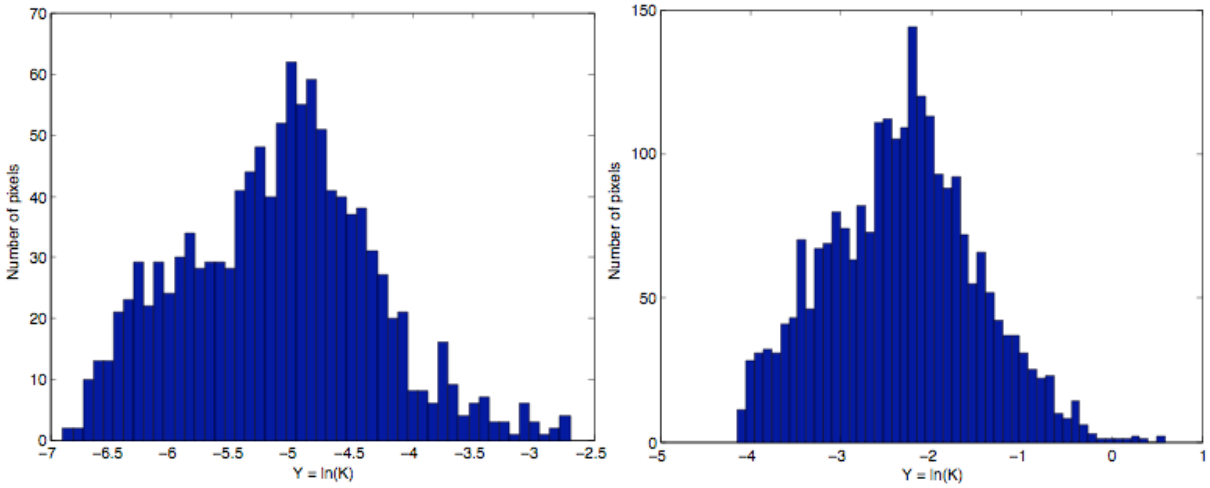

FIGURE 3. Distributions of log hydraulic conductivity $Y = \ln K$ within the silty-sand (left) and clean-sand (right) facies. Note that these distributions are non-Gaussian.

We use our implementation of SVMs to reconstruct the boundary between the two materials from a few (randomly selected) data points. Sampling densities ranging from approximately 0.25% (10 data points) to 5% (80 data points) were considered. For each sampling density, we randomly generated 20 realizations of the locations of data points and counted the number of elements on the grid that were misclassified by the SVMs. The results reported below represent the averages over 20 realizations.

When one deals with poorly differentiated data, a reliable classification of the measurement points is not possible, so that the classifications must be estimated. We accomplish this by employing the *k*-means algorithm described above. In Figure 4, the line with crosses displays the average error in *k*-means estimation of this classification for each sampling density, measured as the fraction of points misclassified with respect to the ground-truth classification. Note that *k*-means estimation error increases with sampling density, which may seem counter-intuitive. 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 Example 1 can be perfectly classified by a threshold of (-4.2 -3.4)/2, the data in Example 2 has a minimum threshold-based classification error of 2/6. The line with circles in Figure 4 shows the smallest possible error that can be obtained by estimating the classification with thresholding the hydraulic conductivity values. We observe the increase of the minimum possible threshold-based estimation error as the number of samples grows, the number of ways of classifying the points grows much faster than the number of ways of partitioning the points based on a threshold.

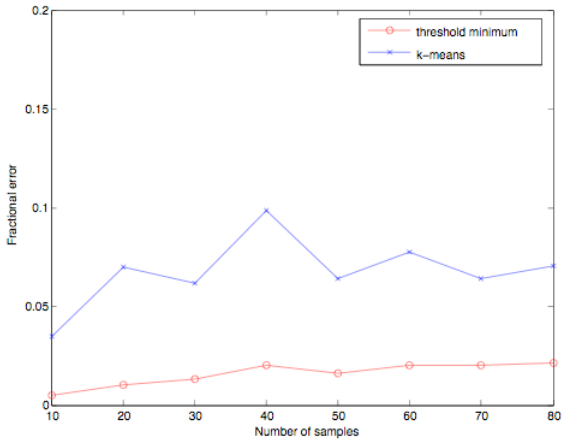

FIGURE 4. The number of misclassified data points (reported as a fractional error, i.e., a fraction of the misclassified data points relative to the total number of sample points) as a function of the number of samples.

TABLE 1. Examples of threshold-based minimum classification errors.

Example 1						
Log-conductivity	- 5.1	- 4.5	- 4.2	- 3.4	- 2.0	- 1.9
Indicators	- 1	- 1	- 1	+ 1	+ 1	+ 1
Example 2						
Log-conductivity	- 5.4	- 5.2	- 4.3	- 4.1	- 3.6	- 2.1
Indicators	- 1	- 1	+ 1	- 1	+ 1	+ 1

After identifying the membership of the data points in either of the facies, i.e., after assigning the values of the indicator function to each data point, we used our implementation of SVMs with the Gaussian kernel to estimate the boundaries between the two facies. To find the SVM parameters, we used the leave-one-out approach, as described in *Wohlberg et al.* [2006]. Figure 5 quantifies the boundary reconstruction errors introduced by this procedure when applied to the indicator function data inferred from the *k*-means clustering algorithm (the line with squares) and the threshold minimum (the line with circles). The errors are reported as a number of misclassified pixels relative to the total number of pixels. As can be expected, the misclassification errors decrease as a number of samples (data points) increases. Since the poor differentiation of data introduces the interpretive errors (Figure 4) when the *k*-

means clustering algorithm is used to assign the values of the indicator function to such data, the reconstructed boundary between the two facies is more prone to errors than its counterpart that could be reconstructed from well-differentiated data (see Figure 7 in *Wohlberg et al.* [2006]). However, considering the challenges posed by poorly differentiated data, the performance of SVMs is remarkable.

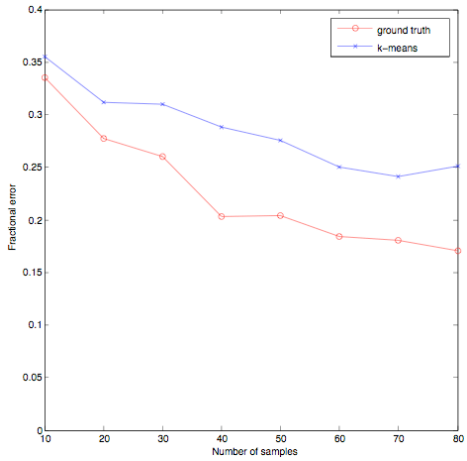

FIGURE 5. The boundary reconstruction errors (reported as a fraction of the misclassified pixels relative to the total number of pixels) as a function of the number of samples.

5. SUMMARY

We employed support vector machines (SVMs) to delineate geologic facies from a small set of poorly differentiated data. This was accomplished (i) by reconstructing, from a few data points, a synthetic randomly generated porous medium consisting of two heterogeneous materials. The challenges posed by the poor differentiation of data stem from the fact that some measurements of, say, hydraulic conductivity do not allow one to determine with a required degree of certainty the membership of sampling locations in a given geologic facies. To assign values of an indicator function to such data, we used the *k*-means clustering algorithm. Our analysis leads us to conclude that the SVMs combined with the *k*-means clustering algorithm provide an attractive, fully automated tool for identification of boundaries between geologic facies from poorly differentiated data.

REFERENCES

Cristianini, N., and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, UK: Cambridge University Press, 2000.

Freeze, R. A., and J. A. Cherry, *Groundwater*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey 07632, 1979.

Guadagnini, L., A. Guadagnini, and D.M. Tartakovsky, Probabilistic Reconstruction of geologic facies, *J. of Hydrol.*, 294, 57-67, 2004.

McQueen, J., Some methods for classification and analysis of multivariate observations. In: *5th Berkeley Symposium on Mathematics, Statistics and Probability*, 1, 281-298, 1967.

Neuman, S. P., Relationship between juxtaposed, overlapping, and fractal representations of multimodal spatial variability, *Water Resour. Res.*, 39(8), 1205, doi:10.1029/2002WR001755, 2003.

Ptak, T., and R. Liedl, Modeling of reactive contaminant transport in hydraulically and hydrogeochemically heterogeneous aquifers using an upscaling approach based on sedimentological facies, In: *Bridging the Gap*

between Measurements and Modeling in Heterogeneous Media (ed by Findikakis, A. N.) (International Groundwater Symposium IAHR, IAHS and ASCE/EWRI, Berkeley, California, IAHR, Madrid, Spain) 425-429 (full paper on accompanying CD-ROM, 9 pages), ISBN 90-805649-4-X, 2002.

Rajaram, H., and D. McLaughlin, Identification of large-scale spatial trends in hydrologic data, *Water Resour. Res.*, 26 (10), 2411-2423, 1990

Riva, M., L. Guadagnini, A. Guadagnini, E. Martac, and T. Ptak, "*A composite medium approach for probabilistic modelling of contaminant travel time distribution to a pumping well in a heterogeneous aquifer*", in Pre-Published Proceedings of the Fifth Int. Conference on Calibration and Reliability in Groundwater Modelling (ModelCARE 2005), 460-466, 2005.

Tartakovsky, D. M., and B. E. Wohlberg, Delineation of geologic facies with statistical learning theory, *Geophys. Res. Lett.*, **31**(18), L18502, 2004, doi:10.1029/2004GL020864, 2004.

Vapnik, V. N., *Statistical Learning Theory*. New York: John Wiley & Sons, Inc., 1998.

Wohlberg, B., D.M. Tartakovsky, and A. Guadagnini, Subsurface Characterization with Support Vector Machines, *IEEE Trans. on Geoscience and Remote Sensing*, 44(1), 47-57, doi: 10.1109/TGRS.2005.859953, 2006.

Winter, C. L., D. M. Tartakovsky, and A. Guadagnini, Numerical solution of moment equations for flow in heterogeneous composite aquifers, *Water Resour. Res.*, 38(5), 1055, doi:10.1029/2001WR000222, 2002.

Winter, C.L., D.M. Tartakovsky, and A. Guadagnini, Moment differential equations for flow in highly heterogeneous porous media, *Surveys in Geophysics*, 24(1), 81-106, 2003.