Mathematical Geology, Vol. 38, No. 7, November 2006 (© 2006)
DOI: 10.1007/s11004-006-9052-x

Statistical Classification of Different Petrographic Varieties of Aggregates by Means of Near and Mid Infrared Spectra[1]

Vera Hofer,[2] Juergen Pilz,[3] and Thorgeir S. Helgason[4]

The increasing interest of the construction aggregates industry in reducing production costs and the costs resulting from improper use of construction materials leads to the question whether it is possible to statistically identify some rock variants by their reflectivity of near-infrared and mid-infrared light. Infrared spectroscopy allows quantitative and qualitative analysis of minerals in a reliable manner, whereas the classification of rocks is complicated by the fact that the optic behavior of minerals forming the rock often appears muted. In addition, minor constituents may dominate the spectrum. Furthermore the relevant spectra form high dimensional data, which are extremely difficult to analyse statistically, especially when curves are very similar. Common methods of multivariate statistics for this type of data, used in chemometric studies, followed by linear discriminant analysis, do not lead to acceptable classification error rates. In this paper wavelets are used in order to reduce dimensionality. As wavelets are better able to mirror local behavior of curves, they are more suitable for selecting characteristic features. The approximation is analyzed in terms of its classification properties using Mahalanobis distance or flexible discriminant analysis.

KEY WORDS: rocks; wavelets; FDA.

INTRODUCTION

Aggregates (sand, gravel and crushed rock) represent the most frequently used construction materials worldwide, employed for example in concrete, asphalt, pavements and dams. The properties of all construction materials need to be appropriate for their intended purpose. The question of whether the rock will resist physical and chemical loads, is of great importance. For instance, specific imperfections in granite result from the transformation of feldspar to kaolinite, or

[1]Received 4 April 2005; accepted 3 February 2006; Published online: 8 February 2007
[2]Department of Statistics and Operations Research, Universitätsstraße 15/E3, Karl-Franzens University, 8010 Graz, Austria; e-mail: vera.hofer@uni-graz.at
[3]Department of Mathematics, Universitätsstraße 65 – 67, University of Klagenfurt, 9020 Klagenfurt, Austria, e-mail: juergen.pilz@uni-klu.ac.at
[4]Petromodel ehf, Sidumula 1, IS-108, Reykjavik, Iceland; e-mail: thorgeir.helgason@petromodel.is

the decay of biotite, and may lead to reduced strength (Müller, 1987, p. 61). For some igneous rocks aggregate strength falls up to 40 % with increasing moisture content, which is a great problem for aggregates in unbound pavement construction (Smith and Collis, 2001, p. 253).

As rocks are increasingly being used up to the limits of their mechanical strength, material tolerances are decreasing (Müller, 1987, p. 158). This leads to the demand for ever more careful assessment of the rock particles making up the aggregates. In order to decrease the costs of damage arising from improper use of aggregates, and to substantially reduce production costs, the industry is interested in an effective and fast method of quality control. As petrological composition influences engineering properties, a reliable method for classification of aggregates is called for. Automatic means for identifying suitable rock characteristics are thus highly desirable.

The need for petrographic (petrological, lithological) discrimination and identification of aggregates has been recognize for decades and a method for this was already described in a publication by the American Society for Testing and Materials (now ASTM International) in USA in 1954 (Smith and Collis, 2001, p. 159). Typical standard test methods of today talk of petrographic examination (ASTM International, 2003) or petrographic description (CEN, 1996). Here one identifies and discriminates between different rock types of the igneous, metamorphic and sedimendary suites of rocks and to some extent the different variants of the rock types (e.g. fresh or unaltered basalt vs. hydrothermally altered basalt).

Petromodel ehf., the company working together with University of Klagenfurt on the topic presented here, is concerned with materials science, and product development is based on the hypothesis that it is possible to predict the engineering properties of aggregates, if their fundamental characteristics/properties are known. Based on various authors (Griffiths, 1967; Helgason, 1990; Mitchell, 1993), it is stated that the engineering properties P of unbound aggregates—mechanical, thermal and durability properties—are governed by fundamental properties FP of the particles and that of the surrounding pore fluid pf. The fundamental properties are taken as the petrographic composition pc, size s and shape sh. Mathematically, this context can be expressed as

$$P = f(FP, pf) = f(pc, s, sh, pf).$$

The function f can in rare cases be explained by a physical model, i.e. a causal connection. More often, the properties have to be related by statistically derived equations based on experimental data. This gives rise to a new approach for predicting and thus measuring or testing the engineering properties of construction aggregates, a method one can call statistical testing or perhaps virtual testing. This is the main role of the Techmodel® software, prepared by University of Klagenfurt for Petromodel (Petromodel, 2002).

And then, for this same materials science logic, another product, the measuring equipment Petroscope®, has been prepared in prototype or alpha version by the company in order to measure the fundamental properties pc, s and sh fast and automatically (Petromodel, 2004). The problem with today's test methods for pc, s and sh is that for some of them one needs highly trained or educated staff. Furthermore, these methods are time consuming and tedious and therefore they are much less used than their importance, as stated above and rooted in materials science, would call for. The present paper has to be seen against the background of these difficulties.

The development of the Petroscope is practiced to a large degree under the umbrella of a project called PETROSCOPE started in 2001 (EUREKA, 2001), and a second one, PETROSCOPE II started in 2005 (EUREKA, 2005), within the pan-European "network for market oriented research and development". Today most of the testing of aggregates is performed at the end of production rather than at the source of the rock or sediment extraction. With a more efficient test method, as the Petroscope measuring equipment promises to offer, testing would be expected to be used increasingly for the analysis of the raw material.

Stemming from the PETROSCOPE project, the current investigation deals with the identification of aggregates by means of their reflectivity of infrared light and using statistical classification for the discrimination of the different rock types and variants. In the first phase presented here, only a few rock types and variants have been studied. This will be extended to cover all the important rock types listed in the pan-European standard on petrographic description (CEN, 1996), i.e. at least 19 different rock types, and some different varieties of the same with different textural properties and possibly different stage of weathering of the surface. Still, with the ever increased use of crushed rock and less use of sediments, the analysis for different degree of weathering becomes less important. At the same time different frequency of light will be used, visible and near infrared.

Spectroscopic data can serve as a proper basis for classification (Hunt, 1973, 1982, 1997), but in contrast to the large amount of work done an minerals, only relatively few studies of the spectroscopic behavior of rock aggregates exist, especially with respect to the mid-infrared region (Logan and others, 1973; Thomson and Salisbury, 1993), perhaps due to the need of more laborious sampling. The problem with spectroscopic data of rocks is that the characteristic optical behavior of certain minerals that constitute the rock may appear muted, and in some cases the contributions from minor constituents dominate the spectrum (Hunt, 1982, p. 286). This causes high variability in the spectra. The appearance of the spectra is also determined by surface conditions, partical size, distribution, illumination angle, and temperature (Hunt, 1982, p. 286), which makes the interpretation of spectra challenging.

DESCRIPTION OF THE DATA

In a first study, for six rock variants, basalt slightly altered, basalt highly altered, rhyolite, two types of granite, called granite 1 (migmatite-granite) and granite 2, and gabbro, ten particles of sizes within the range of 8–32 mm from each of the six samples mentioned were collected in quarries or gravel pits in Iceland and in Finland by Lohja Rudus OY, the Geological Survey of Finland and Petromodel ehf. These samples (particles) were irradiated equidistantly with infrared light from 400 cm^{-1} to 5000 cm^{-1} (mid-infrared, MIR) and from 3500 cm^{-1} to 10000 cm^{-1} (near-infrared, NIR). The spectra were measured in reflectance mode. Due to variation in the particles, such as variable mineralogical texture, and its influence on the appearance of the spectra, three measurements of each particle were taken from different positions. These measurements, performed at VTT Electronics in Finland, resulted in three curves per particle, and therefore 180 curves all together.

Figures 1 and 2 show the spectral lines of the first of the three MIR and NIR measurements, respectively, performed on each of the 10 particles for each class (i.e. rock variant). Figure 3 then shows the mean and the standard deviation of the spectral lines for the first out of the three MIR and NIR measurements, respectively.

In general, the spectral lines of the curves of the MIR and NIR measurements seem to be very similar, although there are also some specific characteristics. This raises the statistical question, whether the differences in the shape of the curves are systematic or just random.

STATISTICAL MODEL

According to the functional data approach (Ramsay and Silverman, 1997, 2001), the data observed are considered as continuous curves, not single observations of scalars, even though the curves were measured at discrete knots and therefore represented by data vectors $\mathbf{x}_i = (x_{i1}, \ldots, x_{in})$, where n indicates the number of observation knots. In general, spectra form high dimensional data, which causes problems in applying common techniques of multivariate statistics, such as classification. As the number of samples compared to the number of observation knots is very small, the observations are highly correlated and an excess of variables is likely to cause a substantial deterioration in the classification performance (Mallet, Coomans, and de Vel, 1996, p. 158). Parameter estimates of a discriminant model become highly variable (imprecise), and in some instances may not be obtained due to numerical instability (Hastie, Tibshirani, and Buja, 1995, p. 74).

To overcome the problem of multicollinearity and diminish spurious effects, variable selection/feature reduction techniques, such as stepwise selection, basis

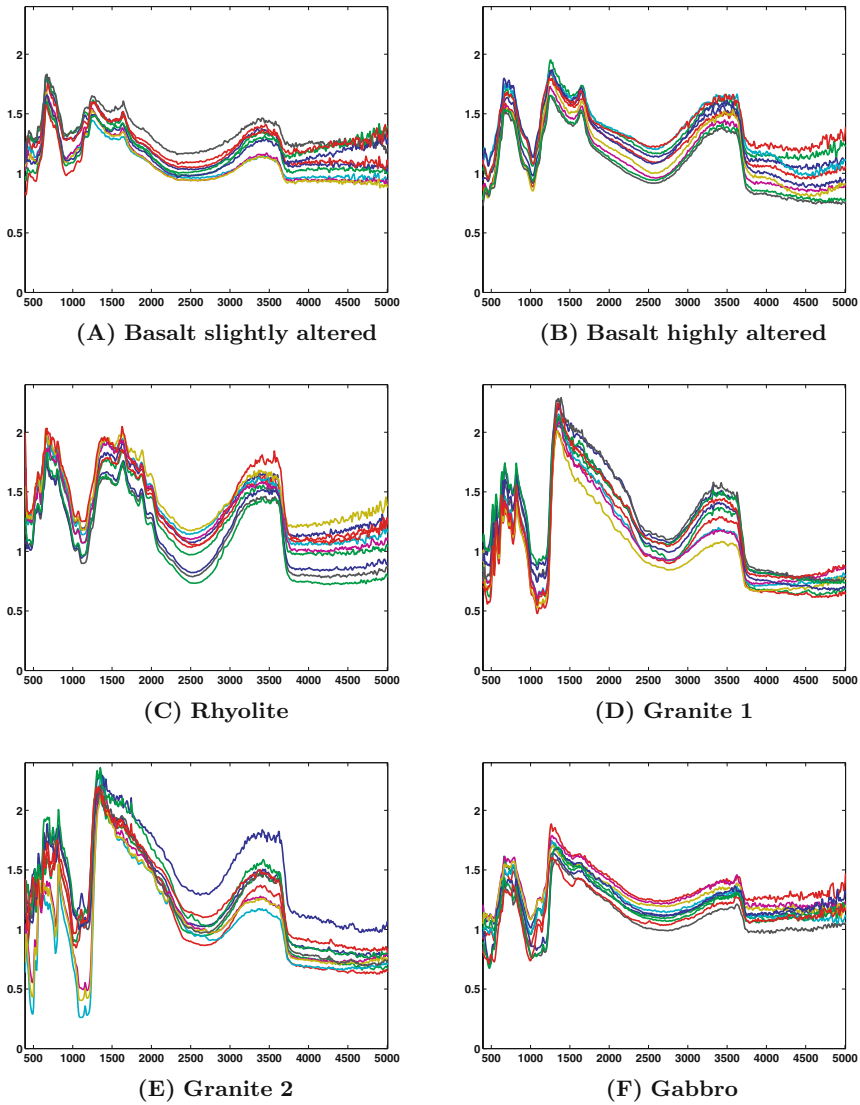

Figure 1. Reflectivity of six rock aggregates in mid-infrared (MIR) between wavenumber $400\,\mathrm{cm}^{-1}$ to $5000\,\mathrm{cm}^{-1}$, first measurement or position for each particle.

representation, principle component analysis (PCA) or partial least square estimation (PLS) (e.g. Abrahamsson and others, 2003), followed by a low-dimensional classifier such as Fisher's linear discriminant analysis (LDA) or flexible discriminant analysis (FDA) can be applied (Mallet, Coomans, and de Vel, 1996, p. 158;

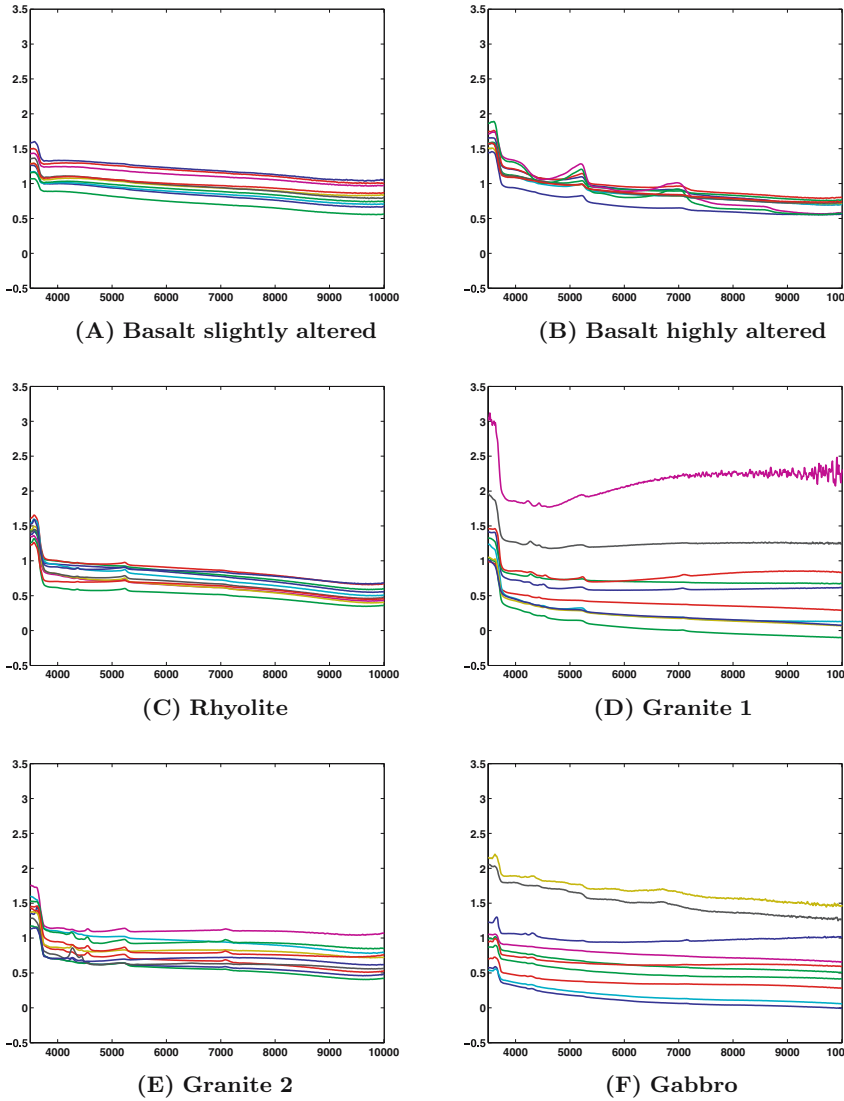

Figure 2. Reflectivity of six rock aggregates in near-infrared (NIR) between wavenumber $3500\,\mathrm{cm}^{-1}$ to $10000\,\mathrm{cm}^{-1}$, first measurement or position for each particle.

Hastie, Tibshirani, and Buja, 1994). Recently, articles on dimension reduction by means of sliced inverse regression have also been published (Li and others, 2003, Cook and Lee, 1999). As an alternative, there also exists high-dimensional

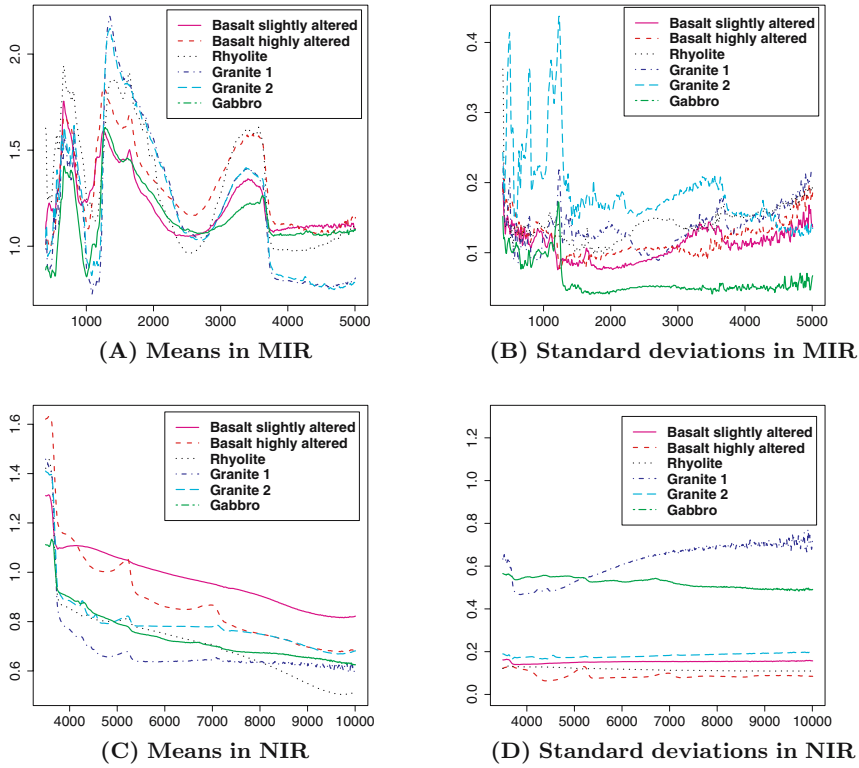

Figure 3. Mean curves and standard deviation curves from first measurement of reflectivity of mid-infrared light ((A) and (B)) and of near infrared light ((C) and (D)).

classifiers such as penalized or regularized discriminant analysis (Hastie, Tibshirani, and Buja, 1995) or support vector machines (Hastie and others, 2004).

In the present paper dimensionality is reduced using basis representations and, if necessary, PCA or PLS estimation. In addition flexible discriminant analysis (FDA) and penalized discriminant analysis (PDA) are applied to the untransformed data. A subsequent paper deals with support vector machines (Hofer, Pilz, and Helgason, submitted). Classification using support vectors is based on finding a seperating hyperplane in a high dimensional feature space by maximizing the margin between the classes. Support vector machines have become very popular due to their flexibility in determining nonlinear bounderies by constructing a linear boundery in a large, transformed version of the feature space (Hastie, Tibshirani, and Friedman, 2001). Yet, they avoid overfitting by controlling the margin between classes and sparsely representing the margin by the support vectors. A further advantage is that support vectors are derived from a convex optimization problem.

The curves observed may be described by the following model:

$$f_i(t) = g_i(t) + \varepsilon_i(t) \qquad i = 1, \ldots, n,$$

where $g(t)$ represents the systematic part, i.e. the characteristic feature that should be investigated, being determined by the petrographic composition of the particle. $\varepsilon(t)$ stands for white noise.

Dimension reduction by use of basis functions allows for the choice of a proper basis, i.e. one spanning a subspace $V \subseteq L^2(\mathbb{R})$ that contains the characteristic function $g(t)$. A projection of the curves observed onto this subspace V separates the systematic components of the signal observed from the random ones.

As information contained in the data "is lost" by projection, the choice of an appropriate subspace V, or an appropriate basis, is crucial. In the present paper wavelets are used. They possess adequate local properties and have turned out to be appropriate for statistical modeling of high dimensional data because the characteristic rock features are summarized by a few basis coefficients.

WAVELETS

Wavelets are functions that are received from a mother wavelet $\psi \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$, which satisfies the *admissibility condition*

$$C_\psi = \int_{\mathbb{R}} \frac{|\Psi(\omega)|^2}{|\omega|}\, d\omega < \infty,$$

where $\Psi(\omega)$ is the Fourier transform of ψ (Vidakovic, 1999, p. 44). The admissibility condition implies $\int \psi(x)\, dx = 0$, or in other words, $\psi(x)$ must have a zero mean. This property of the function ψ motivates the name wavelet. The deminutive comes from the fact that ψ is well localized, and by appropriate scaling, such localization can be made arbitrarily fine (Vidakovic, 1999, p. 44, Vidakovic and Müller, 1999, p. 1).

By translation and dilation of the mother wavelet ψ, we obtain the family

$$\psi_{jk}(x) = 2^{\frac{j}{2}}\, \psi(2^j x - k), \tag{1}$$

where the integers j and $k \in \mathbb{Z}$ are the spatial parameters of the wavelet transform, i.e. they control the wavelet dilation and translation, respectively. In the present paper Mallat's indexing is used (Vidakovic, 1999, p. 52). (Opposed to Mallat's indexing, Daubechies' indexing would lead to the family of functions $\psi_{jk}(x) = 2^{-\frac{j}{2}}\, \psi(2^{-j} x - k)$.)

Table 1. Relation Between Level, Scale and Resolution

Level	-2	-1	0	1	2
Scale	4	2	1	$\frac{1}{2}$	$\frac{1}{4}$
Resolution	$\frac{1}{4}$	$\frac{1}{2}$	1	2	4

The parameter j is responsible for delation and therefore plays an important role in signal processing. Two technical terms are used in this area: the factor of stretch or compression that is called scale, and the inverse of scale that is called resolution. The higher the resolution the better the approximation and the lower the scale. The relation among level j, resolution and scale is shown in Table 1. The scale shows the degree of fininess of the dyadic grid at which a signal is scanned.

The family of functions in equation (1) constitutes a basis of $L^2(\mathbb{R})$. To get a decomposition of a signal from low to high resolution, we define a scaling function, the father wavelet ϕ, that contains all low frequency components up to a specified level J. By dilation and translation we obtain the functions

$$\phi_{J_0 k} = 2^{\frac{J_0}{2}} \phi(2^{J_0} x - k) \qquad k \in \mathbb{Z},$$

which constitute an orthonormal basis of

$$V_{J_0} = span(\{\phi_{J_0 k} | k \in \mathbb{Z}\}) = span(\{\psi_{j k} | \, j < J_0 \wedge k \in \mathbb{Z}\}).$$

Thus, the family of functions $\{\phi_{J_0 k}, \, \psi_{j k} \, | j \geq J_0 \wedge k \in \mathbb{Z}\}$ constitute a new orthonormal basis of $L^2(\mathbb{R})$. Defining $W_j = span\,(\{\psi_{j k} | k \in \mathbb{Z}\})$ for a fixed level j, the space V_{j+1} can be decomposed into two orthonormal subspaces $V_{j+1} = V_j \oplus W_j$. This concept leads to a multiresolution of $L^2(\mathbb{R})$, i.e. a decomposition of $L^2(\mathbb{R})$ into orthogonal subspaces

$$L^2(\mathbb{R}) = V_J \oplus W_J \oplus W_{J+1} \oplus W_{J+2} \oplus \cdots,$$

where V_J contains low frequency components of a function below an approximation level J, and the subspaces W_{J+i} for some i contain frequency components of the signal on a corresponding higher level. A function $f \in L^2(\mathbb{R})$ can therefore be decomposed into orthogonal components

$$f(t) = A_J(t) + \sum_{j \geq J} D_j(t),$$

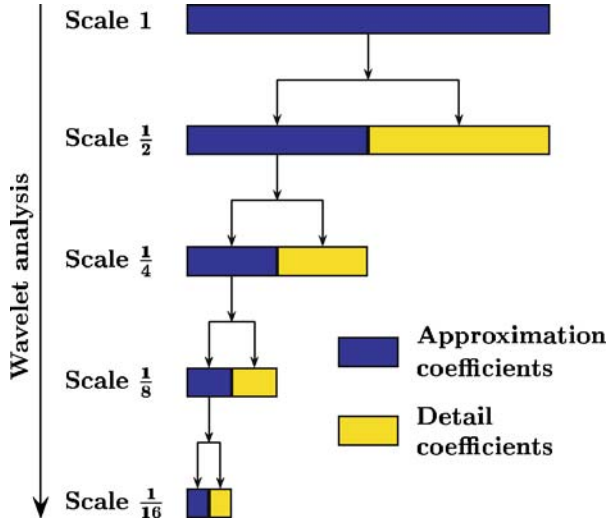

Figure 4. Principle of Wavelet and multiscalar analysis.

where

$$D_j(t) = \sum_{k \in \mathbb{Z}} d_{jk}\, \psi_{jk}(t) \qquad \text{and} \qquad A_J(t) = \sum_{j<J} D_j(t)$$

are the detail on level j, i.e. on scale 2^{-j}, and the approximation of f in V_J, respectively. Figure 4 provides an overview of the decomposition of a signal according to multiresolution analysis.

Wavelets have several advantages that make them so popular (Vidacovic and Müller, 1999, p. 1):

(i) they constitute an orthonormal basis;
(ii) they are local in time via translation, and in space via dilation, which, for example, is not true for Fourier transform;
(iii) the coefficients are a measure of the local behavior, depending on the parameters j and k;
(iv) it is very easy to apply them to functions of more than one variable.

Choice of Wavelet Basis and Parameters

No rule of thumb exists that prescribes the proper basis for the current problem. The choice depends on the data being analyzed. In the present paper, Daubechies wavelets and Symmlets with different numbers of vanishing moments

are used. The Daubechies family as well as the Symmlet family contain members ranging from highly localized to highly smooth ones. Besides this, they have some vanishing moments, which improves computational efficiency. The higher the number of vanishing moments, the more information will be concentrated in a smaller number of wavelet coefficients, since the fine scale wavelet coefficients will be essentially zero where the function is smooth. Symmlets are more symmetric than Daubechies wavelets, which makes interpretation easier (Teppola and Minkkinen, 2000).

Having chosen the wavelet type, the parameters have to be fixed. Theoretical studies regarding which parameters should be chosen to get the optimal wavelet transform or to get a robust wavelet transform already exist (e.g. Ferrando and Kolasa, 2001; Vidakovic, 1999).

In the present investigation the approximation level $J = -5$ was chosen since it turned out that PCA models following the basis transformation had low error rates in this case. The wavelet basis, the level N, corresponding to the highest resolution considered, and the components of the signal (approximation and/or detail(s)) that enter the classification method are chosen adaptively such that the classification error rate is a minimum.

CLASSIFICATION METHOD

Linear discriminant analysis (LDA) and multiple regression techniques are used for classification because they are the most frequently used methods for designing a rule to predict class membership of an item based on p measurements of predictors or features $\mathbf{X} \in \mathbb{R}^p$.

In LDA, it is assumed that the predictors have a multivariate Gaussian distribution with different means, but a common covariance matrix among the classes. According to Bayes decision rule, an observation is assigned to the class having the maximum posterior class probability. In the case of equal prior class probabilities, this rule results in assigning an observation with predictor $\mathbf{X}_0$ to the class with centroid closest to $\mathbf{X}_0$, where distance is measured in the Mahalanobis distance using the pooled within-group covariance matrix.

Altough LDA enjoys a number of favorable properties, such as reasonable robustness with respect to non-normality and to mildly different class covariances, there are two deficiencies: first, LDA is too flexible and therefore tends to overfit the data in the case of large numbers of highly correlated predictor variables. Secondly, LDA is too rigid and therefore underfits the data in situations where the class boundaries in predictor space are complex and nonlinear (Hastie, Tibshirani, and Buja, 1995).

To overcome these problems, modifications of LDA, called penalized discriminant analysis and flexible discriminant analysis, were developed (Hastie,

Tibshirani, and Buja, 1994, 1995). In penalized discriminant analysis (PDA), penalized least square regression is used to cope with high dimensional predictors. In this paper PDA is carried out using ridge regression, i.e. the squared norm of the parameter vector is added to the regression criterion. This method will be referred to as PENRIDGE below.

In flexible discriminant analysis (FDA), nonparametric regression procedures are used to estimate nonlinear class boundaries for classification. In this paper FDA is carried out using MARS and BRUTO. MARS is a procedure for adaptive, nonparametric regression (Friedman, 1991) and is well suited for high-dimensional problems (Hastie, Tibshirani, and Friedman, 2001, p. 283). The regression function is

$$f(\mathbf{x}) = \sum_{m=1}^{M} \beta_m \prod_{k=1}^{K_m} h_{k\,m}(x_{v(k,\,m)}),$$

where $x_1, \ldots, x_p$ are the predictor variables, and $v(k,\ m)$ is the index of the predictor used in the k-th term of the m-th product, M denotes the number of basis functions, and K_m stands for the number of terms in the product, defining the m-th basis function. The basis functions $h_{l\,m}$ are defined in pairs:

$$h_{k\,m}(x) = [x - t_{k\,m}]_+ \qquad h_{k\,m+1}(x) = [t_{k\,m} - x]_+,$$

for m an odd integer, where the knot value $t_{k\,m}$ is one of the unique values of $x_{v(k,\,m)}$, and "+" means positive part. The regression function is a sum of products of piecewise linear functions, constructed in a forward stepwise manner. Starting with the constant function, the basis functions are chosen to cause the greatest decrease in residual sum of squares, until a maximum model size is reached. In a backward "pruning" procedure the least important terms are removed. The fit in the stepwise procedure is measured by a generalized cross-validation criterion.

On the other hand, BRUTO is supposed to provide a set of basis functions for better class separation. BRUTO is an algorithm for estimating class membership by an additive model with adaptive selection of terms and spline smoothing parameters. It is more restrictive than the MARS model, but it deals with the predictors in a smoother fashion (Hastie, Tibshirani, and Buja, 1994, p. 15; 26). The regression function is

$$f(\mathbf{x}) = \sum_{j=1}^{p} f_j(x_j) = \sum_{j=1}^{p} \boldsymbol{\beta}_j' \mathbf{h}_j(x_j),$$

where $\mathbf{h}_j$ denotes a vector of up to N natural spline basis functions defined on the set $\{x_{1j}, \ldots, x_{Nj}\}$, $\boldsymbol{\beta}_j$ are the corresponding coefficients.

RESULTS

As preparation for classification, some basic transformations are necessary. As each particle was measured from three positions, the mean of these measurements is calculated, and a baseline correction is carried out. Petrological examination of the samples should ensure that the particles consist only of one rock type. The classifier is designed using *mean measurements*, because standard deviation declines by a factor of $\frac{1}{\sqrt{n}}$.

Classification is carried out by means of the coefficients in the basis representation of the spectra. The components of the basis representation (approximation and details), entering the classification method, are chosen adaptively such that the classification error rates are minimized. For this purpose various possible combinations of the approximation and the details on different levels are considered. The error rates are estimated using the *leave-one-out* method, i.e. for each sample a classifier is determined from the training sample that contains all samples except the sample considerd. The sample left out is then assigned to one of the classes according to the classification rule designed before. The classification error rates for assigning the three *single measurements* of the particle left out and for assigning the *mean measurement* of this particle are compared.

Details on a low scale are assumed to only contain noise and are therefore dropped. This makes sense, because the peak-to-peak magnitude of details on a low scale are in the order of 10^{-3}, whereas those on a high scale or the approximation are in the order of 10^{-1} or 1, respectively.

A further dimension reduction is achieved by a principle component analysis (PCA) or partial least square estimation (PLS) of the coefficients of the components chosen. The scores found in such a way are classified using Mahalanobis distance, FDA or PDA. The results obtained are compared with the classification results using the coefficients after a wavelet transform but without PCA or PLS (rows named *without* in Tables 2 to 4). In addition, FDA and PDA are also applied to the untransformed spectra (rows named *original* in Tables 2 to 4). As the covariance matrix estimation is not regular, LDA is additionally applied only to the original data after a PCA or PLS dimension reduction (rows named *ORIGPCA* or *ORIGPLS* in Tables 2 to 4).

Initially, observations were carried out over an extended mid-infrared region. However, as a more narrowly defined region exhibits fundamental vibrations (Guenzler and Heise, 1996, p. 29; 345), the extended region was made progressively narrower in order to better capture curve characteristics. A first restriction is to use the spectra up to 4300 cm^{-1} in order to cover the mid-infrared region

Table 2. Classification Error Rates in Different Intervals of Wavenumbers, Depending on the Method Applied to the MIR Spectra and the Optimal Basis Approximation

Method			400 cm^{-1} to 5000 cm^{-1}				400 cm^{-1} to 4300 cm^{-1}			
			5 groups		Granite		5 groups		Granite	
			Mean	Single	Mean	Single	Mean	Single	Mean	Single
LDA	Mahalanobis	PCA	0.00	0.00	0.05	0.10	0.00	0.00	0.00	0.10
		PLS	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.07
		ORIGPCA	0.00	0.00	0.15	0.22	0.00	0.00	0.15	0.22
		ORIGPLS	0.00	0.00	0.15	0.17	0.00	0.00	0.15	0.22
FDA	MARS	PCA	0.00	0.01	0.10	0.22	0.00	0.00	0.20	0.28
		PLS	0.00	0.02	0.10	0.20	0.00	0.01	0.15	0.33
		without	0.00	0.00	0.20	0.27	0.00	0.00	0.25	0.27
		original	0.00	0.02	0.25	0.30	0.02	0.01	0.20	0.30
	BRUTO	PCA	0.00	0.00	0.05	0.18	0.00	0.00	0.25	0.28
		PLS	0.00	0.00	0.05	0.18	0.00	0.00	0.10	0.23
		without	0.00	0.00	0.30	0.32	0.00	0.00	0.40	0.38
		original	0.00	0.01	0.30	0.35	0.00	0.01	0.30	0.35
PDA	PENRIDGE	PCA	0.00	0.00	0.45	0.38	0.00	0.00	0.10	0.17
		PLS	0.00	0.02	0.10	0.20	0.00	0.01	0.05	0.22
		without	0.00	0.00	0.45	0.47	0.00	0.00	0.30	0.38
		original	0.00	0.00	0.40	0.42	0.00	0.00	0.45	0.43

The classifier is designed using average measurements of the training sample. Both types of granite are combined for the *five-group* case and in a *two-group* classification both types of granite are separately classified. Classification error for assigning the mean and the single measurements are reported in the columns *Mean* and *Single*.

Table 3. Classification Error Rate in Different Intervals of Wavenumbers, Depending on the Method Applied to the MIR Spectra and the Optimal Basis Approximation

Method			400 cm^{-1} to 4000 cm^{-1}				460 cm^{-1} to 4000 cm^{-1}			
			5 groups		Granite		5 groups		Granite	
			Mean	Single	Mean	Single	Mean	Single	Mean	Single
LDA	Mahalanobis	PCA	0.00	0.00	0.05	0.05	0.00	0.00	0.00	0.05
		PLS	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.03
		ORIGPCA	0.00	0.00	0.15	0.25	0.00	0.00	0.10	0.18
		ORIGPLS	0.00	0.00	0.20	0.23	0.00	0.00	0.10	0.13
FDA	MARS	PCA	0.00	0.01	0.30	0.37	0.00	0.01	0.20	0.25
		PLS	0.00	0.01	0.20	0.23	0.00	0.01	0.10	0.18
		without	0.00	0.01	0.35	0.40	0.00	0.00	0.10	0.15
		original	0.00	0.01	0.20	0.32	0.00	0.00	0.25	0.30
	BRUTO	PCA	0.00	0.00	0.25	0.40	0.00	0.00	0.10	0.20
		PLS	0.00	0.00	0.15	0.23	0.00	0.00	0.10	0.20
		without	0.00	0.01	0.45	0.47	0.00	0.01	0.25	0.25
		original	0.00	0.01	0.30	0.35	0.00	0.01	0.25	0.32
PDA	PENRIDGE	PCA	0.00	0.01	0.50	0.53	0.00	0.00	0.25	0.23
		PLS	0.00	0.01	0.10	0.22	0.00	0.00	0.10	0.20
		without	0.00	0.01	0.40	0.48	0.00	0.00	0.25	0.25
		original	0.00	0.00	0.45	0.42	0.00	0.00	0.40	0.40

The classifier is designed using average measurements of the training sample. Both types of granite are combined for the *five-group* case and in a *two-group* classification both types of granite are separately classified. Classification error for assigning the mean and the single measurements are reported in the columns *Mean* and *Single*.

Table 4. Classification Error Rate Depending on the Method Used for NIR Spectra

Method			5 groups		Granite	
			Mean	Single	Mean	Single
LDA	Mahalanobis	PCA	0.00	0.01	0.00	0.00
		PLS	0.00	0.02	0.00	0.00
		ORIGPCA	0.03	0.05	0.00	0.00
		ORIGPLS	0.03	0.05	0.00	0.00
FDA	MARS	PCA	0.01	0.04	0.05	0.05
		PLS	0.01	0.05	0.05	0.07
		without	0.00	0.02	0.05	0.05
		original	0.03	0.04	0.05	0.08
	BRUTO	PCA	0.01	0.04	0.00	0.02
		PLS	0.00	0.03	0.00	0.00
		without	0.03	0.04	0.00	0.03
		original	0.03	0.08	0.45	0.42
PDA	PENRIDGE	PCA	0.01	0.04	0.15	0.20
		PLS	0.03	0.06	0.00	0.00
		without	0.02	0.05	0.15	0.18
		original	0.00	0.04	0.20	0.22

The classifier is designed using average measurements of the training sample. Both types of granite are aggregated for the *five-group* case and in a *two-group* classification both types of granite are separately classified. Classification for assigning the mean and the single measurements are reported in the columns *Mean* and *Single*.

and to obtain 2^n observation knots, as is required in discrete wavelet transform. In this case no padding is necessary (Vidakovic, 1999, p. 112). A further restriction is to consider a range up to 4000 cm^{-1} which according to the literature is the boundary between NIR and MIR (Guenzler and Heise, 1996, p. 345). In order to improve the classification error rates when classifying granite using LDA, and to keep the observation range as large as possible, the interval was reduced. This makes sense, because the variablity of the three measurements is high for high frequencies. This resulted in an interval starting at 560 cm^{-1}.

Among the models selected by this algorithm the most stable one is chosen, i.e. stable in the sense that the model has the least number of PC scores as possible and adding further scores does not deteriorate the error rates.

Various calculations showed that apart from the two types of granite, the classes can be identified without any error by most of the methods. This result suggests combining the samples of granite to one class, which leads to a five-group classification problem. An additional error reduction can be reached by a two step classification scheme such that first a classification of five groups is carried out. If a sample is assigned to granite, in a second step it is classified by use of a classification method appropriate for granite (Fig. 5).

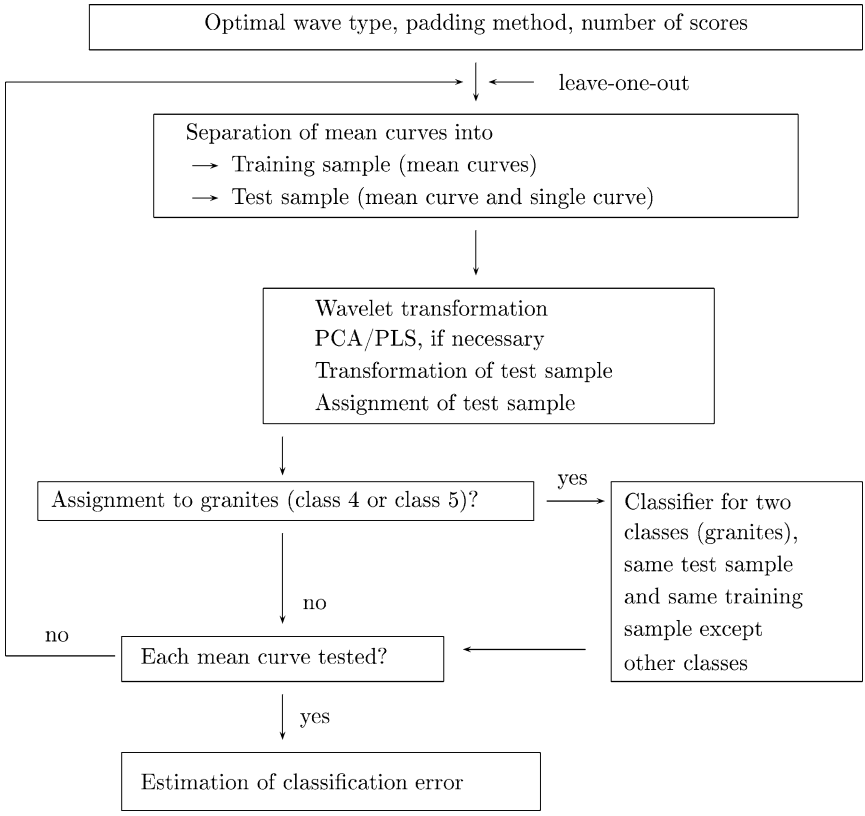

Figure 5. Classification scheme for the two-step method in the case of wavelet transform.

According to Tables 2 and 3 in the five-group classification problem (with the two granites combined) for MIR spectra low error rates are reached, no matter which classification method is applied. The reduction of the spectral range hardly affects the results. Furthermore, a principle component analysis/ partial least square estimation (PCA/PLS) is not necessary when using flexible discriminant analysis (FDA). Using a wavelet transform reduces classification error rates for granite, even in the case of FDA or PDA. When applying PCA/PLS about 8 scores suffice to obtain complete classification.

For granite, FDA leads to much greater error rates than linear discrimiant analysis (LDA). But to obtain a low classification rate with LDA 15 or 16 scores are necessary. This could be seen as an overfit. However, scores corresponding to low eigenvalues could be important when the direction of separation is orthogonal to the first PCs (Jolliffe, 1986, p. 160). Calculations showed that smaller intervals make a separation of granite more difficult. In contrast to the five-group separation,

the classification of granite using log-spectra instead of the original spectra leads to lower error rates.

In contrast to MIR, the two types of granite considered here can be identified well in NIR (Table 4). The other groups show similarities in NIR, which do not exist in MIR. This suggests that the ideal classification method is based on both, MIR and NIR measurements. Thus, in a first step a five-group classification would be carried out using MIR spectra, and if a sample is assigned to granite, a further classification can then be carried out using NIR spectra.

SUMMARY

Aggegates can be classified by means of their reflectivity of mid-infrared (MIR) and near-infrared (NIR) light in a reliable manner, especially when using several measurements of a particle. As the data investigated are curves, they can be represented by a proper wavelet basis. Wavelets mirror local behavior and are appropriate for modelling spectral lines. A wavelet-based reduction of the dimensionality of the data, combined with principle component analysis (PCA) or partial least square estimation (PLS) results in a regular covariance matrix estimation so that linear discriminant analysis (LDA) can be applied. As flexible discriminanat analysis (FDA) can cope with high dimensional data, a basis representation is not necessary to make FDA applicable, but it can entail a reduction of the classification error. However, LDA seems to be superior to FDA. Penalized discriminant analysis (PDA) turns out to be an inappropriate method owing to the unacceptably high error rates in classification. The use of PLS for estimation of the principle components appeared to be superior to the traditional PCA in the sense that the variability in measurements led to less distortion of the classification results. The wavelet model is superior to the PLS-based dimensionality reduction of the original data.

A two step classification method on the basis of MIR as well as NIR spectra leads to a very low classification error. Basalt slightly altered, basalt highly altered, rhyolite and gabbro, as well as granite as a whole, can be identified well in MIR. Having found out that a sample belongs to granite, a further identification can be obtained using NIR light.

REFERENCES

Abrahamsson, C., Johansson, J., Sparén, A., and Lindgren, F., 2003, Comparison of different variable selection methods conducted on NIR transmission measurements on intact tablets: Chemometrics and Intelligent Laboratory Systems, v. 69, p. 3–12.

ASTM International, 2003, C295-03 Standard Guide for Petrographic Examination of Aggregates for Concrete: Annual Book of ASTM Standards, v. 4, no. 2, p. x–y.

CEN, European Committee for Standardization, 1996, EN 932-3-Tests for general properties of aggregates—Part 3—Procedure and terminology for simplified petrographic description: CEN, Brussels, 12 p.

Cook, R. D., and Lee, H., 1999, Dimension Reduction in Binary Response Regression: American Statistical Association, v. 94, p. 1187–1200.

EUREKA 2001, PETROSCOPE—An Optical Analyser for Construction Aggregates and Rocks, Project no. 2569, Announced 28 June 2001, EUREKA, Brussels.

EUREKA, 2005, PETROSCOPE II, Project no. 3665, Announced 01 June 2005, EUREKA, Brussels.

Ferrando, S. E., and Kolasa, L. A., 2001, Averages of best wavelet basis estimates for denoising: Journal of Computational and Applied Mathematics, v. 136, p. 357–367.

Griffiths J. C., 1967, Scientific method in analysis of sediments: McGraw-Hill, New York, 508 p.

Guenzler, H., and Heise, H. M., 1996, IR-Spektroskopie, Eine Einführung: VCH Weinheim, 397 p.

Friedman, J., 1991, Multivariate adaptive regression splines (with discussion): Annals of Statistics, v. 19, no. 1, p. 1–141.

Hastie, T., Tibshirani, R., and Buja, A., 1994, Flexible Discriminant Analysis By Optimal Scoring: Journal of the American Statistical Association, v. 89, p. 1255–1270.

Hastie, T., Tibshirani, R., and Buja, A., 1995, Penalized Discriminant Analysis: Annals of Statistics, v. 23, p. 73 –102.

Hastie, T., Tibshirani, R., and Friedman, J., 2001, The Elements of Statistical Learning: Springer, New York, 533 p.

Hastie, T. J., Rosset, S., Tibshirani, R., and Zhu, J., 2004, The Entire Regularization Path for the Support Vector Machine: Journal of Machine Learning Research v. 5, p. 1391–1415.

Helgason, Th. S., 1990, Characteristics, properties, and quality rating of Icelandic volcanic aggregates: 43rd Canadian Geotechnical Conference, St. Foy, Université Laval, 1, p. 339–345.

Hofer, V., Pilz, J., and Helgason, T. S., 2005, Support Vector Machines for Classification of Aggregates by Means of IR-Spectra: Mathematical Geology (submitted).

Hunt, G. R., 1982, Spectroscopic properties of rocks and minerals, *in*: Carmichael, R.S., ed., Handbook of Physical Properties of Rocks, CRC Press, Boca Raton, p. 295–385.

Hunt, G. R., 1973, Visible and Near-Infrared Spectra of Minerals and Rocks VII, Acidic Igneous Rocks: Modern Geology, v. 4, p. 217 – 224.

Hunt, G. R., 1997, Spectral signatures of particulate minerals in the visible and near-infrared: Geophysics, v. 42, no. 3, p. 501–513.

Jolliffe, I. T., 1986, Principle Component Analysis: Springer Series in Statistics, New York, 257 p.

Li, K.-C, Aragon, Y., Shedden, K., and Agnan, C. T., 2003, Dimension Reduction for Multivariate Responses: American Statistical Association, v. 98, no. 461, p. 99–109.

Logan, L. M., Hunt, G. R., Salisbury, J. W., and Salvatore, R. B., 1973, Compositional Implications of Christiansen Frequency Maximums for Infrared Remote Sensing Applications: Journal of Geographical Research, v. 78, no. 23, p. 4983–5003.

Mallet, Y., Coomans, D., and Vel, O. de, 1996, Recent developments in discriminant analysis on high dimensional spectral data: Chemometrics and Intelligent Laboratory Systems, v. 35, p. 157–173.

Mitchell, J. K., 1993, Fundamentals of soil behavior: 2nd edition, Wiley, New York, 437 p.

Müller, F., 1987, Gesteinskunde, Lehrbuch und Nachschlagewerk über Gesteine für Hochbau, Innenarchitektur, Kunst und Restauration: Ebner Verlag, Ulm, 216 p.

Petromodel ehf., 2002, PM Techmodel 1.0 - User manual: Petromodel, Reykjavik, 43 p.

Petromodel ehf., 2004, Apparatus and method for analysis of size, form and angularity and for compositional analysis of mineral and rock particles, Application submitted to the Icelandic Patent Office, Reykjavik, 07 September 2004.

Ramsay, J. O., and Silverman, B. W., Functional Data Analysis, 1997: Springer Series in Statistics, New York, 310 p.

Ramsay, J. O., and Silverman, B. W., 2002, Applied Functional Data Analysis: Springer, New York, 190 p.

Smith, M. R., and Fookes P. G., 2001, Aggregates - Sand, Gravel and Crushed Rock Aggregates for Construction Purposes: The Geological Society London, 3rd ed., 339 p.

Teppola, P., and Minkkinen, 2000, Wavelet-PLS regression models for both exploratory data analysis and process monitoring: Journal of Chemometrics, v. 14, p. 383–399.

Thomson, J. L., and Salisbury, J. W., 1993, The mid-infrared reflectance of mineral mixtures (7–14 μm): Remote Sensing of Environment, v. 45, p. 1 –13.

Vidakovic, B., 1999, Statistical Modeling by Wavelets: Wiley, New York, 382 p.

Vidakovic, B., Müller, P., 1999, An Introduction to Wavelets, *in*: Müller, P., Vidakovic, B., eds., Bayesian Inference in Wavelet-Based Models, Springer, New York, 394 p.