Mathematical Geology, Vol. 34, No. 2, February 2002 (© 2002)

A New Nonparametric Discriminant Analysis Algorithm Accounting for Bounded Data Errors[1]

P. Nivlet,[2] F. Fournier,[2] and J. J. Royer[3]

In a statistical pattern recognition context, discriminant analysis is designed to classify, when possible, objects into predefined categories. Because this method requires precise input data, uncertainties cannot be propagated in the classifying process. In real case studies, this could lead to drastic misinterpretations of objects. A new nonparametric algorithm based on interval arithmetic has thus been developed to propagate interval-form data. They consist in calculating interval conditional probability density functions and interval posterior probabilities. Objects are eventually assigned to a subset of classes, consistent with the data and their uncertainties. The classifying model is thus less precise, but more realistic than the standard one, which we prove on a real case study.

KEY WORDS: pattern recognition, interval arithmetic, rock typing, borehole data.

INTRODUCTION

Supervised pattern recognition covers a large range of applications in Earth Sciences. In reservoir characterization for petroleum exploration and production, it is commonly applied to the interpretation of borehole records and to the analysis of seismic data (Dequirez and others, 1995). In both cases, a set of indirect measurements of the reservoir characteristics is recorded over a wide spatial domain. Conversely, few direct data are available mainly in the vicinities of wells, such as cores for lithological and sedimentological description, or laboratory petrophysical data. The observed information of the reservoir characteristics is used to train pattern recognition algorithms for classifying an "unknown" set of indirect data to predefined categories of reservoir properties.

[1] Received 18 September 2000; accepted 28 March 2001.

[2] Division Géophysique et Instrumentation, Institut Français du Pétrole, 92500 Rueil-Malmaison, France; e-mail: philippe.nivlet@ifp.fr or frederique.fournier@ifp.fr

[3] Computer Science Department, CNRS/CRPG/ENSG, 54501 Vandoeuvre-Les-Nancy Cedex, France; e-mail: jean-jacques.royer@ensg.inpl-nancy.fr

0882-8121/02/0200-0223/1 © 2002 International Association for Mathematical Geology

For example, different rock types are defined from cores in wells. A classification function is trained, if possible, to identify these rock types from the borehole data. If this classification function performs satisfactorily, it will be applied to predict the rock type from the recorded borehole data at depth in uncored wells. In seismic interpretation, a standard approach is to analyse the seismic character of the traces at the reservoir level, and to map its spatial variations over the entire field. In many cases, the variation in seismic facies (or character) is related to significant variation in the reservoir characteristics. When geological variation of interest is identified at wells, such as porous versus tight reservoir materials, or shale-dominated versus sand-dominated reservoir materials, seismic traces associated with typical wells are used to define the categories to train a classification function. The traces are characterized by a set of seismic attributes, and the classification function aims to predict the reservoir type from the seismic attributes values in the traces of the interwell spaces.

In this context, discriminant analysis is a powerful technique (Fisher, 1936; Fukunaga, 1972; Hand, 1981). Firstly, because it works in a probabilistic frame, probabilities of good assignment can be associated with predicted categories. These probabilities are valuable for assessing the reliability of the interpretation. Secondly, discriminant analysis provides a guide for feature selection. This is useful since many records are available, and numerous attributes can be extracted from the seismic information. Criteria based on the performance of the discriminant function help in selecting the most relevant parameters with respect to the prediction problem that should be addressed. Lastly, discriminant analysis through nonparametric algorithms allows a proper identification of patterns, even if they are nonlinear, which is common in geosciences.

In this paper, we present an extension of nonparametric discriminant analysis to account for uncertain feature measurements. This new algorithm is based on interval arithmetic. The theoretical aspects will be developed, followed by an application to electrofacies analysis from borehole measurements.

THEORETICAL ASPECTS

Discriminant Analysis

Let X be a random vector in $\mathbb{R}^p$ and $C = \{C_1, \ldots, C_N\}$, a predefined set of classes. Discriminant analysis (Fisher, 1936; Fukunaga, 1972; Hand, 1981) aims at calibrating—and estimating the efficiency of—a statistical relationship between C and X

$$C = R(X) \tag{1}$$

The classifying process is based on the Bayes rule, which estimates the posterior probability to assign an observation $\boldsymbol{x}$ to the class C_i

$$p(C_i \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid C_i)p(C_i)}{\sum_{j=1}^{N} p(\boldsymbol{x} \mid C_j)p(C_j)} \tag{2}$$

with $p(\boldsymbol{x} \mid C_i)$, the conditional probability density function (CPDF) of C_i, and $p(C_i)$, prior probability of C_i.

The CPDFs are often assumed to follow a normal distribution, whose parameters are estimated on the training sample. The method is then parametric. In that context, two approaches are possible.

- Quadratic approach: the mean vectors and variance–covariance matrices are calculated independently on each training subset C_i.
- Linear approach: the variance–covariance matrices are assumed to be the same for each training subset C_i.

In many case studies little is known about the true distribution of X. When the size of training sample is sufficient, it is preferable to estimate the CPDFs with a nonparametric method. Silverman (1986) reviews various approaches that are available. Among the most popular methods, is the k-nearest-neighbor method, in which the CPDFs are inversely proportional to the distance between $\boldsymbol{x}$ and its kth nearest neighbor in the training population. The method has two drawbacks: the estimated CPDFs are nondifferentiable and they have heavy tails. An alternative method that we concentrate on here is the kernel method for estimating the CPDFs

$$p(\boldsymbol{x} \mid C_i) = \frac{1}{n_i h^p} \sum_{j=1}^{n_i} K\left(\frac{\boldsymbol{x} - \boldsymbol{x_{ij}}}{h}\right) \tag{3}$$

with h, the smoothing parameter, K, kernel function, $\boldsymbol{x_{ij}}$, jth observation of the training sample belonging to class C_i, and n_i, size of the subpopulation in the training sample belonging to class C_i.

A particularly attractive kernel shape for minimizing the mean square error on the CPDF estimate is given by Epanechnikov (1969)

$$\begin{cases} K(\boldsymbol{u}) = \frac{p+2}{2N_p}(1 - \boldsymbol{u}^t\boldsymbol{u}) & \text{if } |\boldsymbol{u}| \leq 1 \\ K(\boldsymbol{u}) = 0 & \text{otherwise} \end{cases} \tag{4}$$

with N_p, a normalization coefficient depending on p such that $\int_{\boldsymbol{u}} K(\boldsymbol{u})\, d\boldsymbol{u}$.

Equation (4) can be rewritten with the change in variable $\boldsymbol{u} = \frac{x-x_{ij}}{h}$

$$\begin{cases} K\left(\frac{x-x_{ij}}{h}\right) = \frac{p+2}{2N_p}\left[1 - \frac{(x-x_{ij})^t(x-x_{ij})}{h^2}\right] & \text{if } |\boldsymbol{x} - \boldsymbol{x_{ij}}| \leq h \\ K\left(\frac{x-x_{ij}}{h}\right) = 0 & \text{otherwise} \end{cases} \tag{5}$$

In the following, any vector y in $\mathbb{R}^p$ will be denoted as $\boldsymbol{y} = (y^{(1)}, y^{(2)}, \ldots, y^{(p)})^t$. Thus,

$$\begin{cases} K\left(\frac{x-x_{ij}}{h}\right) = \frac{p+2}{2N_p}\left[1 - \frac{\sum_{k=1}^{p}\left(x^{(k)}-x_{ij}^{(k)}\right)^2}{h^2}\right] & \text{if } |\boldsymbol{x} - \boldsymbol{x_{ij}}| \leq h \\ K\left(\frac{x-x_{ij}}{h}\right) = 0 & \text{otherwise} \end{cases} \tag{6}$$

Once the CPDFs and posterior probabilities have been computed, the maximum likelihood rule is applied. It means that an observation $\boldsymbol{x}$ will be assigned to the class C_i which has maximum posterior probability $p(C_i \mid \boldsymbol{x})$, also called probability of good assignment. This standard discriminant analysis algorithm provides a reliable classifying model, but it fails to account for data errors both in the calibration and in the assignment phases. In this paper, we address the propagation of the uncertainties on the measurement $\boldsymbol{x}$ through the discriminant analysis process.

Uncertainties have been widely studied in the literature. Efron (1981) has compared various algorithms based on the bootstrap principle to estimate the variance of the result. He emphases errors resulting from the limited size of the calibration population. The Monte-Carlo approach, which enables the distribution of the result to be estimated, also propagates uncertainties (Ripley, 1987). However, both methods are known to underestimate the errors on the final result (Doser and others, 1998). Here, we aim to restrict the errors in the results of discriminant analysis that are due to errors in the measurement $\boldsymbol{x_i}$ of the training sample. In reality, these errors are not well known, and the standard assumption that they follow a gaussian distribution is often unrealistic. The limits of these uncertainties are usually available, however.

Incorporating these limits to a discriminant analysis algorithm means by replacing real observation $\boldsymbol{x_i}$ by interval observations $[\boldsymbol{x}_{\boldsymbol{ij}}^{-}; \boldsymbol{x}_{\boldsymbol{ij}}^{+}]$ in computing Eqs. (2)–(6). Interval arithmetic, whose principles are explained below, enables this to be done.

Interval Arithmetic

Interval arithmetic was first developed by Moore (1966) to compute interval data. Recent developments of this theory can be found in Jaulin (2000). By convention, in the following, $I(\mathbb{R})$ designates the set of real intervals, and $I(\mathbb{R}^p)$ the set of real p-dimension arrays. The minimum of the real interval $x_{[\ ]}$ is denoted x^-, and the maximum, x^+.

The four standard arithmetic operations for interval computations are: Let $x_{[\ ]} = [x^-; x^+]$ and $y_{[\ ]} = [y^-; y^+]$ be two intervals in $I(\mathbb{R}^p)$. Then,

$$\begin{cases} x_{[\ ]} + y_{[\ ]} = [x^- + y^-; x^+ + y^+] \\ x_{[\ ]} - y_{[\ ]} = [x^- - y^+; x^+ - y^-] \\ \quad x_{[\ ]}y_{[\ ]} = [\min\{x^-y^-; x^-y^+; x^+y^-; x^+y^+\};\ \max\{x^-y^-; x^-y^+; \\ \qquad\qquad x^+y^-; x^+y^+\}] \\ \quad \dfrac{1}{x_{[\ ]}} = \left[\dfrac{1}{x^+}; \dfrac{1}{x^-}\right] \quad \text{if } 0 \notin x_{[\ ]} \\ \quad \dfrac{y_{[\ ]}}{x_{[\ ]}} = y_{[\ ]} \cdot \dfrac{1}{x_{[\ ]}} \quad \text{if } 0 \notin x_{[\ ]} \end{cases} \tag{7}$$

Comparisons between two intervals are possible, but there only exists a partial order in $I(\mathbb{R}^p)$, defined by Eq. (8).

$$x_{[\ ]} \succ y_{[\ ]} \Leftrightarrow x^- > y^+ \tag{8}$$

The inclusion function $f_{[\ ]}$ of any real function f of p real variables is defined as follow:

$$\begin{aligned} f_{[\ ]} &: I(\mathbb{R})^p \rightarrow I(\mathbb{R}) \\ \boldsymbol{x}_{[\ ]} &\mapsto y_{[\ ]} \supseteq \{y = f(x) \in \mathbb{R} \mid x \in \boldsymbol{x}\} \end{aligned} \tag{9}$$

It is said to be optimal when $f_{[\ ]}(\boldsymbol{x}_{[\ ]}) = \{y = f(\boldsymbol{x}) \in \mathbb{R} \mid \boldsymbol{x} \in \boldsymbol{x}_{[\ ]}\}$.

For elementary monotonous functions, such as the exponential, the optimal inclusion function is easy to find

$$\exp_{[\ ]}(x_{[\ ]}) = [\exp(x^-);\ \exp(x^+)] \tag{10}$$

It is not so for more complex ones. However, it is possible to compute the interval extension of any function by combining elementary interval functions

[Eq. (10)] with the basic arithmetic operations [Eq. (7)]. This interval extension is called the natural inclusion function. Alefeld and Herzberger (1983) showed that the width of this interval function is usually large, as in the following simple example.

Let $f_{[\,]}(x_{[\,]}) = x_{[\,]} \cdot x_{[\,]}$; then, after Eq. (7), $f_{[\,]}([-1; 1]) = [-1; 1]$ which is true in the sense of Eq. (9), but is far from optimal. This arises because interval arithmetic computes each occurrence of a single interval variable, x, as if it were an independent variable. To overcome this, the analytical expression of the function f has to be transformed, if possible, to avoid the presence of redundant variables in its mathematical formulation. For example, the previous quadratic function would return the optimal bounds if it were written $f_{[\,]}(x_{[\,]}) = x_{[\,]}^2$.

These basic definitions are applied to Eqs. (2)–(6) to integrate errors in variables in the discriminant analysis model. These interval probabilistic objects are closely linked with the imprecise probability theory developed below.

Imprecise Probabilities

Interval probabilities, computed with the interval arithmetic rules, apparently violate Kolmogorov's total probability axiom.

Let A and B be two disjoint parts of $\boldsymbol{\Omega}$ such as $p(A \cup B) = p(\boldsymbol{\Omega}) = 1$. If $p_{[\,]}$ is an interval extension of p, the total probability axiom ($p_{[\,]}(A) + p_{[\,]}(B) = [1; 1]$) is not verified in the general case. Walley (1991) has developed the imprecise probability theory to handle such objects. The basic axioms that an interval probability $\mathrm{p}_{[\,]}$ should verify are

1. $p_{[\,]}$ is a positive-definite measure: $\forall A \in \boldsymbol{\Omega}$, $0 \leq p^-(A) \leq p^+(A) \leq 1$;
2. $p_{[\,]}$ is coherent: $\forall A_i$ independent events in $\boldsymbol{\Omega}$, there exists a probability p^* verifying the standard Kolmogorov's axioms, such as $\forall A_i \in \boldsymbol{\Omega}$, $0 \leq p^*(A_i) \leq p^*(A_i) \leq 1$

Interval probabilities generated with interval function extensions are called credal sets (Cozman, 1997a, b; Zaffalon, 1999), or envelopes of probabilities.

Interval Density Function

To develop an interval arithmetic based discriminant analysis, we first have to compute an interval extension for expression (6), which is a weighted sum of quadratic terms. Formula (11) gives the expressions for the optimal bounds of a quadratic term $Q_{ij[\,]}^{(k)}(x_{[\,]}^{(k)}, h) = (x_{[\,]}^{(k)} - x_{ij[\,]}^{(k)})^2/h^2$.

$$\left\{\begin{array}{l}
Q_{ij}^{(k)-}\left(x_{[\,]}^{(k)}, h\right) = \left\{\begin{array}{ll}
\left(\frac{x^{(k)0}-x_{ij}^{(k)-}}{h}\right)^2 & \text{if } x_{ij}^{(k)0} \leq x^{(k)0} \leq x_{ij}^{(k)-} - \left(x^{(k)0} - x^{(k)-}\right) + h \\
\left(\frac{x^{(k)0}-x_{ij}^{(k)+}}{h}\right)^2 & \text{if } x_{ij}^{(k)+} + \left(x^{(k)+} - x^{(k)0}\right) - h \leq x^{(k)0} \leq x_{ij}^{(k)0} \\
1 & \text{otherwise}
\end{array}\right. \\
\\
Q_{ij}^{(k)-}\left(x_{[\,]}^{(k)}, h\right) = \left\{\begin{array}{ll}
\left(\frac{x^{(k)-}-x_{ij}^{(k)-}}{h}\right)^2 & \text{if } x_{ij}^{(k)-} - \left(x^{(k)0} - x^{(k)-}\right) - h \leq x^{(k)0} \leq x_{ij}^{(k)-} \\
 & \quad - \left(x^{(k)0} - x^{(k)-}\right) \\
0 & \text{if } x_{ij}^{(k)-} - \left(x^{(k)0} - x^{(k)-}\right) \leq x^{(k)0} \leq x_{ij}^{(k)+} \\
 & \quad + \left(x^{(k)+} - x^{(k)0}\right) \\
\left(\frac{x^{(k)+}-x_{ij}^{(k)+}}{h}\right)^2 & \text{if } x_{ij}^{(k)+} + \left(x^{(k)+} - x^{(k)0}\right) \leq x^{(k)0} \leq x_{ij}^{(k)+} \\
 & \quad + \left(x^{(k)+} - x^{(k)0}1\right) + h \\
1 & \text{otherwise}
\end{array}\right.
\end{array}\right. \tag{11}$$

where $x^{(k)0}$ denotes the center of the interval $x_{[\,]}^{(k)}$.

Each component $x_{ij}^{(k)}$ is not independent, because of the constraint $|x - x_{ij}| < h$. We thus need the following algorithm, which returns the exact interval bounds for the interval kernel function.

```
1. $D_{\max} = h; k = 1; \boldsymbol{r} = [0;\ 0]$
2. While $D_{\max} > 0$ and $k \leq p$
        compute $Q_{ij}^{(k)-}\left(x_{[\,]}^{(k)},\ D_{\max}\right)$ and $Q_{ij}^{(k)+}\left(x_{[\,]}^{(k)},\ D_{\max}\right)$
        $\left\{\begin{array}{l} r^- \leftarrow \max\left\{r^- + Q_{ij}^{(k)-}\left(x_{[\,]}^{(k)},\ D_{\max}\right);\ 1\right\} \\ r^+ \leftarrow \max\left\{r^+ + Q_{ij}^{(k)+}\left(x_{[\,]}^{(k)},\ D_{\max}\right);\ 1\right\} \end{array}\right.$
        $\mathrm{D}_{\max} \leftarrow \mathrm{D}_{\max} - \min\left\{\left|x^{(k)0} - x_{ij}^{(k)-}\right|;\ \left|x^{(k)0} - x_{ij}^{(k)+}\right|\right\}$
        $k \leftarrow k + 1$
3. return $K_{[\,]}(\frac{\boldsymbol{x}_{[]}-\boldsymbol{x}_{ij[]}}{h}) = \frac{p+2}{2N_p}(1 - r)$
```

The CPDF calculations are then straightforward, for Eq. (3) is a sum of independent kernel functions. The optimal interval extensions of the CPDFs are then

$$\begin{aligned} p_{[\,]}(\boldsymbol{x}_{[\,]} \mid C_i) &= \frac{1}{n_i h^p} \sum_{j=1}^{n_i} K_{[\,]}\left(\frac{\boldsymbol{x}_{[\,]} - \boldsymbol{x_{ij}}_{[\,]}}{h}\right) \\ &= \left[\frac{1}{n_i h^p} \sum_{j=1}^{n_i} K^{-}\left(\frac{\boldsymbol{x}_{[\,]} - \boldsymbol{x_{ij}}_{[\,]}}{h}\right); \frac{1}{n_i h^p} \sum_{j=1}^{n_i} K^{+}\left(\frac{\boldsymbol{x}_{[\,]} - \boldsymbol{x_{ij}}_{[\,]}}{h}\right)\right] \end{aligned} \tag{12}$$

Interval Posterior Probability

As for the previous computations, implementation of intervals into Bayes rule is straightforward, once Eq. (2) is transformed so that each term only appears once

$$p(C_i \mid \boldsymbol{x}) = \left(1 + \frac{\sum_{j \neq i} p(C_j) p(x \mid C_j)}{p(C_i) p(x \mid C_i)}\right) \tag{13}$$

Using the interval arithmetic basic rules, the interval posterior probabilities are

$$\begin{aligned} p_{[\,]}(C_i \mid \boldsymbol{x}_{[\,]}) = \Bigg[&\left(1 + \frac{\sum_{j \neq i} p^{+}(C_j) p^{+}(\boldsymbol{x}_{[\,]} \mid C_j)}{p^{-}(C_i) p^{-}(\boldsymbol{x}_{[\,]} \mid C_i)}\right)^{-1}; \\ &\left(1 + \frac{\sum_{j \neq i} p^{-}(C_j) p^{-}(\boldsymbol{x}_{[\,]} \mid C_j)}{p^{+}(C_i) p^{+}(\boldsymbol{x}_{[\,]} \mid C_i)}\right)^{-1} \Bigg] \end{aligned} \tag{14}$$

The validity of formula (14), called the Generalized Bayes Rule, is proved within the scope of imprecise probabilities in the Appendix.

Assignment

Assignment follows an interval extension of the maximum likelihood rule, which we call interval dominance criterion. This means that we have to find which of the different interval posterior probabilities $p_{[\,]}(C_i \mid \boldsymbol{x}_{[\,]})$; $i = 1, \ldots, N$ is dominant. This is equivalent to comparing the intervals $p_{[\,]}(C_i) p_{[\,]}(\boldsymbol{x}_{[\,]} \mid C_i)$; $i = 1, \ldots, N$, using Eq. (8). Beforehand, we sort the values of $p^{+}(C_i) p^{+}(\boldsymbol{x}_{[\,]} \mid C_i)$ by decreasing order

$$p^{+}(C_{i1}) p^{+}(\boldsymbol{x}_{[\,]} \mid C_{i1}) \geq p^{+}(C_{i2}) p^{+}(\boldsymbol{x}_{[\,]} \mid C_{i2}) > \cdots > p^{+}(C_{iN}) p^{+}(\boldsymbol{x}_{[\,]} \mid C_{iN})$$

Then, using the basic definition for comparing intervals [Eq. (8)], if $p^-(C_{i1})$ $p^-(\boldsymbol{x}_{[\,]} \mid C_{i1}) \geq p^+(C_{i2})p^+(\boldsymbol{x}_{[\,]} \mid C_{i2})$, $\boldsymbol{x}_{[\,]}$, is assigned to C_{i1}. Otherwise, $\boldsymbol{x}_{[\,]}$ cannot be assigned to either C_{i1} and C_{i2}, and we repeat the former test with C_{i1} and $C_{i3}, \ldots, C_{i\{p\}}$, until $p^-(C_{i1})p^-(\boldsymbol{x}_{[\,]} \mid C_{i1}) \geq p^+(C_{i\{p\}})p^+(\boldsymbol{x}_{[\,]} \mid C_{i\{p\}})$.

In conclusion, the uncertainties in the observations of the training sample have been shown to be propagated in a discriminant analysis algorithm. These uncertainties in the data generate uncertainties in the CPDFs and in the posterior probabilities, which cause uncertain assignment of the observations to the different categories.

CASE STUDY

This study distinguishes between rock types for developing a detailed reservoir model in the Brent formation of the North Sea. Many wells are available to provide borehole measurements. However, only a few wells, such as Well C, provided sample cores with detailed high resolution sequential stratigraphic analysis. It is used to calibrate electrofacies automatically defined from the borehole data in terms of rock types. The calibration procedure is done conventionally with discriminant analysis. If it is satisfactory, the standard discriminant function is used to extend description of rock types to the remaining unsampled wells. First we describe the results of the standard discriminant analysis. Following, we show how the interval discriminant algorithm accounts for the uncertainties in the borehole data measurements, in predicting the rock types.

The Data

Three borehole attributes are used to describe the rock types (Fig. 1)

- Neutron porosity (NPHI);
- Bulk density (RHOB);
- Total gamma ray (SGR).

The training set comprises six types of rocks (Fig. 1), whose characteristics (mean value for each attribute and size of the training subset) are given in Table 1. On crossplots Neutron–density (Fig. 2(A)), and neutron–gamma ray (Fig. 2(B)), marine shales (diamonds), and silty shales (circles) are clearly associated with large gamma-ray values. The three classes of sandstone correspond to different porosity characteristics, from low (crosses for Sandstone II) to medium (squares for Sandstone I) and high (triangles for Sandstone III) porosities. A radioactive sandstone (inverse triangles) was calibrated as micaceous sandstone. On these crossplots, the remaining points correspond to uncored samples.

Table 1. Summary of the Training Set Characteristics

Class	Size of the training subset	NPHI (p.u.) (mean values)	RHOB (g/cm^3) (mean values)	SGR (API) (mean values)
Marine shales	42	0.310	2.594	90.18
Silt shales	71	0.229	2.554	60.71
Sandstones I	34	0.141	2.352	37.99
Sandstones II	49	0.106	2.373	23.03
Sandstones III	62	0.183	2.251	40.67
Micaceous sandstones	35	0.199	2.351	59.21

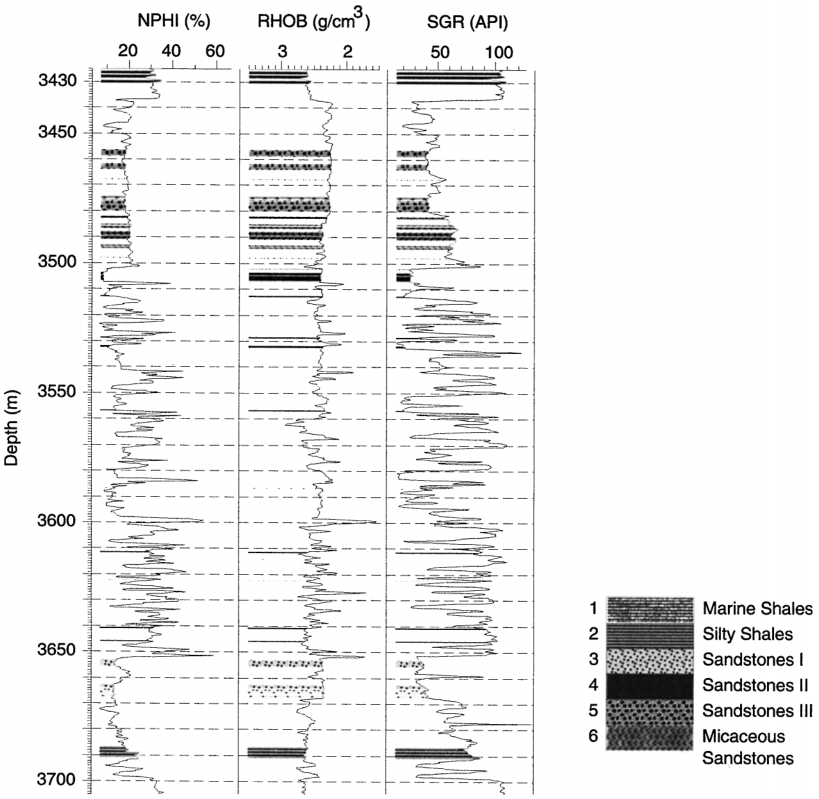

Figure 1. NPHI–RHOB–SGR measurements for well C highlighting the training sample.

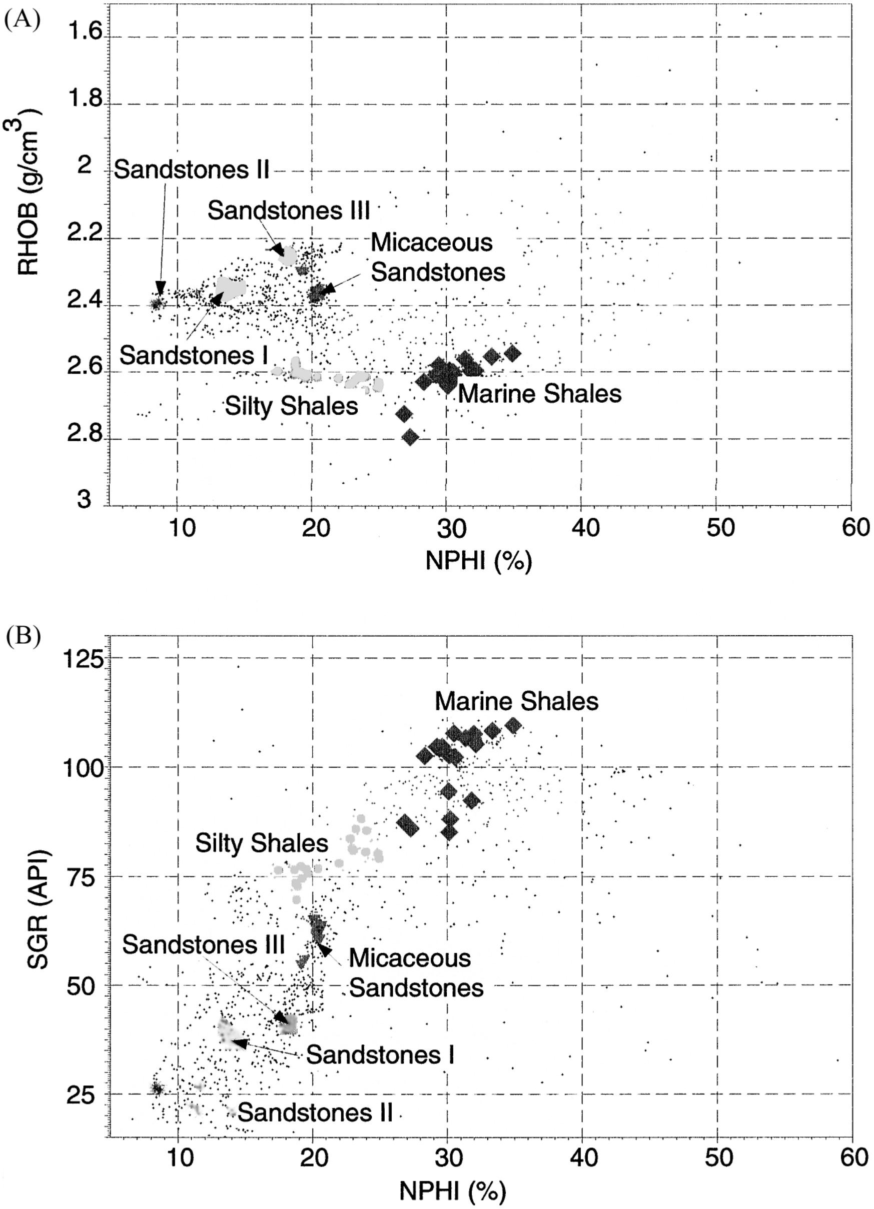

Figure 2. (A) NPHI–RHOB and (B) NPHI–SGR crossplots for well C highlighting the training sample.

Table 2. Results of the Reassignment Test on the Training Sample

	To					
			Sandstones			Micaceous
From	Marine shales	Silty shales	I	II	III	sandstones
Marine shales	100%	0%	0%	0%	0%	0%
Silty shales	0%	100%	0%	0%	0%	0%
Sandstones I	0%	0%	100%	0%	0%	0%
Sandstones II	0%	0%	0%	100%	0%	0%
Sandstones III	0%	0%	0%	0%	98%	2%
Micaceous sandstones	0%	0%	0%	0%	0%	100%

Interpretation of the Logs With Standard Discriminant Analysis

The training classes do not appear to be particularly Gaussian (see the slightly curved shape on the NPHI–RHOB diagram in Fig. 2(A)). Moreover, there are enough training points for each class. Therefore, a nonparametric approach, as the kernel method, has an advantage over the parametric one.

A reassignment test was used (Table 2) to assess the performance of the discrimination on the training set. For each training class, represented in rows, we noted the percentage of training samples reassigned to the different rock-types by the calibrated discriminant function (in columns). The performance of the discrimination, corresponding to the diagonal percentages in Table 2, is good, with an average error rate of less than 1%. The calibrated function was then used to assign all the depth samples of well C to the training classes.

Figure 3 shows as a function the predicted rock types in relation with their associated probabilities of good assignment. Between 3437 and 3500 m, a zone with thick layers of sandstone is predicted with a low to medium probability of good assignment. At the bottom of this “reservoir zone,” a series of thinner interbedded sandstone and shales is predicted to a depth of 3600 m. This zone is characterized by highly varying probability of good assignment. Above 3437 m and below 3600 m, there is mainly marine to silty shale predicted facies, corresponding to the nonreservoir formation, with a high probability of good assignment.

On the NPHI–RHOB and NPHI–SGR crossplots (Fig. 4), the symbols are coded according to the predicted rock type, but some samples (corresponding to high neutron porosities and low bulk densities) are not assigned to any class. This is due to the finite size of the kernel used in the CPDF estimation: their characteristics are far from any class centroid, and their CPDFs are null. Thus, it is not possible to evaluate the posterior probabilities for these points.

The samples assigned to marine and silty shales have high posterior probabilities. This is because these classes are well separated from the others in the

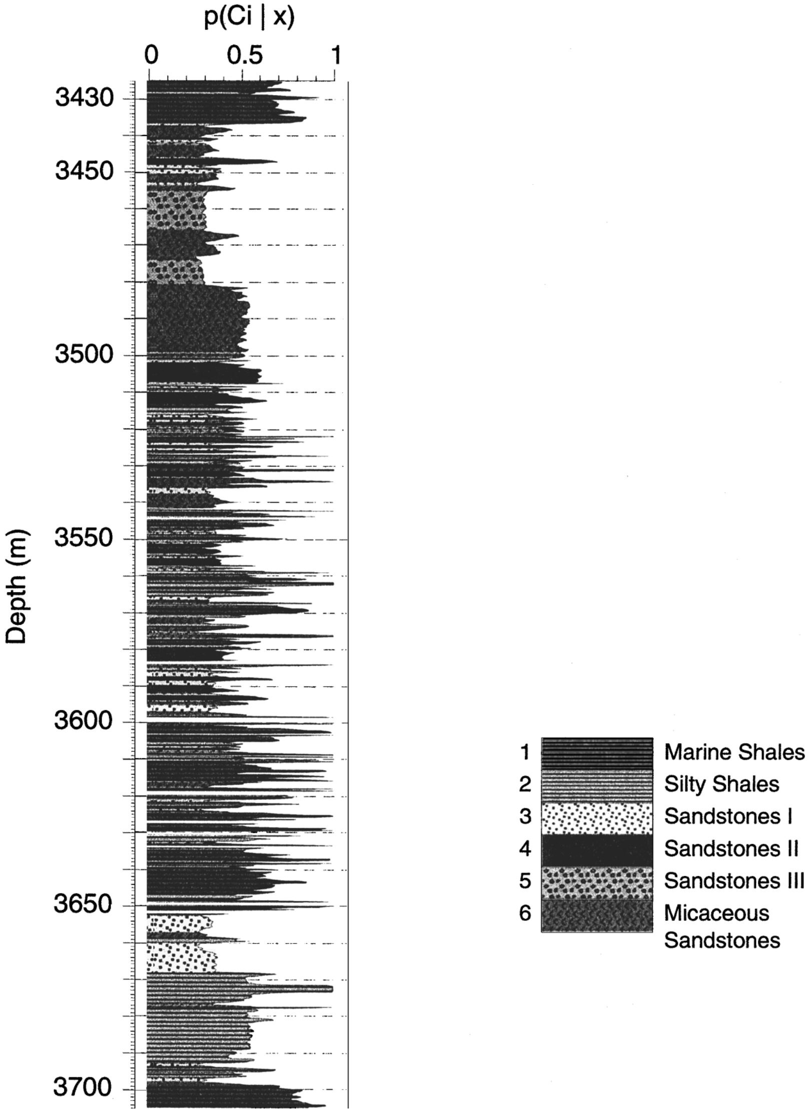

Figure 3. Probability of good assignment shown as a log with associated predicted rock type.

training sample. For the other classes, which have closer characteristics (Table 1), the posterior probabilities are lower, indicating that the assignment is less sure.

In the following, the results from the interval arithmetic discriminant analysis are compared with those from the standard discriminant analysis.

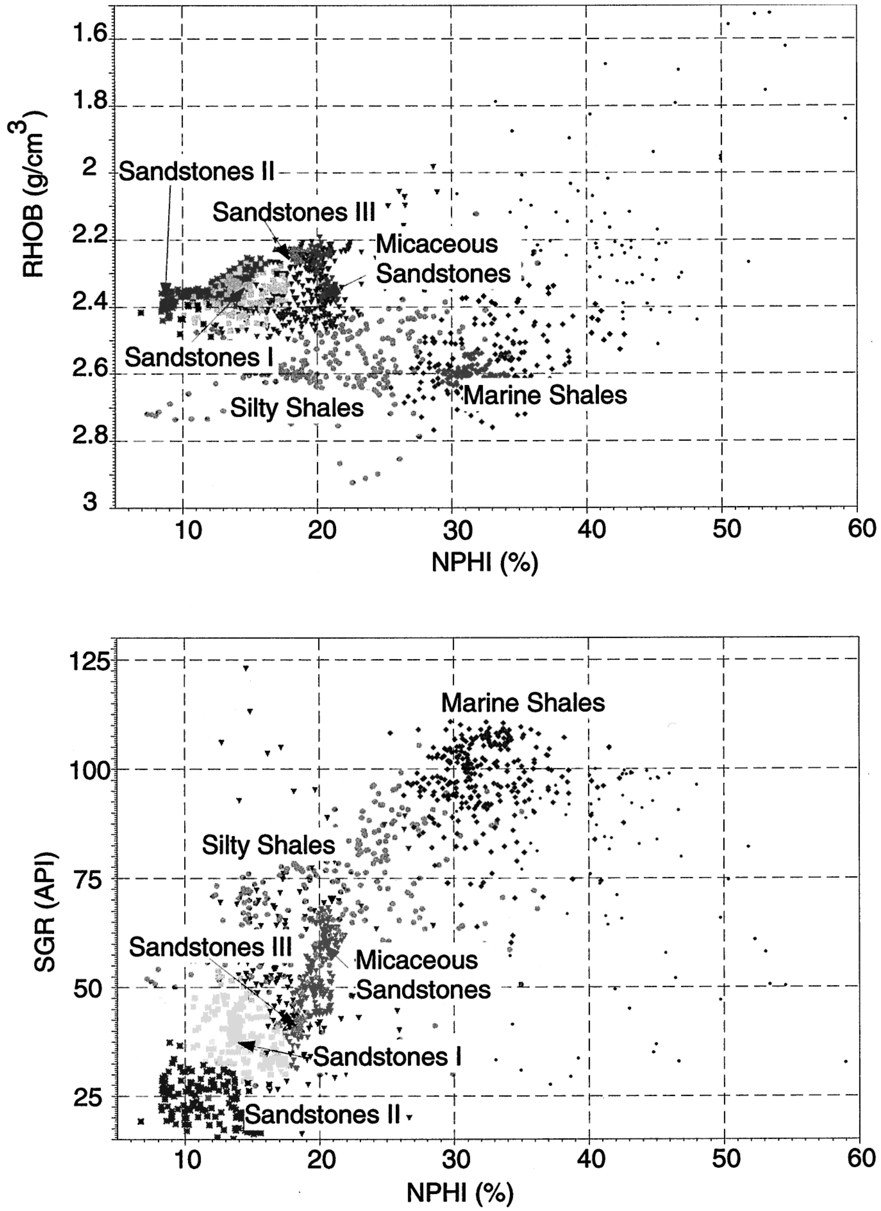

Figure 4. Predicted rock type NPHI–RHOB (top) and NPHI–SGR (bottom) crossplots for well C.

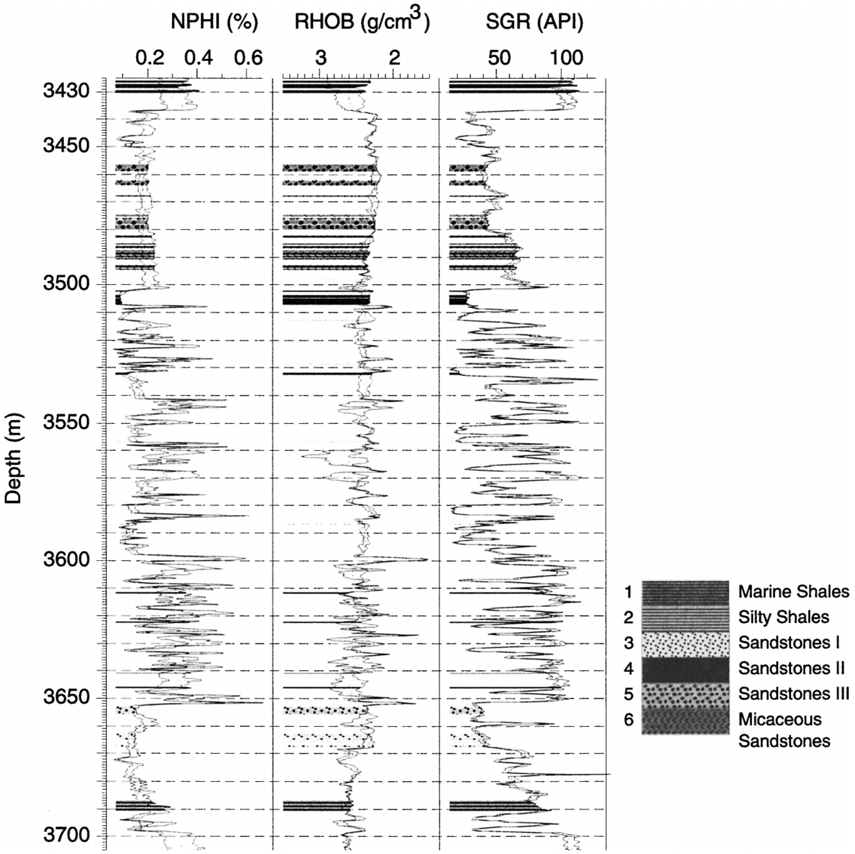

Figure 5. Measurement errors associated with NPHI–RHOB–SGR logs for well C.

Interpretation of the Logs With Interval Discriminant Analysis

Uncertainties in the three attributes (Fig. 5) integrate the nominal precision of the measuring tools which can be considered as constant for the whole data set. Typical values for these uncertainties are ±1–2% for the bulk density and neutron porosity measurement, and ±7% for gamma ray counting rates (Allen and others, 1989). In some zones (such as between 3600 and 3650 m), the in situ measuring conditions deteriorate, and this error is large.

The imprecise discriminant function is calibrated with the same depth samples as in the standard discriminant analysis. The performance of the interval function is then tested with a reassignment method. Table 3 shows the results of this test procedure, crossing calibration classes (in rows) with the predicted classes

Table 3. Results of the Reassignment Test on the Training Sample Using the Interval Calibration Process

	To					
			Sandstones			Micaceous
From	Marine shales	Silty shales	I	II	III	sandstones
Marine shales	[100, 100]%	[0, 0]%	[0, 0]%	[0, 0]%	[0, 0]%	[0, 0]%
Silty shales	[0, 3]%	[82, 100]%	[0, 0]%	[0, 0]%	[0, 0]%	[0, 15]%
Sandstones I	[0, 0]%	[0, 0]%	[100, 100]%	[0, 0]%	[0, 0]%	[0, 0]%
Sandstones II	[0, 0]%	[0, 0]%	[0, 0]%	[100, 100]%	[0, 0]%	[0, 0]%
Sandstones III	[0, 0]%	[0, 0]%	[0, 84]%	[0, 0]%	[0, 100]%	[0, 100]%
Micaceous sandstones	[0, 0]%	[0, 0]%	[0, 0]%	[0, 0]%	[0, 0]%	[100, 100]%

(in columns) for the training set. For example, the cell at the intersection of the row "silty shales" and of the column "marine shales" ([0; 3]%) indicates that between 0 and 3% of the silty shales are assigned incorrectly to marine shales. As for the standard algorithm, the highest percentages on the diagonal indicate the best performances. The width of the intervals reflects imprecision propagated through the calibration process. Two major changes are noted comparing these results with Table 2. The imprecision for the reassignment to high porosity Sandstones III, is the maximum. These samples are assigned to micaceous sandstones, or medium porosity Sandstones I. The lack of precision in reassigning the silty shales is less spectacular. For the other classes, the performance of the calibrated function remains high. At this stage of the analysis, the definition of the Sandstone III class and its use in the analysis is questionable because it is poorly identified. It was retained however, so that we can compare these with these results with those from the standard algorithm.

The first step of the assignment phase involves estimating the (interval) CPDFs for each a priori class, and each depth sample to be assigned. Envelopes of posterior probabilities (shown as interval borehole records on Fig. 6) are then calculated for the different training classes with the Generalized Bayes Rule [Eq. (14)]. Following the interval domination criterion, we carry out the interval assignment procedure. Figure 7(A) illustrates the facies imprecise prediction: it comprises of six columns corresponding to the six predefined rock types. For each depth, it indicates the one or more facies predicted by the interval analysis. As expected, the class predicted by standard discriminant analysis (Fig. 7(B)) matches at least one of these predicted by the interval algorithm. This result is due to the inclusion property [Eq. (8)] of interval arithmetic: the point sample is included in the interval sample and results from any interval arithmetic algorithm includes results from the associated point algorithm. Black colored samples in Figure 7(B) correspond to the data for which the estimated CPDF array is null. These points are outliers, and cannot be assigned; they comprise only 2% of the whole population.

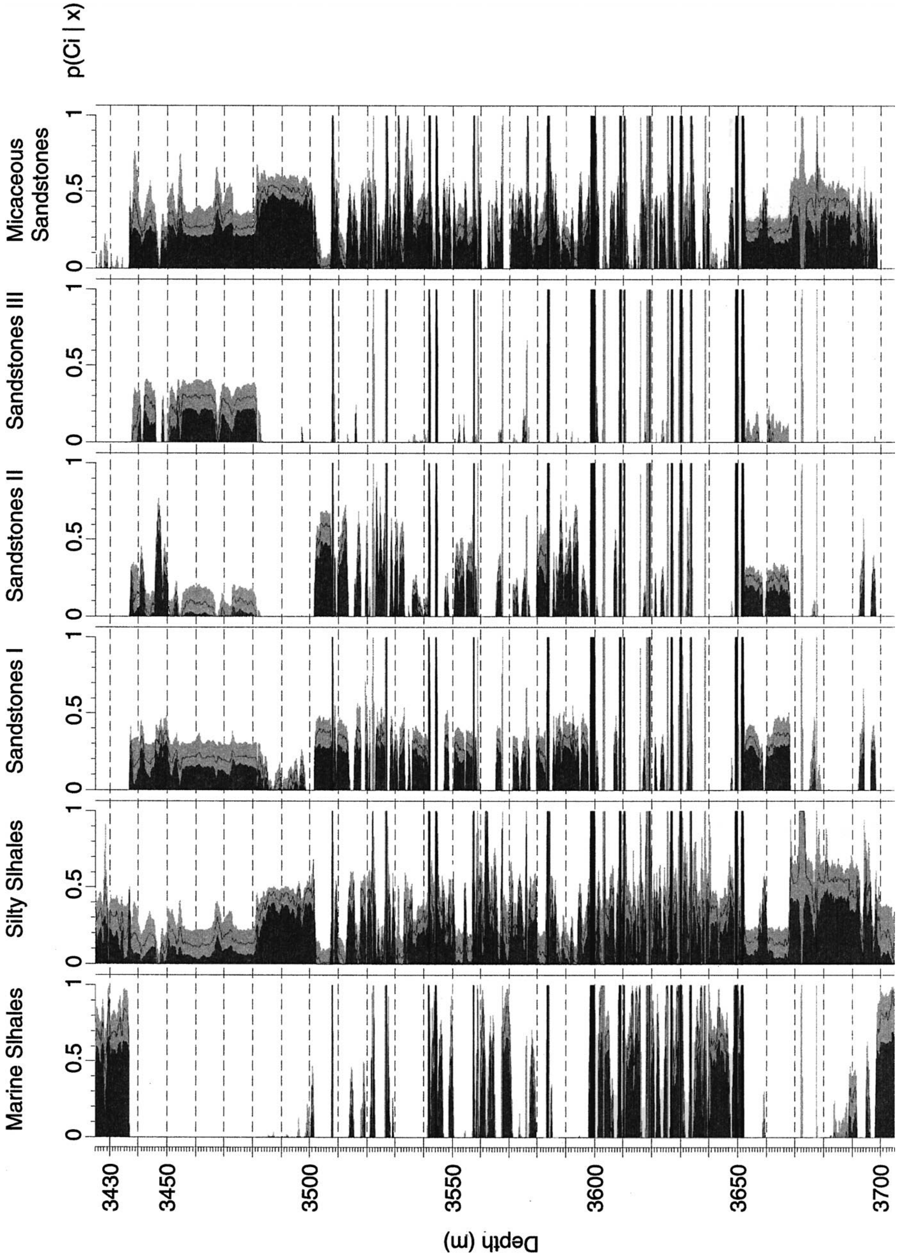

Figure 6. Interval posterior probabilities for each rock type.

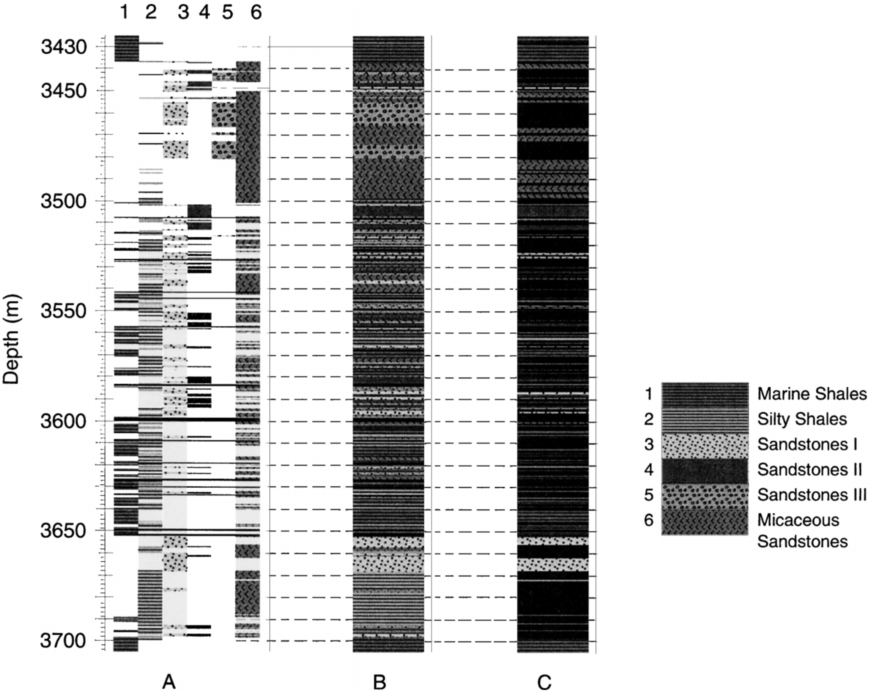

Figure 7. (A) Multiple predicted rock type by interval analysis compared with (B) predicted rock type by standard analysis, and (C) error invariant predicted rock type log.

Comparing Figure 7(A) and (B) shows that at some depths, the prediction of the rock type is not modified by the measurement errors. For example, most of the samples assigned to marine or silty shales are also assigned to the same electrofacies with the interval algorithm. On the other hand, the interval algorithm predicts multiple electrofacies for samples that were assigned to the Sandstone III class by the standard algorithm (see the interval between 3455 and 3465 m, for example, in the reservoir zone). This class is not discriminated sufficiently to generate a precise output by the interval algorithm. Figure 7(C) is another representation of the predicted facies, where black colored intervals correspond to samples that could be assigned to either no class or to several classes. A detailed analysis of this figure shows that the imprecision on the final assignment can be explained by two main factors:

- The amplitude of measurement errors: for example, in the interbedded zone, the uncertainties are higher on the measurements, and the assignment, less stable.

Table 4. Comparison Between Electrofacies Predicted by the Standard Discriminant Analysis (Displayed in Columns) and the Interval Algorithm (Displayed in Rows)

	To					
			Sandstones			Micaceous sandstones
From	Marine shales	Silty shales	I	II	III	
Marine shales	[69, 100]%	[0, 31]%	[0, 0]%	[0, 0]%	[0, 0]%	[0, 1]%
Silty shales	[0, 24]%	[19, 100]%	[0, 2]%	[0, 1]%	[0, 0]%	[0, 68]%
Sandstones I	[0, 0]%	[0, 11]%	[46, 100]%	[0, 28]%	[0, 2]%	[0, 40]%
Sandstones II	[0, 0]%	[0, 2]%	[0, 36]%	[63, 100]%	[0, 0]%	[0, 15]%
Sandstones III	[0, 0]%	[0, 1]%	[0, 90]%	[0, 1]%	[0, 100]%	[0, 100]%
Micaceous sandstones	[0, 0]%	[0, 47]%	[0, 18]%	[0, 4]%	[0, 12]%	[41, 100]%

- The measurement array to be assigned: when the measurement array is close to a boundary zone in the attribute space, a small uncertainty can cause unstability in the final result. For example, in the interval 3550–3555 m, measurement errors are quite small, but the interval algorithm cannot discriminate between the Sandstone I and III classes.

Comparison between assignments from the interval and the standard discriminant analysis is summarized in Table 4. As in Table 3, electrofacies predicted by interval analysis are in the columns, whereas rows refer to the class predicted by the standard discriminant algorithm. This table shows the extent to which the uncertainties in the initial data propagate in the final prediction. Rough Set theory (Beaubouef, Petry, and Arora, 1998; Pawlak, 1991) defines a simple way to analyze it. This theory was developed to handle databases where information is either true, false, or uncertain. The core knowledge is defined as the number of information that are true.

In discriminant analysis, the core knowledge is the minimum number of samples that remain unaffected by uncertainties (and are assigned to a single facies by the interval algorithm). It corresponds to the sum of the minimum diagonal interval elements in Table 4. This core includes 625 samples, or 45% of the whole population. The 55% remaining have imprecise assignment. The degree of imprecision of assignment to each class is also evaluated and represented in Figure 8. The percentages below represent then the proportion of samples that are precisely assigned to a given electrofacies by the interval analysis, among all the depths samples assigned to the same electrofacies by the standard analysis.

- The core of the Sandstone III class contains no element. Consequently, this class is not well-defined, which was noted in the test phase. The silty shale class is also quite imprecise, with a core proportion of 19%.

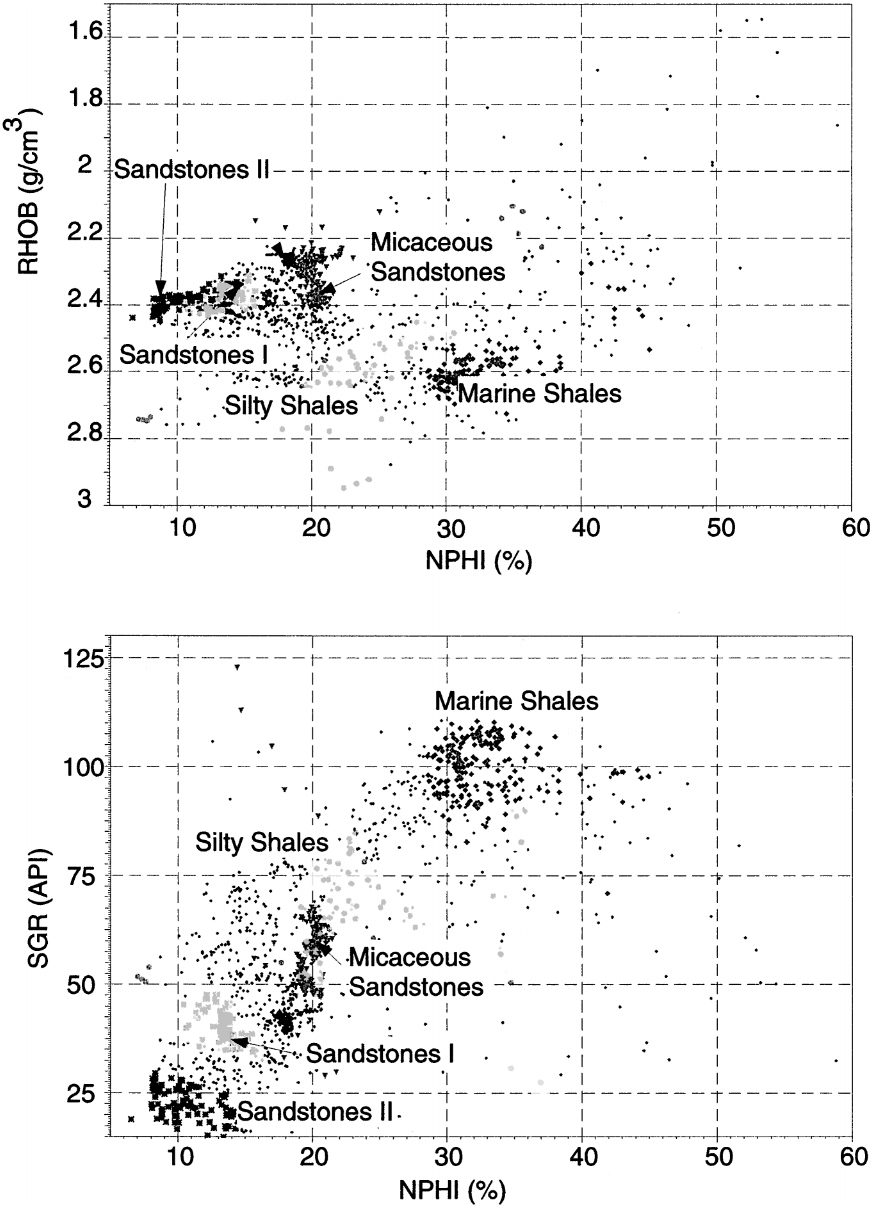

Figure 8. Extensions of the cores of knowledge for each rough rock type in the attribute space.

- The Marine shale class is defined precisely, and has a core proportion of 69%. Only 31% of these samples are possibly assigned to silty shales.
- Other classes are defined more or less precisely with a core proportion between 40 and 60%.

This analysis has shown that some classes were sensitive to the uncertainties in the raw data. The three attributes considered are not sufficient to define a stable rock type. A further analysis could be done by redefining the prior classes, or by adding new attributes in the analysis.

CONCLUSION

In the petroleum exploration context, evaluation of the quality of the reservoir is a major challenge. Within that scope supervised statistical pattern recognition techniques such as discriminant analysis are proved to be efficient. This paper has described a more general method based on interval arithmetic, taking into account measurement errors. It provides the reservoir engineer with an efficient tool to characterize the reservoir zone and the associated uncertainties. More specifically, it provides a direct quantitative assessment of the stability of the predicted reservoir pattern, given by the standard algorithm. The method does not need any hypothesis on the distribution of the errors to propagate errors. The solution given by interval analysis is always reliable because of the inclusion property. As a consequence, it gives a much more realistic interpretation of the reservoir characteristics. The example described has shown that the stability of the assignment was explained by two factors: the closeness of the measures to class boundaries and the amplitude of errors.

REFERENCES

Alefeld, G., and Herzberger, J., 1983, Introduction to interval computations (Computer Science and Applied Mathematics no. 42): Academic Press, New York, 333 p.

Allen, D., Bergt, D., Best, D., Clark, B., Falconer, I., Hache, J. M., Kienitz, C., Lesage, M., Rasmus, J., Roulet, C., and Wraight, P., 1989, Logging while drilling: Oilfield Rev., v. 1, no. 1, p. 4–19.

Beaubouef, T., Petry, F. E., and Arora, G., 1998, Information—theoretic measures of uncertainty for rough sets and rough relational databases: J. Inf. Sci., v. 109, p. 185–195.

Cozman, F. G., 1997a, An informal introduction to quasi-Bayesian theory (and lower probability, lower expectations, choquet capacities, robust Bayesian methods, etc) for artificial intelligence: Robotic Institute, Carneghie Mellon University Report. Available at http//: www.cs.cmu.edu/∼fgcozman.

Cozman, F. G., 1997b, Robustness analysis of Bayesian networks with local convex sets of distributions, *in* electronic proceedings of the 13th Annual Conference on Uncertainty and Artificial Intelligence. Available at http//: www.cs.cmu.edu/∼fgcozman.

Dequirez, P. Y., Fournier, F., Blanchet, C., Feuchtwanger, T., and Torriero, D., 1995, Integrated stratigraphic and lithologic interpretation of the East Senlac heavy oil pool, *in* SEG 65th Annual International Society of Exploration Geophysicists Meeting, Houston, October 8–13, 1995, Expanded abstracts, Chap. 1.4, p. 104–107.

Doser, D. I., Crain, K. D., Baker, M. R., Kreinovich, V., and Gerstenberger, M. C., 1998, Estimating uncertainties for geophysical tomography: Reliable Comput., v. 4, no. 3, p. 241–268.

Efron, B., 1981, Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods: Biometrika, v. 68, no. 3, p. 589–599.

Epanechnikov, V. A., 1969, Nonparametric estimate of a multivariate probability density: Theor. Probab. Appl., v. 14, p. 179–188.

Fisher, R. A., 1936, The use of multiple measurements in taxonomic problems: Ann. Eugenics, v. 7, p. 153–158.

Fukunaga, K., 1972, Introduction to statistical pattern analysis: Academic Press, New York, 591 p.

Hand, D. J., 1981, Discrimination and classification: Wiley, New York, 333 p.

Jaulin, L., 2000, Le Calcul Ensembliste par Analyse par Intervalles et ses Applications. Habilitation Thesis Disertation. Available at http://www.istia.univ-angers.fr/~jaulin.

Moore, R. E., 1966, Interval analysis: Prenctice-Hall, Englewood Cliffs, 145 p.

Pawlak, Z., 1991, Rough sets; Theoretical aspects of reasoning about data: (Series D: System theory, knowledge engineering and problem solving): Kluwer Academic Publishers, Dordrecht, 229 p.

Ripley, B. D., 1987, Stochastic simulation, Wiley series in probability and mathematical statistics: Wiley, New York, 233 p.

Silverman, B. W., 1986, Density estimation for statistical and data analysis (Monographs on Statistics and Applied Probabilities no. 26): Chapman and Hall, London, 175 p.

Walley, P., 1991, Statistical reasoning with imprecise probabilities (Monographs on Statistics and Applied Probabilities no. 42): Chapman and Hall, London, 706 p.

Zaffalon, M., 1999, A credal approach to naive classification, *in* electronic proceedings of the 1st International Sympsosium of Imprecise Probabilities and Their Applications. Available at http//:www.ensmain.rug.ac.be.

APPENDIX: COHERENCY MATCHING OF THE INTERVAL PROBABILITIES

We prove now that the interval posterior probabilities meet the coherency axioms

1. $p_{[\,]}$ is a positive definite measure: $\forall A \in \boldsymbol{\Omega}, 0 \leq p^-(A) \leq p^+(A) \leq 1$.
2. $p_{[\,]}$ is coherent: $\forall \{A_1, A_2, \ldots, A_n\} \in \boldsymbol{\Omega}$ independent, there exists a probability p^* verifying the standard Kolmogorov's axioms, such as $\forall A_i$, $p^-(A_i) \leq p^*(A_i) \leq p^+(A_i)$.

The first axiom is verified because by definition, $0 \leq p^-(x \mid C_i) \leq p^+(x \mid C_i)$. Thus,

$$\forall x, C_i, 0 \leq \left(1 + \frac{\sum_{j \neq i} p^+(C_j) p^+(x \mid C_j)}{p^-(C_i) p^-(x \mid C_i)}\right)^{-1}$$

$$\leq \left(1 + \frac{\sum_{j \neq i} p^-(C_j) p^-(x \mid C_j)}{p^+(C_i) p^+(x \mid C_i)}\right)^{-1} \leq 1,$$

which proves

$$\forall x, C_i, 0 \leq p^-(C_i \mid x) \leq p^+(C_i \mid x) \leq 1$$

To check the coherency axiom, one has to prove that for any x, their exists an N-dimensional array $(p^*(C_1 \mid x), p^*(C_2 \mid x), \ldots, p^*(C_N \mid x))^t$ such as

$$\begin{cases} p^-(C_j \mid x) \leq p^*(C_j \mid x) \leq p^+(C_j \mid x), \quad \forall j = 1, \ldots, N \\ \sum_j p^*(C_j \mid x) = 1 \end{cases} \tag{A1}$$

and which is generated by a CPDF array $(p^*(x \mid C_1), p^*(x \mid C_2), \ldots, p^*(x \mid C_N))^t$, that is, verifying the condition

$$0 \leq p^-(x \mid C_j) \leq p^*(x \mid C_j) \leq p^+(x \mid C_j), \quad \forall j = 1, \ldots, N \tag{A2}$$

To prove this, we first consider the CPDF array which generates the maximum posterior probability $p(\mathrm{C}_i \mid x)$, i being fixed

$$\begin{cases} p^*(x \mid C_i) = p^+(x \mid C_i) \\ p^*(x \mid C_j) = p^-(x \mid C_j), \quad \forall j \neq i, \end{cases} \tag{A3}$$

weighted by the N-dimensional prior probability array

$$(p^-(C_1), \ldots, p^-(C_{i-1}), p^+(C_i), p^-(C_{i+1}), \ldots, p^-(C_N))^t.$$

Let $(p^*(C_1 \mid x), p^*(C_2 \mid x), \ldots, p^*(C_N \mid x))^t$ be the posterior probability array generated by this CPDF array through the Bayes rule

$$\begin{cases} p^*(C_i \mid x) = p^+(C_i \mid x) \\ p^*(C_j \mid x) = \left(1 + \frac{\sum_{k \neq j} p^*(C_k) p^*(x|C_k)}{p^*(C_j) p^*(x|C_j)}\right)^{-1}, \quad \forall j \neq i \end{cases} \tag{A4}$$

Hence, putting (A3) into (A4), we get

$$\begin{cases} p^*(C_i \mid x) = p^+(C_i \mid x) \\ p^*(C_j \mid x) = \left(1 + \frac{p^+(C_i) p^+(x|C_i) + \sum_{k \neq j, i} p^-(C_k) p^-(x|C_k)}{p^-(C_j) p^-(x|C_j)}\right)^{-1}, \quad \forall j \neq i \end{cases} \tag{A5}$$

As $0 \leq p^{-}(x \mid C_j) \leq p^{+}(x \mid C_j)$ and $0 \leq p^{-}(C_j) \leq p^{+}(C_j)$, $\forall j = 1, \ldots, N$, Eq. (A5) can be changed to

$$\left(1 + \frac{\sum_{k \neq j} p^{+}(C_k) p^{+}(x \mid C_k)}{p^{-}(C_j) p^{-}(x \mid C_j)}\right)^{-1} \leq p^{*}(C_j \mid x) \leq \left(1 + \frac{\sum_{k \neq j} p^{-}(C_k) p^{-}(x \mid C_k)}{p^{+}(C_j) p^{+}(x \mid C_j)}\right)^{-1}, \quad \forall j \neq i$$

or, $p^{-}(C_j \mid x) \leq p^{*}(C_j \mid x) \leq p^{+}(C_j \mid x), \quad \forall j = 1, \ldots, N.$

The coherency axiom is thus verified for the maximum posterior probability $p(C_i \mid x)$. The same proof could be given to check the coherency of the minimum posterior probability $p(C_i \mid x)$, and this would remain true, which proves the coherency of the posterior probabilities.