

Water Resources Research

RESEARCH ARTICLE

10.1029/2021WR029622

Key Points:

- We propose probabilistic Support Vector Machines (pSVM) to reconstruct hydrofacies from sparse data and to quantify predictive uncertainty
- pSVM generate smoother probability maps than those produced by indicator Kriging (IK), that is, pSVM provide more conservative estimates
- pSVM are preferable to IK because the latter has more tunable parameters and higher data requirements

Correspondence to:


D. M. Tartakovsky,
tartakovsky@stanford.edu

Citation:

Dendumrongsup, N., & Tartakovsky, D. M. (2021). Probabilistic reconstruction of hydrofacies with support vector machines. *Water Resources Research*, 57, e2021WR029622. <https://doi.org/10.1029/2021WR029622>

Received 16 JAN 2021
 Accepted 17 APR 2021

Probabilistic Reconstruction of Hydrofacies With Support Vector Machines

Nutchapol Dendumrongsup¹ and Daniel M. Tartakovsky¹ 

¹Energy Resource Engineering Department, Stanford University, Stanford, CA, USA

Abstract Delineation of geological features from limited hard and/or soft data is crucial to predicting subsurface phenomena. Ubiquitous sparsity of available data implies that the reliability of any delineation effort is inherently uncertain. We present probabilistic support vector machines (pSVM) as a viable method for both hydrofacies delineation from sparse data and quantification of the corresponding predictive uncertainty. Our numerical experiments with synthetic data demonstrate an agreement between the probability of a pixel classifier predicted with pSVM and indicator Kriging. While the latter requires manual inference of a variogram (two-point correlation function) from spatial observations, pSVM are highly automated and less data intensive. We also investigate the robustness of pSVM with respect to its hyper-parameters and the number of measurements. Having investigated these features of pSVM, we deploy them to delineate, from lithological data collected in a number of wells, the spatial extent of an aquitard separating two aquifers in Southern California.

1. Introduction

The need to delineate geological features, for example, lithofacies, is ubiquitous in subsurface application. Understanding subsurface geology is crucial to exploration of mineral resources and oil and gas reservoirs. It is also of central importance in subsurface hydrology since geologic makeup of the subsurface plays a crucial role in fluid flow and contaminant transport. A typical example is a problem of locating permeable zones in an aquiclude that separates two aquifers, the upper aquifer contaminated with industrial and/or agricultural pollutants, and the lower aquifer used for municipal water supplies (Guadagnini et al., 2004).

Geostatistics has long been used to gain insight into spatial distributions of physical properties of geologic formations (Isaaks & Srivastava, 1990). By adapting the probabilistic framework it accounts for inherent predictive uncertainty that arises from subsurface heterogeneity and data sparsity. In doing so, geostatistics relies on the ergodicity hypothesis, which postulates the equivalence between spatial statistics and its ensemble counterpart. This hypothesis cannot be proven, but it does require a subsurface environment to be (weakly) stationary, that is, relevant subsurface properties to have both constant means and variances and translation-invariant correlation functions. Tools of the statistical learning theory (Vapnik, 1998), such as support vector machines (SVM) (Tartakovsky & Wohlberg, 2004), provide an attractive alternative to geostatistics (indicator Kriging). Specifically, SVM do not invoke ergodicity, are highly automated, and require fewer measurements to remain viable (Wohlberg et al., 2006). Like all machine learning-based approaches, the SVM training consists of identification of the model parameters by minimizing the discrepancy between the SVM's predictions and the training data; in contrast to many machine learning techniques, such as neural networks, the minimization problem is quadratic and, hence, has an easily computable unique solution. Once trained, the SVM model is verified by evaluating its performance on the test (unseen) data. Unlike Kriging, which also minimizes the difference between available data and model predictions, SVM minimize the difference between unseen data and model predictions (the so-called generalization error). The former does interpolation, while the latter performs regression (Tartakovsky & Wohlberg, 2004).

Relatively low data requirements are a key advantage that distinguishes SVM from other machine learning techniques, for example, deep neural networks, used for facies delineation (Zeng et al., 2018). Unlike neural networks, SVM possess rigorous performance guarantees and error bounds (Vapnik, 1998). A related attractive feature of SVM is that they result in a straightforward quadratic optimization problem whose unique solution is trivially obtained; this is in contrast with neural networks whose minimization functions have multiple local minima. Like many other machine learning techniques, such as k -nearest neighbors

(Tartakovsky et al., 2007) and deep neural networks, standard SVM provide only a “best” estimate of the spatial arrangement of geological features consistent with available data (Wohlberg et al., 2006). To the best of our knowledge, geostatistical techniques (Guadagnini et al., 2004) are the only means to quantify predictive uncertainty in subsurface delineation from sparse data.

We develop probabilistic SVM (pSVM) in order to quantify uncertainty inherent in such reconstructions and to identify facies with a required degree of fidelity. While the original pSVM (Platt, 2000) were designed to quantify the probability of SVM misclassifying pixels of a complete image, our method is designed to cope with the sparsity of subsurface data, that is, with the task of reconstructing an image from a few pixels. Unlike the geostatistical approach of Guadagnini et al. (2004) used for this purpose, the pSVM approach does not require the construction of a variogram and, hence, has significantly lower data requirements. Other advantages of the SVM framework over its geostatistical counterparts are discussed in detail by Wohlberg et al. (2006).

In Section 2, we formulate the problem of subsurface facies delineation from sparse data. Section 3 contains a brief overview of the standard SVM approach to this problem and introduces pSVM. The latter is used in Section 4 to probabilistically reconstruct facies from a synthetic data set, which enables us to study the method’s accuracy, robustness, and data requirements. In Section 5, we deploy pSVM to analyze lithological data collected in multiple wells at a Southern California site. Main conclusions drawn from this study are summarized in Section 6.

2. Problem of Facies Delineation

Consider a two-dimensional subsurface environment D consisting of two lithofacies M_1 and M_2 ($D = M_1 \cup M_2$), for example, high- and low-permeability heterogeneous geologic materials. Our goal is to reconstruct a (single- or multi-connected) boundary between these facies from continuous parameter data $K_i = K(\mathbf{x}_i)$ collected at N locations, $\mathbf{x}_i = (x_i, y_i)^\top$ with $i \in \{1, \dots, N\}$, throughout the domain D . These locations form a set $T = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$.

To transform this task into a classification problem, we convert values K_i of the continuous function $K(\mathbf{x}_i)$ into values of an indicator function (aka categorical variable)

$$J(\mathbf{x}_i) = \begin{cases} 1 & \mathbf{x}_i \in M_1, \\ -1 & \mathbf{x}_i \in M_2. \end{cases} \quad (1)$$

This step assumes that the data $\{K_i\}_{i=1}^N$ are well differentiated, that is, each measurement K_i unambiguously identifies the measurement location \mathbf{x}_i as belonging either to facies M_1 or M_2 . In the case of poorly differentiated data, this step could be preceded by a nearest neighbor classifier (Wohlberg & Tartakovsky, 2009).

3. Support Vector Machines

SVMs are often considered one of the best “out of the box” classifiers that yield great performance in a variety of settings (James et al., 2014). In their simplest form, SVMs are applicable to linearly separable data, for example, data collected from perfectly stratified geologic media in which different geologic facies are separated by either planes in three dimensions or straight lines in two dimensions. Nonlinear SVMs enable one to deal with general subsurface environments by projecting them into higher-dimensional space in which the data are linearly separable by a hyperplane. Linear and nonlinear SVMs are briefly reviewed in Sections 3.1 and 3.2 for the sake of completeness.

3.1. Linear SVM

A linearly separable data set $\{J_i\}_{i=1}^N$ implies the existence of a straight line, $\mathbf{a} \cdot \mathbf{x} + b = 0$, that separates the locations \mathbf{x}_i at which $J = -1$ from those at which $J = 1$ (Figure 1a). The unknown constants $\mathbf{a} = (a_1, a_2)^\top$ and b are computed by maximizing the distance (margin) between $\mathbf{a} \cdot \mathbf{x} + b = 0$ and the locations at which $J_i = -1$ and $J_i = 1$.

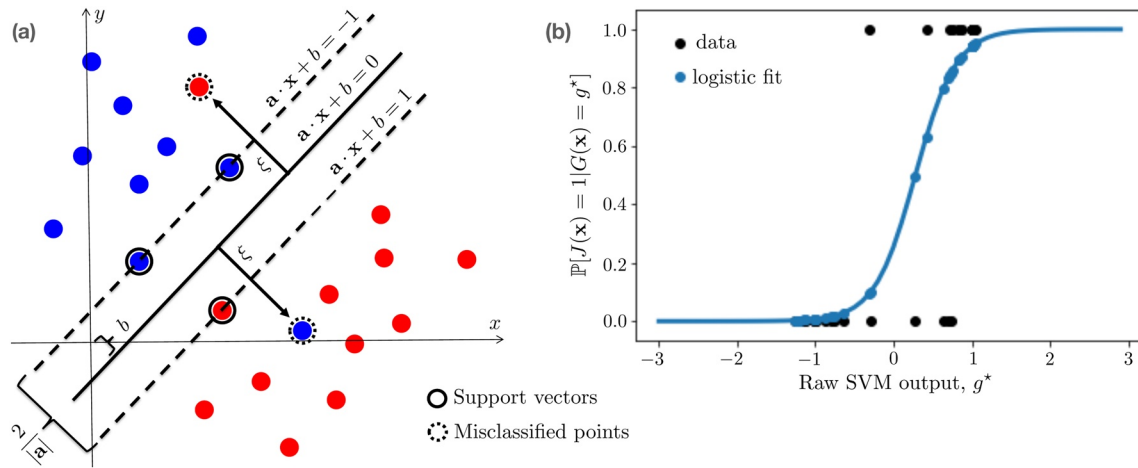


Figure 1. (a) A linear SVM classifier maximizes the margin between the reconstructed boundary (straight line), $\mathbf{a} \cdot \mathbf{x} + b = 0$, and the locations \mathbf{x}_i at which $J_i = -1$ (blue circles) and $J_i = +1$ (red circles). (b) The logistic fit of raw SVM output represents a probability $\mathbb{P}[\mathbf{x} \in \mathcal{M}_1]$ conditioned on the SVM raw output being g^* , that is, $\mathbb{P}[J(\mathbf{x}) = 1 | G(\mathbf{x}) = g^*]$.

The solid line in Figure 1a represents the optimum classifier line $\mathbf{a} \cdot \mathbf{x} + b = 0$, while the dotted lines indicate the extent of the margin, the region within which the boundary could be shifted orthogonally while preserving the perfect classifying accuracy (James et al., 2014) (The points used to construct the margin are called the support vectors.) These two lines have the same slope as the classifier line (the vector \mathbf{a}) but differ by the intercepts, that is, their equations are written as $\mathbf{a} \cdot \mathbf{x} + b = \pm 1$.

All the data points \mathbf{x}_i ($i = 1, \dots, N$) satisfy either $\mathbf{a} \cdot \mathbf{x} + b \geq 1$ or $\mathbf{a} \cdot \mathbf{x} + b \leq -1$. These inequalities are combined into one,

$$(\mathbf{a} \cdot \mathbf{x}_i + b)J_i \geq 1, \quad i = 1, \dots, N. \quad (2)$$

These inequalities become an equality for the support vectors \mathbf{x}_i . Let $a = |\mathbf{a}|$ denote the Euclidean length of \mathbf{a} . The margin d is given by $d = 2/a$ (Wohlberg et al., 2006). The SVM identifies the values of \mathbf{a} and b by maximizing the margin d or, equivalently, by minimizing $a/2$ subject to the linear constraints (Equation 2). Introducing Lagrange multipliers $\boldsymbol{\gamma} = \{\gamma_1, \dots, \gamma_N\}$, this yields an optimization problem $\{\mathbf{a}^*, b^*\} = \arg \min_{\mathbf{a}, b, \boldsymbol{\gamma}} L$, where the objective function $L(\mathbf{a}, b, \boldsymbol{\gamma})$ is given by

$$L = \frac{a}{2} - \sum_{i=1}^N \gamma_i [(\mathbf{a} \cdot \mathbf{x}_i + b)J_i - 1]. \quad (3)$$

The indicator function at points \mathbf{x} where measurements are absent is given by

$$J(\mathbf{x}) = \text{sgn}(\mathbf{a}^* \cdot \mathbf{x} + b^*). \quad (4)$$

It is usually referred to as a decision function in the SVM literature.

The linear SVM can be augmented to account for slight deviations from a perfectly linear classification boundary by introducing slack variables $\xi_i \geq 0$ ($i = 1, \dots, N$). The linear SVM minimization problem is replaced with the problem of minimizing the objective loss function $a/2 + C \sum_{i=1}^N \xi_i$ subject to the constraints $(\mathbf{a} \cdot \mathbf{x}_i + b)J_i \geq 1 - \xi_i$ with $i = 1, \dots, N$. Magnitude of the constant C determines the strength of the slack penalty. Introducing Lagrange multipliers $\boldsymbol{\gamma} = \{\gamma_1, \dots, \gamma_N\}$ and $\boldsymbol{\delta} = \{\delta_1, \dots, \delta_N\}$ for $i = 1, \dots, N$ gives an objective function similar to (3). This optimization problem is rewritten as $\{\mathbf{a}^*, b^*, \boldsymbol{\xi}^*\} = \arg \min_{\mathbf{a}, b, \boldsymbol{\xi}, \boldsymbol{\gamma}, \boldsymbol{\delta}} L_{\xi}$, where $L_{\xi}(\mathbf{a}, b, \boldsymbol{\xi}, \boldsymbol{\gamma}, \boldsymbol{\delta})$ is defined as

$$L_{\xi} = L - \sum_{i=1}^N (\gamma_i + \delta_i - C)\xi_i, \quad (5)$$

with L given by (3). To facilitate the solution of this optimization problem, one converts it into its dual,

$$\boldsymbol{\gamma}^* = \operatorname{argmax}_{\boldsymbol{\gamma}} \left[\sum_{i=1}^N \gamma_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \gamma_i \gamma_j J_j J_i \mathbf{x}_i \cdot \mathbf{x}_j \right] \quad (6)$$

subject to constraints $0 \leq \gamma_i \leq C$ and $\sum_{i=1}^N \gamma_i J_i = 0$. Once $\boldsymbol{\gamma}^*$ is obtained, the solution of $\partial L_{\ell} / \partial a_k = 0$ ($k = 1, 2$) in (5) is

$$\mathbf{a}^* = \sum_{i=1}^N \gamma_i^* J_i \mathbf{x}_i. \quad (7)$$

Let $\mathbf{x}_n = \mathbf{x}_+$ and $\mathbf{x}_k = \mathbf{x}_-$, for some n and k , denote support vectors for which $J = 1$ and $J = -1$, respectively. For these support vectors, the SVM inequality constraints $(\mathbf{a} \cdot \mathbf{x}_i + b) J_i \geq 1 - \xi_i$ turn into equations, $\pm(\mathbf{a} \cdot \mathbf{x}_{\pm} + b) = 1 - \xi_{\pm}$ with $\xi_n = \xi_+$ and $\xi_k = \xi_-$. Their solution is

$$b^* = -\frac{1}{2}(\mathbf{a}^* \cdot (\mathbf{x}_+ + \mathbf{x}_-) - \xi_- + \xi_+). \quad (8)$$

Thus, a solution for the indicator function in (4) is

$$J(\mathbf{x}) = \operatorname{sgn} \left(\sum_{i=1}^N \gamma_i^* J_i \mathbf{x}_i \cdot \mathbf{x} + b^* \right). \quad (9)$$

3.2. Nonlinear SVM

Boundaries of lithofacies in the subsurface are rarely, if ever, planes (straight lines). Hence, parameter data $\{K_i\}_{i=1}^N$ or its indicator counterpart $\{J_i\}_{i=1}^N$ belonging to different lithofacies cannot be separated by the line $\mathbf{a}^* \cdot \mathbf{x} + b^* = 0$ in $d = 2$ or 3 spatial dimensions. Fortunately, it has been proven (Vapnik, 1998) that there exists a higher-dimensional space (whose dimension m is generally unknown) in which the data become linearly separable. Let $\mathbf{F} : \mathbb{R}^d \rightarrow \mathbb{R}^m$ denote a map of the d -dimensional physical space onto that m -dimensional space (known as a feature space). In other words, every point $\mathbf{x} \in D$ corresponds to a point $\hat{\mathbf{x}} \in \mathbb{R}^m$, such that $\hat{\mathbf{x}} = \mathbf{F}(\mathbf{x})$. The linear SVM in \mathbb{R}^m separates the data by a hyperplane $\hat{\mathbf{a}}^* \cdot \hat{\mathbf{x}} + b^* = 0$, whose coefficients $\hat{\mathbf{a}}^* \in \mathbb{R}^m$ and $b^* \in \mathbb{R}$ are determined from the transformed data set $\{\hat{\mathbf{x}}_i, J_i\}_{i=1}^N$. This is accomplished by solving the quadratic optimization of the linear SVM in (3) and (5), in which \mathbf{a} and \mathbf{x} are replaced with $\hat{\mathbf{a}}$ and $\hat{\mathbf{x}}$. Similar to (4), the indicator function is given by $J(\mathbf{x}) = \operatorname{sgn}(\hat{\mathbf{a}}^* \cdot \mathbf{F}(\mathbf{x}) + b^*)$.

While this indicator function is linear in the m -dimensional feature space, it corresponds to a nonlinear function in the physical space, whose specific form is determined by the mapping \mathbf{F} . The latter is proven to exist, but its form is generally unknown and, hence, $J(\mathbf{x})$ is not directly computable. Instead, one solves the dual constrained optimization problem in (6) with \mathbf{x}_i and \mathbf{x}_j replaced by $\hat{\mathbf{x}}_i = \mathbf{F}(\mathbf{x}_i)$ and $\hat{\mathbf{x}}_j = \mathbf{F}(\mathbf{x}_j)$. The resulting inner product of the mapping functions, $\mathbf{F}(\mathbf{x}_i) \cdot \mathbf{F}(\mathbf{x}_j)$, remains uncomputable and is replaced with an empirical function called a Mercer kernel, $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) \equiv \mathbf{F}(\mathbf{x}_i) \cdot \mathbf{F}(\mathbf{x}_j)$. Examples of Mercer kernels include polynomials, sigmoid functions (e.g., hyperbolic tangent), and Gaussian functions (James et al., 2014). Based on our experiments, the exponential radial basis function kernel,

$$\mathcal{K}_{\text{ERB}}(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\| / \ell), \quad (10)$$

where ℓ denotes the kernel's width or the radius of influence of the samples selected to be support vectors, yields the best results in terms of the error on the test set. A nonlinear kernel is generally expected to outperform its linear counterpart in capturing a nonlinear boundary between two classes. Although not done here, one can cycle through multiple kernels and select the one that has the smallest error on the test data set. Once a functional form for the Mercer kernel $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j)$ has been selected, the dual optimization problem

$$\boldsymbol{\gamma}^* = \operatorname{argmax}_{\boldsymbol{\gamma}} \left[\sum_{i=1}^N \gamma_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \gamma_i \gamma_j J_j J_i \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) \right] \quad (11)$$

is solved subject to constraints $0 \leq \gamma_i \leq C$ and $\sum_{i=1}^N \gamma_i J_i = 0$.

In analogy to (9), the indicator function is written as

$$J(\mathbf{x}) = \text{sgn}g(\mathbf{x}), \quad g(\mathbf{x}) = \sum_{i=1}^N \gamma_i^* J_i K(\mathbf{x}_i, \mathbf{x}) + b^*. \quad (12)$$

Combining (7) and (8), both written for their counterparts in \mathbb{R}^m , yields a computable expression for the constant b^* ,

$$b^* = -\frac{1}{2} \sum_{i=1}^N \gamma_i^* J_i [\mathcal{K}(\mathbf{x}_i, \mathbf{x}_+) + \mathcal{K}(\mathbf{x}_i, \mathbf{x}_-)]. \quad (13)$$

3.3. Uncertainty Quantification for SVM Predictions

To quantify uncertainty in SVM reconstruction, we repurpose the pSVM originally developed in the context of classification of nonperfectly separable data (Platt, 2000). The original SVM classifier J is binary, defined by the sign of the function $g(\mathbf{x})$ in (12). Instead, we use a value of $g(\mathbf{x})$ to estimate probability $\mathbb{P}[\mathbf{x} \in M_1]$ of the point \mathbf{x} belonging to the material M_1 . Let $g_i = g(\mathbf{x}_i)$ with $i = 1, \dots, N$ constitute a training set. These numbers are thought of as realizations of the corresponding random variables G_1, \dots, G_N , which are characterized by the (unknown) class-conditioned probability $\mathbb{P}[G_i \leq g^* \mid J(\mathbf{x}_i) = 1]$ where g^* is a value of $g(\mathbf{x}_i)$ at the point of interest \mathbf{x}_i . Instead of estimating this probability, we estimate the conditional probability $\mathbb{P}[J(\mathbf{x}_i) = 1 \mid G_i = g^*]$ and extend this probability to any point \mathbf{x} where measurements are not available, $\mathbb{P}[J(\mathbf{x}_i) = 1 \mid G_i = g(\mathbf{x})]$.

Our parametric estimation strategy relies on the assumed functional form of $\mathbb{P}(J(\mathbf{x}) = 1 \mid G(\mathbf{x}) = g^*)$. By way of example, we consider a sigmoidal function in Figure 1b,

$$\mathbb{P}[J(\mathbf{x}) = 1 \mid G(\mathbf{x}) = g^*] = \frac{1}{1 + \exp(Ag^* + B)}. \quad (14)$$

The fitting parameters A and B are found by minimizing the negative log-likelihood of the training data. First, we map the training set $\{\mathbf{x}_i, J_i\}_{i=1}^N$ onto a training set $\{g_i, t_i\}_{i=1}^N$, where $t_i = (J_i + 1)/2$ are target probabilities. Then, the negative log-likelihood function or a “cross-entropy error function” is minimized to find optimal values A^* and B^* ,

$$\{A^*, B^*\} = \arg \min_{A, B} \left[-\sum_i^N t_i \ln p_i + (1 - t_i) \ln(1 - p_i) \right], \quad (15)$$

where $p_i = \mathbb{P}(J(\mathbf{x}_i) = 1 \mid G(\mathbf{x}_i) = g^*)$. We use the trust-region Newton algorithm (Fan et al., 2008; Lin et al., 2007) to solve this two-parameter minimization problem. Finally, (14) with $A = A^*$ and $B = B^*$ is used in place of $\mathbb{P}[\mathbf{x} \in M_1]$.

4. Performance Analysis on Synthetic Data

4.1. Synthetic Data Set

Our goal is to reconstruct probabilistically the lithofacies defined by, for example, the hydraulic conductivity field $K(\mathbf{x})$ in Figure 2 from N measurements $K_i = K(\mathbf{x}_i)$ collected at randomly selected locations \mathbf{x}_i ($i = 1, \dots, N$). The field $K(\mathbf{x})$, originally used for similar purpose by Wohlberg et al. (2006), is constructed by superimposing two autocorrelated, weakly stationary, normally distributed random fields, representing two distinct spatial distributions of log-conductivity with the ensemble means of 0.1 and 7.0. When hydraulic conductivities are expressed in centimeters per day, this corresponds to clayey and sandy materials, respectively. The two log-conductivity distributions are mutually uncorrelated, have unit variance and Gaussian autocorrelation with unit correlation scale. SGSIM software (Deutsch & Journel, 1998) is used to generate both fields on a 60×60 grid, using a grid spacing of $1/5$ of the log-conductivity correlation length. Next, the composite porous medium is constructed by randomly choosing the shape of the internal boundary. The corresponding indicator field $J(\mathbf{x})$ is constructed by assigning to each pixel either $+1$ or -1 , that is,

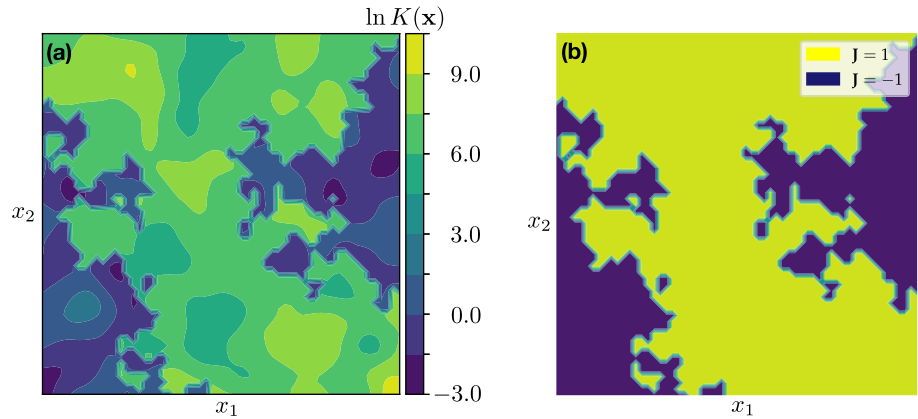


Figure 2. (a) Log-hydraulic conductivity field $\ln K(\mathbf{x})$ and (b) the corresponding indicator function $J(\mathbf{x})$ used in our numerical experiments. The two heterogeneous facies are sufficiently distinct for the mapping $\ln K(\mathbf{x}) \rightarrow J(\mathbf{x})$ not to introduce a classification error.

identifying its membership in either facies M_1 or M_2 , using a threshold value of 4.0. Given the vast difference between the means, this assignment is free of classification error.

4.2. Probabilistic Facies Reconstruction

The boundaries between materials M_1 and M_2 reconstructed by the standard (deterministic) SVM from $N = 50$ data points (out of the total of $N_{\text{tot}} = 3,600$ pixels) are shown in Figure 3a. Even with this relatively sparse sampling, SVM mislabel some the data points in order to prevent the overfitting of the model (blue dots are corresponding to geological facie $J = -1$ that end up in the area in the middle area). That is because, similar to many other machine learning techniques (e.g., neural networks), SVM minimize the generalization error rather than the observation error. The misclassification seen in Figure 3a is the result of a trade-off between the classification error on the training set and the minimal classification error on the unseen data. Such mislabeling of pixels of a full image gave impetus to the original pSVM (Platt, 2000). Here we use it as a probabilistic classifier of incomplete images with the majority of pixels missing.

Figure 3b exhibits a representative probability map of facies M_1 reconstructed by pSVM from $N = 50$ measurements. It indicates the confidence in identifying each pixel as a member of M_1 . The dark blue areas represent subdomains where the probability $\mathbb{P}[\mathbf{x} \in M_1]$ is close to zero, that is, these areas are highly likely

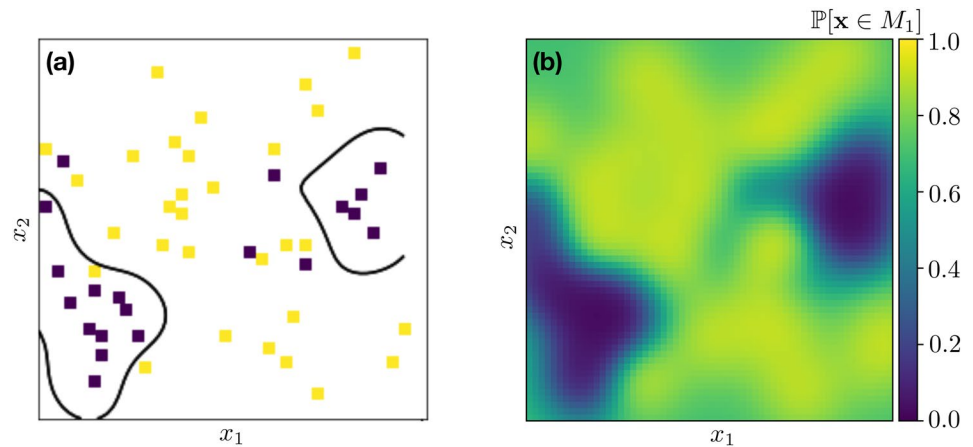


Figure 3. (a) Boundaries drawn by deterministic SVM with the slack penalty constant $C = 1$ and the kernel width $\ell = 2$; blue and orange pixels represent samples from materials M_1 ($J = 1$) and M_2 ($J = -1$), respectively. (b) Probability map of facies M_1 reconstructed by pSVM with the slack penalty constant $C = 1$ and the kernel width $\ell = 2$ from $N = 50$ pixels (out of the total of $N_{\text{tot}} = 3,600$).

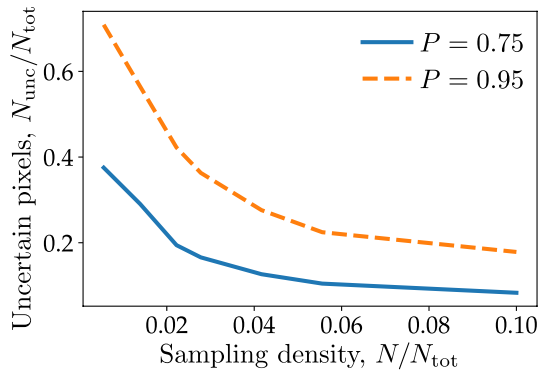


Figure 4. Relative number of uncertain pixels, N_{unc}/N_{tot} , as function of the sampling density, N/N_{tot} , for two degrees of certainty, $P = 0.75$ and 0.95 , that is, for correspondingly wider confidence intervals.

$P = 0.75$ and 0.95 ; this result represents an average over 20 realizations of the set $T = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ of randomly selected measurement locations, each of which yields a probability map similar to the one in Figure 3b. (The negligible computational cost of SVM training—less than a second on a laptop—facilitate their use in the ensemble setting.) As expected, the number of uncertain pixels increases with the probability threshold P and decreases with the sampling density.

4.3. Comparison with Indicator Kriging

Indicator Kriging (IK) (Isaaks & Srivastava, 1990) provides an alternative means for probabilistic reconstruction of hydrofacies (Guadagnini et al., 2004). This method defines the indicator function as $I(\mathbf{x}_i) = 1$ for $\mathbf{x}_i \in M_1$ and $= 0$ for $\mathbf{x}_i \in M_2$, and treats it as a stationary random field. The best linear unbiased estimator (aka Kriging) interpolates between the measurement points, yielding $\mathbb{E}[I(\mathbf{x})] = \mathbb{P}[\mathbf{x} \in M_1]$. The correlation function (variogram) of $I(\mathbf{x})$, which is inferred from the spatial data $I_i = I(\mathbf{x}_i)$ with $i = 1, \dots, N$, determines the interpolation weights.

Figure 5 exhibits realizations of the probability maps of M_1 alternatively identified with pSVM and IK from $N = 50$ measurements. (The measurement locations are chosen at random, varying between realizations shown in Figure 5 and differing from a realization depicted in Figure 3b; this allows us to illustrate the impact of measurement locations on the quality and reliability of facies reconstruction.) The probability maps generated with pSVM and IK are qualitatively similar, even though pSVM generates a smoother map than that produced by IK. This suggests that pSVM provides a more conservative facies reconstruction (larger areas with probabilities other than 0 and 1). This finding is consistent among all the realizations of measurement locations we have analyzed, including those shown in Figure 5.

The qualitative similitude between pSVM and IK argues in favor of the former. That is because IK is more complex and possesses more tunable parameters, such as lag, lag separation, lag tolerance, azimuth, dip, tolerance, and bandwidth. Manual fitting of data to an experimental variogram is highly subjective, requiring one to visually identify an appropriate nugget effect and sill. Finally, construction of a variogram requires a large number of samples collected at various degrees of spatial separation, while SVM theoretically work with as few as two data points (support vectors).

Figure 6 serves to quantify the discrepancy between the two methods for constructing the probability maps in Figure 5, and to compare them with their empirical counterpart for the ground truth in Figure 2. A metric for this assessment is computed as follows. First, for each map in Figure 5 corresponding to a given realization of the measurement locations, we construct a histogram of pixels whose probabilities fall within bins of size $\Delta p = 0.1$. Second, we compute the average probability for each bin: For example, let $\{\mathbf{x}_k\}_{k \in \mathcal{I}_{0.1,0.2}}$ denote a set of $N_{0.1,0.2}$ pixels with probabilities $\{p_k\}_{k \in \mathcal{I}_{0.1,0.2}}$ that fall within the bin $[0.1, 0.2)$; then the average probability for that bin is

to consist of material M_2 . On the other hand, the yellow and light green areas represent subdomains that likely belong to material M_1 . SVM classification of the remaining parts of the simulation domain is highly uncertain. Although not shown here, and as expected, this transition zone increases as the sampling density, N/N_{tot} , decreases. We also found that the measurement locations play smaller role on the confidence maps as the sampling density increases. To convert the probabilistic map into a deterministic one providing the “best” guess of the spatial extent of the hydrofacies, one can define the boundary between facies M_1 and M_2 as the probability isoline corresponding to the relative number of samples of facies M_1 in the total number of samples. That is the strategy used in the geostatistical approach of Guadagnini et al. (2004).

One metric of pSVM output is the fractional number of uncertain pixels, N_{unc}/N_{tot} , defined for a given probability threshold $P > 0.5$. A pixel \mathbf{x} is deemed uncertain with confidence P if its membership probability $\mathbb{P}[\mathbf{x} \in M_1]$ falls within the interval $[1 - P, P]$. Figure 4 shows N_{unc}/N_{tot} as function of the sampling density N/N_{tot} for two degrees of certainty,

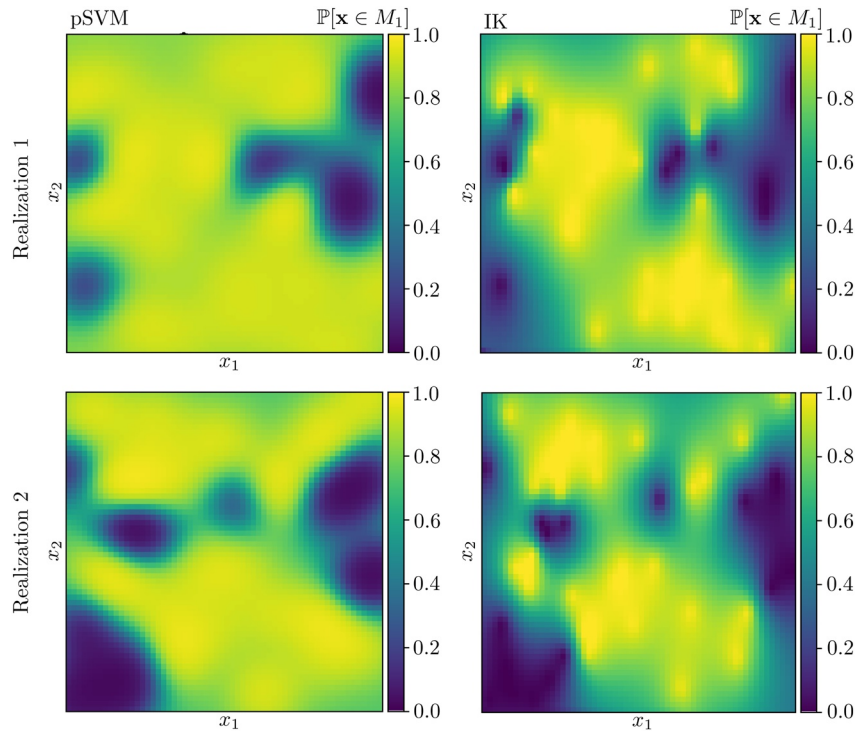


Figure 5. Probability maps of facies M_1 constructed by pSVM (left column) and indicator Kriging (right column) from $N = 50$ measurements. The measurement locations are selected at random, giving rise to different realizations of the reconstructed fields.

$$\bar{p}_{\{0,1,0,2\}} = (1 / N_{0,1,0,2}) \sum_{k \in \mathcal{I}_{0,1,0,2}} p_k.$$

The values of \bar{p} for each bin, inferred from the probability maps designated by Realization 1 in Figure 5, are plotted in Figure 6 against the corresponding probabilities p^{true} . (Other realizations yield similar results.) These are computed as the fraction of the pixels in a given bin labeled as $J = 1$ in Figure 2. For example, if the number of the $J = 1$ pixels in the set $\{\mathbf{x}_k\}_{k \in \mathcal{I}_{0,1,0,2}}$ is $N_{0,1,0,2}^{J=1}$, then $p_{\{0,1,0,2\}}^{\text{true}} = N_{0,1,0,2}^{J=1} / N_{0,1,0,2}$. The 45° line in Figure 6 corresponds to the perfect agreement between the average probability \bar{p} and the corresponding

empirical probability p^{true} inferred from the ground truth. The average probabilities \bar{p} predicted by either pSVM or Kriging exhibit comparable deviations from the 45° line; this suggests that, according to this metric, the two methods for probabilistic reconstruction of geologic facies have comparable accuracy.

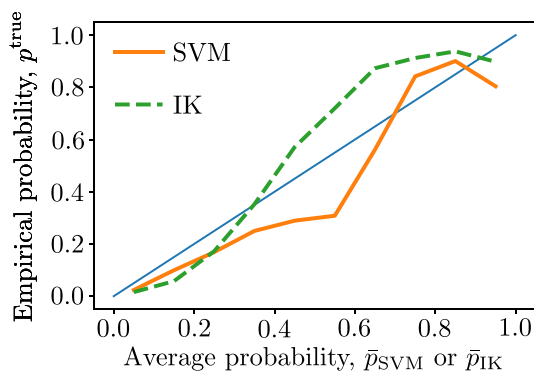


Figure 6. Comparison of the average probability \bar{p} predicted by either pSVM (\bar{p}_{SVM}) or indicator Kriging (\bar{p}_{IK}) and the corresponding empirical probability p^{true} inferred from the ground truth. The 45° line corresponds to the perfect agreement between the two.

4.4. Sensitivity to SVM Parameters

Performance of SVM is controlled by two parameters: The slack penalty constant C in (5) and the kernel width ℓ in (10). The regularization parameter C provides a trade-off between the correct classification of training data and the minimization of generalization error. Larger values of C allow smaller margins if the decision function is better at correctly classifying all training points. Smaller values of C promote larger margins and, hence, a simpler decision function at the cost of training accuracy.

Small values of the parameter ℓ indicate a long-range influence of each observation, while its high values limit the overall impact of each data

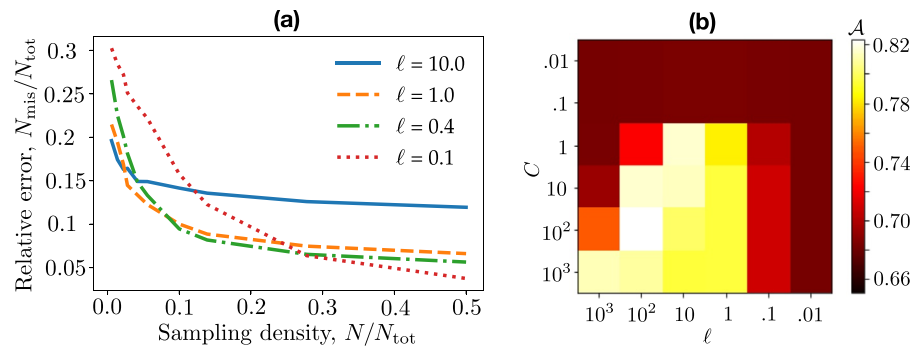


Figure 7. (a) Relative number of misclassified pixels, $N_{\text{mis}}/N_{\text{tot}}$, as function of sampling density N/N_{tot} , for several values of the kernel width ℓ . (b) Classification accuracy \mathcal{A} computed from the five-fold cross-validation. Light color elements correspond to the combinations of the SVM parameters C and ℓ that lead to well-performing models.

point. If ℓ is too small, the radius of influence of the support vectors includes only the support vector itself and no amount of regularization with C would prevent overfitting. When ℓ is very large, the model is too constrained and cannot capture the complexity or “shape” of the data. The region of influence of any selected support vector would include the whole training set. The resulting model will behave similarly to a linear model with a set of hyperplanes separating the centers of two classes.

Figure 7a shows the impact of ℓ on the SVM accuracy, that is, on the relative number of misclassified pixels, $N_{\text{mis}}/N_{\text{tot}}$. Small values of ℓ are beneficial when sampling density is high, while its large values yield a better performance when data are very sparse. In the latter case, small values of ℓ lead to overfit and result in low prediction accuracy.

Figure 8 exhibits representative reconstructions of geological facies obtained with two choices of the parameter ℓ (and $C = 1.0$) for two sampling densities. Large values of ℓ yield boundaries that are too smooth, while small values of ℓ cause boundaries to follow the training points too closely. Selecting a right value for ℓ is more crucial for low sample density ($N/N_{\text{tot}} = 50/3600$). As expected, this situation also gives rise to appreciable variation between realizations (different sample locations). Both the importance of selecting

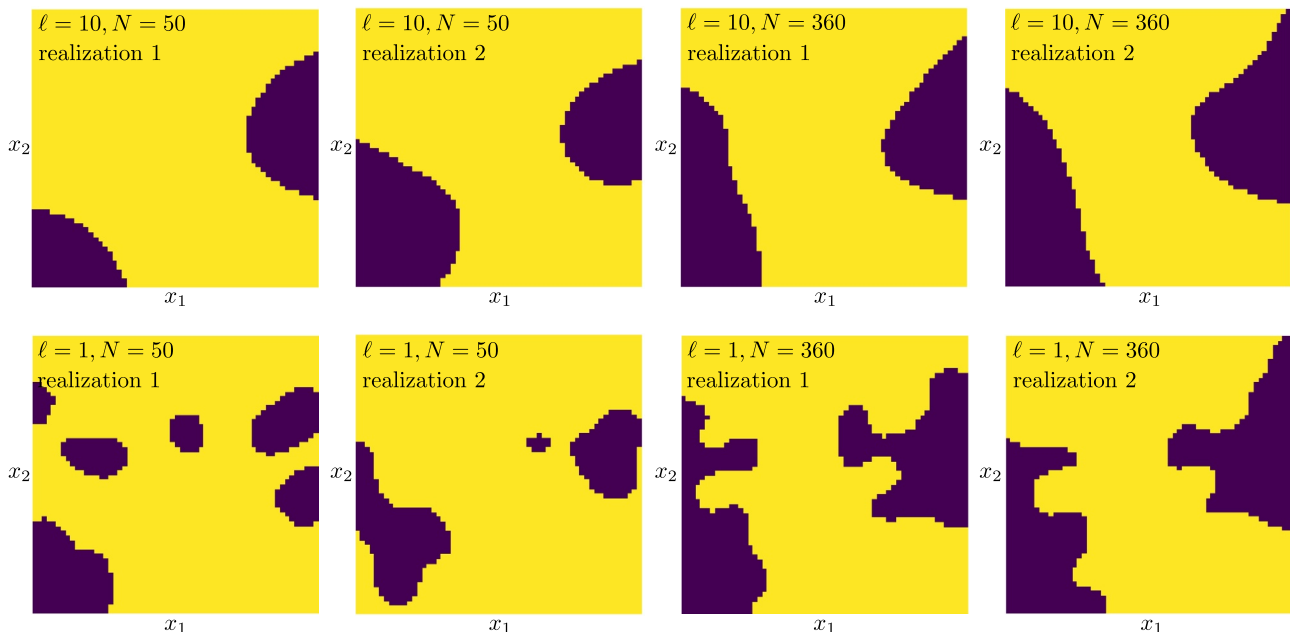


Figure 8. Representative lithofacies reconstructions via SVM with $C = 1$ and either $\ell = 10.0$ (top row) or $\ell = 1.0$ (bottom row), for two sets of N samples (realizations 1 and 2).

a value for ℓ and the between-realizations variability diminish when the sampling density increases to $N/N_{\text{tot}} = 360/3600$.

4.5. Strategy for Parameter Selection

Unlike IK, SVM possess a well-established framework for tuning its hyper-parameter. Specifically, optimal values of C and ℓ are chosen with the following algorithm.

- Identify a region in the two-dimensional SVM parameter space (spanned by parameters C and ℓ), over which C and ℓ are allowed to vary. In our computational examples, we chose the intervals $C \in [10^{-2}, 10^3]$ and $\ell \in [10^{-2}, 10^3]$ to exclude parameters for which poor performance is expected a priori
- Discretize this region with a regular grid and carry out the SVM reconstruction for all pairs of the parameter values. We used the mesh size $\Delta_{\log C} = 1$ and $\Delta_{\log \ell} = 1$, which results in 36 reconstructions
- Perform five-fold cross-validation to evaluate the test error of each parameter combination. The data are split into $k = 5$ subsets or “folds” \mathcal{F}_i with $i = 1, \dots, k$. For every i , the model is trained on all folds except for the i th fold. The test error on the i th fold is computed as

$$\mathcal{E}_i = \frac{1}{k} \sum_{n \in \mathcal{F}_i} \mathbb{1}(J_n \neq \hat{J}_{n/i}),$$

where J_n is a true label of pixel \mathbf{x}_n , and $\hat{J}_{n/i}$ is a prediction for the pixel \mathbf{x}_n obtained without using the fold \mathcal{F}_i of the data set to fit the model. Next, cross-validation error is obtained by averaging the test errors of individual folds,

$$\mathcal{E}_{\text{cv}} = \frac{1}{k} \sum_{i=1}^k \mathcal{E}_i.$$

- Finally, the classification accuracy is defined as $\mathcal{A} = 1 - \mathcal{E}_{\text{cv}}$

Figure 7b exhibits the classification accuracy \mathcal{A} for the 36 combinations of the SVM parameters C and ℓ . The best-performing SVM models are parameterized with the values of C and ℓ that lie on the diagonal of the plot. Possible spurious variations of the classification accuracy \mathcal{A} between different chosen parameters can be smoothed out by increasing the number of folds, k , used in cross-validation at the expense of computational time. While the analysis was not performed on the most optimal hyper-parameters from cross-validation, the cross-validation accuracy obtained with the chosen parameters $C = 1$ and $\ell = 2$ is not far from that of the optimum.

5. Application to Field Data

To demonstrate the applicability of pSVM to real-world groundwater problems, we consider a data set from a site in Southern California. The raw lithological data, collected from 107 cone penetrometer tests (CPTs) spread over approximately $3 \text{ km} \times 3 \text{ km}$ area, can be found in the file rawdata.txt available for download with this submission. The CPT data consist of vertical profiles of 1s and 0s, labeling geologic materials of low and high conductivity, respectively (Figure 9a). We say that a CPT location lies in the aquitard (facies M_1) if the vertical average of these measurements exceeds a certain threshold δ_{th} ; otherwise, it is said to belong to a high-permeability inclusion (facies M_2). The data points corresponding to threshold $\delta_{\text{th}} = 0.5$, marked as “low” for facies M_1 and “high” for facies M_2 , are shown in Figure 9b.

A visual comparison of Figures 3a and 9b reveals a qualitative similarity between the synthetic and field data. In order to select the most appropriate hyper-parameters, we perform cross-validation, whose results are depicted in Figure 10. In addition, the performance analysis is made possible by the availability of the ground truth in Section 4 guides our selection of the hyper-parameters C and ℓ here. In particular, we set the slack penalty constant $C = 1$ and the kernel width $\ell = 0.1$ for the pSVM approach (Otherwise, prior knowledge about the geological structure of an aquifer could assist in choosing appropriate values of hyper-parameters: The presence of large continuous lithofacies calls for the use of small values of the parameters C and ℓ , while an aquifer with a large number of inclusions is represented by large values of C and ℓ .) Fitting the boundary requires two hyper-parameters, A and B in (14). Yet they need not be pre-determined

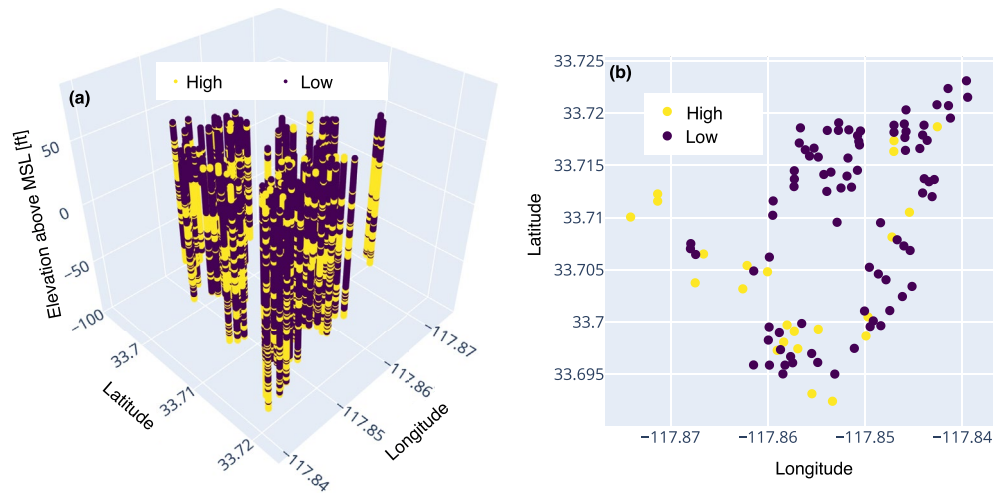


Figure 9. (a) Raw binary data collected with 117 cone penetrometer tests. (b) Vertically averaged CPT data with threshold $\delta_{th} = 0.5$; these data are used to delineate high-permeability inclusions (facies M_2) in the aquitard (facies M_1) in the horizontal plane of a groundwater model.

for the probability calculation in the pSVM, being determined from data by solving the optimization problem in (15).

While pSVM involve only two hyper-parameters (C and ℓ), IK has seven tunable parameters: lag (set to 6.0), lag separation (0.005), lag tolerance (0.001), azimuth (20.0), dip (0.0), tolerance (70.0), and bandwidth (0.1). Additional IK parameters (range, sill, and nugget) are obtained from the variogram fitting, in a procedure akin to determination of the parameters A and B in pSVM.

Figure 11 exhibits probability maps of the aquitard (facies M_1) generated by pSVM and IK from the PCT data in Figure 9b. The predicted probability, $\mathbb{P}[\mathbf{x} \in M_1]$, of a point \mathbf{x} far removed from data locations belonging to the aquitard is close to 1. That is to be expected since the data set in Figure 9b contains many more points from M_1 . This emerging feature of pSVM is consistent with the built-in characteristic of the geostatistical approach to probabilistic facies delineation (Guadagnini et al., 2004). The pSVM-based classification is more conservative since, being a regression rather than interpolation (as Kriging is), it aims to minimize the generalization error rather than the interpolation error. Normalization before fitting pSVM is crucial in this

case because the distribution of each feature (x and y values) is far from the standard Gaussian distribution. IK provides a more aggressive prediction, which is controlled by such unobservable hyper-parameters as the size of the search ellipsoid. If the latter is set to be small, a large portion of the space is left unclassified because the ellipsoid would not contain any data points. On the other hand, too large of an ellipsoid would result in a prediction that is unrealistically uniform. A more aggressive prediction provided by IK is likely to result in an unjustifiably overconfident interpretation of the field data.

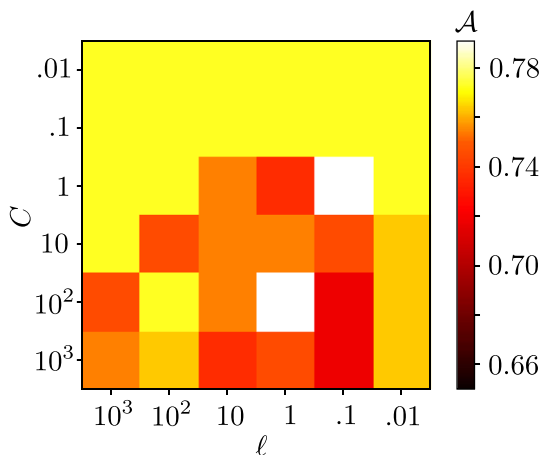


Figure 10. Classification accuracy A computed from the five-fold cross-validation. Light color elements correspond to the combinations of the SVM parameters C and ℓ that yield well-performing models.

6. Conclusions

We introduced probabilistic support vector machines (pSVM) as a means of delineation of subsurface lithofacies from sparse data. The method replaces the binary classifier with its continuous counterpart that is constructed by fitting a logistic curve to observations. The result is a probability map that provides the likelihood of a pixel belonging to a facies, rather than a deterministic pixel label provided by standard SVM. Our numerical experiments lead to the following major conclusions.

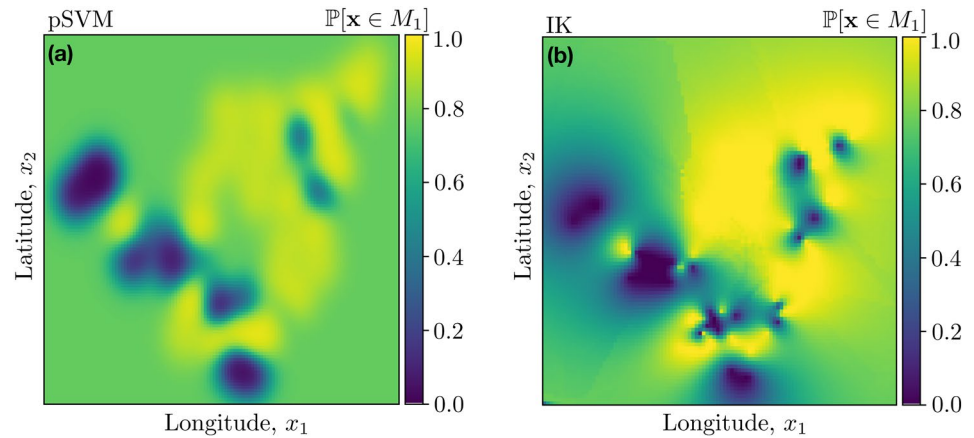


Figure 11. Probability maps of the aquitard (facies M_1) generated by (a) pSVM and (b) indicator Kriging from the PCT data shown in Figure 9b.

- (1) Probability maps generated with pSVM and indicator Kriging (IK), the current method of choice for probabilistic forecasting, are qualitatively similar
- (2) pSVM generate smoother probability maps than those produced by IK, suggesting that pSVM provide a more conservative facies reconstruction (larger areas with probabilities other than 0 and 1) to ensure more reasonable classification of the unseen space
- (3) The qualitative similitude between pSVM and IK argues in favor of the former, because IK is more complex, has more tunable parameters, and has higher data requirements
- (4) Performance of SVM is controlled by two parameters: The slack penalty constant C and the kernel width ℓ
- (5) Small values of ℓ are beneficial when sampling density is high, while its large values yield a better performance for sparse data

More work remains to be done in the area of probabilistic image reconstruction from sparse data. In addition to their use as a classifier, SVM can be deployed as a regressor to estimate the parameter values between points where the parameter is sampled (Wohlberg et al., 2006). One line of future research is to develop pSVM for quantification of uncertainty in the estimates of a parameter of interest (hydraulic conductivity, in our examples).

Another venue is to explore the conformal prediction (Hechtlinger et al., 2018) as an alternative to pSVM for quantification of uncertainty in subsurface delineation from sparse data. This approach uses past experience to determine precise levels of confidence in new predictions. It is designed for an on-line setting in which labels are predicted successively, each one being revealed before the next is predicted (Shafer & Vovk, 2008). Such a strategy might indicate regions in space where a prediction with a required degree of certainty is not possible due to the lack of information.

One drawback of SVM in general and pSVM in particular is the sampling bias problem. SVM algorithm could be modified by assigning different weights to each sampled point (Yang et al., 2007). Such weighted classification could be easily achieved by scikit-learn (Pedregosa et al., 2011). However, unlike in Kriging, a principled way to determine the weights for SVM is still lacking. We leave this challenge for a follow-up study.

In our experiments, pSVM have outperformed the IK-based approach by being highly automated and producing consistent results. When training images (either from raw outcrop data or processed technique) are available, multi-point geostatistics (Tahmasebi, 2018) might be a more suitable approach than its two-point counterpart (IK). Incorporation of training images into pSVM is needed to provide a fair comparison between these two strategies.

Data Availability Statement

The field data used in Section 5 are provided in the file rawdata.txt accompanying this article.

Acknowledgments

The first author thanks A. Pradhan and A. Pollack for their help with Stanford Geostatistical Modeling Software (SGEMs). This research was supported in part by Total and ADNOC. There are no data sharing issues since all of the numerical information is provided in the figures produced by solving the equations in the paper.

References

- Deutsch, C. V., & Journel, A. G. (1998). *GSLIB geostatistical software library and user's guide* (2nd ed.). Oxford Univ. Press.
- Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., & Lin, C.-J. (2008). LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9, 1871–1874.
- Guadagnini, L., Guadagnini, A., & Tartakovsky, D. M. (2004). Probabilistic reconstruction of geologic facies. *Journal of Hydrology*, 294, 57–67. <https://doi.org/10.1016/j.jhydrol.2004.02.007>
- Hechtlinger, Y., Póczos, B., & Wasserman, L. (2018). *Cautious deep learning*.
- Isaaks, E. H., & Srivastava, R. M. (1990). *An introduction to applied geostatistics*. Oxford Univ. Press.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2014). *An introduction to statistical learning: With applications in R*. Springer.
- Lin, C.-J., Weng, R. C., & Keerthi, S. S. (2007). Trust region newton methods for large-scale logistic regression. *Proceedings of the 24th international conference on machine learning* (pp. 561–568). ACM. <https://doi.org/10.1145/1273496.1273567>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Platt, J. C. (2000). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In A. J., Smola, P. L., Bartlett, B. Schölkopf, & D. Schuurmans (Eds.), Eds., *Advances in large margin classifiers* (pp. 61–74). MIT Press.
- Shafer, G., & Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9, 371–421.
- Tahmasebi, P. (2018). Multiple point statistics: A review. In B. Daya Sagar, Q. Cheng, & F. Agterberg (Eds.), Eds., *Handbook of mathematical geosciences: Fifty years of IAMG* (pp. 613–643). Springer. https://doi.org/10.1007/978-3-319-78999-6_30
- Tartakovsky, D. M., & Wohlberg, B. E. (2004). Delineation of geologic facies with statistical learning theory. *Geophysical Research Letters*, 31(18), L18502. <https://doi.org/10.1029/2004GL020864>
- Tartakovsky, D. M., Wohlberg, B. E., & Guadagnini, A. (2007). Nearest neighbor classification for facies delineation. *Water Resources Research*, 34, L05404. <https://doi.org/10.1029/2007GL029245>
- Vapnik, V. N. (1998). *Statistical learning theory*. Wiley-Interscience.
- Wohlberg, B., & Tartakovsky, D. M. (2009). Delineation of geological facies from poorly differentiated data. *Advances in Water Resources*, 32(2), 225–230. <https://doi.org/10.1016/j.advwatres.2008.10.014>
- Wohlberg, B., Tartakovsky, D. M., & Guadagnini, A. (2006). Subsurface characterization with support vector machines. *IEEE Transactions on Geoscience and Remote Sensing*, 44(1), 47–57. <https://doi.org/10.1109/TGRS.2005.859953>
- Yang, X., Song, Q., & Cao, A. (2007). Weighted support vector machine for data classification. *Proceedings of 2005 IEEE international joint conference on neural networks*, 21, 859–864. <https://doi.org/10.1109/IJCNN.2005.1555965>
- Zeng, Y., Jiang, K., & Chen, J. (2018). *Automatic seismic salt interpretation with deep convolutional neural networks*. arXiv e-prints. arXiv:1812.01101.