Machine Learning Methods for Inverse Modeling

Daniel M. Tartakovsky[1,3], Alberto Guadagnini[2], and Brendt E. Wohlberg[3]

[1] Department of Mechanical and Aerospace Engineering, University of California, San Diego, La Jolla, CA 92093, USA dmt@ucsd.edu
[2] Politecnico di Milano, Dipartimento di Ingegneria Idraulica, Ambientale, Infrastrutture viarie, Rilevamento, 20133 Milan, Italy alberto.guadagnini@polimi.it
[3] Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA brendt@lanl.gov

Summary. Geostatistics has become a preferred tool for the identification of lithofacies from sparse data, such as measurements of hydraulic conductivity and porosity. Recently we demonstrated that the support vector machine (SVM), a tool from machine learning, can be readily adapted for this task, and offers significant advantages. On the conceptual side, the SVM avoids the use of untestable assumptions, such as ergodicity, while on the practical side, the SVM outperforms geostatistics at low sampling densities. In this study, we use the SVM within an inverse modeling framework to incorporate hydraulic head measurements into lithofacies delineation, and identify the directions of future research.

1 Introduction

Heterogeneous aquifers typically consist of multiple lithofacies, the spatial arrangement of which significantly affects flow and transport in the subsurface. The estimation of these lithofacies is complicated by the sparsity of data and by the lack of a clear correlation between identifiable geologic indicators and attributes (e.g. hydraulic conductivity and porosity). This so-called zonation problem has been studied by [1, 2, 3, 4], among others.

Data which are used in geomaterials classification procedures are typically obtained from core samples that often disturb soils and are by necessity sparse, thus contributing to predictive uncertainty associated with the location of different geomaterials. Within a stochastic framework, this uncertainty is quantified by treating a formation's properties as random fields that are characterized by multivariate probability density functions or, equivalently, by their joint ensemble moments. Geostatistics has become an invaluable tool for estimating facies distributions at points in a computational domain where

data are not available, as well as for quantifying the corresponding uncertainty [5].

Recently we [6, 7, 8] demonstrated that Support Vector Machine (SVM) techniques, a subset of machine learning algorithms, provide a viable alternative to geostatistical frameworks by allowing one to delineate lithofacies in the absence of sufficient data parameterization, without treating geologic parameters as random and, hence, without the need for the ergodicity assumption. This has been done by using both well and poorly differentiated parameter data. For additional information on the use of the SVM and other machine learning techniques in environmental applications, we refer the interested reader to [9].

In this study, we use machine learning within an inverse modeling framework to incorporate hydraulic head measurements into lithofacies identification. We apply the approach to a synthetic case of steady-state flow through a domain consisting of two materials separated by highly irregular boundaries (see Fig. 1). For simplicity, the hydraulic conductivity of each material is assumed to be constant.

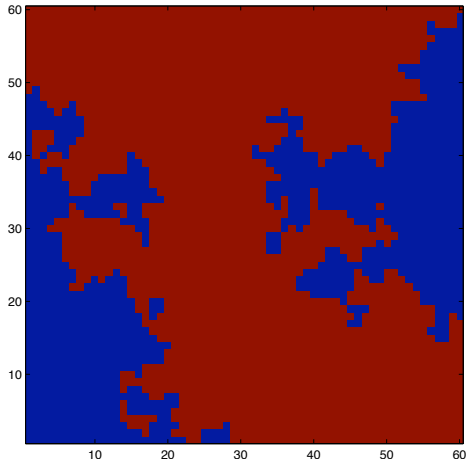

Fig. 1. Flow domain consisting of two contrasting geologic facies. A highly conducting material is shown in red and a low conducting material in blue.

2 A Problem of Facies Delineation

Consider the problem of reconstructing a boundary between two homogeneous geologic facies from either parameter data $K_i = K(\mathbf{x}_i)$ or system state data $h_i = h(\mathbf{x}_i, t)$ or both. Without loss of generality, we assume that both data sets are collected at the same N locations $\mathbf{x}_i = (x_i, y_i)^T$, where $i \in \{1, \dots, N\}$.

Such problems are ubiquitous in subsurface hydrology since the geologic structure of the subsurface plays a crucial role in fluid flow and contaminant transport. A typical example is the problem of locating permeable zones in the aquiclude that separates two aquifers, the upper aquifer contaminated with industrial pollutants, and the lower aquifer used for municipal water supplies [5].

Parameter data can include measurements of hydraulic conductivity, electric resistivity, cumulative thickness of relevant geologic facies, and grain sizes. We will call the problem of estimating the internal boundaries between geologic lithofacies from such data *a forward facies delineation problem*. System state data consist of measurements of hydraulic head, flow rate, concentration, etc. We will call the problem of estimating the internal boundaries between geologic lithofacies from such data *an inverse facies delineation problem*. Often both data types are available at the same locations, e.g., when observation and/or pumping wells are outfitted with flow-meters.

2.1 Forward Facies Delineation Problem

Since parameter data characterize geologic materials, they allow one to label the points where they are taken by mean of the indicator function

$$I_i \equiv I(\mathbf{x}_i) = \begin{cases} +1 & \mathbf{x}_i \in M_1 \\ -1 & \mathbf{x}_i \in M_2, \end{cases} \tag{1}$$

where M_1 and M_2 are the two facies. This step typically involves an analysis of a data histogram, which is often nontrivial, since a typical geologic facies is heterogeneous. Here we assume that the available parameter data $\{K(\mathbf{x}_i)\}_{i=1}^N$ are well differentiated, so that the process of assigning the values of the indicator functions to points $\{\mathbf{x}_i\}_{i=1}^N$ does not introduce interpretive errors. This assumption can be relaxed to account for poor differentiation of data [5].

While it is customary to employ geostatistics for facies delineation, we [6, 7] showed that the SVM, a tool from the statistical learning theory, can be readily adapted for this task and offers significant advantages. On the conceptual side, the SVM avoids the use of untestable assumptions, such as ergodicity. On the practical side, the SVM outperforms geostatistics at low sampling densities.

The SVM also has an advantage over neural networks, another tool from the machine learning theory. This is because the SVM solves a convex optimization by minimizing the quadratic functional

$$\max_{\boldsymbol{\gamma}} \left\{ \sum_{i=1}^{N} \gamma_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \gamma_i \gamma_j I_i I_j \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) \right\}, \tag{2}$$

where $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j)$ is a given Mercer kernel, subject to the constraints

$$0 \le \gamma_i \le C \quad \text{and} \quad \sum_{i=1}^{N} \gamma_i I_i = 0. \tag{3}$$

This optimization problem has a well defined global minimum that is influenced by the choice of the fitting parameter C. If $\{\gamma_i^\star\}_{i=1}^N$ denote a solution of the optimization problem (2), then the indicator function $I(\mathbf{x})$ at any point $\mathbf{x}$, and hence the boundary separating two materials, is given by [10, p. 203]

$$I(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{N} \gamma_i^\star I_i \mathcal{K}(\mathbf{x}, \mathbf{x}_i) + b^\star \right), \tag{4}$$

where

$$b^\star = I_j - \sum_{i=1}^{N} \gamma_i^\star I_i \mathcal{K}(\mathbf{x}_j, \mathbf{x}_i) \tag{5}$$

for some j such that $\gamma_j > 0$.

2.2 Inverse Facies Delineation Problem

The incorporation of system state data into the SVM framework is challenging, since (i) one cannot assign the indicator function to such data and (ii) the relationship between the two data types is nonlinear. To be concrete, we consider steady-state saturated flow, so that the parameter K stands for hydraulic conductivity and the system state h is hydraulic head. We propose the following SVM-based algorithm to delineate geologic facies from parameter and system state data.

1. Use an SVM to reconstruct facies from parameter data $\{K(\mathbf{x}_i)\}_{i=1}^N$.
2. Use the resulting parameter field as an initial guess for the optimization problem

$$\min_{\boldsymbol{\gamma}} \left\{ -\sum_{i=1}^{N} \gamma_i + \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \gamma_i \gamma_j I_i I_j \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) + \lambda \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left[h_i - h_s(\mathbf{x}_i) \right]} \right\}, \tag{6}$$

subject to the constraints (3) and fixed $\lambda > 0$. Here $h_s(\mathbf{x})$ is a computed system state, e.g., a numerical solution of the steady-state flow equation $\nabla \cdot (K \nabla h) = 0$ subject to appropriate boundary conditions. The hydraulic conductivity $K(\mathbf{x})$ is determined by the current state of $\{\gamma_i\}_{i=1}^N$.

The proposed approach aims to retain the maximization of the SVM margin based on conductivity data (Step 1), while minimizing the difference between the measured and computed heads. This balance is controlled by the choice of the parameter λ in (6). The higher its value, the more weight one

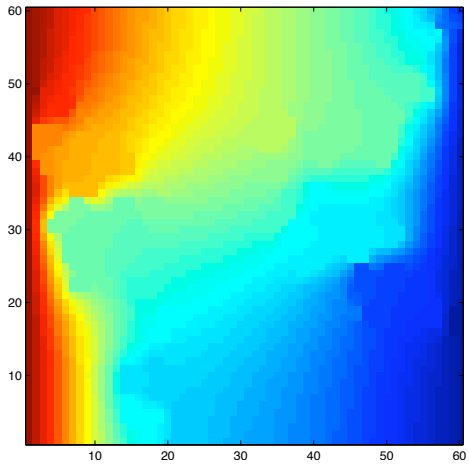

Fig. 2. Hydraulic head distribution in the flow domain shown in Fig. 1.

assigns to the head measurements relative to the conductivity measurements, and vice versa.

In the proposed approach, hydraulic head data affect only the radial functions weights, which, in principle, might provide too few degrees of freedom. Indeed, one can expect the estimated facies boundary to be overly smooth, when a conductivity data set is small. However, if such a data set is complemented by a few hydraulic head measurements, fewer degrees of freedom are necessary to obtain a good correspondence.

Also, it is important to note that the proposed approach replaces the quadratic optimization in the SVM (2) with the nonlinear optimization (6). This nonlinearity arises from the nonlocal relationship between hydraulic conductivity K and hydraulic head h. This raise important questions of whether or not the SVM parameterization of the boundaries is adequate and the use of SVMs for facies delineation is appropriate. We begin to address these questions by analyzing the computational example provided below.

3 Computational Example

We employ the proposed SVM-based algorithm to reconstruct the boundaries between two geologic facies shown in Fig. 1 from N randomly selected data points $\{\mathbf{x}_i\}_{i=1}^N$. At these data points, both hydraulic conductivity K and hydraulic head h are sampled. The values of hydraulic head are obtained by solving the steady-state flow equation $\nabla \cdot (K \nabla h) = 0$ with hydraulic conductivity $K(\mathbf{x})$ distribution shown in Fig. 1. Flow is driven by the hydraulic heads $h = H_1$ and $h = H_2$ prescribed along the left and right vertical boundaries, respectively. The lower and upper horizontal boundaries are impermeable. This results in the reference hydraulic head distribution shown in Fig. 2.

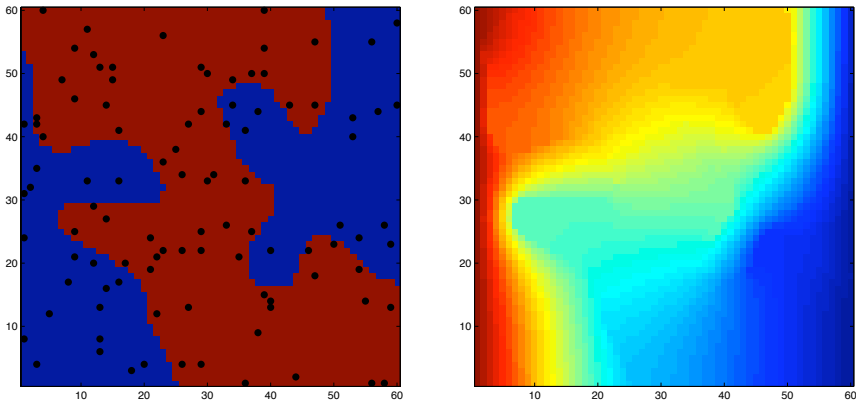

Fig. 3. The conductivity field reconstructed from $N = 100$ conductivity measurements, whose locations are shown by the black dots (a) and the corresponding hydraulic head distribution (b).

The first step in the proposed algorithm consists of the use of an SVM to reconstruct the boundaries from $N = 100$ conductivity measurements. The location of these measurements, the reconstructed boundaries, and the corresponding hydraulic head distribution are shown in Fig. 3. We used the Gaussian kernel $\mathcal{K}(\mathbf{x}, \mathbf{x}') = \exp\left[-||\mathbf{x} - \mathbf{x}'||^2/(2\sigma^2)\right]$ with parameter $\sigma = 5$, and set SVM parameter $C = 1000$. (These fixed values were chosen for good facies delineation performance based on our previous experience [7], but in a more realistic setting these would be chosen via a cross-validation method.)

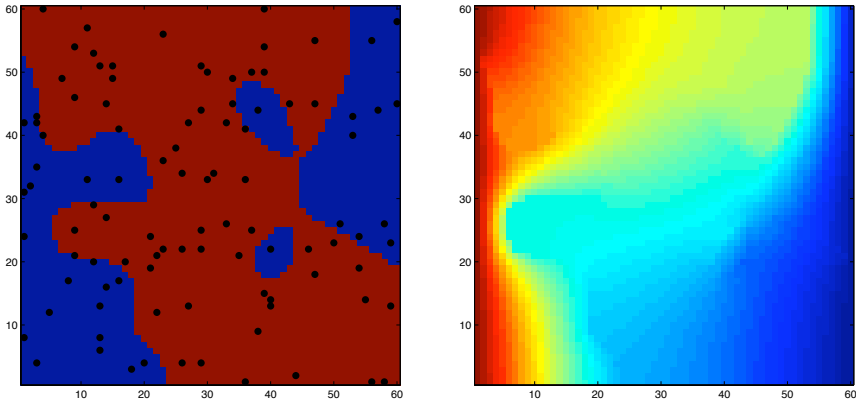

Fig. 4. The conductivity field reconstructed from $N = 100$ measurements of hydraulic conductivity and head, whose locations are shown by the black dots (a) and the corresponding hydraulic head distribution (b).

Using this hydraulic conductivity field as an initial guess in the second step, and choosing $\lambda = 1000$ to provide appropriate weighting to each term, we obtain the reconstructed boundaries and the corresponding hydraulic head distribution in Fig. 4. Table 1 provides a quantitative comparison of these reconstructions. While the incorporation of hydraulic head measurements leads to only a marginal reduction in the boundary reconstruction errors, it significantly reduces the error in predictions of hydraulic head distribution. Moreover, the L^2 norm used in this comparison does not tell the whole story. A visual comparison of the reconstructed conductivity fields (Figs 3a and 4a) with the reference conductivity field (Fig 1) reveals that the joint use of hydraulic conductivity and head data also noticeably improves the boundary reconstruction.

Table 1. Reconstruction errors (L^2 norm with respect to the reference field) resulting from the reliance on K data only and on combined K and h data.

	Conductivity K	Head h
K data	111.83	52.04
K and h data	106.86	25.67

4 Summary and Discussion

Support Vector Machines (SVM), a tool from machine learning, provides a number of advantages over geostatistics. Previous applications of SVM focused on the delineation of lithofacies from measurements of properties of geologic materials (parameter data). Such data allow one to determine the membership of spatial locations where the measurements are made in a relatively straightforward fashion.

The task of identifying geologic units from system state data is significantly more challenging, since the membership of such data in a given unit is not identifiable from data analysis alone. We proposed an SVM-based approach that allows one to combine both types of data with the aim of improving the accuracy and robustness of the facies delineation. The preliminary results reported here demonstrate the potential of the proposed approach.

A number of issues remain open and remain the focus of our ongoing research. These include

- Since the proposed approach relies on a nonlinear optimization with many degrees of freedom, its utility and reliability depends critically on the selection of the optimization strategy. The nonlinear constrained optimization algorithm (function *fmincon* in the MATLAB Optimization Toolbox) used in the present analysis is known to converge to local, rather than global,

minima and requires careful adjustment of optimization parameters. This poses the question of selection of optimal optimization strategies.

- As the sampling density (the number of elements in the numerical grid where data are available relative to the total number of elements) decreases, the last term in the optimization functional might become flat. Consequently, the SVM parameterization of boundaries might not guarantee an optimal performance with respect to hydraulic heads. This calls for a detailed analysis of the influence of sampling density on the performance of the proposed approach.
- Locations of data points are expected to play a significant role in the accuracy of facies delineation. To provide an unbiased assessment of the performance of the proposed approach, it is desirable to average the reconstruction errors resulting from several alternative placements of data points for the same sampling data. This task is impossible without a robust optimization algorithm (see above).

These issues and limitations might explain a relatively modest reduction (50%) of the reconstruction errors obtained with the SVM inversion, while geostatistical inverse methods, e.g., the inversion of stochastic moment equations based on the pilot point method [11, 12], often yield an order of magnitude error reductions. However, it is important to emphasize that the success of this and other geostatistical inversion techniques depends heavily on the number and location of pilot points, the quality and quantity of data, the presence/absence of priors, and the way used to regularize the objective function [13]. The results presented here reduce the bias by averaging over twenty possible locations of data points.

Finally, the proposed SVM inversion procedure might suffer from an inadequate number of degrees of freedom. To alleviate this problem, we are working on its extension that incorporates the ideas behind pilot point methods (but not their geostatistical implementation) into the SVM framework.

References

1. Sun, N.Z., Yeh, W.W.G.: Identification of parameter structure in groundwater inverse problem. Water Resour. Res. **21** (1985) 869 – 883
2. Carrera, J., Neuman, S.P.: Estimation of aquifer parameters under transient and steady sate conditions, 3. Application to synthetic and field data. Water Resour. Res. **22** (1986) 228 – 242
3. Eppstein, M.J., Dougherty, D.E.: Simultaneous estimation of transmissivity values and zonation. Water Resour. Res. **32** (1996) 3321 – 3336
4. Tsai, F.T.C., Yeh, W.W.G.: Characterization and identification of aquifer heterogeneity with generalized parameterization and Bayesian estimation,. Water Resour. Res. **40** (2004) W10102 doi:10.1029/2003WR002893.
5. Guadagnini, L., Guadagnini, A., Tartakovsky, D.M.: Probabilistic reconstruction of geologic facies. J. of Hydrol. **294** (2004) 57 – 67

6. Tartakovsky, D.M., Wohlberg, B.E.: Delineation of geologic facies with statistical learning theory. Geophys. Res. Lett. **31** (2004) L18502 doi:10.1029/2004GL020864.
7. Wohlberg, B.E., Tartakovsky, D.M., Guadagnini, A.: Subsurface characterization with Support Vector Machines. IEEE Trans. Geosci. Remote Sens. **44** (2006) 47 – 57 doi:10.1109/TGRS.2005.859953.
8. Guadagnini, A., Wohlberg, B.E., Tartakovsky, D.M., De Simoni, M.: Support Vector Machines for delineation of geologic facies from poorly differentiated data. In: Proceedings of the CMWR XVI Conference, Copenhagen, June 2006. (2006)
9. Kanevski, M., Maignan, M.: Analysis and Modelling of Spatial Environment Data. Marcel Dekker (2004)
10. Schölkopf, B., Smola, A.J.: Learning with Kernels. The MIT Press, Cambridge, MA, USA (2002)
11. Hernandez, A.F., Neuman, S.P., Guadagnini, A., Carrera, J.: Conditioning mean steady state flow on hydraulic head and conductivity through geostatistical inversion. Stoch. Environ. Res. Risk Assess. **17** (2003) 329 – 338
12. Hernandez, A.F., Neuman, S.P., Guadagnini, A., Carrera, J.: Inverse stochastic moment analysis of steady state flow in randomly heterogeneous media. Water Resour. Res. **42** (2006) W05425, doi:10.1029/2005WR004449
13. Alcolea, A., Carrera, J., Medina, A.: Pilot points method incorporating prior information for solving the groundwater flow inverse problem. Adv. Water Resour. **29** (2006) 1678 – 1689