Journal of Data Science **6**(2008), 219-229

Bayesian Wavelet Regression for Spatial Estimation

G. Ávarez[1] and B. Sansó[2]
[1] *Universidad del Rey Juan Carlos and* [2] *University of California Santa Cruz*

Abstract: We consider the problem of estimating the properties of an oil reservoir, like porosity and sand thickness, in an exploration scenario where only a few wells have been drilled. We use gamma ray records measured directly from the wells as well as seismic traces recorded around the wells. To model the association between the soil properties and the signals, we fit a linear regression model. Additionally we account for the spatial correlation structure of the observations using a correlation function that depends on the distance between two points. We transform the predictor variable using discrete wavelets and then perform a Bayesian variable selection using a Metropolis search. We obtain predictions of the properties over the whole reservoir providing a probabilistic quantification of their uncertainties, thanks to the Bayesian nature of our method. The cross-validated results show that a very high accuracy can be achieved even with a very small number of wavelet coefficients.

Key words: Bayesian variable selection, reservoir exploration, wavelet regression.

1. Introduction

Predicting the properties of a reservoir using the information provided by data collected from wells is a fundamental issue in petroleum management and exploration. Data collected from wells usually consists of core analyses, performed, on actual samples of the soil, and recordings of different electromagnetic, physical, chemical or radioactive properties of the soil obtained by inserting various tools into the well. These are usually refer to in the industry as well logs or well profiles, (see Hearst and Nelson, 2000; Tiab and Donaldson, 2004). Core and well log data are expensive and are available only at the specific locations were wells have been drilled. To obtain a global description of the reservoir, petroleum engineers perform an array of geostatistical techniques for interpolation (see Sheldon, 1995; Doyen, 1988). In an exploration scenario, log and core data are usually scarce, as they are only available at the few locations where wells have been drilled, whether seismic data are usually available for the whole reservoir. The scarcity of well

data presents a problem for most geostatistical methods, since they are generally imprecise when the number of locations is small. Moreover, seismic information can be difficult to incorporate. An example of the data that we consider in this paper is presented in Figure 1.

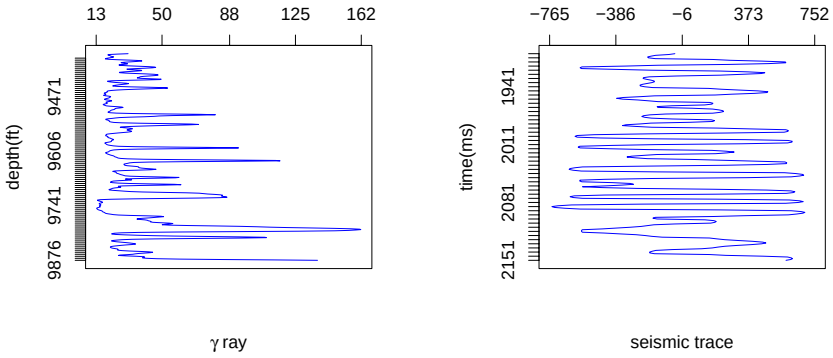

Figure 1: Left panel shows a typical γ-ray. Right panel corresponds to a typical seismic trace.

The method we present in this paper consists on regressing the observations obtained from the wells on the wavelet decomposition of a signal, either well logs or seismic traces. Wavelet transformations provide a parsimonious representation of the information in the signal. Their multiresolution properties have been successfully applied to quantify the decay of energy from large to small scales in well logs and seismic traces, see Álvarez *et al.* (2003). The method uses a Bayesian approach to estimate the property of interest on a location in a reservoir and quantify the uncertainty associated with the estimation. This includes a stochastic variable selection model to reduce the number of wavelet coefficient needed for accurate prediction of the reservoir properties.

We present the model in Section 2. In Section 3 we discuss the methods used to make inferece on model parameters and produce estimates for the soil properties at unobserved sites. In Section 4 we present the results followed by some discussion in Section 5.

2. The Model

The motivating example of this paper consists of data from 14 wells located in a reservoir in Southwestern Venezuela. For these wells we analyze γ-ray logs obtained at a depth of 9000 feet.

The γ-ray log is one of the most useful log data available. The γ-ray log can be summarized as the continuous measurement of the natural, or in some cases

introduced, radioactivity in a well. It can be recorded in open and cased holes, separately or in conjunction with virtually any other log, or with perforating guns. It is a measurement of the natural radiation of various formations penetrated by a well (or in some cases artificially placed sources of radiation). Dolomites, limestones, sandstones, and salts typically exhibit a low level of radiation, while shales, clays, and rocks of igneous origin typically have higher levels of radiation. It is these differences that make the γ-ray log very useful in determining lithologies and in the evaluation of the shale volume, porosity and other rock properties in zones of interest, see for example Robinson (2000) and Hearst and Nelson (2000).

Additionally, data from a grid of seismic traces covering an area of 100 Km^2 around the wells is also available. The reservoir is a three dimensional domain. In our analysis we consider a fixed depth that is of petrophysical interest. It consist of a window of 150 feet, equivalent to about 32 seconds for the seismic traces. For such stratum we have available the average values of porosity and clay volume at the sites of the oil wells. In other words, the response variables in our problem consist of the set of two reservoir properties observed at 14 different locations. These locations are irregularly scattered over the whole area of the reservoir. Figure 2 shows the location of the 14 wells.

The explanatory variables consist of well logs and seismic traces. These are series of, respectively, 512 and 128 readings. The Data were provided by Intevep, the research branch of PDVSA, the Venezuelan state-own oil company.

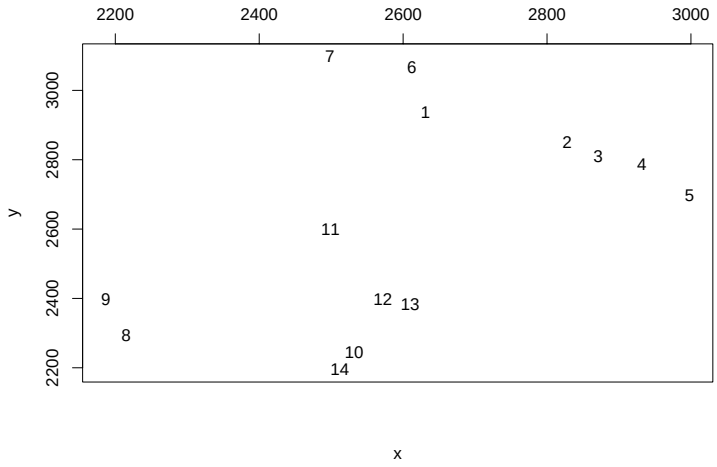

Figure 2: Location of the 14 well in the reservoir under study. Wells are numbered so that they can be referenced in the analysis.

In order to relate the reservoir properties to the well logs or the traces, we

consider the regression model

$$\mathbf{y} = \alpha\mathbf{1} + \boldsymbol{X}\boldsymbol{b} + \varepsilon \tag{2.1}$$

where $\boldsymbol{y} \in \mathbb{R}^n$ is the vector of properties at the locations of the wells. In our case $n = 14$. $\alpha \in \mathbb{R}$ is an intercept, $\mathbf{1}$ is an n-dimensional vector of ones. $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is the matrix of the signals, either well logs or seismic traces and $p = 512$ or 128. $\boldsymbol{b} \in \mathbb{R}^p$ is a vector of coefficients and $\varepsilon \in \mathbb{R}^n$ is the error term. Notice that the model in (2.1) has a larger number of regression coefficients than data. Thus direct estimation of the $\boldsymbol{b}$ using traditional regression methods is unfeasible. We either have to impose some restrictions or consider prior information.

Motivated by the work of Brown, Fearn and Vannucci (2001) we consider a wavelet transformation of the signals. The idea of such a transformation is that a reduced number of wavelet coefficients should be able to capture the information in the signals needed to predict the value of the reservoir property at a given location. A discrete wavelet transformation is given by an orthogonal matrix (see for example Vidakovic 1999, chapter 4), say $\boldsymbol{W} \in \mathbb{R}^{p\times p}$, such that $\boldsymbol{W}\boldsymbol{W}' = \boldsymbol{I}$, where $\boldsymbol{I}$ denotes the identity matrix. We then have that

$$\mathbf{y} = \alpha\mathbf{1} + \boldsymbol{X}\boldsymbol{W}\boldsymbol{W}'\boldsymbol{b} + \varepsilon = \alpha\mathbf{1} + \boldsymbol{Z}\beta + \varepsilon \tag{2.2}$$

where $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{W}$ is the matrix of wavelet coefficients corresponding the series in $\boldsymbol{X}$ and $\beta = \boldsymbol{W}'\boldsymbol{b}$ the new regression vector.

Since we expected that locations in the reservoir that are close to each other will have similar properties, ε should have a non-diagonal correlation structure. We model such spatial correlation by assuming that the error term ε corresponds to an isotropic random field with an exponentially decaying correlation function. Thus

$$\operatorname{cov}(\varepsilon_i, \varepsilon_j) = \sigma^2 \exp\left(-\frac{1}{\lambda}||s_i - s_j||\right),$$

where s_k denotes the location of the kth site. The proposed correlation can easily be substituted for any other parametric family of correlations offering wider flexibility, like the Mat 廨 n class, as described for example in Handcock and Stein (1993).

We take a Bayesian approach to estimate the parameters in the model and, following Brown Fearn and Vannucci (2001), we specify a prior for the original regression coefficients $\boldsymbol{b}$ as a p-variate normal with mean 0 and covariance matrix $\boldsymbol{H}$, denoted as $N_p(0, \boldsymbol{H})$, for $\boldsymbol{H}$ corresponding to the covariance matrix of an autoregressive process of order one. Such a distribution is used to guarantee that the components of $\boldsymbol{b}$ vary smoothly and that the variances of the transformed coefficient β show the typical decay of wavelet coefficients. The selection of the

relevant wavelet coefficients is achieved by considering a prior distribution for the parameter β_i given by

$$p(\beta_i) \propto (1-\gamma_i)\delta_0 + \gamma_i N_1(0, \tilde{h}_i) \quad ,$$

where δ_0 is a point mass at 0 and γ_i is a binary variable that indicates whether the i-th coefficient is 0 or not. $\tilde{h}_i$ corresponds to the diagonal of the matrix $\tilde{\boldsymbol{H}} = \boldsymbol{WHW'}$. So, a priori, the i-th coefficient has probability $1-\gamma_i$ of being equal to zero and probability γ_i of being distributed as a normal with zero mean and variance $\tilde{h}_i$. To complete our model we consider the following prior distributions: $\alpha \sim N_1(0, h\sigma^2)$, $\gamma_i \sim Ber(\omega)$, $i = 1, 2 \cdots, p.$, where $Ber(p)$ denotes a Bernoulli with parameter p, σ^2 follows an inverse gamma with parameters a_σ and b_σ and λ follows a gamma with parameters a_λ and b_λ. a_σ and b_σ were chosen to obtain a diffuse prior on σ^2. a_λ was taken as 0.4 for clay volume and 0.5 for porosity, b_λ was set to 4 in both cases. These values are compatible with the covariograms of the observations, but reflect a fair level of uncertainty.

3. Estimation and Prediction

To obtain inferences on the parameters in our model we explore their joint posterior distribution using a Markov chain Monte Carlo method (MCMC) as proposed, for example, in Gamerman and Lopes (2006). The MCMC that we use consists of sampling iteratively from the distributions of each of the parameters or blocks of parameters conditional on all the remaining ones. Thus, samples of α are obtained from a univariate normal. Samples of β are obtained from a multivariate normal and samples of σ^2 correspond to an inverse gamma. The spatial range λ is sampled by considering a Metropolis-Hastings step that consists of accepting or rejecting a proposed value with a probability that depends on the conditional density of λ evaluated at the current state of the chain. A more detailed description of the full conditionals is presented in Álvarez (2003).

Generating samples of $\gamma = (\gamma_1, \ldots, \gamma_p)$ presents the challenge of dealing with a highly multivariate distribution, since in our case p is 512 when using the γ-ray logs as predictor and 128 when using the seismic traces. We proceed by considering a random initial configuration $\gamma^{(0)}$. Then, at each iteration, one the following two ways of choosing a candidate configuration is chosen with (fixed) probability ϕ: (a) Generate a new candidate by choosing at random a component. This component is deleted if it is part of the current configuration and added if it is not; (b) Select two components i and j such that $\gamma_i = 0$ and $\gamma_j = 1$ and swap their values. The proposed configuration is rejected or accepted following a Metropolis-Hastings rule. Experience shows that good predictions can be obtained with about 20 wavelet coefficients. Thus our prior distribution for γ is such that the

prior expected number of coefficients, $p\omega$, is equal to 20. Thus $\omega = 0.16$ when the seismic traces are used, and $\omega = 0.04$ when the well logs are used.

The posterior predictive density of a new observation y_N, given the observed data $\boldsymbol{y}_{obs}$, can be estimated from the m simulated values from the MCMC for the joint parameter vector, say $\theta^{(j)}$, using the approximation

$$\hat{p}(y_N|\boldsymbol{y}_{obs}) = \frac{1}{m}\sum_{j=1}^{m} p(y_N|\theta^{(j)}).$$

In our case, in order to predict the value of a property for a specific location s_N, we use the information provided by the wavelet transformation of the signal $z_N \in \mathbb{R}^p$ corresponding to s_N. Thus, using the j-th iteration from the MCMC,

$$p(y_N|\theta^{(j)}) = N\left(\alpha^{(j)} + \boldsymbol{z}'_N\beta^{(j)} + \frac{\boldsymbol{V}^{(j)}}{v^{(j)}_{N,2}}(\mathbf{y} - \alpha^{(j)}\mathbf{1} - \boldsymbol{Z}\beta^{(j)}), \boldsymbol{V}^{(j)}_r\right), \qquad (3.1)$$

where

$$\boldsymbol{V}^{(j)}_N = \begin{pmatrix} \boldsymbol{V}^{(j)} & \mathbf{v}^{(j)}_{1,N} \\ \mathbf{v}^{(j)}_{N,1} & v^{(j)}_{N,2} \end{pmatrix} \in \mathbb{R}^{(n+1)\times(n+1)}$$

and

$$\boldsymbol{V}_r = \boldsymbol{V}^{(j)} - \frac{\mathbf{v}^{(j)}_{N,1}\mathbf{v}^{(j)}_{1,N}}{v^{(j)}_{N,2}} \in \mathbb{R}^{n\times n},$$

$\mathbf{v}_{1,N} \in \mathbb{R}^{(n\times 1)}$, $\mathbf{v}_{N,1} \in \mathbb{R}^{(1\times n)}$ and $[\boldsymbol{V}^{(j)}_N]_{k,l} = \sigma^2 \exp(-\frac{1}{\lambda^j}||s_l - s_k||)$. In words, to obtain a prediction at a location s_N we calculate the wavelet decomposition of the signal at that location and then, for each set of simulated values from the the MCMC, we calculate the spatial correlation matrix $\boldsymbol{V}^{(j)}_N$ and sample the normal distribution specified in (3.1). The result is a set of samples $y^{(1)}_N, \ldots, y^{(m)}_N$ from the posterior predictive distribution of y_N.

4. Results

We fitted model (2.2) separately for each property using first the γ-ray logs and then the seismic traces. We considered a wavelet transformation based on the Haar basis. We present results that were obtained from 5,000 iterations of a MCMC after a burn in period of 500 iterations. To explore the predictive capability of the model we adopted a "leave one out" approach, consisting on obtaining the posterior predictive distribution for each of the 14 locations using the remaining 13.

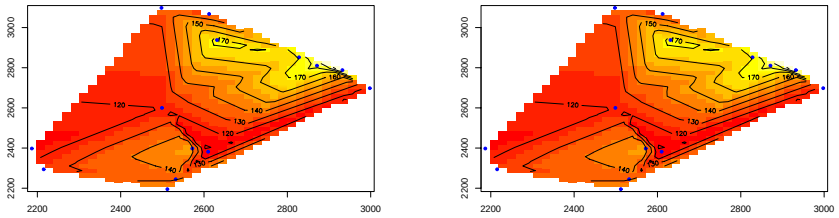

Figure 3: Linear interpolations of the predicted properties obtained using a "leave one out" estimation. The predictions at each well are obtained using the γ-ray logs. Top panel corresponds to clay volume and bottom panel corresponds to porosity.

Figure 3 shows the predicted values of porosity and clay volume interpolated over the convex hull of the locations of the wells. The predictions for each well are given by the medians of the simulated values obtained from the 4,500 samples from the MCMC. In Figure 4 we compare the predictive distribution of each of the 14 wells, based on the remaining 13 wells to the actual observed values of clay volume. Similar results are obtained for porosity. Notice that the observations are very central to the predictive densities. The former shows that the method has a very high level of predictive accuracy. Also, given the Bayesian nature of the method, we are not only providing an estimate of the properties at each location, but a precise assessment of the uncertainties involved in such estimation, given by the predictive distribution.

The number of non-zero coefficients can vary from one iteration to the other of the MCMC. Nevertheless we observed that no more than 10 coefficients where different from zero at any given iteration. This implies that less than 2% of the wavelet coefficient contain enough information to accurately predict the values of porosity and clay volume.

Clearly, the predictive ability of the model depends on the number of wells that are used. To assess the robustness of the method with respect to the number of locations used for prediction we chose a well located at the center of the field and predicted its porosity using the remaining 13 locations. We then deleted one location at random at a time and obtained the prediction with the remaining locations. The results using γ-ray logs as predictors are shown in Figure 5 for porosity. A similar behavior is observed for clay volume. As expected, we observe that the width of the predictive intervals increases as the number of wells decreases. Nevertheless observations and point-wise predictions are fairly close even for as little as six location for the clay volume and four for the porosity.

We repeated the whole analysis using the seismic traces as a predictor of both, clay volume and porosity. Figure 6 shows the interquartile ranges for the leave

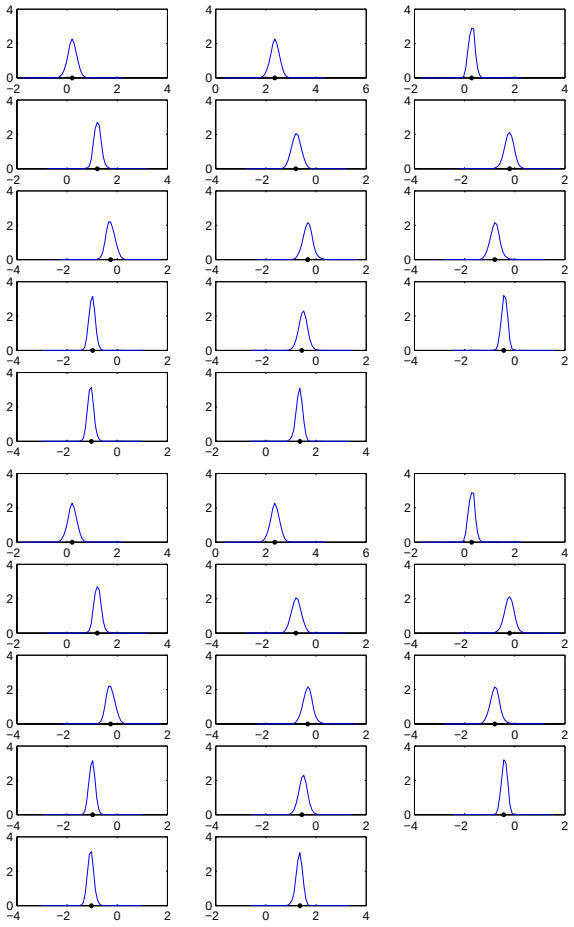

Figure 4: Predictive densities for porosity, obtained using γ-ray logs, at each of the 14 wells using the remaining 13. Actual observations are marked as a dot.

one out predictions. The accuracy of the predictions is lower than the one obtained when the γ-ray logs are used, but the performance of the model is remarkably good in this case as well.

5. Discussion

We have proposed a wavelet regression with spatially correlated errors for the prediction of soil properties using petrophysical or seismic signals from an oil reservoir. We convert an over-parameterized model into a parsimonious one by imposing appropriate priors on the parameters and using a stochastic variable selection approach. The results are assessed using the posterior predictive distribution of the property of interest at a location that has been left out of the

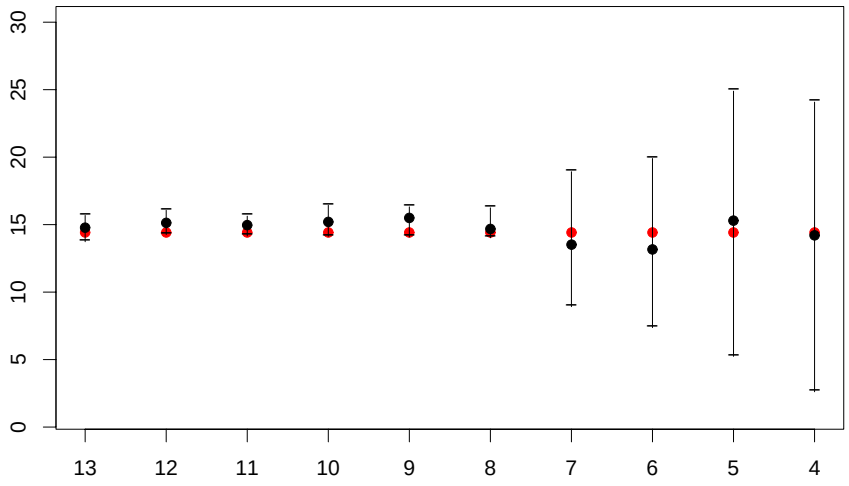

Figure 5: Predictive 95% probability intervals obtained for a centric well, using γ-ray logs, as a function of the number of predictive wells. Left panel corresponds to clay volume and right panel corresponds to porosity.

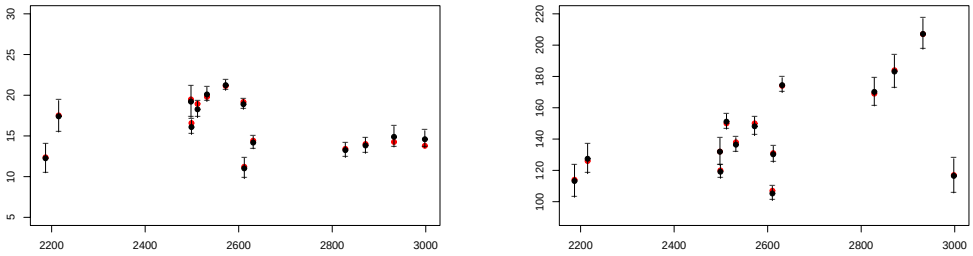

Figure 6: Interquartile intervals of each of the 14 wells using the remaining 13 wells. Predictions are obtained using seismic traces. Left panel corresponds to clay volume and bottom panel corresponds to porosity.

analysis and the procedure is repeated for each location at a time. For both soil properties considered, we obtain predictions that are compatible with the observations. Moreover, by providing a predictive distribution, the estimation uncertainties are fully accounted for. These results are achieved with a very small number of wavelet coefficients. This is an important byproduct of the model that can be used for data compression.

Acknowledgments

The authors are grateful to Dr Reinaldo Michelena for proposing the problem and providing the data considered in this paper. The second author was partially supported by the National Science Foundation grant DMS 0504851.

References

Álvarez, G. (2003). *Clasificación de Litologías y Estimación de Propiedades de Yacimientos a Partir de Datos Sísmicos*. Ph.D thesis, Universidad Simón Bolívar.

Álvarez, G, Sansó, B., Michelena, R. and Jiménez, J. R. (2003). Lithologic characterization of a reservoir using continuos wavelet transforms. *IEEE Transaction on Geoscience and Remote Sensing* **41**, 59-65.

Brown, P.J., Fearn, T. and Vannucci, M. (2001). Bayesian wavelet regression on curves with application to a spectroscopic calibration problem. *Journal of the American Statistical Association* **96**, 398-408.

Doyen, P. (1988). Porosity from seismic data: A geostatistical approach. *Geophysics* **53**, 1263-1274.

Gamerman, D. and Lopes, H. F. (2006). *Markov Chain Monte Carlo – Stochastic Simulation for Bayesian Inference, 2nd ed.*. Chapman and Hall.

Handcock, M. S. and Stein, M. L. (1993). A Bayesian Analysis of Kriging. *Technometrics* **35**, 109-122.

Hearst, J. and Nelson, P. H. (2000). *A Handbook for Geophysicists, Geologists and Engineers, 2nd ed.*. Wiley.

Robinson, E. A. (2000). *Geophysical Signal Analysis*. Society of Exploration.

Sheldon, B. (1995). Using geostatistics to aid in reservoir characterization. *The Leading Edge* **14**, 967-974.

Tiab, D. and Donaldson, C. D. (2004). *Petrophysics: Theory and Practice of Measuring Reservoir Rock and Fluid Transport Properties, 2nd ed.*. Gulf Pub Co.

Vidakovic, B. (1999). *Statistical Modeling by Wavelets*. John Wiley and Sons.

Received December 31, 2006; accepted March 7, 2007.

G. Ávarez
Department of Statistics and Operation Reseaarch
Universidad del Rey Juan Carlos
Tulipan Street, CA 28942
Madrid, Spain
giselle.alvarez@urjc.es

B. Sans
Departmeent of Applied Mathematics and Statistics
University of California Santa Cruz
1156 High Street
Santa Cruz, CA 95064, USA
bruno@ams.ucsc.edu