University of Alberta

Support Vector Classification for Geostatistical Modeling of Categorical Variables

by

Enrique Carlos Gallardo Vizcaino

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Science
in
Mining Engineering

Department of Civil and Environmental Engineering

Fall 2009
Edmonton, Alberta

Examining Committee

Dr. Oy Leuangthong (Supervisor), Civil and Environmental Engineering

Dr. Jozef Szymanski (Chair and Examiner), Civil and Environmental Engineering

Dr. Nilanjan Ray (Examiner), Computing Science

To my wife Sonia and
my sons Enrique and Esteban.

I am glad that my life is not mine anymore,
it belongs to all three of you

Abstract

Subsurface geological characterization often requires solving a classification problem to obtain a model of facies that is later populated with continuous properties like porosity or permeability. The classification problem, which consists of assigning a single category to any unsampled location based on observed data, is analyzed and solved in this thesis using geostatistical and machine learning tools.

This research proposes an easy-to-implement heuristic technique that uses geostatistical criteria, such as correct classification of the observed data and good reproduction of the global proportions of categories, to obtain from the SVC algorithm a boundary classifier. This boundary is used to generate the facies model.

The case studies show that the implementation of the proposed technique is highly automatic. The responses are comparable in terms of prediction accuracy to those obtained by the conventional geostatistical approach. They also show how simple information from SVC allows for an improvement in the response of conventional geostatistical indicator simulation models.

Acknowledgements

I want to thank my supervisor Dr. Oy Leuangthong. Her guidance and support have been an essential component of my academic experience in the University of Alberta. I deeply appreciate the time she spend reviewing my research or meeting me to discuss it.

Discussions with my colleagues at the Centre for Computational Geostatistics (CCG), especially Miguel Cuba, have enriched this research. Thank you my friends.

The Colombian Oil Petroleum Company - ECOPETROL S.A. provided financial support for my studies in Canada.

Table of Contents

List of Tables

List of Figures

List of Symbols

$\mathbf{u}$	Location in space
$\mathbf{h}$	Distance or lag vector
$S(\mathbf{u})$	Categorical random variable S at location $\mathbf{u}$
$s_1, s_2 \ldots, s_K$	K facies or categories
$I(\mathbf{u}; k)$	Indicator random variable at location $\mathbf{u}$
$E\{I(\mathbf{u}; k)\}$	Expected value of $I(\mathbf{u}; k)$
$[I(\mathbf{u}; k)]^*$	Estimate of $I(\mathbf{u}; k)$
$i(\mathbf{u})$	Indicator code at location $\mathbf{u}$
λ_α	Kriging weight assigned to sample α
$C_I(\mathbf{h})$	Indicator covariance function at lag $\mathbf{h}$
$\gamma_I(\mathbf{h})$:	Indicator semivariogram function at lag $\mathbf{h}$
$\mathbf{w}$	SVM vector of weights
$\|\mathbf{w}\|$	Euclidean norm of the vector $\mathbf{w}$
b	SVM bias term
g	Measure of the complexity of a function
η_α	Lagrange multiplier assigned to sample α in SVC
ξ_α	Slack variable assigned to sample α in SVC
Nsv	Number of support vectors
P	SVC penalty parameter
γ	Parameter of the Grbf kernel
$\mathcal{F}$	High dimensional space
Φ	Mapping function
$k(\mathbf{u}, \mathbf{u}')$	Kernel function between locations $\mathbf{u}$ and $\mathbf{u}'$
$K(\mathbf{u}, \mathbf{u}')$	Symmetric matrix with the values of the kernel function
Grbf	Gaussian radial basis function kernel
IK	Indicator kriging
OIK	Ordinary indicator kriging
SIK	Simple indicator kriging
SIS	Sequential indicator simulation

SRM	Structural risk minimization
SVC	Support vector classification
SVR	Support vector regression
SVM	Support vector machine

1 Introduction

Investment decisions in the petroleum and mining industries rely on resource characterization of sites of economic interest based on observed data. In the petroleum industry, it is necessary to predict the prevalence of net and non-net facies (e.g. sand or shale) at locations to be drilled. Similarly, in the mining industry, the classification of ore or waste is determined before mining an area. In the simplest case, resource characterization can be broken into three distinct phases: geology modeling, petrophysics or grade modeling, and decision modeling.

In many cases, the geology model is first constructed. This can be accomplished using expert geological information to manually draw or digitize the geologist's vision of the subsurface, or it can be done using cell-based, object-based or multiple point statistics modeling approaches.

Once the geology model is constructed, continuous properties are then populated in this same model. In the case of a reservoir, these properties are petrophysical in nature such as porosity and permeability. In mining, these properties are usually metal and/or mineral grades. The values associated to any one location are highly dependent on the corresponding geology.

Finally, the resource model is processed through some type of response function for decision making. For a petroleum reservoir, the flow characteristics are of primary importance to production so the response function is usually a flow simulation; while, in the mining industry, production is dependent on the classification of ore and waste which is based on economics of mining ore and moving waste.

In the first phase of modeling, the type of modeling is considered to be discrete or categorical modeling, while the second phase is clearly the construction of a continuous model. Depending on the field of application, the third phase can result in continuous and/or discrete models. This thesis is concerned with categorical modeling, in particular

the modeling of facies or rock types. It considers both geostatistical and machine learning tools to solve the problem of assigning a single facies/rock type to a location.

Characterizing a site is not only the prediction of the value of geological variables at multiple locations; it also includes the quantification of the uncertainty associated to those predictions. This thesis shows that conventional geostatistical models that take information from a machine learning algorithm may improve uncertainty quantification.

1.1 Problem Setting

This thesis is aimed at the analysis of the classification problem, which is defined as the problem of assigning a single category to an unsampled location based on a limited set of observed data. Consider the case of two facies shown in Figure 1.1. A single facies, white for sand or black for shale, must be assigned to the unsampled location **u** (**u** represents a vector of coordinates) based on 10 observed data. Figure 1.1 shows how information about the facies at specific coordinates is collected by drilling wells.

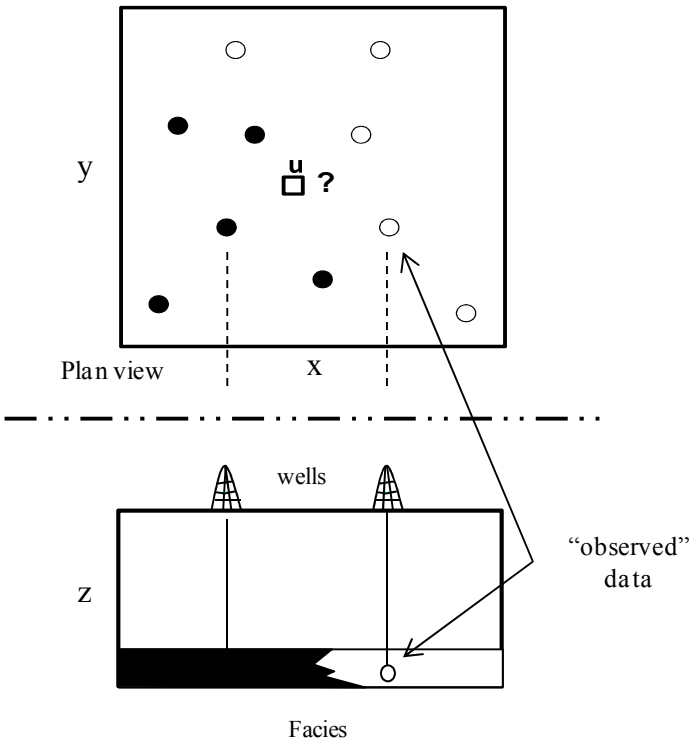

Figure 1.1: Classification problem. A single facies, white for sand or black for shale, must be assigned to the unsampled location **u** based on 10 observed data.

Several approaches can be applied to solve this problem, Figure 1.2 illustrates a method in which a selected subset of neighbouring data is linearly combined to estimate the facies at the location **u**. Seeing that the result of combining the observed data is neither white nor black, a grey scale rule is used to assign a single facies to the unsampled location **u**.

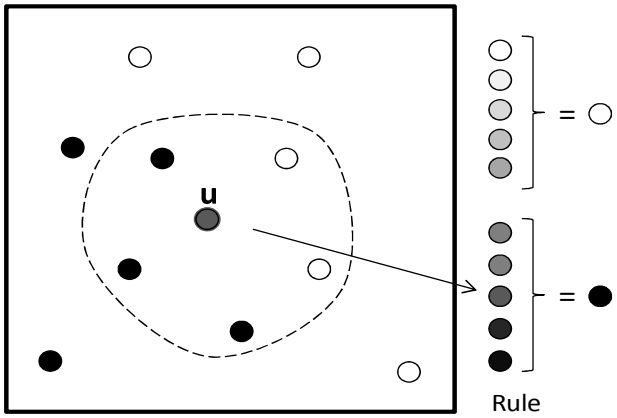

Figure 1.2: A linear combination of neighbouring data plus a rule is used to assign a single facies to the unsampled location **u**.

Figure 1.3 shows a method in which the observed data is used to generate a boundary that separates sand (white) and shale (black) data. The boundary is then used as a classifier to assign a single facies to the unsampled location **u**.

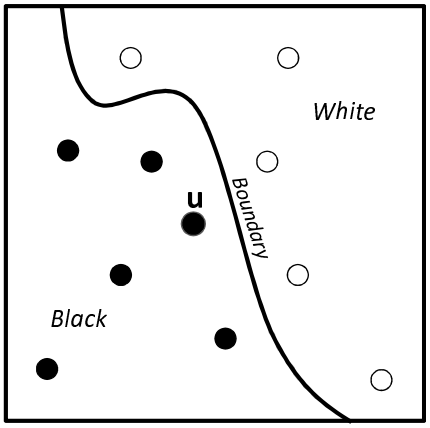

Figure 1.3: A boundary classifier is used to assign a single facies to the unsampled location **u**.

The first approach presented belongs to the geostatistics field; the weights to linearly combine the data are calculated by kriging and a classification rule is defined based on conditional probabilities. The second approach belongs to the machine learning field; the parameters that define the boundary classifier are calculated using the support vector classification (SVC) algorithm.

Although the geostatistical approach to solve the classification problem is mainly data driven, the results are by no means objective; that is, the subjectivity of the modeler greatly influences the algorithmic responses. To put it in another way, the responses from kriging are not only influenced by the quantity and quality of the observed data, but also by the expertise of the modeler. Furthermore, the ability of the modeler to make appropriate geostatistical modeling decision might significantly impact the time to obtain a response. On the other hand, the machine learning approach, specifically, the SVC algorithm has the ability to generate almost fully data driven responses with limited influence of the modeler. Contrary to the geostatistical approach, the time to get a response from the SVC algorithm is little influenced by the modeler because it is mostly computational time.

The use of support vector machines (SVM) to model spatially distributed data is not new, but the studies published have reached diverse conclusions. For instance, Wohlberg, Tartakovky and Guadagnini (2006) use SVC for lithofacies delineation and concluded that SVC slightly outperform geostatistic in reconstructing the boundaries between two geological facies. Kanevski et al. (2001) compare SVM to geostatistical approaches to analyze continuous and categorical reservoir data. They concluded that the responses of both approaches are quite good. Kanevsky, Wong and Canu (2000) explore the use of a hybrid machine learning-geostatistical method for the analysis of pollution data. They reported the results as promising. In contrast, Gilardi and Bengio (2005) discuss the use of machine learning for environmental data and concluded that kriging is more reliable than SVM to estimate continuous data. Gilardi and Dubois (2000) also using environmental data reported that the results obtained using SVM are worse than those obtained using kriging.

Unlike most of the literature reviewed, this research does not try to demonstrate the superiority of one approach over the other. The basic idea is that under certain conditions, the solution to the classification problem from both approaches should tend to converge to a unique response that balances the flexibility of kriging to incorporate subjective prior

knowledge with the less subjective data-driven response of the SVC algorithm. If the responses tend to converge, then it should be possible to use valuable information obtained from the SVC model to improve the response of conventional geostatistical models (or hybrid geostatistical-machine learning models).

1.2 Objectives and Scope

The goals of this thesis are: (1) to analyze geostatistical and machine learning (SVC) algorithms to solve the classification problem, (2) to discuss the practical application of those algorithms and their responses, and (3) to develop new concepts to solve the classification problem.

To achieve the above goals, the following objectives are set: (1) to translate the SVC algorithm to classical geostatistical notation, (2) to prepare a synthetic classification problem that allows the assessment and comparison of the responses of the different approaches, and (3) to propose a new algorithm to solve the classification problem using a hybrid of geostatistical and machine learning tools.

The scope of this thesis is limited to the analysis of binary classification problems in two dimensions. Geostatistical simulation is addressed using cell-based techniques.

1.3 Proposed Approach

This thesis proposes an easy-to-implement technique to solve the problem of facies classification. The proposed technique obtains from the SVC algorithm a boundary classifier with good prediction property, correct classification of the sampled data set and good reproduction of the global proportions of facies. The novel technique adopts a heuristic approach to select the parameters required to implement SVC. It is based on the performance of the boundary classifier on the observed data and the unsampled locations.

The practical implementations show how information obtained from SVC can be used to construct a conventional (hybrid) geostatistical model that better classifies unsampled locations compared to the model built without the SVC information. They also show that a set of simulated realizations generated with the model, using information from the SVC response, reproduces better the semivariogram and the global facies proportions. Further,

the simulated realizations yield better prediction properties compared to a set of realizations that does not use the SVC information.

GSLIB (Deutsch and Journel, 1998) and LIBSVM (Chang and Lin, 2001) are used to perform the classification and simulation tasks required for this research. MATLAB code was written to implement the proposed technique and to manipulate, visualize and analyze the information.

1.4 Outline of the Thesis

Chapter 2 contains a literature review of geostatistical and machine learning concepts. The geostatistical section includes the description of fundamental concepts as random variables and random functions, and the algorithms for modeling categorical variables such as indicator kriging and sequential indicator simulation. The machine learning section presents the concepts of statistical learning theory and the structural risk minimization principle. The novelty is the translation of the description of the SVC algorithm to classical geostatistical notation.

Chapter 3 introduces a heuristic technique to select the parameters of a SVC machine. A practical implementation example is presented using a synthetic reference data and a large sampled data set. Results are analyzed and issues discussed.

Chapter 4 presents another application of the proposed technique but using a small data set sampled from the same reference data. It also shows how information obtained from the SVC response is used to generate conventional geostatistical classification and simulation models with improved responses.

Chapter 5 recapitulates the conclusions of the thesis and provides other ideas to complement and/or to expand the presented research.

Appendix A contains the MATLAB code written to develop this research.

2 Fundamental Concepts

The technique proposed in this thesis is based on well known algorithms within geostatistics and machine learning. Specifically, kriging and simulation algorithms for categorical data on the geostatistics side, and the support vector classification (SVC) algorithm on the machine learning side.

Only fundamental concepts needed to describe the aforementioned algorithms will be presented. The reader interested in geostatistics is referred to the introductory books of Isaaks and Srivastava (1989) and Goovaerts (1997). Chiles and Delfiner (1999) and Journel and Huijbregts (1978) offer a more theoretical treatment of the subject, while Deutsch (2002) offers a more practical approach. On machine learning, essential references are Vapnik (1998, 1999). A good introductory book is Hastie, Tibshirani and Friedman (2001). Burges (1998) and Kecman (2001) contain an excellent description of the SVM algorithms and its variants.

2.1 Geostatistics

Geostatistics refers to a field of applied statistics, wherein a set of spatial statistical tools are used to make inferences, usually related to naturally occurring phenomena. While it originated in the mining industry (Krige, 1951; Matheron, 1970), it has since found use in other natural resource sectors including the petroleum industry, environmental, forestry and agriculture.

There are two main goals in geostatistics: (1) the estimation of the value of spatial variables at unsampled locations over an area of interest, and (2) the modeling of the uncertainty about these inferred values. The former goal can be obtained by kriging, while the latter requires a (sequential) simulation approach.

Before describing the relevant geostatistical algorithms needed for this research, the foundation concept of random functions is introduced. The material presented in sections

2.1.1 and 2.1.2 follows Goovaerts (1997). Note that all the equations presented in these sections are provided only for categorical data.

2.1.1 Random Functions

The random function (RF) concept allows the interpretation of a set of unknown values as a set of spatially dependent random variables (RV).

"A RV is a variable that can take a series of outcome values according to some probability distribution" (Goovaerts, 1997, p. 63). By convention, a categorical RV at a location **u** that can take any one of *K* mutually exclusive discrete and non-ordered values is denoted by $S(\mathbf{u})$ and its outcome by s_k.

$S(\mathbf{u})$ is characterized by its conditional cumulative distribution function (ccdf) which provides for any ordering of the *K* outcomes s_k, "the probability for any one of the categories $s_{k'}$ ordered lesser or equal to s_k to prevail at **u** "(Goovaerts, 1997, p. 64). conditional to *n* surrounding data:

$$F(\mathbf{u}; s_k \mid (n)) = \sum_{k'=1}^{k} p(\mathbf{u}; s_{k'} \mid (n)) = Prob\left\{\bigcup_{k'=1}^{k} S(\mathbf{u}) = s_{k'} \mid (n)\right\} \tag{2.1}$$

where the probability for the category s_k to prevail at **u** conditional to *n* surrounding data was defined as:

$$p(\mathbf{u}; s_k \mid (n)) = Prob\{S(\mathbf{u}) = s_k \mid (n)\} \tag{2.2}$$

with the condition that the *K* probabilities $p(\mathbf{u}; s_k|(n))$ must be in the range [0,1] and sum to one.

A RF is a set of RV in a finite domain *A* characterized by a multivariate ccdf. The multivariate ccdf of $S(\mathbf{u})$ represents the spatial law of the RF and is denoted by:

$$\begin{aligned} &F(\mathbf{u}_1,...,\mathbf{u}_N; s_{k_1} \mid (n),...,s_{k_N} \mid (n)) = \\ &Prob\left\{\bigcup_{k'=1}^{k} S(\mathbf{u}_1) = s_{k'_N} \mid (n),...,\bigcup_{k'=1}^{k} S(\mathbf{u}_N = s_{k'_N} \mid (n)\right\} \end{aligned} \tag{2.3}$$

where N represent the total number of locations.

Indicator random variable I(u;k): To make possible the implementation of geostatistical algorithms that perform mathematical operations, the categorical RV $S(\mathbf{u})$ (e.g. sand and shale) is transformed to a binary indicator RV (e.g. sand = 1 and shale = 0) by:

$$I(\mathbf{u};k)=\begin{cases}1 & if\ S(\mathbf{u})=s_k \\ 0 & otherwise\end{cases} \qquad k=1,\ldots,K \tag{2.4}$$

Decision of stationarity. To perform geostatistical predictions, the multivariate ccdf is assumed to be invariant under translation within the domain A. Often this requirement is too strict and only second order stationarity is assumed; this decision of stationarity implies that the expected value is independent of the locations and the indicator covariance $C_I(\mathbf{h})$ between two locations only depends on a vector $\mathbf{h}$.

The relationship that exists between the indicator covariance and the indicator semivariogram $\gamma_I(\mathbf{h})$ is used to obtain $C_I(\mathbf{h})$:

$$C_{_I}(\mathbf{h})=C_I(0)-\gamma_I(\mathbf{h}) \tag{2.5}$$

where $C_I(0)$ is the variance of the data. In practice, to assure a licit indicator semivariogram, a mathematical model is adopted based on the experimental indicator semivariogram calculated by:

$$\gamma_I(\mathbf{h};k)=\frac{1}{2N(\mathbf{h})}\sum_{i=1}^{N(\mathbf{h})}\left[I(\mathbf{u}_i;k)-I(\mathbf{u}_i+\mathbf{h};k)\right]^2 \qquad k=1,...,K \tag{2.6}$$

where $N(\mathbf{h})$ is the total number of data pairs separated by the vector $\mathbf{h}$. Description of licit semivariogram models can be found in Deutsch (2002 , p. 133).

2.1.2 Classification of Categorical Variables

Kriging describes a set of linear regression tools used to estimate the value of a variable at any location **u** over an area of interest. As such, the estimator is a linear combination of the data surrounding the location **u**. It takes the form shown on (2.9) for an indicator RV.

Categorical RV cannot be estimated as a linear combination of neighboring data; kriging applied to the indicator RV provides at each location **u** the conditional probability of occurrence of each category. This conditional probability can be used to allocate a single category to the location **u** in two ways: (1) in conjunction with some applied rule or (2) as input for a sequential simulation algorithm as presented in section 2.1.3. To illustrate both ways, consider the problem of estimating the category s at any location **u** over the domain A, based on the following set of n observed data:

$$\{s_k(\mathbf{u}_\alpha); \alpha = 1,\ldots,n; k = 1,\ldots,K\} \tag{2.7}$$

where $k = 1, \ldots, K$ indexes the number of mutually exclusive categories $s_1, \ldots, s_K$.

Using a rule, the problem is addressed in three steps: (1) preprocessing of data, (2) indicator kriging and (3) classification rule.

Preprocessing of data. As explained in (2.4), the categorical RV must be transformed to an indicator RV $I(\mathbf{u}_\alpha; k)$ by:

$$I(\mathbf{u}_\alpha; k) = \begin{cases} 1 & if\ S(\mathbf{u}_\alpha) = s_k \\ 0 & otherwise \end{cases} \qquad k = 1,\ldots,K \tag{2.8}$$

Indicator kriging. Kriging applied to the indicator RV produces the conditional probability of occurrence of each category k at the location **u**. The estimated value of the indicator RV is:

$$\left[I(\mathbf{u};k)\right]^* = \sum_{\alpha=1}^{n} \lambda_\alpha \left[I(\mathbf{u}_\alpha;k) - E\{I(\mathbf{u}_\alpha;k)\}\right] + E\{I(\mathbf{u};k)\} \quad ; \quad k = 1,\ldots,K \tag{2.9}$$

where * denotes an estimate of the indicator RV and λ_α represents the weight assigned to $I(\mathbf{u}_\alpha;k)$. The simple indicator kriging (SIK) estimator is obtained when the expected value $\mathrm{E}\{I(\mathbf{u};k)\}$ is considered constant and known throughout the domain A.

The weights λ_α are determined to minimize the expected value of the error in the estimated under the constraint of unbiasedness of the estimator. The optimization problem:

$$\begin{aligned} &\underset{\lambda}{minimize} \quad E\left\{\left[I(\mathbf{u};k)^* - I(\mathbf{u};k)\right]^2\right\} \\ &subject\ to \quad E\left\{I(\mathbf{u};k)^* - I(\mathbf{u};k)\right\} = 0 \end{aligned} \tag{2.10}$$

leads to the simple kriging system of equations or the normal equations:

$$\sum_{\beta=1}^{n} \lambda_\beta \cdot C_I(\mathbf{u}_\alpha, \mathbf{u}_\beta) = C_I(\mathbf{u}, \mathbf{u}_\alpha) \qquad ; \qquad \alpha, \beta = 1, ..., n \tag{2.11}$$

Other types of kriging are available, most of which are variants of simple kriging involving the imposition of constraints on the system of equations or the relaxation of stationarity of the mean. For instance, ordinary kriging requires the weights to sum to one, and universal kriging imposes a deterministic model of the mean to the system.

An important property of kriging that will be recalled in the next chapters is the exactitude property. It means that kriging as an exact interpolator, reproduces the observed data at their locations.

Classification rule. The most practical and simplest classification rule is to allocate the unsampled locations to the category with the largest conditional probability of occurrence (Goovaerts, 1997, p.356). The rule becomes:

$$s(\mathbf{u}) = s_k \quad if \quad I(\mathbf{u};k)^* > I(\mathbf{u};k')^* \quad \forall k, k' \quad ; \quad k = 1, ..., K \tag{2.12}$$

Note that this rule classifies correctly all the observed data due to the exactitude property of kriging. A different classification procedure can be found in Soares (1992). The alternative to the classification rule is using the kriged conditional probabilities to implement the sequential simulation principle as described below.

2.1.3 Simulation by Sequential Indicator Simulation (SIS)

Consider now the problem of generating multiple realizations of the spatial distribution of $S(\mathbf{u})$ in the gridded domain A, based on the set (2.7).

The cell-based modeling technique known as sequential indicator simulation (SIS) is used to generate the realizations. The SIS algorithm (Journel and Alabert, 1988; Alabert and Massonat, 1990; Deutsch and Journel, 1998) is based on the sequential simulation paradigm (Isaaks, 1990). It requires coding the categorical data as indicator RV and it proceeds as follow:

1. Define a random path visiting each node of the gridded domain A only once
2. At each node **u'**:

    a. Estimate the K conditional probability values $p(\mathbf{u};s_k|(n))$ using indicator kriging. The conditioning data is the set (2.7) and previously simulated values.
    b. Ensure that each conditional probability value is in the range [0, 1], and that their summation adds up to unity.
    c. The K conditional probability values define a probability distribution function for the indicator variable at location **u'**. Draw a simulated value from this function and add it to the conditioning data.

3. Proceed to the next node along the random path and repeat steps above. Visit all the nodes to obtain a complete realization.
4. Set a different random path and repeat steps 2 and 3 to generate a new realization.

Implementation details of the SIS algorithm can be found in Deutsch and Journel (1998).

Other algorithms are available to generate simulated realizations of categorical variables. The cell-based truncated gaussian simulation algorithm (Matheron et al., 1987; H. Beucher et al., 1993; Xu and Journel, 1993) "generates realizations of a continuous gaussian variable and then truncates them at a series of thresholds to create" (Deutsch, 2002, p.204) realizations of the categorical variable. Object-based models that insert geo-objects into a simulated field can be found in Deutsch and Wang (1996), and Holden, Hauge, Skare and Skorstad (1998). More recently, multiple-points statistics (MPS)

methods (Journel and Alabert, 1989; Guardiano and Srivastava, 1993; Strabelle and Journel, 2001) that use spatial moments greater than two have been used to generate simulated realizations.

2.2 Machine Learning

Machine learning refers to a branch of statistics and computer science (Gilardi and Bengio, 2005), whose goal is to learn a dependence between variables using a set of observed data.

Since the first learning machine, the perceptron of Rosenblatt (1962), and the theoretical works of Vapnik and Chervonenkis (1974) about statistical learning theory, there has been a boom of machine learning algorithms, such as: neural networks (NN), support vector machines (SVM), linear discriminant analysis (LDA), etc. The machine learning algorithms have found application in many fields of science and engineering: speech and handwriting recognition, natural language processing, stock market analysis, bioinformatics, etc (Kecman, 2001, p.2-3).

In the context of spatial data, the SVM algorithms allow to estimate the value of continuous (support vector regression - SVR) or categorical variables (support vector classification - SVC) at unsampled locations on areas of interest. The next section describes the SVC algorithm and its fundamental concepts.

2.2.1 Structural Risk Minimization Principle (SRM)

The SVC algorithm implements the structural risk minimization (SRM) principle developed in statistical learning theory. SRM tries to minimize the bound of the generalization error defined as:

$$R(\mathbf{w}) \leq R_{emp}(\mathbf{w}) + R_{conf}(\mathbf{w}, g/n) \tag{2.13}$$

where R is a bound of the generalization error, R_{emp} is the empirical error on the observed data and R_{conf} is a confidence term that depends on the number of observed data (n), the complexity (g) of the modeling function, and the weights $\mathbf{w}$ that define the approximating function.

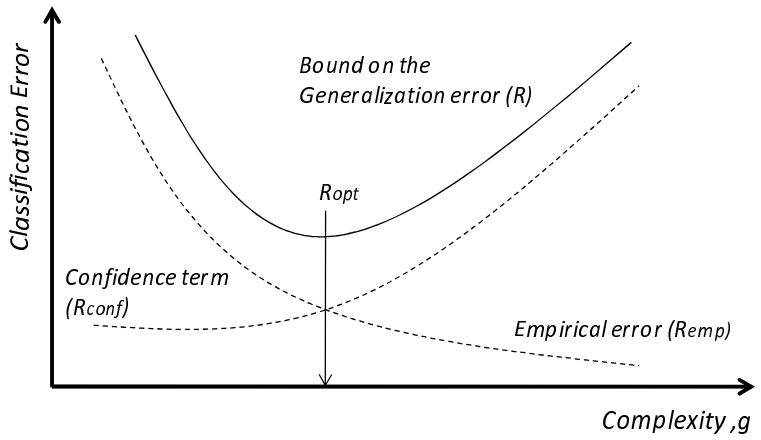

Figure 2.1: Bound of the generalization error. The optimum value corresponds to an appropriate level of empirical error and confidence (Modified redraw from Vapnik 1995, p.96).

Figure 2.1 illustrates how the SRM principle defines a "trade-off between the quality of the approximation of the observed data and the complexity of the approximating function" (Vapnik, 1995, p. 95).

2.2.2 Support Vector Classification (SVC)

The support vector algorithm (Boser, Guyon and Vapnik, 1992) was initially developed for solving classification problems. Soon after, it was extended to deal with regression problems (Muller et al. 1997). The discussion presented in this section follows Kecman (2001, Chapter 2). The SVC algorithm seeks for the weight parameters **w** and the bias term *b* of a decision boundary (a hyperplane) of the form:

$$\mathbf{w}^{\mathrm{T}}\mathbf{u} + b = 0 \tag{2.14}$$

The boundary will separate the categories given on the observed data with a maximum margin as illustrated in Figure 2.2. Vapnik (1999) shows that the SVC implements the SRM principle by controlling the model complexity through the width of the margin. The decision boundary will classify the binary category *s* at unsampled locations **u** according to the rule:

$$s(\mathbf{u}) = \begin{cases} s_1 & if \quad \mathbf{w}^{\mathbf{T}}\mathbf{u} + b > 0 \\ s_2 & if \quad \mathbf{w}^{\mathbf{T}}\mathbf{u} + b < 0 \end{cases} \tag{2.15}$$

To highlight the differences between the geostatistical and the machine learning approaches to classify unsampled locations, SVC is used to solve the problem stated in section 2.1.2. SVC has the following steps: (1) preprocessing of data, (2) SVC training and (3) SVC testing.

Preprocessing of data. Without loss of generality, suppose that the set (2.7) has two categories, s_1 and s_2. To perform SVC the set is coded as:

$$i(\mathbf{u}_\alpha) = \begin{cases} 1 & if \ s(\mathbf{u}_\alpha) = s_1 \\ -1 & \text{if } s(\mathbf{u}_\alpha) = s_2 \end{cases} \tag{2.16}$$

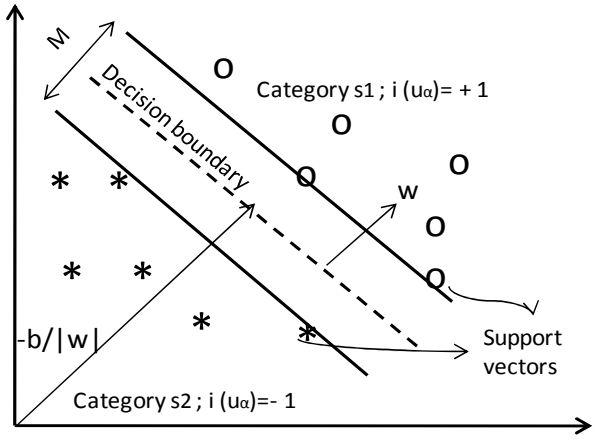

Figure 2.2: Linear separable case. Basic SVC concepts: codification (±1), margin (*M*), weights (**w**) and support vectors.

SVC training. Finding the weighting parameters **w** and the bias term *b* of the decision boundary (2.14) using the observed data is referred to as training the SVC. In machine learning jargon the observed data is called the training set. The percentage of observed data misclassified by the decision boundary is called training error or empirical error. The complement of the empirical error to add to unity is called in this thesis *empirical accuracy*.

The boundary (2.14) is determined to maximize the margin of separation between the categories s_1 and s_2 (Figure 2.2). If the data (2.7) is linearly separable, the optimization problem is expressed as:

$$\begin{aligned} &\underset{\mathbf{w},b}{minimize} \quad \frac{1}{2}\|\mathbf{w}\|^2 \\ &subject\ to \quad i(\mathbf{u}_\alpha)[\mathbf{w}^{\mathrm{T}}\mathbf{u}_\alpha + b] \geq 1 \quad ; \quad \alpha = 1,...,n \end{aligned} \tag{2.17}$$

where $\|\mathbf{w}\|$ represents the Euclidean norm of the vector $\mathbf{w}$. This nonlinear optimization problem with inequality constrains is solved using the Lagrange formalism and leads to the following results for $\mathbf{w}$ and b:

$$\mathbf{w} = \sum_{\alpha=1}^{n} \eta_\alpha i(\mathbf{u}_\alpha)\mathbf{u}_\alpha \tag{2.18}$$

$$b = \frac{1}{Nsv}\left(\sum_{\alpha=1}^{Nsv}\left(\frac{1}{i(\mathbf{u}_\alpha)} - \mathbf{u}_\alpha^{\mathrm{T}}\mathbf{w}\right)\right) \quad ; \quad \alpha = 1,...,Nsv \tag{2.19}$$

where η_α are Lagrange multipliers and Nsv is the numbers of support vectors; that is, training data whose η_α are not zero. Substituting (2.18) into (2.14) the boundary becomes:

$$\sum_{\alpha=1}^{n} i(\mathbf{u}_\alpha)\eta_\alpha \mathbf{u}^{\mathrm{T}}\mathbf{u}_\alpha + b = 0 \tag{2.20}$$

An overlap of the categories may indicate that a plane that separates them does not exist. To deal with this case, the linear SVC was adapted (Cortes, 1995; Cortes and Vapnik, 1995) by the introduction of slack variables ξ_α ($\alpha = 1,\ldots,n$) in the optimization problem. The slack variables ξ_α relax the constraints in (2.17), so, some classification errors are permitted but at a certain cost. Now, the optimization problem is:

$$\begin{aligned} &\underset{\mathbf{w},b,\xi}{minimize} \quad \frac{1}{2}\|\mathbf{w}\|^2 + P\sum_{\alpha=1}^{n}\xi_\alpha \\ &subject\ to \quad i(\mathbf{u}_\alpha)[\mathbf{w}^{\mathrm{T}}\mathbf{u}_\alpha + b] \geq 1 - \xi_\alpha \\ &\qquad\qquad \xi_\alpha \geq 0 \quad ; \quad \alpha = 1,...,n \end{aligned} \tag{2.21}$$

Here, P is a user-defined penalty parameter. The optimization problem has the same solution shown in (2.18), (2.19) and (2.20), the only difference is the bounds of the multipliers η_α that appear in the Lagrange formalism.

To cope with data that is not linearly separable, the vectors **u** are mapped into a higher-dimensional space $\mathcal{F}$ by a map function Φ. In the space $\mathcal{F}$, the linear SVC algorithm is applied. The linear classifier in the space $\mathcal{F}$ will create a non-linear decision boundary in the original input space (Figure 2.3).

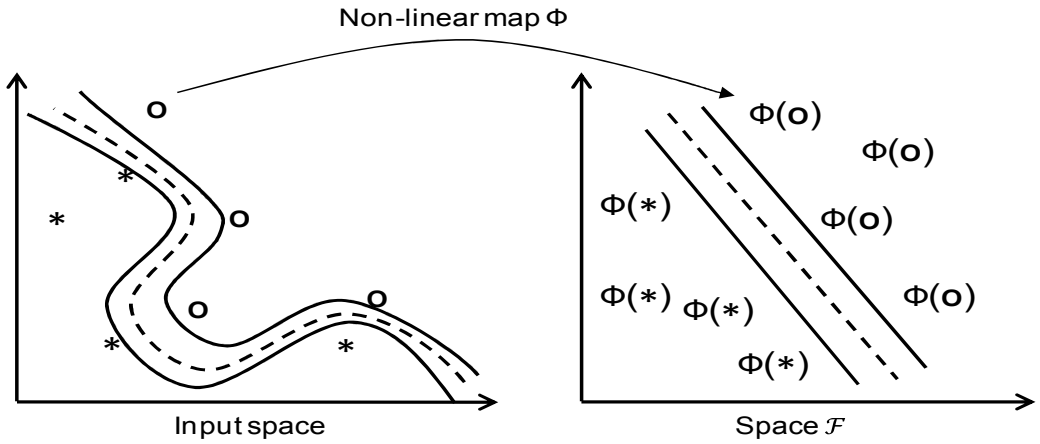

Figure 2.3: The SVC algorithm is applied in a high dimensional space (Modified redrawn from Cristianini and Schölkopf , 2002, p. 40)

The implementation of the SVC algorithm in the space $\mathcal{F}$ is done by using kernels; this consists of replacing the scalar product between training data with a kernel function in the formulation of the SVC algorithm. The kernel is a function in the input space of the vectors **u**, which returns the dot products of the images in some space $\mathcal{F}$, without even knowing the form of the map Φ:

$$k(\mathbf{u}_\alpha, \mathbf{u}_\beta) = \left\langle \Phi(\mathbf{u}_\alpha), \Phi(\mathbf{u}_\beta) \right\rangle \tag{2.22}$$

SVC testing. SVC testing means to use the decision boundary (2.20) to allocate a single category s to the unsampled location **u** according to the rule:

$$s(\mathbf{u}) = \begin{cases} s_1 & if \quad \left(\sum_{\alpha=1}^{n} i(\mathbf{u}_\alpha)\eta_\alpha K(\mathbf{u}, \mathbf{u}_\alpha) + b \right) > 0 \\ s_2 & if \quad \left(\sum_{\alpha=1}^{n} i(\mathbf{u}_\alpha)\eta_\alpha K(\mathbf{u}, \mathbf{u}_\alpha) + b \right) < 0 \end{cases} \tag{2.23}$$

where $K(\mathbf{u}, \mathbf{u}_\alpha)$ is a symmetric positive-definite matrix that contains the values of the kernel function.

In the machine learning jargon the unsampled locations taken together are called the testing set. The percentage of unsampled locations misclassified by the decision boundary is called testing error or generalization error. The complement of the generalization error to add to unity is called *generalization accuracy*.

2.2.3 Model Selection

Model selection refers to the task of choosing a kernel function with its parameters and the penalty parameter *P* to train the SVC algorithm. Model selection can be considered part of the training SVC step, but it deserves an independent section due its importance in the performance of the classifier.

Kernel selection. Basic licit kernels found in practice due its simplicity and good performance are the linear, the polynomial and the Gaussian radial basis function (Grbf) kernels. An extensive and in depth description of these and others more complex kernels can be found in (Cristianini and Shawe-Taylor, 2004). The special interest for this research is the Grbf kernel. It has the form:

$$k(\mathbf{u}, \mathbf{u}') = \exp\left(-\gamma \left\| \mathbf{u} - \mathbf{u}' \right\|^2\right); \gamma > 0 \tag{2.24}$$

where **u** and **u'** represent any two different locations and γ is a kernel parameter that must be selected by the user. The reasons for choosing the Grbf kernel are:

- The SVC algorithm trained with a Grbf kernel can separate correctly any arbitrary number of data (Burges, 1998, p.151). This is a very convenient property for

geostatistical applications, where it is desirable to classify correctly all the observed data.

- Compared to the other basic kernels, the Grbf kernel has the advantage of having less numerical difficulties than the polynomial kernel (Hsu, Chang, and Lin, 2008), while the linear kernel can be considered a special case of the Grbf kernel (Keerthi and Lin, 2003), and
- The Grbf kernel has shown good performance in geostatistical applications, e.g. Wohlberg, Tartakovsky and Guadagnini (2006), Pozdnoukhov and Kanesky (2006) and Kanesky et al. (2001).

Parameter selection. Training the SVC algorithm using the Grbf kernel implies simultaneous selection of the pair of parameters (P, γ), so that the boundary classifier can predict unsampled locations with the maximum generalization accuracy (or its complement, the lowest generalization error). Given that the true generalization accuracy is unknown, a proxy value is calculated usually by k-fold cross-validation.

In k-fold cross-validation, the observed data is randomly divided into k equal sized subsets. Then, the SVC algorithm is sequentially trained using the k-1 subsets and tested in the remaining subset. Training is repeated k times and the percentage of data correctly classified for all the k subsets that are not included in the training data is recorded as the cross-validation accuracy (Abe, 2005, p. 73).

To select the optimal pair of parameters (P, γ), the conventional approach calculates the k-fold cross-validation accuracy for every pair (P, γ) on a predefined grid-search and it chooses the one with the maximum value. To explore a wide range of parameter combinations, the grid is designed as an exponentially growing sequence of P and γ values (Hsu, Chang and Lin, 2008), for instance:

$$P = \{2^{-3}, 2^{-2}, ..., 2^{8}, 2^{9}\}$$
$$\gamma = \{2^{-10}, 2^{-9}, ..., 2^{11}, 2^{12}\}$$

The selected pair of parameters (P, γ) is used to train the SVC algorithm with the complete set of observed data. A theoretical description of cross validation along with a benchmark with others techniques for model selection can be found in Anguita, Boni, Ridella, Rivieccio and Sterpi (2005).

2.3 Summary

This chapter showed how IK and SVC algorithms can solve the problem of classifying categorical variables at locations of interest, and how the SIS algorithm can generate multiple realizations of the spatial distribution of categorical variable using a probabilistic approach.

The next chapter highlights some of the differences between the IK and SVC approaches, and a technique for SVC parameter selection that tries to reconcile such differences is proposed. Applications are presented throughout the following chapters.

3 Selection of SVC Parameters (P,γ)

Chapter 3 introduces a heuristic technique to obtain from the SVC algorithm a boundary classifier with good generalization ability and desirable properties from the geostatistical approach, that is, correct classification of all the observed data and reproduction of the global facies proportions.

First, the proposed technique is presented; second, implementation issues are discussed, and then a classification problem for a large data set is solved to illustrate its application. The classification problem was defined in Section 1.1 as the problem of assigning a single category to an unsampled location based on a limited set of observed data. The reference map is available, so the performance of the approach can be directly assessed. These results are compared to the responses of the conventional application of geostatistics (IK) and machine learning (SVC) methods.

3.1 Introduction

To solve a classification problem, the SVC approach normally neglects the empirical accuracy in the process of selecting the pair of parameters (P, γ). The objective of the traditional model selection is to maximize a proxy of the generalization accuracy, therefore, in practice, obtaining an empirical accuracy of 100% is considered a suboptimal result associated to an over fitted boundary. In fact, it is often alleged that a classifier which explains correctly all the training data will not necessarily generalizes well.

The conventional semivariogram based geostatistical approach constructs models that always classifies correctly all the observed data; it is a consequence of the exactitude property of kriging. Additionally, in practical applications, it is accepted that a measure of the quality of the response of a classification model is its ability to reproduce the

statistics (global proportions) of the observed data, assuming of course, that those statistics are representative of the population.

In view of the above concepts, one can argue that a map estimated using SVC should be similar in terms of features and generalization accuracy to the map obtained by using a geostatistical classification model as long as: (1) the SVC algorithm is trained with a pair (P, γ) that produces a classifier with empirical accuracy of 100%, and (2) the categories obtained at the unsampled and sampled locations by testing the SVC algorithm have proportions that reproduces reasonably well the representative proportions of the observed data. The former point is equivalent to the exactitude property of kriging, while the latter point corresponds to reproduction of the input/target global proportions.

3.2 Proposed Technique

The proposed technique differs from the conventional SVC approach in the criteria to select the pair of parameters (P, γ). The new technique does not use the cross-validation accuracy; it makes use of the empirical accuracy and the proportions of categories in the estimated response to select the pair of parameters (P, γ). The heuristic parameter selection has the following steps:

1) Select a grid-search for the pair of parameters (P, γ)

2) Visit once the nodes of the (P, γ) grid-search and at each node:

    a. Train the SVC algorithm using the observed data and calculate the empirical accuracy.

    b. Test the SVC algorithm using all the locations in the domain of interest and the model obtained in step (2.a). Calculate the proportions on the resulting categorical map.

3) Plot contour lines of the empirical accuracy and the proportions over the (P, γ) grid-search. Select the node (P, γ) where the first contour of empirical accuracy of 100% intersects the closest contour to the target proportions.

Figure 3.1 shows the work flow for the proposed technique.

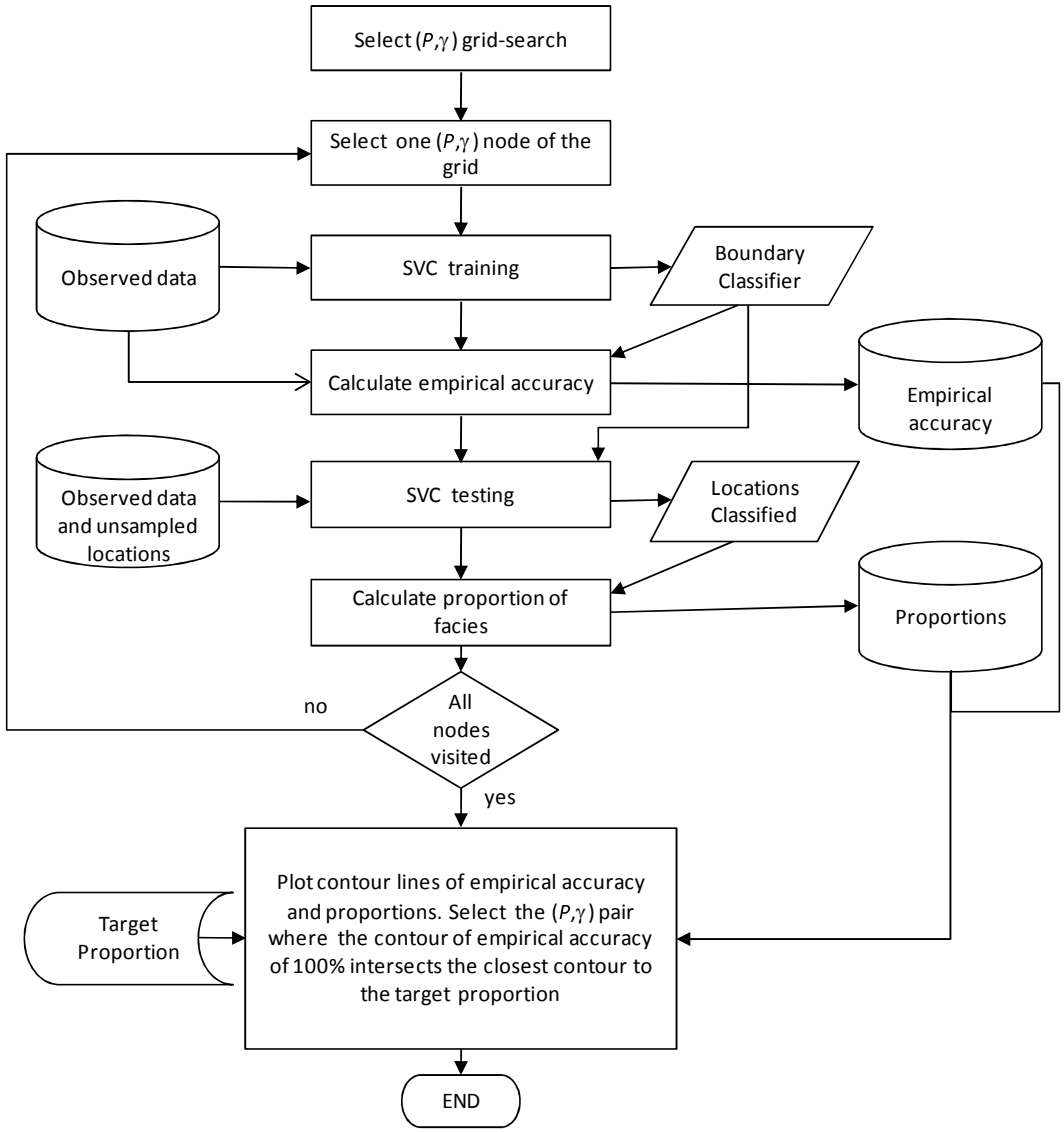

Figure 3.1: Work flow for the proposed technique to select the SVC parameters (P, γ)

3.3 Implementation Aspects

Some implementation decisions to apply the proposed technique to real problems include: the selection of the (P, γ) grid-search and the computational time, the calculation of the empirical and the generalization accuracy, and the selection of the target proportion. These issues are discussed below.

3.3.1 Grid-search Selection and Computational Time

The selection of the (P, γ) grid-search impacts the computational time spent by the proposed technique. Given that the SVC algorithm must be trained and tested at each node of the grid, the selection depends on the number of observed data (training set) and unsampled locations (testing set). For this reason, before discussing the selection of the (P, γ) grid-search, the aspects that influence the computational time for the proposed and the k-fold cross-validation techniques are analyzed and compared.

Computational time. The number of observed data impacts the computational time of training the SVC algorithm. The proposed technique requires training the SVC algorithm once at each node, while the k-fold cross-validation technique requires training the SVC algorithms k times at each node. Therefore, considering only the SVC training, the proposed technique is faster than k -fold cross-validation.

The number of unsampled locations impacts the computational time of testing the SVC algorithm. The proposed technique requires testing the SVC algorithm at each node of the grid-search, while the k-fold cross-validation only requires testing the SVC algorithm at the pair of parameters (P,γ) chosen in the training step. Therefore, considering only the SVC testing, the proposed technique is slower than k-fold cross-validation.

Training the SVC algorithm is a time consuming task; it solves an optimization problem to obtain the parameters that define the boundary classifier. Testing the SVC algorithm is a quick task; it just takes the calculated boundary to classify the unsampled locations. Consequently, as the proposed technique has fewer training calculation than the k-fold cross-validation, one can reasonably conclude that in the overall process, the proposed technique is computationally less time consuming to select the SVC pair of parameters (P,γ).

Currently, there are no practical limitations in the number of data to train the SVC algorithm, the LIBSVM software implements the Sequential Minimization Optimization (SMO) algorithm (Platt, 1998) to solve the optimization problem and can efficiently deal with data sets up to around 10 000 samples. Kecman (2005, p.95) reports that the Iterative Single Data Algorithm (ISDA) for training kernel machines (Kecman, Huang, and Vogt, 2005) can efficiently solve data sets over one million points.

Finally, if the number of unsampled locations is too large, it is possible to reduce the computational time of the proposed technique by calculating the proportions required in step 2.b over a regularly sampled subset of those locations. Once a (P,γ) pair of parameters is selected, the SVC is tested using the complete set of locations in the domain.

(P, γ) Grid-search selection. To explore a wide range of parameter combinations, the grid-search is designed as an exponentially growing sequence of *P* and *γ* values. Common values to explore are:

$$P = \{2^{-3}, 2^{-2}, ..., 2^{8}, 2^{9}\}$$
$$\gamma = \{2^{-10}, 2^{-9}, ..., 2^{11}, 2^{12}\}$$

However, it is not the range but the degree of discretization that ultimately determines the number of nodes in the grid-search, which consequently affects the computational time and the quality of the result. Keeping in mind that the training task has the greatest impact on the computational time, discretization of the exponents in intervals of 0.1 or 0.2 can be used for sets of observed data up to 10000 nodes.

For large sets of observed data (>10000) a two-step (*P*,γ) grid-search selection can be applied (Hsu, Chang, and Lin, 2008). First, a coarse grid (e.g. discretization of the exponents in intervals or 0.5 to 1) is used to identify a "good" region on the grid; and second, a finer grid (e.g. discretization of the exponent in intervals of 0.1 or less) on that region is defined to get the definitive SVC response.

3.3.2 Empirical and Generalization Accuracy

The empirical error is the percentage of observed data correctly classified by the SVC boundary:

$$Empirical\ accuracy = \frac{Number\ of\ sampled\ locations\ correctly\ classified}{Total\ number\ of\ sampled\ locations} \cdot 100 \tag{3.1}$$

Technically, the machine learning field would calculate the generalization accuracy as the percentage of unsampled locations correctly classified by the SVC boundary; however, in the spatial context, it is more natural to assess the performance of the boundary classifier considering all the locations of interest, sampled and unsampled locations. Accordingly, in this thesis the generalization accuracy is calculated as:

$$Generalization\ accuracy = \frac{Number\ of\ sampled\ and\ unsampled\ locations\ correctly\ classified}{Total\ number\ of\ locations\ in\ the\ domain\ of\ interest} \cdot 100 \tag{3.2}$$

3.3.3 Target Proportions

The target proportion in step 3 of the proposed technique is calculated from the observed data. It should be a representative proportion of the entire domain of interest.

The observed data seldom is representative of the entire domain of interest, factors such as the density (or sparseness) and location of the data might affect its representativeness. The influence of these elements in the proposed technique is now discussed.

Dense and regularly spaced data set: It is the most favorable case. The target proportions calculated from the observed data tend to be representative of the entire domain. The proposed technique produces a good response that it is comparable to the conventional geostatistical approach. The application in Section 3.4 illustrates this case.

Dense data set with clusters: The target proportions calculated from the observed data tend to be non-representative of the entire domain. Cell-declustering or polygonal declustering techniques (Deutsch, 2002, p.50-57) can be applied to obtain a representative target proportion. The SVC algorithm is not affected by clusters. Clustered points are considered redundant and they do not appear in the resulting set of support vectors. Even if some clustered points are removed from the observed data, the response of the SVC algorithm does not change. The proposed technique will produce a good response in this case.

Sparse data set with or without out clusters: The target proportions calculated from sparse data may or may not be representative of the entire domain. Declustering techniques can be applied to look for representative target proportions, but the result might be non-conclusive or misleading. The application in Chapter 4 briefly discusses this issue and shows that the proposed technique produces a good response that it is comparable to the conventional geostatistical approach. With ancillary information the response of the proposed technique might be improved by using an anisotropic Gaussian radial basis function kernel and/or a modified version of the SVC algorithm for imbalanced data (Abe, 2005, p.65). The evaluation of these two concepts is beyond the scope of this research.

Observed data located away from the limits of the domain: The SVC algorithm classifies unsampled locations based on a boundary that separates the set of observed data. Regions between the data and the limits of the domain might be incorrectly classified by the proposed technique based on SVC. For cases where the observed data is located far away from the limits of the domain or when those limits are poorly defined, the proposed technique is not recommended.

The following example illustrates the application of the new technique to solve a classification problem using a relatively large set of data. This application shows that the responses of the proposed technique and the conventional geostatistical approach are similar under the favorable conditions offered by a large set of data. If the responses of both approaches did not converge in an advantageous scenario, it would not be reasonable to expect a good result from the proposed technique in solving the more challenging case of sparse data.

3.4 Application

A synthetic classification study for a large data set was prepared to illustrate the application of the proposed methodology. A reference data set of two facies was generated and a subset was sampled to be used as observed data.

Three methods were used to allocate the facies at the unsampled locations of the reference data: (1) Ordinary indicator kriging (OIK) with a classification rule (2) SVC

with k-fold cross-validation, and (3) SVC with the proposed heuristic parameter selection. The results from each method are documented and discussed.

3.4.1 Reference Data Set

The reference data set is generated by unconditional Gaussian simulation using an isotropic spherical semivariogram model without nugget effect and a range of 200 m. The Gaussian field is transformed to a categorical variable of two facies, white (code 1) and black (code 0) using a threshold of 0 normal units. The noise or small scale variability in the map is removed by applying twice a moving window cleaning procedure.

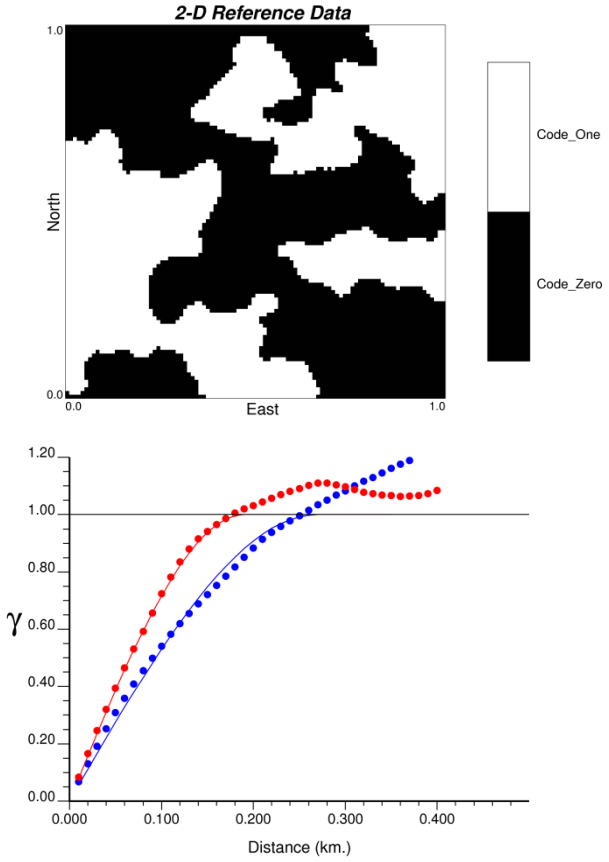

Figure 3.2: Map of reference data (top). Standardized experimental indicator semivariogram represented by dots, and standardized indicator semivariogram models represented by solid lines. Blue and red colors indicate the major and minor directions of anisotropy, respectively (bottom).

The reference data (Figure 3.2) has a resolution of 10 m x 10 m spacing that spans an area of 1km x 1km, so it contains 10000 nodes. The standardized experimental indicator semivariogram calculated with the reference data shows an anisotropic behavior, where the major direction of continuity is chosen as the East-West direction and the minor direction is the orthogonal North-South direction. The proportions for the white and black facies are 45.75% and 54.25%, respectively. For convenience, the Table 3.1 on page 36 summarizes all the numerical results of Chapter 3.

The standardized indicator semivariogram model is:

$$\gamma(\mathbf{h}) = Sph_{\substack{h\max=0.27 \\ h\min=0.19}}(\mathbf{h})$$

The selection of a 1km. x 1km. box domain is not arbitrary; the objective is to avoid having to re-scale the data to the range [0, 1] which is a common practice to apply the SVC algorithm.

3.4.2 Sampled Data (Observed Data or Training Data)

A relatively large (2.25% of the reference data) sample was drawn to reduce the subjectivity in the construction of the geostatistical model, specifically, to facilitate the modeling of the semivariogram and the calculation of the representative global proportions.

The observed data is sampled from the reference map at nominally 70 m x 70 m spacing. Figure 3.3 shows the map of the 225 samples.

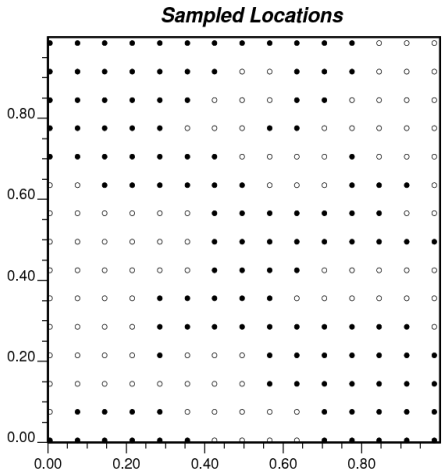

Figure 3.3: Map of 225 samples at nominally 70 m x 70 m spacing.

3.4.3 Classification by Indicator Kriging

As explained in Chapter 2, the geostatistical approach usually has the following steps: Preprocessing of data, exploratory data analysis, variography, kriging and classification.

Preprocessing, analysis and variography of the data. The data does not need to be pre-processed, the facies have been already coded as 1 (white) or 0 (black).

The exploratory data analysis aims for representative statistics of the variables of interest. In this case, the samples are regularly spaced; there is no visual evidence of clusters; therefore, declustering is not required. The sample proportions for the white and black facies of 45.89% and 54.11%, respectively, are considered representative of the population.

The standardized experimental indicator semivariogram (Figure 3.4) shows the major direction of anisotropic in the East-West direction and the minor direction of anisotropic in the North-South direction. The standardized indicator semivariogram was modeled as:

$$\gamma(\mathbf{h}) = Sph_{\substack{h\max=0.28 \\ h\min=0.19}}(\mathbf{h})$$

As expected, there are minor differences between the indicator semivariogram model of the reference data and the indicator semivariogram model of the data sampled.

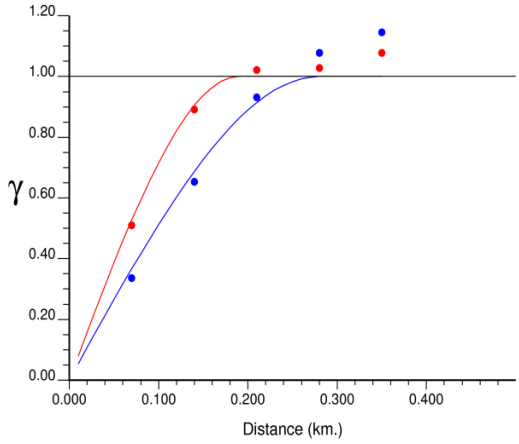

Figure 3.4: Dots represent the experimental standardized indicator semivariogram. Solid lines represent the standardized indicator semivariogram model, blue and red for the major and minor directions of anisotropy, respectively.

Indicator kriging. The white and black facies are categorical variables; they cannot be estimated as a linear combination of neighboring data. The indicator kriging algorithm provides a probability map for the facies. The map, in conjunction with a classification rule, can be used to assign a single facies to each location.

Figure 3.5 shows the map of conditional probabilities of occurrence for the white facies at the resolution of the reference data set obtained by ordinary indicator kriging (OIK).

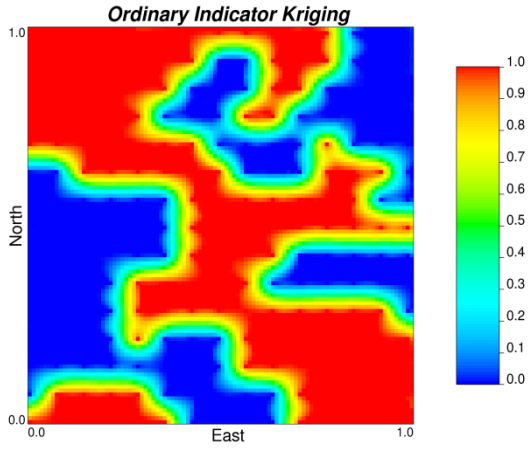

Figure 3.5: Ordinary indicator kriging map for the white facies

Classification. For the binary problem at hand, the classification rule presented in Section 2.1.2 is equivalent to use a threshold rule that allocates the locations to the facies

whose probability of occurrence is greater than 50%. The rule is simple and it provides results that can be compared to that obtained by the SVC algorithm.

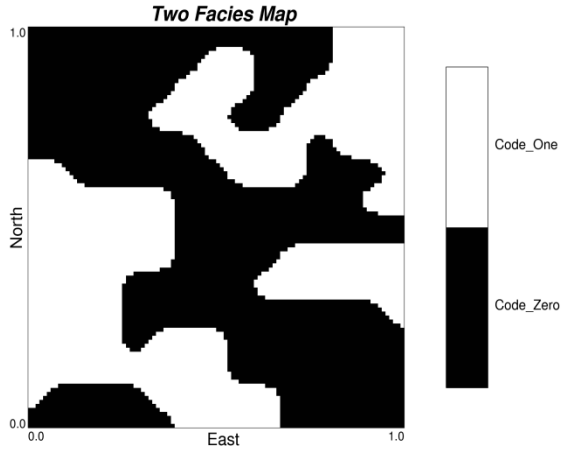

Figure 3.6: Map of facies after applying a threshold rule of 50%.

The threshold rule is applied to each location to get the map of facies shown in Figure 3.6. When this categorical map is compared node to node to the reference map, it has a generalization accuracy of 91.51%. The estimated map has proportions for the white and black facies of 44.86% and 55.14%, respectively.

3.4.4 SVC using Cross-validation

The SVC algorithm has the following steps: Preprocessing of data, model selection (kernel selection and cross-validation), SVC training and testing.

Preprocessing of data. The SVC algorithm requires the facies to be coded as -1 and 1; however, the LIBSVM software can receive the facies as 0 and 1, so there is no need for additional re-coding.

To avoid numerical difficulties due to the use of very large numbers during the calculations of the kernel matrices (Hsu, Chang, and Lin, 2008) the coordinates should be linearly rescaled to the range [0, 1]. The reference data is already in the recommended scale.

Model selection. Model selection implies the definition of the kernel and its parameters, and the definition of the penalty parameter *P*. Since the kernel was chosen to be the Grbf, only the pair of parameters (*P,* γ) must be tuned.

An 81x81 grid-search space for the pair of parameters (*P,* γ) was defined to do 10-fold cross-validation. More specifically, the grid-search has the following sequence of values:

$$P = \{2^{-2}, 2^{-1.9}, 2^{-1.8}, ..., 2^{5.8}, 2^{5.9}, 2^{6}\} \quad ; \quad \gamma = \{2^{3}, 2^{3.1}, 2^{3.2}, ..., 2^{10.8}, 2^{10.9}, 2^{11}\}$$

For each pair (*P,* γ) in the grid-search the cross-validation accuracy is calculated. The pair with the highest accuracy is selected to train the SVC algorithm.

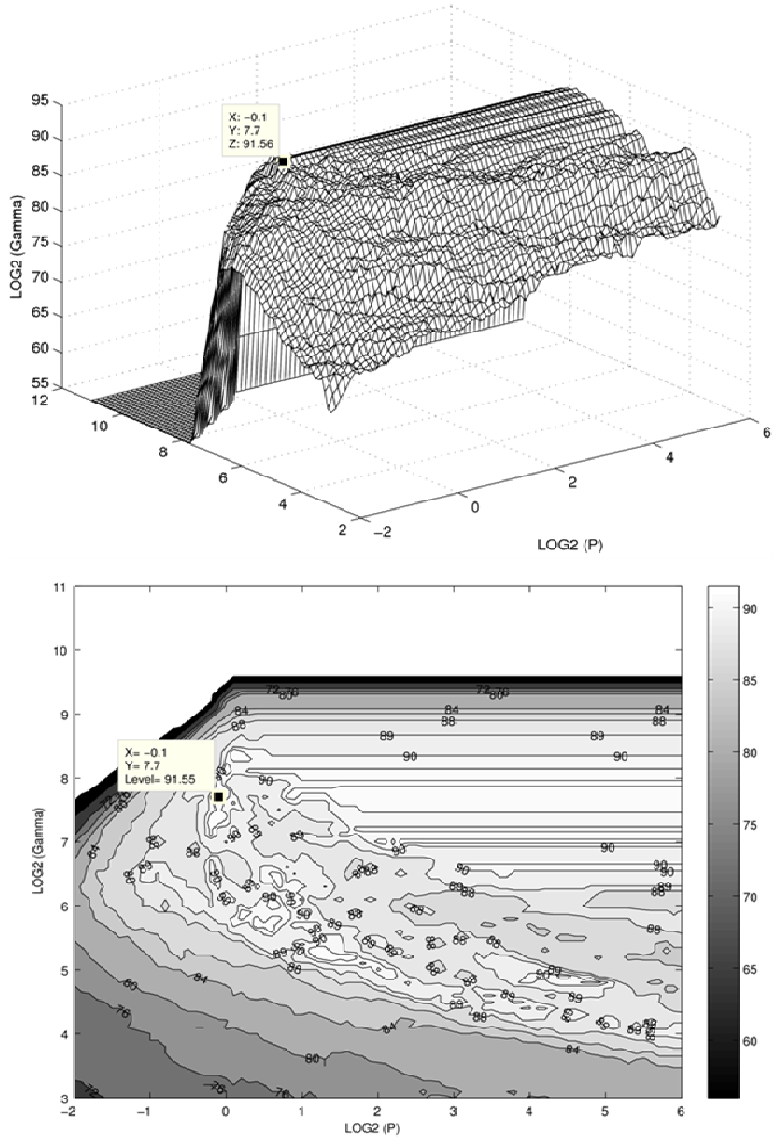

Figure 3.7: Response surface (top) and contour lines (bottom) of the cross-validation accuracy. The flag shows the maximum value of 91.56% at the pair ($\log_2 P$, $\log_2 \gamma$) = (-0.1, 7.7).

Figure 3.7 shows the surface and the contour lines of the cross-validation accuracy, where the maximum value is 91.56% at the pair ($\log_2 P$, $\log_2 \gamma$) = (-0.1, 7.7).

SVC training and testing. Once the pair ($\log_2 P$, $\log_2 \gamma$) = (-0.1, 7.7) is selected, the SVC algorithm is trained using the 225 sampled locations. The result is a boundary classifier with 216 support vectors. The boundary is used to classify all the locations (sampled and unsampled locations) in the domain of interest.

Figure 3.8 shows the categorical map obtained after the testing procedure. The map has a generalization accuracy of 91.37%. The proportions of white and black facies are 44.5% and 55.5%, respectively.

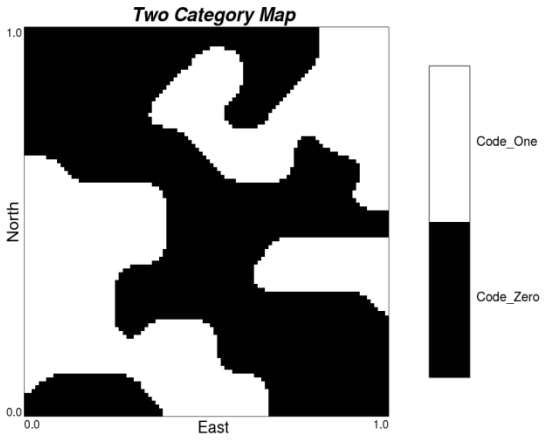

Figure 3.8: Map of facies obtained by k-fold cross-validation technique.

3.4.5 SVC with Heuristic Parameter Selection

The heuristic technique differs from the conventional SVC algorithm in the criteria to select the parameters (P, γ). Using the 81x81 grid-search defined in Section 3.4.4 the key aspects of the procedure proposed in Section 3.2 are sketched in Figure 3.9. First, the empirical accuracy contour map (Figure 3.9 a) and proportion contour map (Figure 3.9 b) are plotted; and then, the maps are combined to find the intersection point between the 100% empirical accuracy contour line and the target proportion of 45% contour line (Figure 3.9 c). In this case, the intersection was found at the pair (1.6, 6.9) which is used to train the SVC algorithm. Figure 3.10 shows the real combined contour map generated for this example. The node flagged at ($\log_2 P$, $\log_2 \gamma$) = (1.6, 6.9) is the intersection

between the empirical accuracy contour of 100% and the contour of 45.15% white proportion, which is the nearest value to the target 45.89% proportion.

As before, once the pair ($\log_2 P$, $\log_2 \gamma$) = (1.6, 6.9) is selected, the SVC algorithm is trained using the 225 sampled locations. The result is a boundary classifier with 137 support vectors. The boundary is used to generate the map of facies shown in Figure 3.11. This estimated map has a generalization accuracy of 91.46%. The proportions of the white and black facies are 45.15% and 54.85%, respectively.

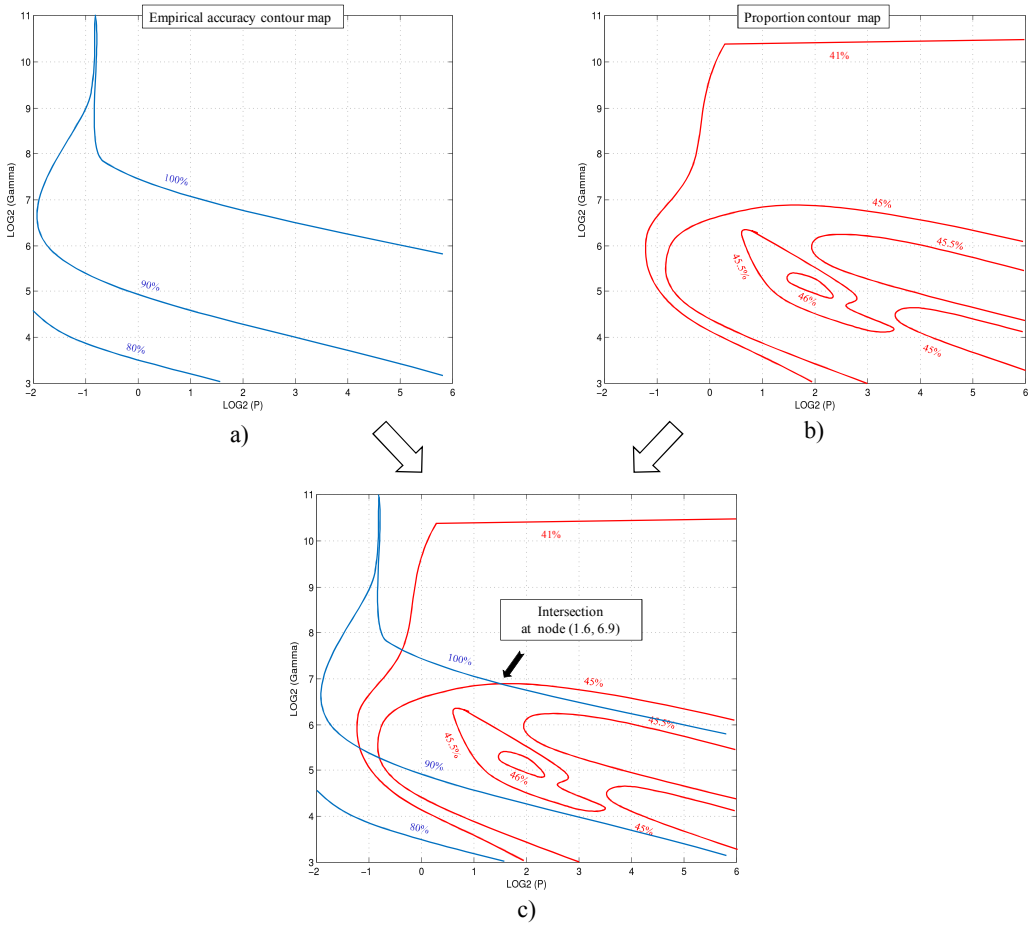

Figure 3.9: Empirical accuracy contour map (a), proportion contour map (b) and combined map (c) to find the intersection point between the 100% empirical accuracy contour line and the target proportion of 45% contour line.

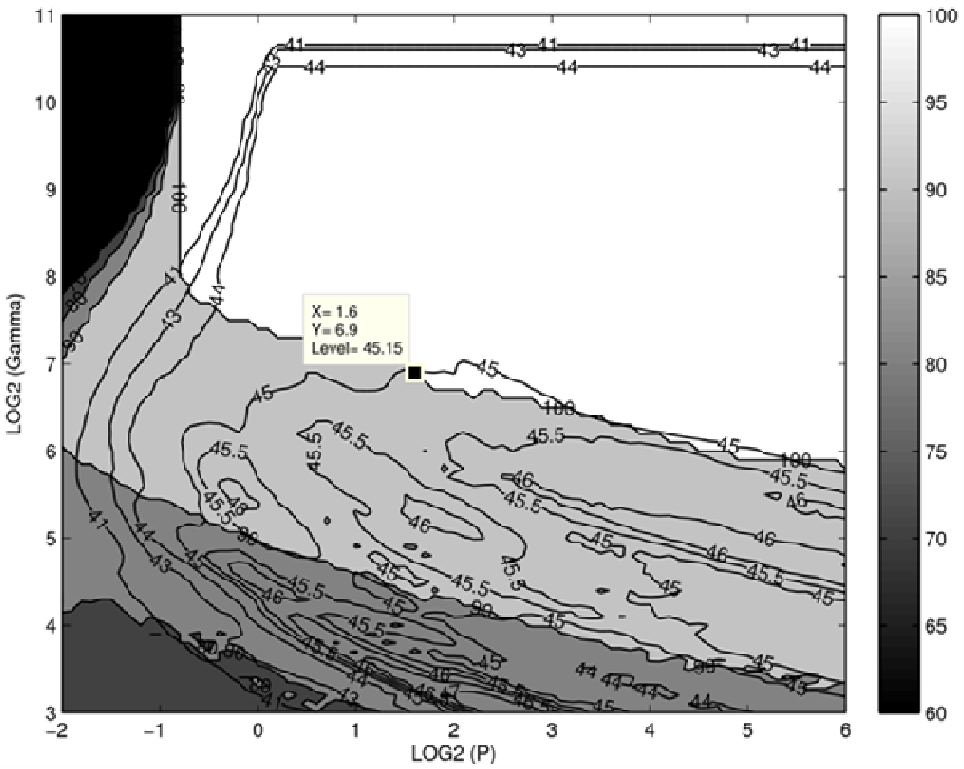

Figure 3.10: Contour lines of the empirical accuracy (grey scale) and the proportions of white facies of the estimated categorical maps. The flag shows the intersection node ($\log_2 P$, $\log_2 \gamma$) = (1.6, 6.9).

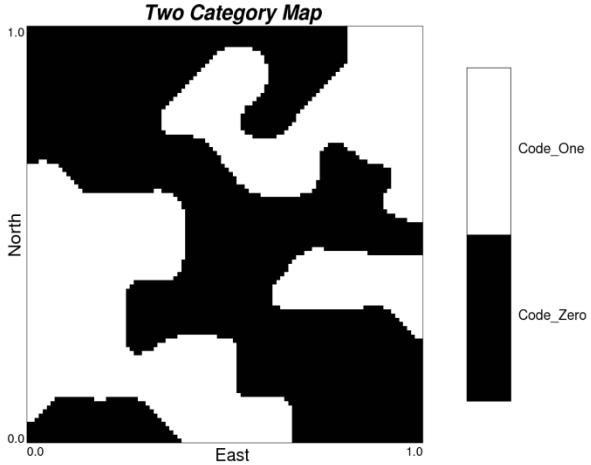

Figure 3.11: Map of facies obtained by the proposed technique.

3.5 Discussion

Numerical results of Chapter 3 are summarized in Table 3.1. The three methodologies produce similar numerical results, which is a direct consequence of having a relatively large amount of data regularly spaced.

Table 3.1: Summary of numerical results of Chapter 3

Technique	Generalization accuracy (%)	Proportions of facies (%)		Number of support vectors
		White	Black	
Reference data	N.A	45.75	54.25	N.A
Observed data (225 samples)	N.A	45.89	54.11	N.A
Geostatistical approach - OIK	91.51	44.86	55.14	N.A
SVC heuristic approach	91.46	45.15	54.85	137
SVC k-fold cross-validation	91.37	44.5	55.5	216

N.A: Not apply

Comparing the conventional SVC to the proposed heuristic SVC approach, the latter not only has slightly better generalization accuracy, but less support vectors which is a desirable sparse property for the SVC classifiers.

The conventional geostatistics OIK technique produces slightly better generalization accuracy than the conventional and the proposed SVC algorithm. However, the SVC algorithms have the advantage of being totally automatic. As they do not require any subjective modeling decisions, the results are fully reproducible based only on the observed data.

The example presented here confirms that the responses of the OIK algorithm and the proposed SVC algorithm tend to converge, at least, for large data sets.

4 Application to Sparse Data

Chapter 4 is concerned with the use of a small set of observed data to classify a much larger set of unsampled locations and to generate simulated realizations. The synthetic case presented shows the major difference between the geostatistical and machine learning approaches to solve the classification problem. It also shows how simple information from SVC allows for an improvement in the response of conventional geostatistical classification and indicator simulation models.

Three sections are developed: the first section implements the proposed technique to solve a classification problem based on a sparse data set. The result is compared to conventional OIK and SVC; the second section explores the effect of using SVC information on the performance of geostatistical realizations; and, the third section summarizes and discusses the results.

4.1 The Misleading Data Point

This case study shows that for small data sets, the best geostatistical model is not necessarily the one that uses all the available data. This statement might seem odd from the geostatistics perspective, but it is sound from the machine learning approach of maximizing the generalization accuracy. The case study has the following steps:

(1) A synthetic reference 2D data set with two facies is generated. A subset of 50 samples is randomly drawn to be used as observed data.

(2) The SVC algorithm is trained to obtain a boundary which is used to assign the facies to all the locations (sampled and unsampled locations) of the domain under study. Information about how the observed data is classified by the SVC boundary is collected.

(3) Two facies maps are generated in parallel by OIK with a classification rule. One map does not account for the information collected in step 2, while the other map uses this information in its construction.

To reduce the influence of subjective decisions on the results, complete access to the reference map is assumed for steps 2 and 3. This allows the true generalization accuracy to be calculated and used to select the SVC parameters (P,γ). In the same way, the semivariogram and statistics of the reference map are available to generate the geostatistical model.

(4) The pair of SVC parameters ($P,\ \gamma$) is selected by the proposed technique without information from the reference data. The SVC is then tested to get an estimated map.

(5) The pair of SVC parameters ($P,\ \gamma$) is selected by conventional k-fold cross-validation. The SVC is then tested to get an estimated map.

4.1.1 Reference and Sampled Data

The same reference data set described in section 3.4.1 is used for this exercise. Figure 4.1 shows the locations of 50 samples randomly drawn from the reference map. The proportions of white and black facies are 36% and 64%, respectively.

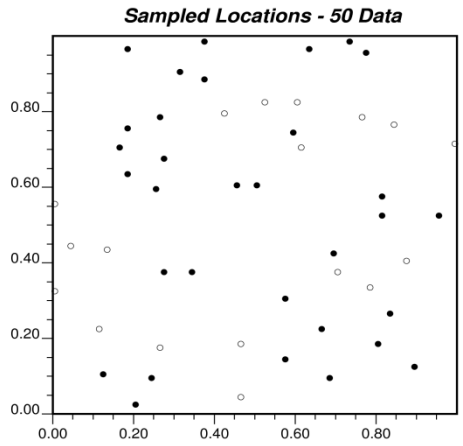

Figure 4.1: Map of 50 randomly sampled locations

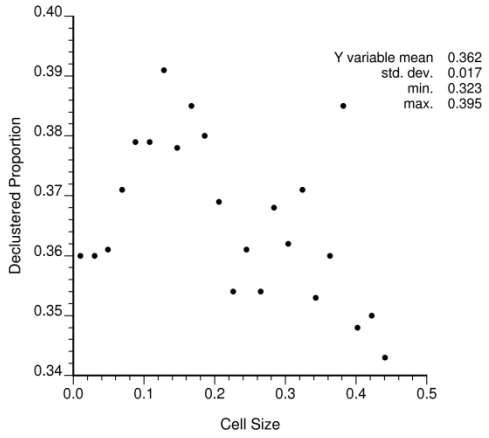

Figure 4.2: The declustered proportions versus the cell size.

The sampled data does not reveal the statistics of the reference map. However, the application of a declustering method does not seem solving this lack of representativeness. The Figure 4.2 shows the cell-declustered proportions (Deutsch, 1998) for the white facies versus the cell size. This plot does not allow choosing a proportion value conclusively; for that reason, the equal-weighted proportion of the sample is chosen as target proportion to apply the proposed technique.

4.1.2 SVC using the True Generalization Accuracy

The SVC algorithm has the following steps: Preprocessing of data, model selection and SVC training and testing.

Preprocessing of data. As was explained in Chapter 3, it is not necessary to code the facies as -1 and 1, the LIBSVM software accepts them as 0 and 1. Neither is required to re-scale the coordinates.

Model selection. This thesis only implements the Grbf kernel, so the task in model selection is to pick the pair of SVC parameters (P, γ). The 81x81 grid-search presented in section 3.4.4 for the pairs (P, γ) was kept.

For each pair (P, γ) on the grid-search, the SVC algorithm is trained using the 50 samples. The boundary classifier obtained assigns the facies to all the locations (50

sampled and 9950 unsampled locations) in the domain. The true generalization accuracy and the empirical accuracy are recorded.

The contour lines of the generalization and empirical accuracies (grey scale) are plotted on the grid-search as shown in Figure 4.3. The arrow points to the maximum value of the generalization accuracy, which is 80.05% at the pair ($\log_2 P$, $\log_2 \gamma$) = (0.5, 6.1). For convenience, Table 4.1 on page 53 summarizes all the numerical results of Section 4.1.

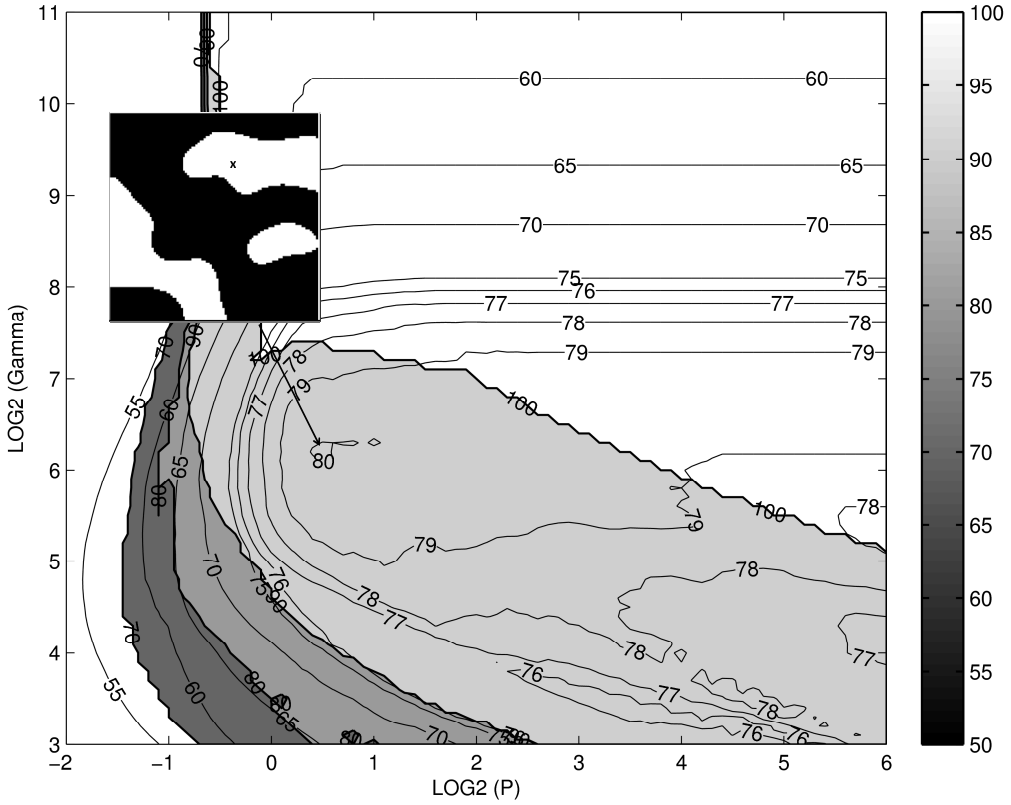

Figure 4.3: Contour lines of the true generalization accuracy and the empirical accuracy (grey scale). Subplot shows the SVC solution map for the pair ($\log_2 P$, $\log_2 \gamma$) = (0.5, 6.1). The "x" identifies a misclassified location.

SVC training and testing. In general, once the pair (P, γ) is chosen, the SVC algorithm is trained with the observed data and the output boundary is used to classify the sampled and the unsampled locations. In this case, all the procedure was done during the model selection step while calculating the true generalization accuracy.

Figure 4.3 shows that the best SVC solution is reached at the pair ($\log_2 P$, $\log_2 \gamma$) = (0.5, 6.1). The subplot in Figure 4.3 shows the estimated map at the scale of the reference data. It has a generalization accuracy of 80.05% and the proportions of white and black facies are 34.86% and 65.14%, respectively. It is worth noting that the pair ($\log_2 P$, $\log_2 \gamma$) = (1.0, 6.3) produces a SVC solution with 80.04% of generalization accuracy, almost the same value reached by the best pair, but the proportions for this set of parameters are 35.69% and 64.31% for white and black, respectively. Incidentally, these proportions are closer to the proportions of the observed data.

Note that the SVC solution with the best generalization accuracy, does not classify correctly all the observed data. The solution lies on an area of the map with 98.0% empirical accuracy, which means that the boundary misclassified one location. The "x" on the subplot identifies the misclassified location.

4.1.3 OIK Classification with Access to the Reference Map

To limit the effect of subjective decisions in the construction of the geostatistical model, the reference semivariogram and global proportions of facies are assumed to be known. Since the SVC solution indicates that the maximum generalization accuracy is reached at the cost of misclassifying one location, two OIK classifiers are built in parallel, one considers the set of 50 samples, and the other, after deleting the sample misclassified by the SVC algorithm, only considers 49 samples (Figure 4.4). Results are contrasted and discussed.

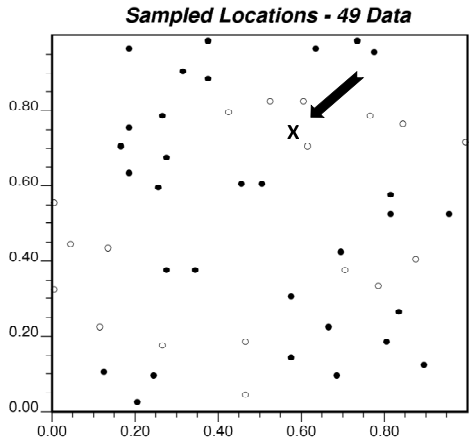

Figure 4.4: Map of 49 randomly sampled locations. The arrow points out the location deleted of the original set of 50 samples.

The geostatistical approach has the following steps: Preprocessing of data, exploratory data analysis, variography, kriging and classification.

Preprocessing, analysis and variography of the data. The facies are coded as 1 (white) or 0 (black).

In this exercise, the facies proportions of the reference map, 45.75% for white and 54.25 for black, are assumed to be known. The reference data and the observed data with 50 samples are described in sections 3.4.1 and 4.1.1., respectively. The set with 49 samples has white and black proportions of 36.73% and 63.27%, respectively.

The standardized indicator semivariogram model of the reference data (see Section 3.4.1) is used.

Indicator kriging. Figure 4.5 shows the maps of conditional probabilities of occurrence for the white category at the resolution of the reference data. The maps were obtained by OIK using the data set of 50 samples (top) and 49 samples (bottom).

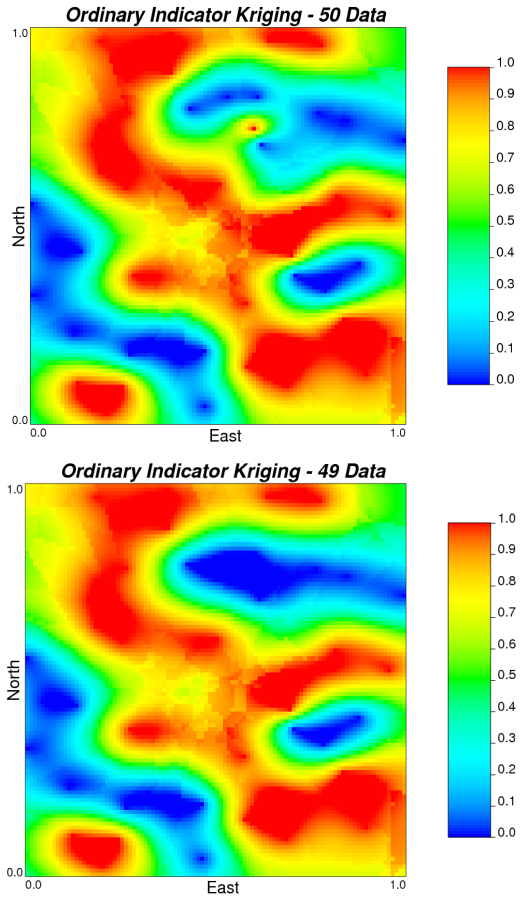

Figure 4.5: OIK maps for the white facies using 50 samples (top) and 49 samples (bottom)

The visual comparison of the maps clearly illustrates the effect of the deleted data, not only at its own location, but on the kriged results of surroundings locations.

Classification. The 50% threshold rule is applied to each location to get the binary maps shown in Figure 4.6. The top map (50 samples) has a generalization accuracy of 80.43% and proportions for white and black facies of 34.98% and 65.02%, respectively. The bottom map (49 samples) has a generalization accuracy of 80.62% and proportions for white and black of 34.99% and 65.01%, respectively.

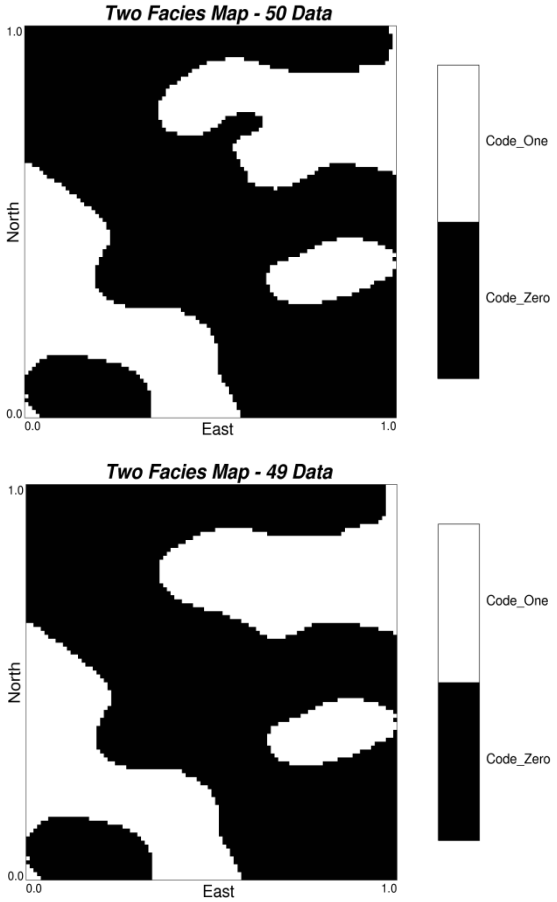

Figure 4.6: Maps of facies obtained by OIK plus a classification rule. Results for 50 samples (top) and 49 samples (bottom).

A comparison of the OIK classifiers to the SVC solution reveals that the former maps have slightly better generalization accuracy (80.4% and 80.6% to 80.1%) and reproduction of the reference facies proportions. It also can be seen that the geostatistical

model built with 49 samples offers slightly better results, in terms of generalization accuracy and reproduction of global proportions, than the model built with 50 samples.

These results suggest that in presence of "perfect" information (that is, knowing the reference data), the geostatistical model generates better classifiers than the SVC algorithm; however, the information collected from the SVC algorithm helps to obtain a better response from the geostatistical model.

In practical applications, the reference data is unknown, and the case of perfect information is nonexistent. The next section shows how the heuristic approach only uses the observed data to select the SVC parameters and to identify the information required to enrich the response of the geostatistical model.

4.1.4 SVC with Heuristic Parameter Selection

As described in Chapter 3, the heuristic parameter selection is based on the analysis of the contour lines of the empirical accuracy and the proportions of facies plotted on the grid-search. Figure 4.7 shows these curves for the white facies for this case study.

The first aspect to check is the intersection point between the contour curves of 36% of proportion (recall that this is the proportion for white facies in the set of 50 samples) and the empirical accuracy curve of 100%. The heuristic approach states that the SVC algorithm trained with the pair of parameters ($\log_2 P$, $\log_2 \gamma$) = (2.8, 6.5) taken from that intersection will generate a map that is similar to the OIK map.

The top-right subplot in Figure 4.7 shows the map obtained with the pair ($\log_2 P$, $\log_2 \gamma$) = (2.8, 6.5) which, as predicted, is similar in features to the map obtained by OIK (Figure 4.6 top). It has a generalization accuracy of 79.67% and the proportions of white and black facies are 36.14% and 63.86%, respectively.

The second aspect to check is the inflection point of the proportion curve of 36%. This point has the interesting characteristic of reproducing the target proportions with the lowest parameter penalty P. The lower the parameter penalty P the larger the margin, and a large margin is related with a boundary that generalizes well. Following this argument, the SVC algorithm trained with the pair of parameters taken from that inflection point should generate a solution with good generalization property. For this particular case, the pair of parameters ($\log_2 P$, $\log_2 \gamma$) = (1.1, 6.1) is very close to the second best pair ($\log_2 P$, $\log_2 \gamma$) = (1.0, 6.3) obtained by using the true generalization accuracy.

The bottom-left subplot in Figure 4.7 shows the map obtained with the pair ($\log_2 P$, $\log_2 \gamma$) = (1.1, 6.1). It has a generalization accuracy of 79.91% and the proportions of white and black facies are 36.12% and 63.88%, respectively. Note that this map misclassified the same location as the map depicted in Figure 4.3. Therefore, without knowing the reference data, it produces exactly the same information required to improve the geostatistical model.

A corollary of the above discussion is that the proposed technique is able to fairly reproduce in one single run, the response of two geostatistical models and to identify the locations that make them different. For instance, compare the two maps in Figure 4.6 to the subplots in Figure 4.7. In practical applications, it would be possible with the proposed technique to anticipate the response of slightly different geostatistical models without even constructing them.

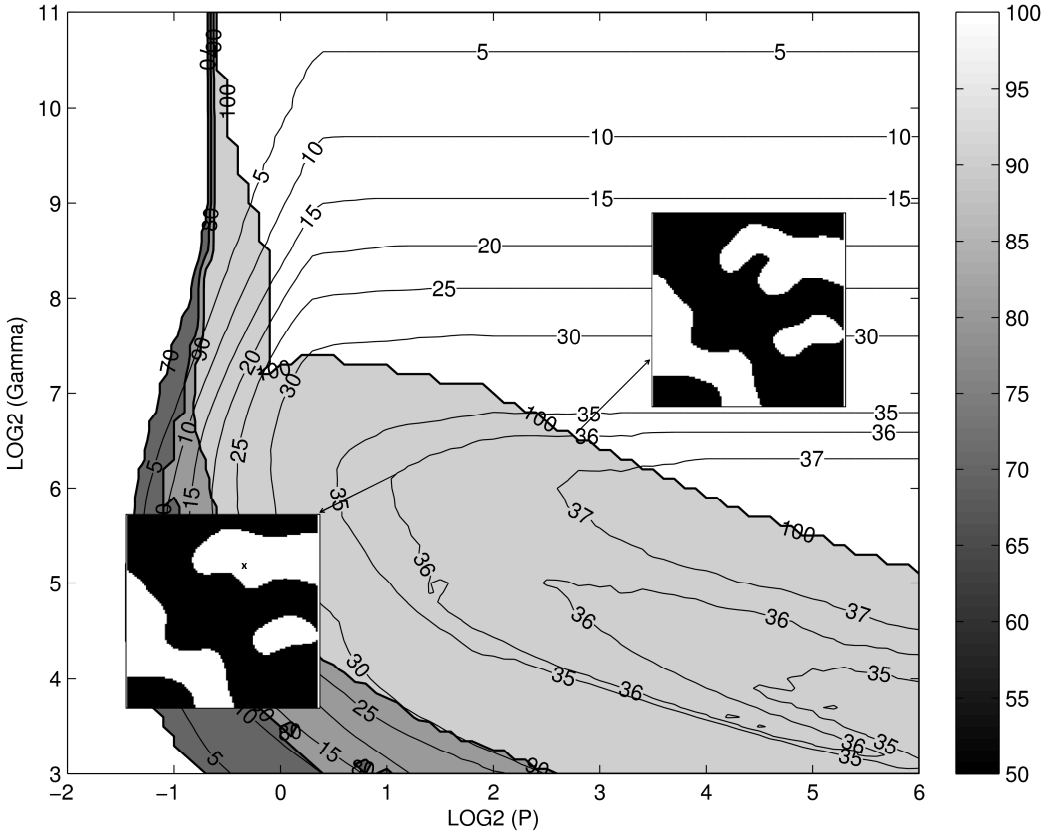

Figure 4.7: Contour lines of empirical accuracy (grey scale) and proportions of white facies. Bottom-left subplot shows the SVC solution map for the pair ($\log_2 P$, $\log_2 \gamma$) = (1.1, 6.1). The "x" identifies the misclassified location. Top-right subplot shows the SVC solution map for the pair ($\log_2 P$, $\log_2 \gamma$) = (2.8, 6.6).

4.1.5 SVC using Cross-validation

This section shows the result of the SVC algorithm trained using cross-validation. The 81x81 grid-search of parameter pairs (P, γ) is kept and 10-fold cross-validation is performed at each node to calculate the cross-validation accuracy.

Figure 4.8 shows the contour lines of the cross-validation accuracy. There are multiple points ($\log_2 P$, $\log_2 \gamma$) that produce a maximum value of 80.0% in the area bounded by the pairs (0.3, 4.8), (0.3, 3.9), (1.0, 4.8) and (1.0, 3.9). To select a single pair (P, γ) it is necessary to re-do the analysis with a finer grid. This clearly would make cross-validation computationally more expensive than the proposed technique. Moreover, when the "best" cross-validation zone is superposed with the true generalization accuracy contour map (Figure 4.3), it is evident that only generalization accuracies between 76% and 78% can be reached. In other words, in the best case the cross-validation result has 2% less generalization accuracy than the result obtained by the heuristic approach.

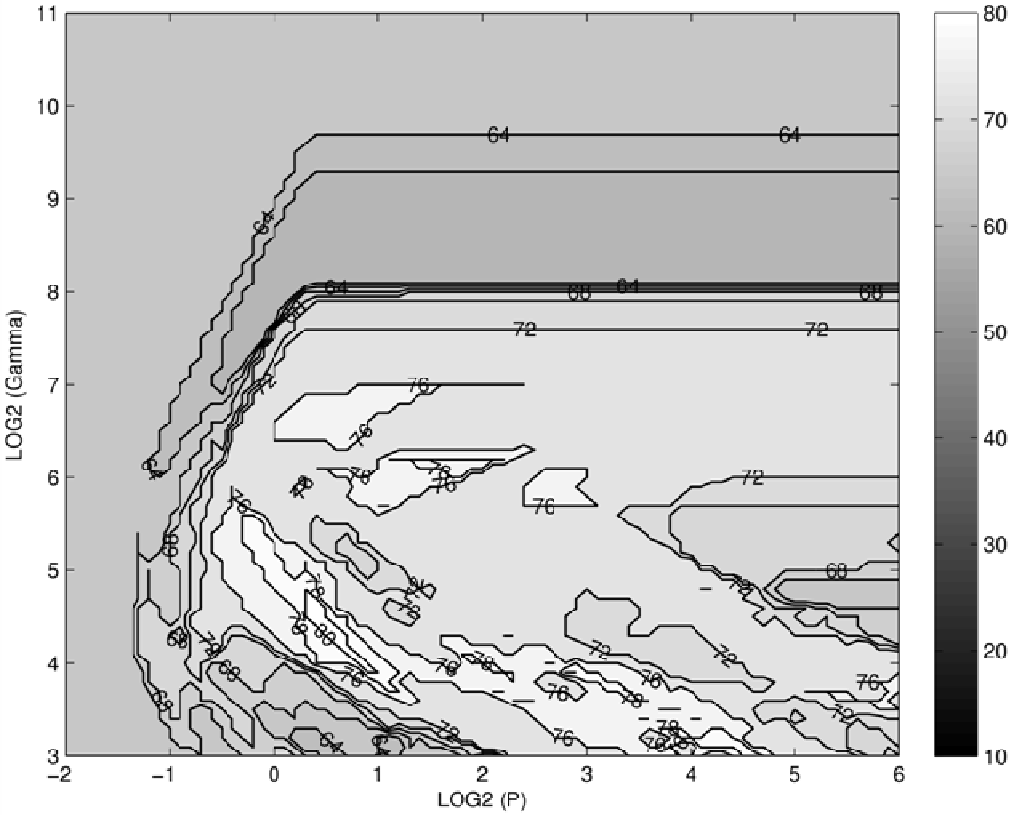

Figure 4.8: Contour lines of cross-validation accuracy. Multiple pairs ($\log_2 P$, $\log_2 \gamma$) have the maximum value of 80.0% on the area bounded by (0.3, 4.8), (0.3, 3.9), (1.0, 4.8) and (1.0, 3.9).

So far, the synthetic case has shown that the generalization accuracy of the response of a geostatistical classification model was increased by using information from the SVC solution. The next section explores the impact of incorporating such information into geostatistical indicator simulation.

4.2 Impact on Geostatistical Indicator Simulations

In the same line of reasoning as section 4.1.3, sequential indicator simulation (SIS) is used to generate two sets of 1000 realizations each. One set considers the observed data set with 50 samples, and the other, after deleting the sample misclassified by the SVC algorithm, only considers 49 samples. Both sets were created using the semivariogram model and the facies proportions of the reference data. Both sets of realizations are validated on the following bases:

- Reproduction of semivariogram model
- Reproduction of global proportions of facies, and
- Generalization accuracy

4.2.1 Sequential Indicator Simulation (SIS)

Figure 5.1 shows some examples of SIS realizations generated following the procedure described in Section 2.1.3. Clearly, it is not possible by visual inspection to conclude about the quality of the realizations, objectives measures must be applied to evaluate them. The next sections are devoted to compare the two sets of realizations.

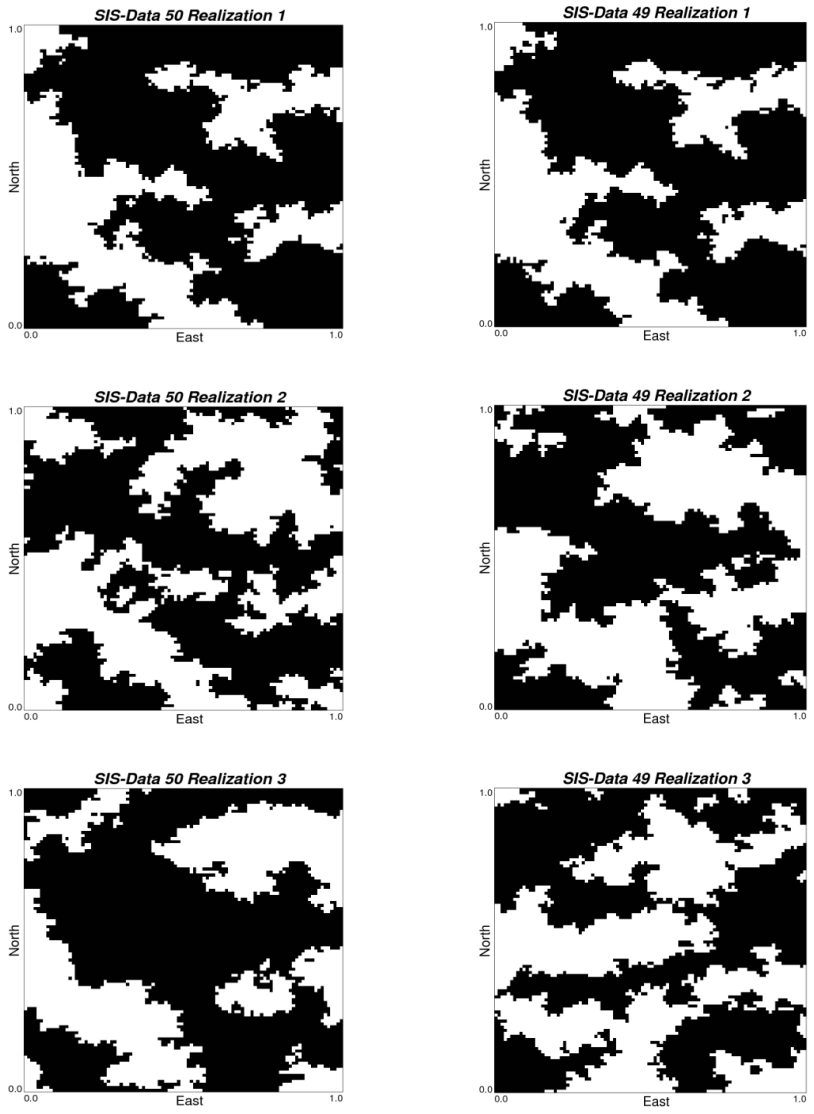

Figure 4.9: Simulated realizations generated by SIS. Realizations using 50 samples (left) and 49 samples (right).

4.2.2 Validation of Realizations

In the practice of geostatistics, the consistency of a set of simulated realizations is validated by checking that on average, the realizations honors (1) the data, (2) the global proportions of facies, and (3) the semivariogram model (Leuangthong, McLennan and Deutsch, 2003). Visually, a check of reasonableness of the models is required.

The first element of the check list is assured by the exact interpolation property of kriging. The analysis of the second and third elements is provided below. The visual

inspection is evidently subjective; instead, an analysis of the generalization accuracy is included.

Semivariogram reproduction. Figure 4.10 shows that both sets of realizations seem to similarly reproduce the indicator semivariogram model. Due to the sparse samples and the tendency of SIS to generate pixelated realizations (Deutsch, 1998), the realizations exhibit a small nugget effect that does not match exactly the indicator semivariogram model.

Arguably, in the minor direction, the set of realizations generated with 49 samples have a slightly better reproduction of the indicator semivariogram model than the set of realizations generated with 50 samples.

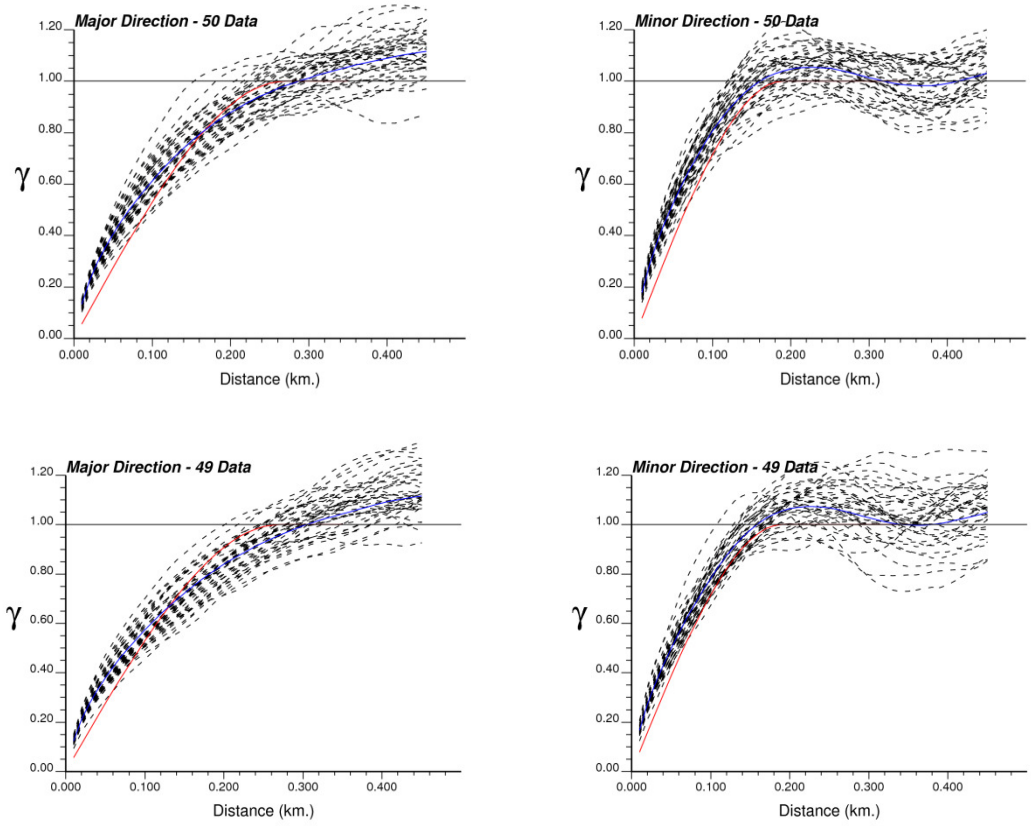

Figure 4.10: Semivariogram reproduction. The indicator semivariogram models are represented by the solid red lines. The experimental indicator semivariograms from the simulations are represented by the dashed black lines and their average indicator semivariograms by the solid blue lines.

Global proportion reproduction. Figure 4.11 shows the histograms of the global proportion of white facies for the two set of realizations under study. The arrows points to the reference proportions of white facies which is 45.75%.

The histogram generated with the model of 49 samples is slightly more accurate and precise than the histogram generated with the model of 50 samples. The former not only has on average a slightly better reproduction of the global proportions (40.4% to 39.5%) but its range is smaller.

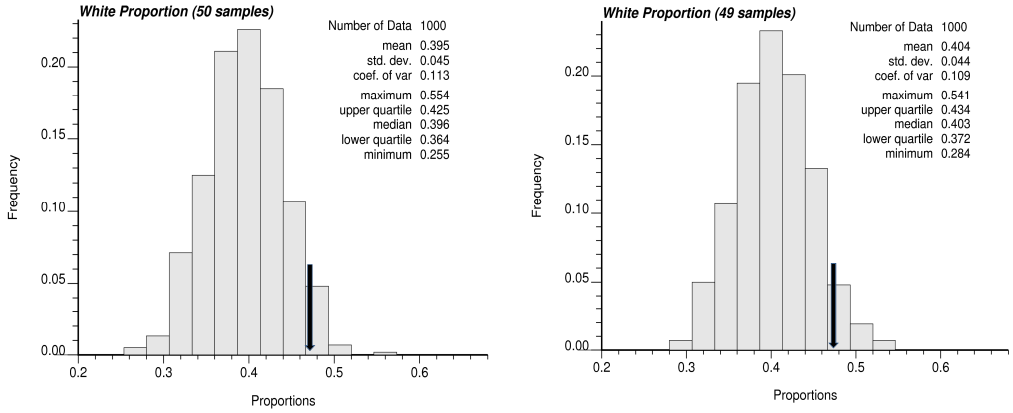

Figure 4.11: Histograms of global proportions for the white facies. The arrows points to the reference proportion of white facies, 45.75%.

Generalization accuracy. The best possible comparison between the two set of simulated realizations is based on how close every realization is to the reference map. A measure of the proximity of one realization to the truth (or reference) is what we have called the generalization accuracy.

Figure 4.12 shows the histograms and the cumulative distribution function (cdf) of the generalization accuracy for the two sets of realizations. The histograms show that the realizations from the set generated with 49 samples have, on average, slightly better generalization accuracy than the realizations generated with 50 samples (70.4% to 70.1%). Moreover, an inspection of the histogram tails indicates that in the overall the best realization belongs to the set generated with 49 samples (maximum generalization accuracy of 79.2%), while the worst realization belongs to the set generated with 50 samples (minimum generalization accuracy of 58.5%).

The superimposed cdfs plainly shows the differences between the two set of realizations. In a quantile to quantile comparison, a realization taken from the set generated with 49 samples will better explain the truth compared to a realization taken from the set generated with 50 samples.

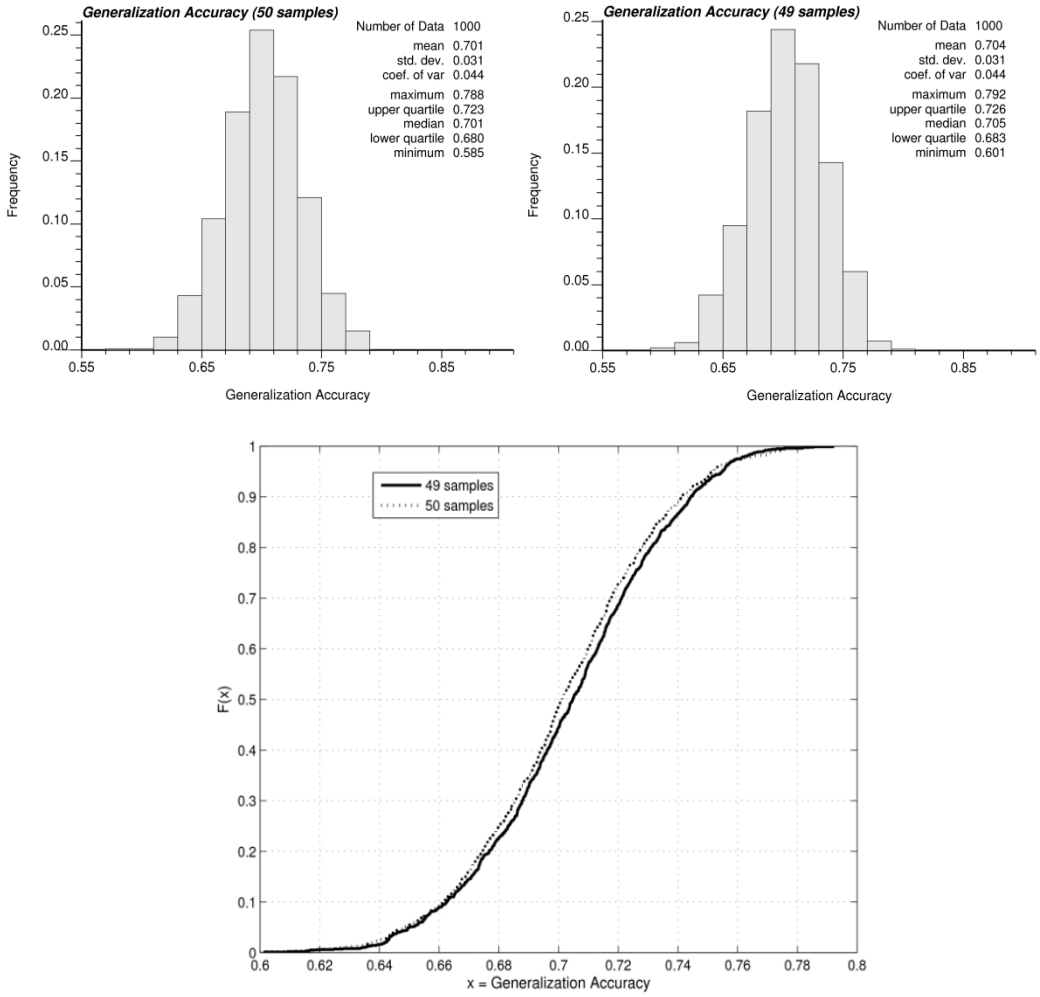

Figure 4.12: Histograms (top) and cumulative distribution functions (bottom) of the generalization accuracy for the two sets of realizations.

4.3 Discussion

Table 4.1 summarizes the numerical results of the Section 4.1. Based on it, the following statements can be made:

With “perfect information” OIK produces maps that generalize better than the maps obtained by SVC. However, the differences in performance are negligible and the SVC algorithm has the advantage of being automatic and almost fully data driven.

A good geostatistical model does not necessarily use all the available data, even when the observed data set is small. This counter-intuitive statement is supported by the results of the synthetic case presented. The responses of the geostatistical model built with 49 samples outperformed the model built with 50 samples. Information collected from the SVC solution was used to improve the response of the original geostatistical model.

The proposed heuristic parameter selection for the SVC algorithm outperformed the conventional SVC with *k*-fold cross-validation. The former produces a classifier with better generalization accuracy and it was able to reconcile the results of the SVC and OIK algorithms.

Finally, all three analyses presented in Section 4.2 suggest that the set of realizations generated with 49 samples is a better representation of the truth than the set generated with 50 samples. The former not only has a better indicator semivariogram and global proportions reproduction, it also has better generalization properties than the latter simulations.

Table 4.1: Summary of numerical results of Section 4.1

Technique	Generalization accuracy (%)	Proportion of facies (%)		Number of support vectors
		White	Black	
Reference data	N.A	45.75	54.25	N.A
Observed data (50 samples)	N.A	36.00	64.00	N.A
Observed data (49 samples)	N.A	36.73	63.27	N.A
OIK classifier (49 samples)	80.62	34.99	65.01	N.A
OIK classifier (50 samples)	80.43	34.98	65.02	N.A
SVC – knowing the reference data	80.05	34.86	65.14	45
SVC – heuristic approach (Inflection point)	79.91	36.12	63.88	43
SVC – heuristic approach (Intersection point)	79.67	36.14	63.86	46
SVC cross-validation	76 to 78	N.C	N.C	N.C

N.C: Not calculated. N.A: No apply.

5 Conclusions and Future Work

Like the responses obtained from the spatial data itself, the conclusions extracted from this research depend on the desired level of analysis. The following three perspectives, from the more general to the more specific, are worth emphasizing: (1) geostatistics and machine learning fields; (2) indicator kriging and support vector classification algorithms, and (3) large data set and small data set problems.

Geostatistic and machine learning fields: Geostatistics is a field of applied statistics that has found a well deserved place in resource sectors where the analysis of spatial data is important, for instance, mining, petroleum and environmental industries. Geostatistics has its foundations on random function theory and it responses are highly influenced by the expertise of the practitioner. The last feature offers flexibility and the opportunity of introducing in the model intangible knowledge beyond the observed data; the cost is the inclusion of subjectivities to the model that might not be positively appreciated.

Machine learning is a branch of statistic and computer science with a very broad area of application, for instance; speech and hand-writing recognition, medical diagnosis, bioinformatics, etc. Machine learning looks for dependence between variables using a limited number of observations. Its mathematical foundation is based on statistical learning theory and its responses are data driven. Once an algorithm has been chosen for the problem at hand, there is little room for subjectivities compared to geostatistics. The last characteristic allows repeatability of the response with little influence of the practitioner; the cost is less flexibility to introduce ancillary information.

The practical examples presented in Chapters 3 and 4 illustrate how the proposed geostatistical-machine learning approach allows the creation of models for spatial data with a reasonable equilibrium between the subjectivity and flexibility of the modeler dependent geostatistical approach and the objectivity and rigidity of the data driven machine learning approach.

Indicator kriging and support vector classification algorithms: Indicator kriging and support vector classification are the basic algorithms implemented in this research to solve the classification problem.

This research shows that if the practitioner has perfect knowledge of the reference data, he can construct an IK model that slightly outperforms the SVC algorithm response in terms of generalization accuracy. The advantage of the IK over SVC resides in its flexibility to directly capture, via the indicator semivariogram model, the spatial patterns that exist in the reference data. The SVC with an isotropic Gaussian radial basic function kernel is an universal classifier that allows easy automation with only two parameters to tune. The cost of such simplification is that capturing information beyond the observe data set is not straightforward.

In spite of the results obtained here, two practical facts must be considered in favour of SVC. First, the practitioner does not know the reference data, so IK will not always outperform SVC. Second, the SVC algorithm is fully data driven and automatic.

Chapter 4 showed that a (hybrid) IK-SVC model outperformed the conventional IK approach. Even more important, the improved classification model was critical to generate sets of geostatistical realizations that, taken altogether, are a better representation of the reality than the set generated without information from the SVC response.

Large and small data sets: Chapter 3 showed that the response from the conventional IK approach and the SVC algorithm tends to converge when the set of observed data is large. A large data set means that the semivariogram and the calculation of representative statistics for the geostatistical model can be more easily determined.

For small data sets, Chapter 4 showed that the IK and the SVC algorithms produce different solutions to the classification problem. The difficulty to calculate reliable experimental indicator semivariograms and representative statistics increases the subjectivities in the geostatistical model. In such cases, the fully data driven SVC algorithm can be considered most robust than its counterpart IK algorithm. Here, robustness means that the response of the algorithm is less dependent on the expertise of the practitioner.

5.1 Summary of Contributions

Two new and complimentary concepts are introduced in this thesis. The first one is a novel technique for selecting the parameters of a SVC machine for spatial categorical data. The second concept is an illustration about how a (hybrid) geostatistical-machine learning model for categorical spatial data might lead to classification and simulation models with improved responses.

Heuristic SVC parameter selection: Based on a conceptual dissection of the IK and SVC algorithms; Chapters 3 and 4 introduced a methodology for SVC parameter selection when a Gaussian radial basis function kernel is used. The method uses geostatistical criteria to select the SVC pair of parameters (P, γ) and it looks for a convergence on the responses of both algorithms. The proposed method is fully automatic and it offers an easy to use tool to the geostatistician for modeling facies.

Geostatistical modeling: The SVC algorithm is data driven, it is totally automatic and its response depends on the set of observed or training data. The IK algorithm is modeler dependent; its responses depend not only on the data but on the ability of the user to extract information from them. Chapter 4 illustrated how simple information from the SVC algorithm response can be used to generate geostatistical classification and simulation models with enhanced responses.

Beyond these two very specific contributions, the author anticipates that this thesis will keep open the door for research on hybrid geostatistics-machine learning modeling that looks for an equilibrium between (1) the objective mathematical concepts of the statistical learning theory implemented in the learning machines, and (2) the solid, well established, flexible but oftentimes subjective approach of geostatistics for the analysis of spatial data.

The above conclusions and contributions allow affirming that the goals and objectives appointed for this research were met.

5.2 Future Work

The scope of this thesis is quite specific. This leaves plenty of room for expanding the research presented here towards practical implementation aspects, such as: the dimension of the problem, the anisotropy of the data, the non-representativeness of the observe data, the multicategory problem and the target proportion issue.

The dimension of the problem: The construction of 2D models has limited applications in real life problems. The practice of geostatistics for mining and petroleum purposes requires the construction of 3D models. Since the SVC algorithm was initially developed to deal with very high dimensional problems, the extension of this thesis to 3D should be straightforward. However, it is work that should be done; the results already obtained should not be extrapolated directly without further evidence.

The anisotropy of the data: Petrophysical properties often are not the same in all directions; there are some degrees of anisotropy in their behavior. Geostatistics deals with the anisotropy through the semivariogram which is modeled for orthogonal directions. The SVC algorithm response might be improved by the introduction of an anisotropic Gaussian radial basis function kernel rather than the isotropic kernel function used in this thesis. Further research would require modifying the LIBSVM software to implement the anisotropic kernel, and exploring the SVC model selection for more than two parameters.

The non-representativeness of the observe data: A preferential sampling campaign in areas with desirable petrophysical properties may lead to non-clustered sparse set of observed data exhibiting proportions that are not representative of the entire domain of interest. If there is ancillary information suggesting this situation, the proposed technique can be applied implementing a modified version of the SVC algorithm called imbalanced SVC (Abe, 2005, p.65). The SVC algorithm for imbalanced data allows the construction of a boundary classifier with asymmetric margins by selecting a different penalty parameter (P) for each category. Further research would require exploring the SVC model selection for more than two parameters.

The multicategory problem: The mathematical description of the SVC algorithm is fine for binary problems but it is limited for multicategory applications. To solve

multicategory problems several techniques have been proposed, like the one-against-one (Kreβel, 1999), the one-against-all (Bottou et al., 1994), the all-at-once (Weston and Watkins, 1998) and the error-correcting output code (Dietterich and Bakiri, 1995). An in-depth analysis of these techniques should be made to determine what is the most suitable for geostatistical applications.

The target proportion issue: The proposed technique requires the selection of a target proportion to be reproduced. This selection is made subjectively by the modeler without considering the construction of the SVC machine. Since the SVC algorithm cannot reproduce any arbitrary target proportion, the existence of the intersection point between the 100% empirical accuracy and the target proportion contour lines is not guaranteed. Further research would require exploring the conditions that should satisfied the selected target proportion in order to guarantee its reproduction by the SVC machine and the existence of the intersection point.

Bibliography

Abe, S. (2005) Support vector machine for patter classification. Springer, USA.

Alabert, F.G., and Massonnat G.J. (1990) Heterogeneity in a complex turbiditic reservoir: Stochastic modelling of facies and petrophysical variability. In 65th Annual Technical Conference and Exhibition, pp. 775–790. SPE paper # 20604.

Anguita, D., Boni, A., Ridella, S., Rivieccio, F., and Sterpi, D. (2005) Theoretical and practical model selection methods for support vector classifiers, StudFuzz 177, pp. 159–179. Springer-Verlag.

Beucher, H.,Galli, A., Le Loc'h, G., and Ravenne, C. (1993) Including a regional trend in reservoir modelling using the truncated Gaussian method. In A. Soares. (Ed.), Geostatistics-Trois, Vol.1. pp.555-566. Kluwer, Dordrecht, Holland.

Boser, B.E., Guyon, I., and Vapnik, V. (1992) A training algorithm for optimal margin classifers. In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, pp. 144-152. ACM Press.

Bottou, L., Cortes, C., Denker, J., Drucker, H., Guyon, I., Jackel, L., LeCun, Y., Muller, U., Sackinger, E., Simard, P., Vapnik, V. (1994) Comparison of classifier methods: a case study in handwriting digit recognition. In International Conference of Pattern Recognition, pp. 77-87. IEEE Computer Society Press.

Burges, C.J. (1998) A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery 2, pp. 121-167. Kluwer Academic Publishers, Boston.

Chang, C.-C. and C.-J.Lin. (2001) LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm. Retrieved 22 December 2008.

Chiles, J.P., and Delfiner, P. (1999) Geostatistics: modeling spatial uncertainty. Wiley-Interscience, New York.

Cortes, C. (1995) Prediction of generalization ability in learning machines. PhD Thesis, Department of Computer Science, University of Rochester, Rochester NY 14627.

Cortes, C., and Vapnik, V. (1995) Support vectors networks. Machine Learning, 20, pp. 273-297.

Cristianini N., and Schölkopf B. (2002) Support vector machines and kernel methods. The new generation of learning machines. AI Magazine Vol. 23 No. 3.

Cristianini, N., and Shawe-Taylor, J. (2004) Kernel methods for pattern analysis. Cambridge University Press.

Deutsch, C.V. (2002) Geostatistical reservoir modeling. Oxford University Press, New York.

Deutsch, C.V., and Journel, A.G. (1998) GSLIB: geostatistical software library and user's guide. Oxford University Press, New York.

Deutsch, C.V., and Wang, L. (1996) Hierarchical object-based modeling of fluvial reservoirs. Mathematical Geology, 28 (7), pp. 857-880.

Dietterich, T. G., and Bakiri, G. (1995) Solving multiclass learning problems via error correcting output codes. Journal of Artificial Intelligence Research, 2, pp. 263–86.

Gilardi, N., and Bengio, S. (2005) Machine learning for automatic environmental mapping: when and how?. Automatic mapping algorithms for routine and emergency

monitoring data. Report on the Spatial Interpolation Comparison (SIC2004) exercise, pp. 123-138.

Gilardi, N., and Dubois, G. (2000) Support vector regression for environmental prediction. European Conference on Geostatistics for Enviromental Applications.

Goovaerts, P. (1997) Geostatistics for natural resources evaluation. Oxford University Press, New York.

Guardiano, F., and Srivastava, M., (1993) Multivariate geostatistics: beyond bivariate moments. Soares, A. (Ed.), Geostatistics Troia '92, Vol.1, pp.133-144.

Hastie, T., Tibshirani, R., Friedman, J. (2001) The elements of statistical learning. Springer Verlag.

Holden, L., Hauge, R., Skare, O., and Skorstad, A. (1998) Modeling of fluvial reservoirs with object models. Mathematical Geology, 30 (5), pp. 473-496.

Hsu C-W., Chang Ch-Ch., and Lin C-J. (2008) A practical guide to support vector classification. Department of Computer Science. National Taiwan University, Taipei 106, Taiwan. http://www.csie.ntu.edu.tw/~cjlin. Last updated: May 21, 2008.

Isaaks, E.H. (1990) The application of Monte Carlo methods to the analysis of spatially correlated data. PhD thesis, Standford University, Standford, CA.

Isaaks, E.H., and Srivastava, R.M. (1989) An introduction to applied geostatistics. Oxford University Press, New York.

Journel, A., and Alabert, F. (1989) Non-Gaussian data expansion in the earth sciences. Terra Nova, 1 (2), pp. 123-134.

Journel, A., and Alabert, F. (1988) Focusing on spatial connectivity of extreme valued attributes: stochastic indicator models of reservoir heterogeneities. SPE paper # 18324.

Journel, A.G., and Huijbregts, C.J. (1978) Mining geostatistics. Academic Press, London.

Kanevski, M., Canu, S., Maignan, M., Wong, P., Pozdnukhov, A., Shibli, S. (2001) Support vector machines for classification and mapping of reservoir data. IDIAP Research Report. IDIAP-RR-01-04.

Kanevski, M., Wong, P., Canu, S. (2000) Environmental data mapping with support vector regression and geostatistics. IDIAP Research Report. IDIAP-RR-00-01.

Kecman, V. (2001) Learning and softcomputing. The MIT Press.

Kecman, V., T.M. Huang, M. Vogt. (2005) Iterative Single Data Algorithm for training kernel machines from huge data sets: theory and performance, StudFuzz 177, pp.255-274. Springer-Verlag, Berlin.

Keerthi, S. S. and C.J. Lin. (2003) Asymptotic behavior of support vector machines with Gaussian kernels. Neural Computation, 15(7), pp. 1667-1689.

Kreßel, U. H.-G. (1999) Pairwise classification and support vector machines. In B. Schölkopf, C. J. C. Burges, and A. J. Smola. (Eds.), Advances in kernel methods: Support vector learning, pp. 255–268. MIT Press, Cambridge, MA.

Krige, D. G. (1951) A statistical approach to some mine valuations and allied problems at the Witwatersrand. Master thesis, University of Witwatersrand.

Leuangthong O., McLennan J., and Deutsch, C. (2003) Minimum acceptance criteria for geostatistical realizations. CCG Report 2003.

Matheron, G. (1970) La theorie des variables regionalisees et ses applications. Fasc. 5, Les Cahiers du Centre de Morphologie Mathematique, Ecole des Mines de Paris, Fontainebleau.

Matheron, G., Beucher, H., De Fouquet, H., Galli, A., Guerillot, D., and Ravenne, C. (1987) Conditional simulation of the geometry of fluvio-deltaic reservoirs. SPE paper #16753.

Muller K.-R., Smola A., Ratsch G., Schölkopf B., Kohlmorgen J., and Vapnik V. (1997) Predicting time series with support vector machines. In Gerstner W., Germond A., Hasler M., and Nicoud J.-D. (Eds.), Artificial Neural Networks ICANN'97, Berlin. Springer Lecture Notes in Computer Science Vol. 1327 pp. 999–1004.

Platt J.C. (1998) Fast training of support vector machines using sequential minimal optimization. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, (Eds.), Advances in kernel methods: Support vector learning, pp. 185–208. MIT Press, Cambridge, MA.

Pozdnoukhov A., Kanevski M. (2006) Monitoring network optimisation for spatial data classification using support vector machines. Int. Journal of Environment and Pollution. Vol. 28. 20 p.

Rosenblatt, F. (1962) Principles of neurodinamics: Perceptron and theory of brain mechanism. Spartan Books, Washington D.C.

Soares, A. (1992) Geostatistical estimation of multi-phase structures. Mathematical Geology, 24(2), pp. 149-160.

Strebelle S., and Journel A.G. (2001) Reservoir modeling using multiple point statistics. SPE Annual Technical Conference and Exhibition, New Orleans, Oct. 2001. SPE paper #71324.

Vapnik, V. (1998) Statistical learning theory. Wiley, New York

Vapnik, V. (1995) The nature of statistical learning theory. Springer-Verlag, New York.

Vapnik, V. and A. Chervonenkis (1974) Theory of pattern recognition (in Russian). Moscow. Nauka.

Weston, J., and Watkins, C. (1998) Multi-class support vector machines. Technical Report CSD-TR-98-04, Royal Holloway, University of London, London.

Wohlberg, B., Tartakovsky, D.M., and Guadagnini, A. (2006) Subsurface characterization with support vector machines. IEEE transactions on geoscience and remote sensing, Vol. 44, No. 1, Jan. 2006.

Xu, W., and Journel, G. (1993) Gtsim: Gaussian truncated simulations of reservoir units in a West Tesax carbonate field. SPE paper #27412

Appendix A Matlab Code

```
% Heuristic Support Vector Classification
% By Enrique Gallardo. MSc Student
% CENTRE FOR COMPUTATIONAL GEOSTATISTICS
% UNIVERSITY OF ALBERTA
% Revised : March 20,2009

%======CLEAN WORKSPACE AND CLOSE ALL OBJECTS==============
clear all
close all
%======LOAD DATA FILES====================================

%======LOAD SAMPLE IN GSLIB FORMAT ========================
newData = importdata('data.dat',' ',5);       %Import the file "data.dat" as structure
data_m=newData.data;                          %Matrix Ax3 with coordinates and data
xapps=data_m(:,1:2);                          %Coordinates
yapps=data_m(:,3);                            %Labels or categories
clear data_m;
%==========================================================

%======LOAD REFERENCE DATA=================================
load ReferenceImage;                        %Load file "ReferenceImage.mat" with the matrix
%C. [e.g. 100x100 matrix]
load XYcoordinates;                           %Load file "XYcoordinates.mat" containing the
%variable xapp with the coordinates in GSLIB format
yapp=reshape(C,10000,1);                      %Reshape the matrix C to GSLIB format [e.g.
%10000 files and 1 column]
xapp=xapp/10000;                            %Rescale the coordinates to [0,1] if needed
%==========================================================
```

```
%=======DEFINITION OF SOME VARIABLES=======================
bestcv_cr=0;                    %Best cross-validation accuracy
bestcvs=0;                      %Best empirical accuracy
bestcv=0;                       %Best Generalization accuracy
LOGC=[];                        %Empty matrices to storage Penalty parameter C
LOGG=[];                        %Empty matrices to storage Kernel parameter Gamma
CVH_cr=[];                      %Empty matrices to storage cross-validation accuracy
CVHs=[];                        %Empty matrices to storage empirical accuracy
CVH=[];                         %Empty matrices to storage generalization accuracy
NSVs=[];                        %Empty matrices to storage Number of Support Vector
propT=0                         %Scalar to calculate global proportions
PROP=[];                        %Empty matrices to storage Global Proportions
n_c=81,                         %Number of C points in the grid-search [e.g. 81 points]
n_g=81,                         %Number of Gamma points in the grid [e.g. 81 points]
%============================================================

%========GENERAL LOOP FOR THE GRID-SEARCH===================
for log2c = -2:.1:6,              %Grid for C. [vector -2:0.1:6 has 81 points]
 for log2g = 3:0.1:11,            % Grid for Gamma. [vector 3:0.1:11 has 81 points]

  %================CROSS-VALIDATION ACCURACY===============
  cmd = ['-v 10 -c ', num2str(2^log2c), ' -g ', num2str(2^log2g)];        % see LIBSVM
  cv_cr = svmtrain(yapps,xapps,cmd);                    % See LIBSVM for use of '-v'
   if (cv_cr > bestcv_cr),                          % Select best crossvalidation accuracy
   bestcv_cr = cv_cr; bestc_cr = 2^log2c; bestg_cr = 2^log2g;
   end
  LOGC=[LOGC log2c];                                  % Storage results
  LOGG=[LOGG log2g];                                  % Storage results
  CVH_cr=[CVH_cr cv_cr];                              % Storage results
  %============================================================

  %================EMPIRICAL ACCURACY=====================
  cmd = ['-c ', num2str(2^log2c), ' -g ', num2str(2^log2g)];        % See Manual LIBSVM
  model = svmtrain(yapps,xapps,cmd);                    % See Manual LIBSVM
  [predicts_label, accuracys, dec_values] = svmpredict(yapps,xapps, model);      %
%Estimating Empirical accuracy
  cvs=accuracys(1,:);                           % Extract empirical accuracy
```

```
   Nsvt=[model.totalSV];                                % Extract number of support vectors
   if (cvs > bestcvs),                                  % Select best empirical accuracy
     bestcvs = cvs; bestcs = 2^log2c; bestgs = 2^log2g;
   end
   CVHs=[CVHs cvs];                                     % Storage results
   NSVs=[NSVs Nsvt];                                    % Storage results
   %============================================================

   %======GENERALIZATION ACCURACY USING REFERENCE DATA======
   [predict_label, accuracy, dec_values] = svmpredict(yapp,xapp, model);          %
%Estimating Generalization accuracy
   cv=accuracy(1,:);                                    % Extract Generalization accuracy
   if (cv > bestcv),                                    % Select best Generalization accuracy
     bestcv = cv; bestc = 2^log2c; bestg = 2^log2g;
   end
   CVH=[CVH cv];                                        % Storage results
   %============================================================

   %======GLOBAL PROPORTIONS FOR LABEL OR CATEGORY 1==========
   propT=mean(predict_label);                           %Proportions for Category 1
   PROP=[PROP propT];                                   %Storage results
%============================================================
  end
end
%===========SAVING OUTPUTS===================================
save Finalresults;                                      %Save workspace
%==========PREPARING PLOTTING ================================
CVHCROSS=reshape(CVH_cr,n_g,n_c);                       %Preparing variables to Plot
CVH1=reshape(CVH,n_g,n_c);                              %Preparing variables to Plot
CVH1s=reshape(CVHs,n_g,n_c);PROP1=reshape(PROP,n_g,n_c);    %Preparing to Plot
LOGG1=reshape(LOGG,n_g,n_c);LOGC1=reshape(LOGC,n_g,n_c); %Preparing to Plot
PROP1=reshape(PROP,n_g,n_c);                            %Preparing variables to Plot
%============================================================
%==========CONTOUR PLOTS=====================================

figure,contour(LOGC1,LOGG1,PROP1);title('Global              Proportions');xlabel('LOG2
(P)');ylabel('LOG2 (Gamma)');
```

```
hold on;
contour(LOGC1,LOGG1,CVH1s);title('Empirical Accuracy');xlabel('LOG2
(P)');ylabel('LOG2 (Gamma)');
figure,contour(LOGC1,LOGG1,CVH1);title('Generalization Accuracy');xlabel('LOG2
(P)');ylabel('LOG2 (Gamma)');
figure,contour(LOGC1,LOGG1,CVHCROSS);title('Crossvalidation
Accuracy');xlabel('LOG2 (P)');ylabel('LOG2 (Gamma)');

%================================================================
```