3D Geological Modeling: Solving as a Classification Problem with the Support Vector Machine

By Alex Smirnoff, Eric Boisvert, and Serge J. Paradis

Geological Survey of Canada
GSC-Quebec
490 de la Couronne
Quebec, Canada G1K 9A9
Telephone: (418) 654-3716
Fax: (418) 654-2615
e-mail: alsmirno@nrcan.gc.ca

ABSTRACT

The process of creating multi-unit 3D geological models by successive unit interpolation may be tedious and time-consuming. Here, we propose to automate this procedure through presenting the problem as a classification task and solving it simultaneously with the Support Vector Machine (SVM), a method known from the field of artificial intelligence. Experiments with various input data and kernel parameters demonstrated that the SVM has great potential in 3D reconstructions from sparse geological information. An extended version of this paper has been accepted for publication in "Computers and Geosciences" (Smirnoff et al., 2008).

INTRODUCTION

Often, geologists are faced with a variety of diverse information that requires generalization and analysis. 3D modeling software packages such as Gocad® of Earth Decision Sciences have proven an excellent means for data presentation and interpretation. The procedure normally requires reconstruction of individual geological units using surfaces interpolated from control points with subsequent fusion of these units into a single model. The popular interpolation techniques include Inverse Distance Weighting (IDW), Discrete Smooth Interpolation (DSI), and various flavors of kriging preceded by semi-variogram analysis.

The above procedure can easily become a tedious and time-consuming task when a complex geomodel is considered. In addition, the traditional interpolation techniques assume reasonable areal coverage of the input data. Therefore, there is a strong need for an algorithm that would automate the process of model creation even in cases when only a few pieces of information on regional geology, (e.g., a few sparse cross-sections) are available. Finding such an algorithm and testing its performance on available data sets was the objective of this study.

Here, we propose the use of the Support Vector Machine (SVM), a tool routinely applied in the field of image analysis and pattern recognition. The SVM is becoming increasingly popular and has been successfully used to solve classification and regression problems in biology (e.g., Noble et al., 2005), hydrology (e.g., Yu et al., 2004), medicine (e.g., El-Naqa et al., 2002), and environmental science (e.g., Gilardi et al., 1999). In this study, we demonstrate that the application of SVM in geology allows sparse data to be efficiently combined in order to reconstruct shape, area, and volume of multiple geological units.

METHODOLOGY

The SVM Algorithm

The SVM algorithm is based on the Statistical Learning Theory developed by V. Vapnik (Vapnik, 1995). It uses a set of examples with known class information to build a hyperplane that separates samples of different classes. In machine learning theory, this is known as supervised learning as opposed to unsupervised learning when no a priori class information is available. This initial dataset is known as a training set, and every sample within it is characterized by features upon which classification is based. Figure 1A demonstrates this for the one-dimensional (single-feature) case. The samples closest to the hyperplane are termed support vectors (filled marks in Figure 1).

In more complicated, non-linear cases, the task of discovering the separator is turned into a linear task by transferring input data into a higher-dimensional space known as the feature space. Figure 1B shows a non-separable one-dimensional data set in the input space. The problem is easily solved through re-mapping data to a higher, two-dimensional space where a linear solution is found (Figure 1C). Functions satisfying certain conditions and known as kernels are normally employed for this

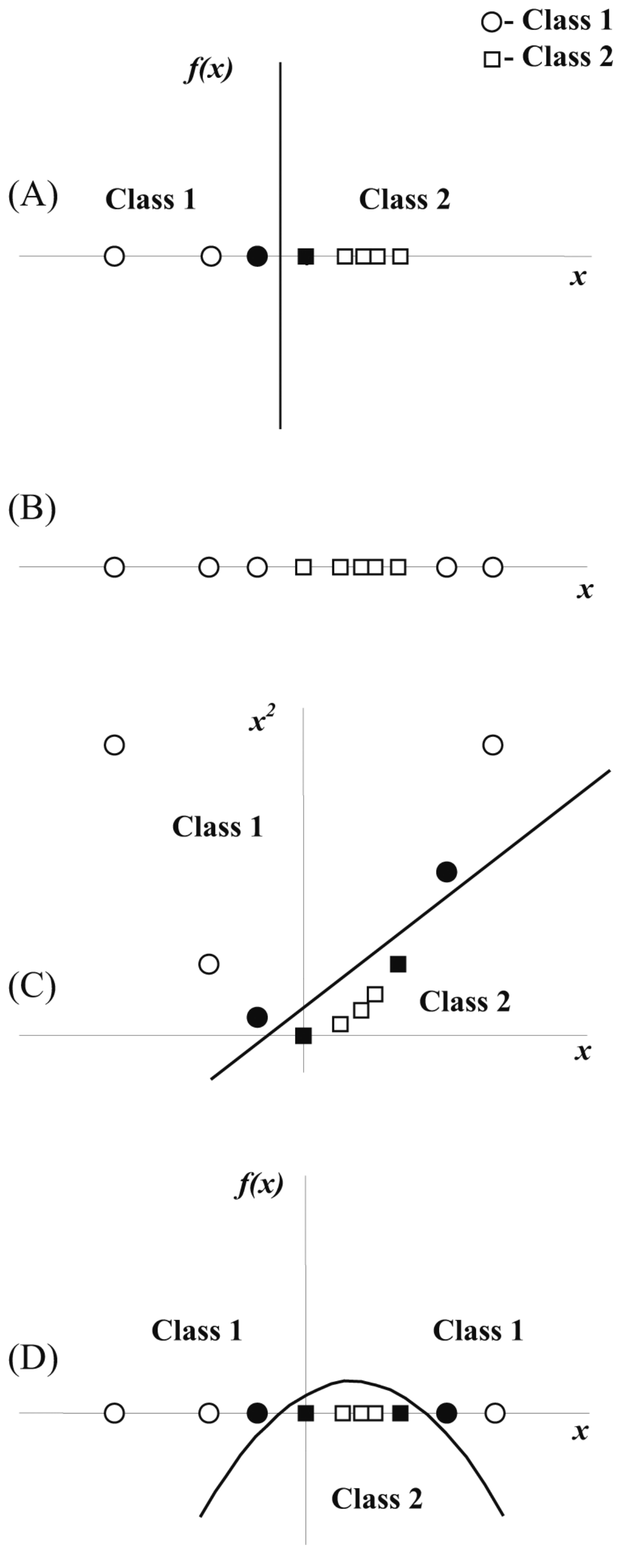

Figure 1. Building the separating hyperplane, in separable and non-separable one-dimensional case. Filled marks represent support vectors. (A) Linearly separable case and decision function in input space; (B) non-separable case in input space. (C) training data re-mapped into two-dimensional feature space using $\phi(x) = (x, x^2)$ and linear solution in this space; (D) solution re-plotted back in input space**.**

transfer (e.g., Abe, 2005). The solution becomes non-linear when shown in the original data space (Figure 1D).

Once the equation for the optimal classifier is found, new data with unknown class information (test samples) can be classified based on the value of this decision function. Unlike most interpolation methods based on the principle that values at points closer in space are more similar, the SVM is a boundary classification method where the boundary is built based on the initial training set among which only a small number of samples (support vectors) are involved in the final decision making.

The classical SVM task is a binary (two-class) classification. However, a number of methods have been developed to support multi-class classification through various combinations of binary methods such as “one-against-all”, “one-against-one”, etc. (see Hsu and Lin, 2002 for references). Therefore, the SVM approach is also applicable for models with more than two classes. More detailed descriptions of the SVM algorithm are available from a number of sources (e.g., Cristianini and Shawe-Taylor, 2000; Abe, 2005).

SVM Application to 3D Modeling

To apply the SVM algorithm to our geological reconstructions, we defined the 3D space-partitioning task as a pure spatial classification problem. Three coordinates uniquely describe every point in the 3D reconstruction space. However, only a limited number of those points possess descriptions or class information that can be identified through well drilling, surficial geology mapping, and seismic profiling. The class information describes the geological unit to which each particular point belongs. Therefore, the points with known class labels become samples in the SVM training set, and point coordinates in the three-dimensional space are used as sample features. Once a classification model based on this training set is built, the rest of the points in the reconstruction space can be classified based on their coordinates (features).

We employed one of the many SVM implementations freely available over the internet, namely LIBSVM developed at the National Taiwan University (see Chang and Lin, 2001 for detailed description). As recommended in Hsu et al. (2004), we used LIBSVM with the radial basis function (RBF) kernel, the most general form of kernel resulting in a prediction model controlled by only two hyperparameters, C and γ. A single solution is obtained for every pair of parameters, and it is sensitive to the choice of their values. However, selecting the appropriate values is a dark art normally done on a try-and-see basis.

For multi-class classification, LIBSVM uses “one-against-one” approach, which was shown to be advantageous to other methods for practical use (Hsu and Lin, 2002). In addition, a set of in-house Java utilities has been developed for scaling, validation, and format conversion purposes.

Data

A 3D geomodel created at the Geological Survey of Canada, Quebec, in the course of the Esker/Abitibi project (Bolduc et al., 2005) was used as the reference dataset in all of the experiments. This model is based on surface geology, well, and cross-section data. Six geological units were sequentially interpolated from the above control points using the Discrete Smooth Interpolation (DSI) technique (Mallet, 1989) in the Gocad® GIS.

Experimental Work

The experimental work was designed to perform the following tasks: (1) investigate whether the SVM can be efficiently used in binary geological reconstructions, (2) test the SVM for multi-unit modeling, and (3) examine how the resulting model depends on the RBF kernel parameters used in the reconstruction.

General Approach

The general approach taken in all of the experiments was as follows:

Prediction

- In Gocad®, create a reconstruction space as a set of volume elements (voxet) of the shape representative of study area geometry. The reconstruction space was defined by a voxet with the following number of volume elements (voxels) in each direction: X-110, Y-240, Z-24.
- Add available data to the reconstruction space. The unit type property for each of the six geological units (SVM classes) was transferred from the stratigraphic grid (SGrid) structure representing the reference model.
- In Gocad®, using a DEM, define all voxet nodes above the surface as air or no-data points.
- Define a training set for the experiment. For the remaining ground points, set the unit type property to zero; these are the points that will be later classified by trained SVM.
- Scale coordinate values for the training set between 0 and 1 as recommended in Hsu et al. (2004).
- Using LIBSVM, build a prediction model based on the training set. A single reference set of kernel parameters ($C = 10^4$ and $\gamma=10^2$) previously determined from 2D experiments was used in all reconstructions except the parameter sensitivity tests.
- Scale coordinate values for the points to be classified and classify them using the prediction model created in the previous step.
- Import the point set with predicted class information back into Gocad®, for visualization and analysis.

Validation

- Based on the available reference data, define the validation set and extract it as a set of points with attached class property.
- Test predicted class labels against the validation set to determine how many original points in each class and overall were adequately classified, a measure also known as the recall rate.

Binary Reconstruction

The training set was composed of the Esker/Abitibi model points located on 11 arbitrarily chosen parallel sections oriented along axis X. Points were grouped into two classes as shown in Table 1. The input data statistics are given in Column 4 of Table 1. As seen from the table, the training set was dominated by points representative of Class 2, which combined all model units except the Esker Unit. The validation set was composed of all the remaining model points (not included in the training set). The number of points used for validation in each of the two classes is shown in Column 1a of Table 3.

Table 1. Geological units, SVM classes, and training set statistics for Esker/Abitibi Binary Model. Total number of points to be classified is 371783.

1. Geological Unit	2. SVM Class	3. All Points (#/%)[b]	4. Training Points (#/%)[c]
Esker	1	20300/5.22	995/0.26
Non-Esker[a]	2	368935/94.78	16457/4.23
All Units	-	389235/100	17452/4.48

[a] Non-Esker unit included Rock, Till, Clay, Littoral and Organic units
[b] Number of all class points and their percentage of all model points
[c] Number of training points per class and their percentage of all model points

Multi-Class Reconstruction

The same training points, arranged into six classes corresponding to the six geological units found in the original model, were used to test the SVM capabilities in multi-class classification (Table 2). For training data statistics, see Column 4 of Table 2. This time, bedrock (Class 6) entirely dominated the training set, with organics (Class 1) being the least representative. The validation set also contained information about all six geological units as shown in Column 2a of Table 3.

Hyperparameters Sensitivity Tests and Multiple Parameter-Set Reconstructions

To analyze the sensitivity of prediction results to the values of hyperparameters, C and γ, we used a simple grid search procedure as proposed in Hsu et al. (2004). The grid search was run for the above training set configurations, and the range of parameters scanned by every search was from 2^{-8} to 2^{15} for C and from 2^{-15} to 2^{12} for γ incrementing parameter values by a power of 2. As in previous experiments, the success rate was determined through direct comparison with the validation set extracted from the reference model. We also examined the dependency of success rate on parameter values for the class with the least number of training points (Class 1–Organics). Finally, binary models were built with combinations of parameters drawn from the margins of the reasonable working range. These included low C (2^{-2}) – high γ (2^{8}), low C (2^{-2}) – low γ (2^{5}), average C (2^{7}) – average γ (2^{6}), high C (2^{14}) – high γ (2^{8}) and high C (2^{14}) – low γ (2^{5}).

RESULTS AND DISCUSSION

Binary Reconstruction

The original esker body, training sections, and the results of binary reconstruction with the reference parameter set are shown in Figure 2. Column 1b of Table 3 describes

Table 2. Geological units, SVM classes, and training set statistics for Esker/Abitibi Multi-Class Model. Total number of points to be classified is 371783.

1. Geological Unit	2. SVM Class	3. All Points (#/%)[a]	4. Training Points (#/%)[b]
Organics	1	1210/0.31	48/0.01
Littoral	2	3819/0.97	193/0.05
Clay	3	13295/3.42	628/0.16
Esker	4	20300/5.22	995/0.26
Till	5	15865/4.08	747/0.19
Bedrock	6	334746/86.00	14841/3.81
All Units	-	389235/100	17452/4.48

[a] Number of class points and their percentage of all model points
[b] Number of training points per class and their percentage of all model points

Table 3. Validation data and results for Esker/Abitibi binary and multi-class model. Number of validation points in original model and percentage of points properly classified by SVM.

SVM Class	1. Binary		2. Multi-Class	
	a. Validation Points (#/%)[a]	b. Success (%)	a. Validation Points (#/%)[a]	b. Success (%)
1	19305/5.19	71.50	1162/0.31	18.76
2	352478/94.81	98.87	3626/0.98	37.20
3	-	-	12667/3.41	57.10
4	-	-	19305/5.19	67.65
5	-	-	15118/4.07	45.72
6	-	-	319905/86.05	95.45
All	371783/100	97.34	371783/100	89.87

[a] Number of validation points per class and their percentage of all model points

the validation results. As seen from Table 3, the success rate of SVM prediction is exceedingly high. Especially remarkable results, 98.87%, are achieved in Class 2. In part, this can be attributed to the fact that points of this class entirely dominate the training set. When the SVM cannot classify a point in binary classification, it tends to attribute it to the predominant class. Considering that bedrock points constitute 94.81% of all points that need to be classified (352478 of 371783 as shown in Column 1 of Table 3), it is no surprise that the overall success of prediction achieves 97.34%.

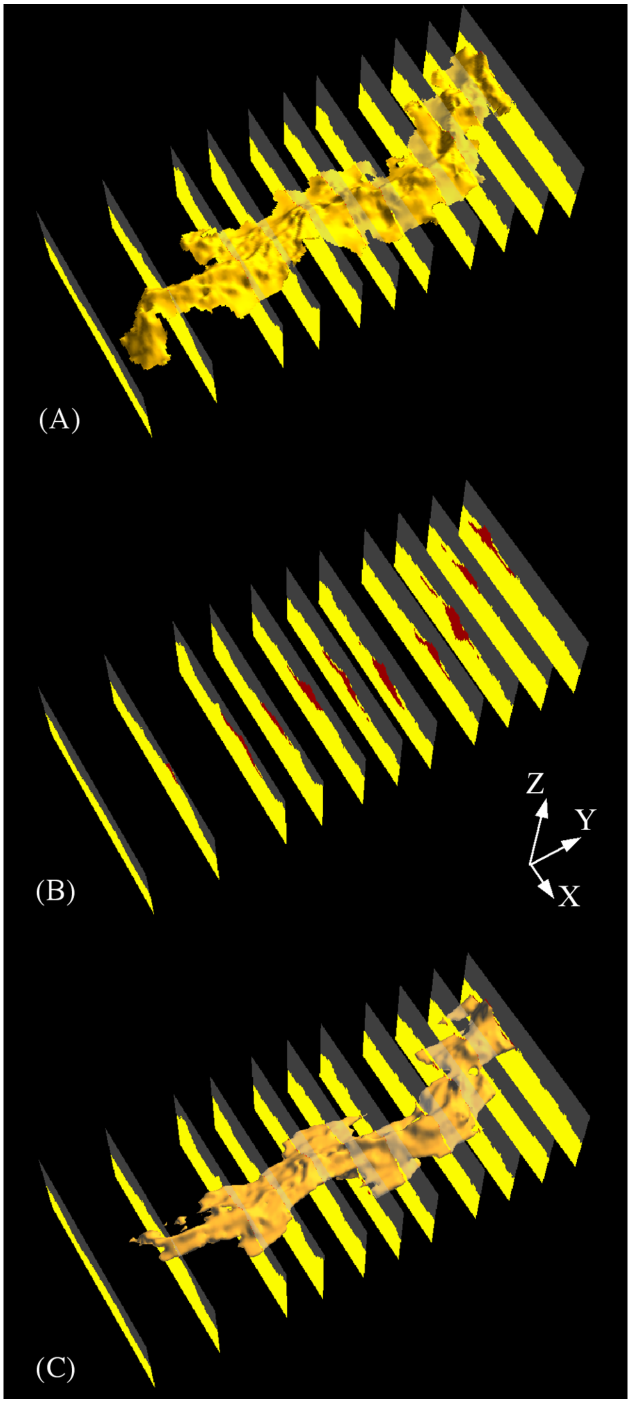

Figure 2. Binary esker reconstruction. (A) Original Esker Unit; (B) training set; (C) reconstructed Esker Unit.

With the above explanation in mind, the classification success in Class 1, which represents only about 6% of the training set, is still as high as 71.50%. This, in our opinion, proves that the SVM can be effectively used for binary (single-unit) reconstructions even with training sets substantially skewed toward one of the classes.

We further analyzed success rate in Class 1 on all model sections where the Esker Unit was present (234 sections). The results are presented in Figure 3. The figure clearly demonstrates that the success of prediction decreases as the distance from a training section increases. As training section # 1 did not intersect the esker body, the success rate on the first 18 sections drops to 0%. Therefore, as could be expected, the overall reliability of prediction is directly proportional to the density of sections with training data.

Multi-Class Reconstruction

The results of this experiment are found in Figure 4 and Column 2b of Table 3. The overall success score is 89.87%, which is mainly controlled by the predominant bedrock class (Class 6). Two other classes, esker and marine clay, demonstrate success rates over 50%. These units are somewhat better represented in the SVM training set than the remaining classes. Figure 5 shows that the success of reconstruction for a particular class in this experiment is almost directly proportional to the number of those class points in the training set, exceeding 75% when the number of training points exceeds 1% of the total.

We also compared area and volume of geological units in the original model and its reconstructed counterpart. The results presented in Table 4 and Figure 6 show that, for both area and volume calculations, the reconstructed and original values for any unit are the same order of magnitude. This suggests that along with single unit modeling, the SVM can efficiently be applied in multi-unit volumetric reconstructions.

Hyperparameter Sensitivity Tests and Multiple Parameter-Set Reconstructions

Figure 7 summarizes the results of grid search for the best pair of hyperparameters in binary and multi-class reconstructions. The best overall success rates, 97.79% and 92.03%, were achieved at [$C=2^1$, $\gamma=2^6$] and [$C=2^{-1}$, $\gamma=2^6$], respectively. The analysis of success rates in the class with the least number of training points yielded 77.38% and 22.46% at [$C=2^2$, $\gamma=2^6$] and [$C=2^9$, $\gamma=2^5$], correspondingly. As seen from the results, parameters are fairly stable, and a single range for C [2^{-3} - 2^{15}] and γ [2^4 -

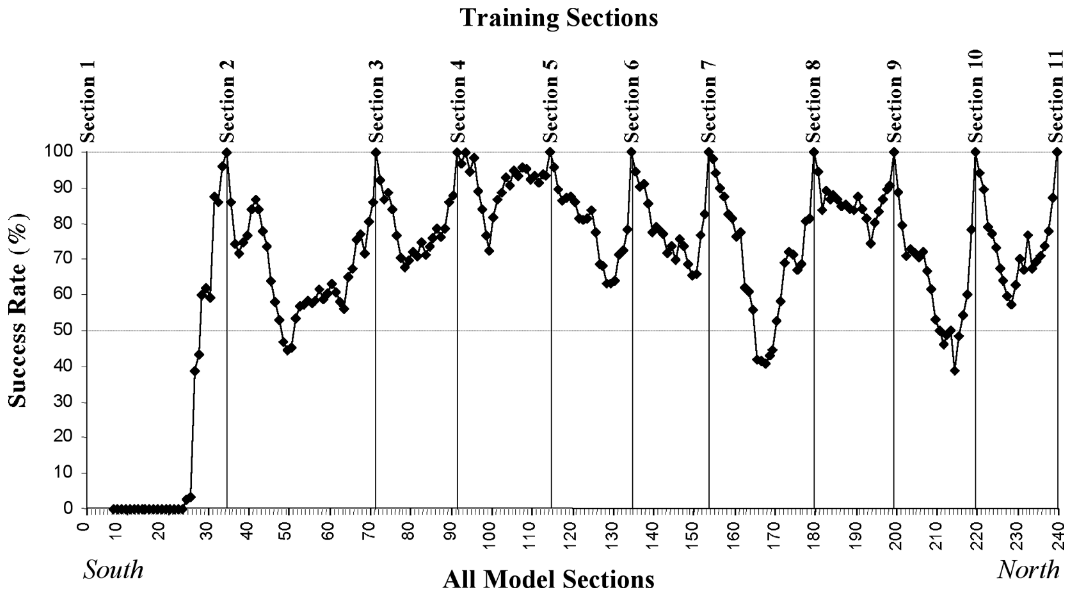

Figure 3. Validation results for binary esker reconstruction from 11 parallel sections. Success in Class 1 (Esker) against section number. Vertical lines indicate training sections. Total length of horizontal axis is 24km and sections are spaced at 100m.

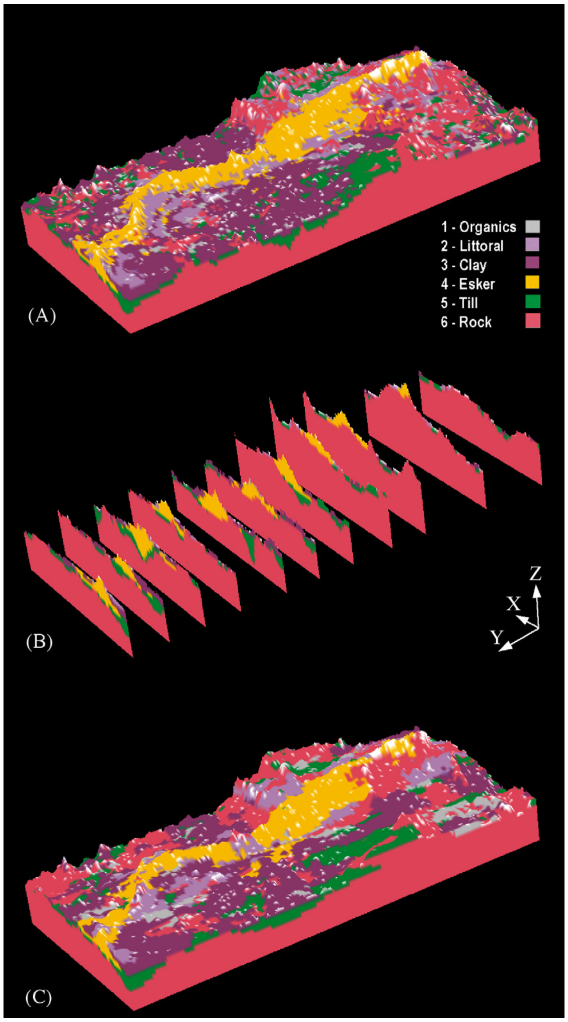

Figure 4. Multi-class esker reconstruction from 11 parallel sections. (A) Original model; (B) training set; (C) reconstructed model.

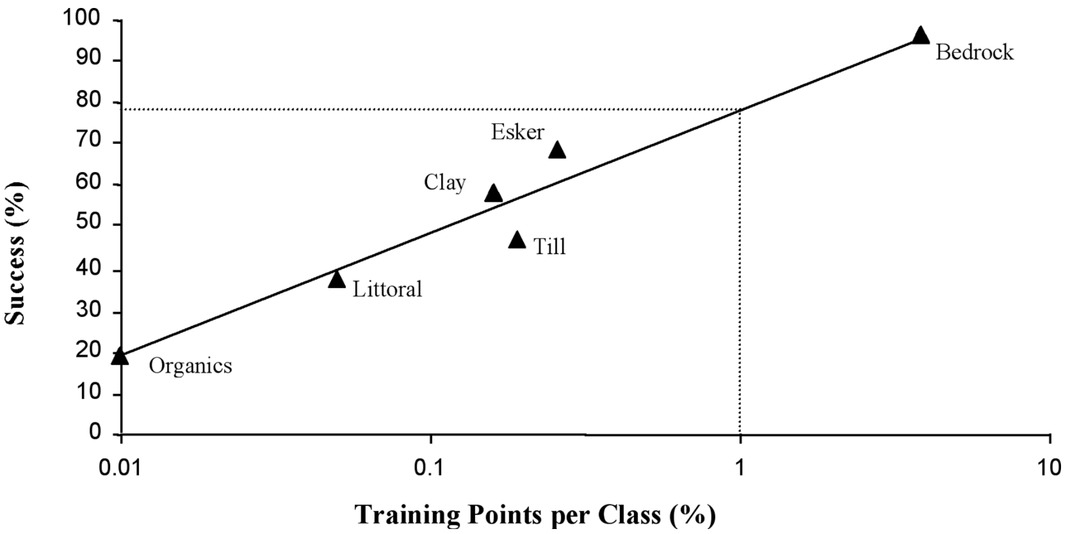

Figure 5. Validation results for multi-class esker reconstruction from 11 parallel sections. Success per class vs. number of training points per class.

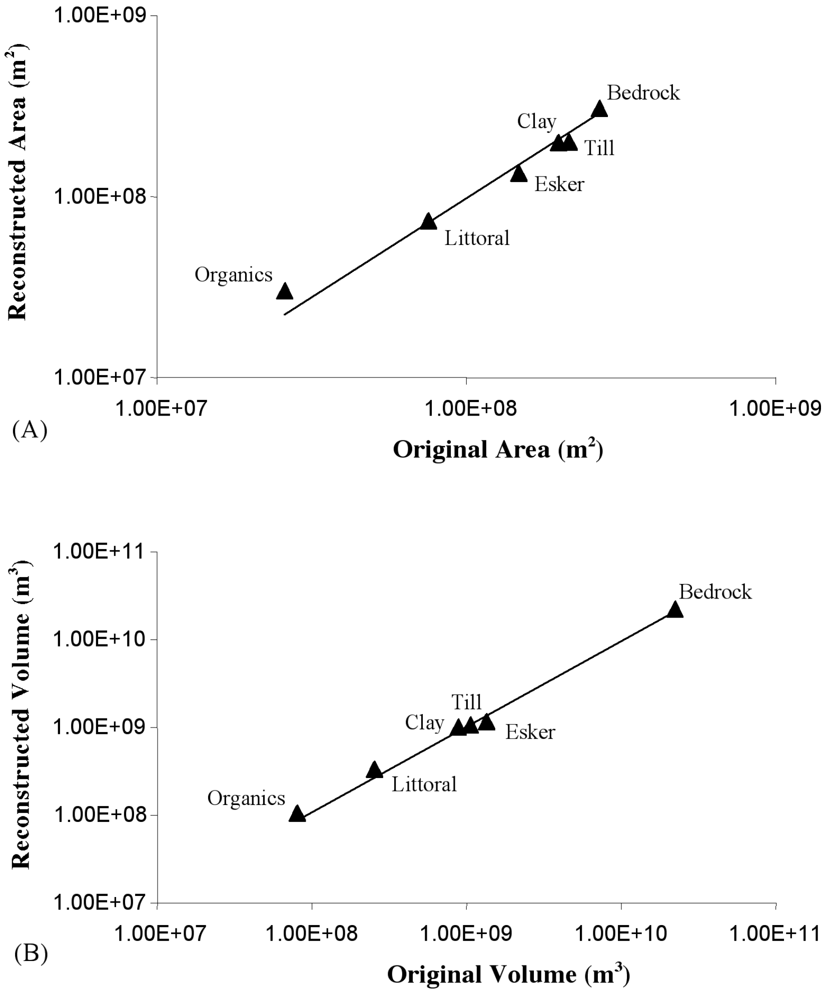

Figure 6. Surface area and volume comparison for original (reference model) and reconstructed geological units. (A) Surface area; (B) volume.

Table 4. Results of SVM reconstruction from 11 parallel sections for Esker/Abitibi. Unit area and volume comparison (original model vs. reconstructed). See Table 2 for training set statistics.

Geological Unit	1. Area (m^2)		2. Volume (m^3)	
	a. Original	b. Reconstructed	a. Original	b. Reconstructed
Organics	2.55E+07	2.99E+07	8.06E+07	1.05E+08
Littoral	7.56E+07	7.34E+07	2.54E+08	3.30E+08
Clay	1.99E+08	1.97E+08	8.86E+08	1.01E+09
Esker	1.48E+08	1.33E+08	1.35E+09	1.16E+09
Till	2.15E+08	1.99E+08	1.06E+09	1.07E+09
Bedrock	2.70E+08	3.06E+08	2.23E+10	2.23E+10

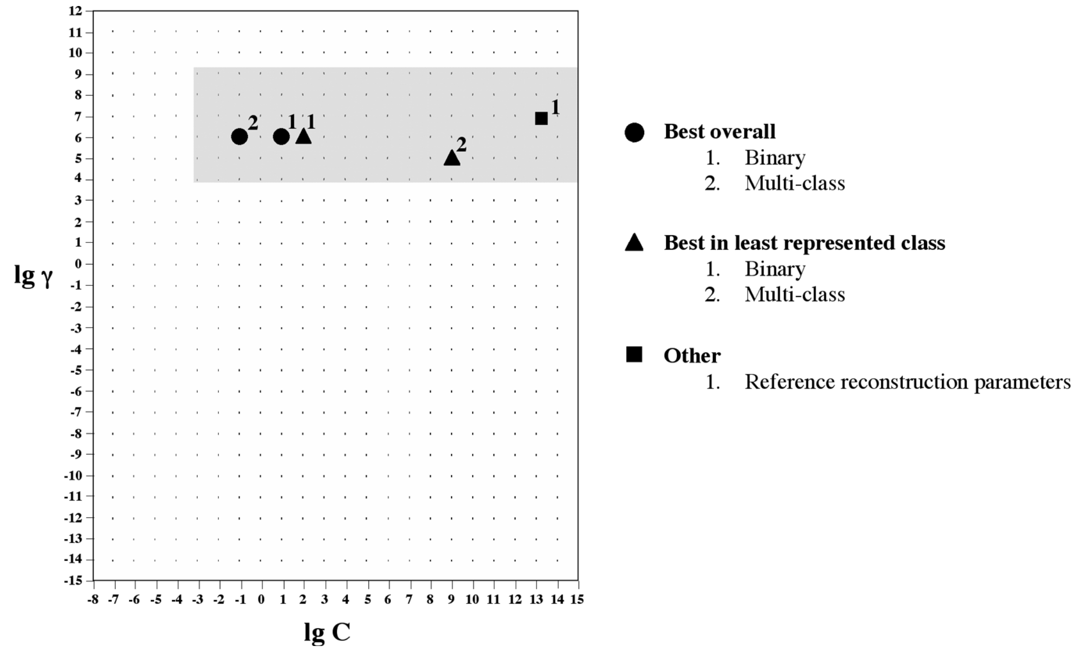

Figure 7. Summary of best results from parameter sensitivity tests for binary and multi-class reconstructions and proposed range for RBF kernel parameters (C [2^{-3}, 2^{15}] and γ [2^4, 2^9]).

2^9] can be recommended. Within this range, higher overall scores and higher scores for over-represented classes are achieved at somewhat lower C values. On the other hand, proper classification of points in the least represented classes requires higher C values. Visual examination of binary models built with combinations of parameters drawn from different corners of the above range also show that a more generalized picture can be achieved at lower C’s (Figure 8a, 8b) while higher values of this parameter promote more detailed interpretation (Figure 8d, 8e). Average C values result in well-balanced models (Figure 8c). The influence of the second parameter (γ) is not as obvious.

CONCLUSIONS

Our experiments clearly showed that the SVM with RBF kernel can be efficiently used for both single- and multi-unit 3D reconstructions. The procedure is performed in a single step, which eliminates the need for unit-by-unit interpolation. Even from a limited training set (e.g., several cross-sections sparsely distributed across the study area) reasonable reconstruction results can be achieved.

It is important, however, that all classes to be reconstructed are reasonably represented in the training set. In the multi-class case, the reconstruction success was

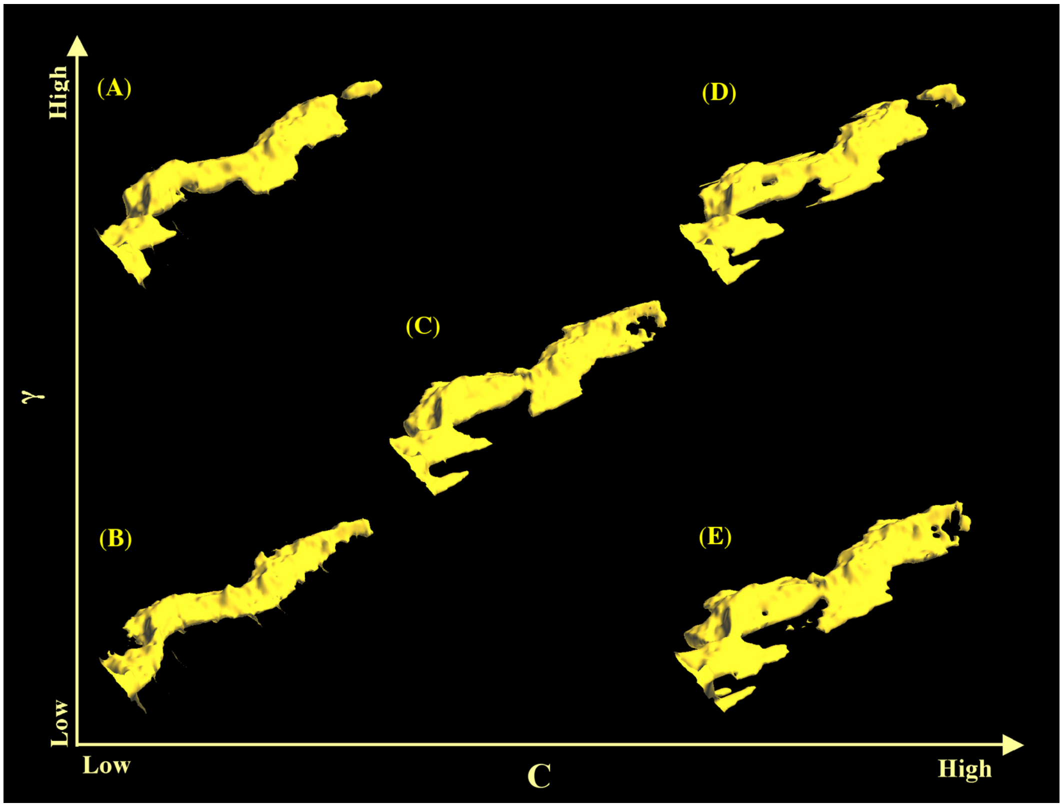

Figure 8. Binary reconstructions with parameters drawn from different corners of the range presented in Figure 7. (A) low C (2^{-2}) – high γ (2^8); (B) low C (2^{-2}) – low γ (2^5); (C) average C (2^7) – average γ (2^6); (D) high C (2^{14}) – high γ (2^8); (E) high C (2^{14}) – low γ (2^5).

shown to be directly proportional to the number of unit samples in the training data. The reliability of prediction is greater in the vicinity of the training data, and therefore, the density of training sections and spatial continuity of lithological units may directly affect the reconstruction results.

The kernel parameters should be chosen from the range 2^{-3}- 2^{15} for C and 2^4- 2^9 for γ. When more model details are required or classes with a small number of training points are involved, higher C values should be considered. Lower C values result in more generalized models with fewer details. This favors classes that dominate the training set.

Finally, our results indicate that when appropriate parameters are chosen, not only the general shape of a geological body, but also such characteristics as its surface area and volume can be reconstructed with results close to those obtained from the application of classical GIS methods.

ACKNOWLEDGEMENTS

The research was conducted at the Quebec Division of the Geological Survey of Canada (GSC), and was funded by the Groundwater Program of the Earth Sciences Sector of Natural Resources Canada (contribution number 20060335). We are especially grateful to Andrée Bolduc (GSC) for her involvement in this work and for having made the Gocad® files with initial data and 3D Esker/Abitibi model used in experiments readily available.

REFERENCES

Abe, S., 2005, Support Vector Machines for Pattern Classification: Springer-Verlag, London, 343 p.

Bolduc A.M., Paradis S.J., Riverin M.N., Lefebvre R., and Michaud Y., 2005, A 3D esker geomodel for groundwater research: The case of the Saint-Mathieu–Berry esker, Abitibi, Quebec, Canada, in Russell, H., Berg, R.C., and Thorleifson, L.H. eds., Three-Dimensional Geological Mapping for Groundwater Applications, Geological Survey of Canada, Ottawa, Ontario, Open File 5048, p.17-20.

Chang, C-C., and Lin, C-J., 2001, LIBSVM: a Library for Support Vector Machines, accessed at http://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf.

Cristianini, N., and Shawe-Taylor, J., 2000, Support Vector Machines: Cambridge University Press, 189 p.

El-Naqa, I., Yang, Y., Wernik, M.N., Galatsanos, N.P., and Nishikawa, R., 2002, Support vector machine learning for the detection of microcalcifications in mammograms: IEEE Transactions on Medical Imaging 21, p. 1552-1563.

Gilardi, N., Kanevski, M., Maignan, M., and Mayoraz, E., 1999, Environmental and Pollution Spatial Data Classification with Support Vector Machines and Geostatistics: Workshop W07 “Intellegent techniques for Spatio-Temporal Data Analysis in Environmental Applications”, ACAI99, July, Greece, p. 43-51.

Hsu, C-W., Chang, C-C., and Lin, C-J., 2004, A Practical Guide to Support Vector Classification, accessed at http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.

Hsu, C-W., and Lin, C-J., 2002, A comparison of methods for multi-class support vector machines: IEEE Transactions on Neural Networks 13(2), p. 415-425.

Mallet, J-L., 1989, Discrete Smooth Interpolation: ACM Transactions on Graphics 8(2), p. 121-144.

Noble, W.S, Kuehn, S., Thurman, R., Yu, M., and Stamatoyannopoulos, J., 2005, Predicting the in vivo signature of human gene regulatory sequences: Bioinformatics 21(1), p. 338-343.

Smirnoff, A., Boisvert, E., and Paradis, S.J., 2008, Support Vector Machine for 3D modelling from sparse geological information of various origins: Computers and Geosciences 34, p. 127-143.

Vapnik, V., 1995, The Nature of Statistical Learning Theory: Springer-Verlag, New York, 311 p.

Yu, X., Liong., S-Y., and Babovic, V., 2004, EC-SVM approach for real-time hydrologic forecasting: Journal of Hydroinformatics 06(3), p. 209-223.