GEOPHYSICS, VOL. 61, NO. 3 (MAY–JUNE 1996); P. 631–638, 2 FIGS.

Monte Carlo simulation of wells

Paul F. M. de Groot*, Albertus H. Bril*, Frans J. T. Floris‡, and A. Ewan Campbell‡

ABSTRACT

We present a method to simulate wells, i.e., 1-D stratigraphic profiles with attached physical properties but without spatial information, using a combination of geological knowledge and Monte Carlo statistics. The simulated data is intended to be used in seismic lateral prediction studies. Our algorithm simulates correlated stochastic variables one by one. There are two major advantages in this approach above the conventional way in which all correlated stochastic vectors are drawn simultaneously. The first advantage is that we can steer the algorithm with rules based on geological reasoning. The second advantage is that we can include hard constraints for each of the stochastic variables. If a simulated value does not satisfy these constraints, it can simply be drawn again.

The input to the simulation algorithm consists of geological rules, probability density functions, correlations, and hard constraints for the stochastic variables. The variables are attached to the entities of a generic integration framework, which consists of acoustic-stratigraphic units organized at three scale levels. The simulation algorithm constructs individual wells by selecting entities from the framework. The order in which the entities occur, and the thickness of each entity, is determined by a combination of random draws and specified geological rules. Acoustic properties and optional user-defined physical properties are attached to the simulated layers by random draws. The acoustic properties are parameterized by top and bottom sonic and density values. The algorithm is capable of simulating acoustic hydrocarbon effects.

The algorithm is demonstrated with a simulated example, describing the stratigraphic and physical variations in an oil field with a fluvial-deltaic labyrinth type reservoir.

INTRODUCTION

The simulation of realistic synthetic reflection sequences or of lithological sequences has been studied by many workers because of the potential benefit in seismic reservoir characterisation applications and in testing new processing algorithms (Sherrif, 1992). For example, Walden and Hosken (1986), Walden (1993) and Kerner and Harris (1994) simulate stochastic models for reflection coefficients with auto-regressive-moving-average (ARMA) processes and non-Gaussian distribution functions and used parameter sets derived from real logs. Barnes and Tarantola (1993) simulate pseudorandom lithological images that follow the statistical distribution of the geological sequences and shapes shown by one or several images of the subsurface considered as reference model. Our approach differs from these methods in that we simulate 1-D stratigraphic profiles (de Groot, 1995a) with attached physical properties (but without spatial information) by a combination of geological reasoning and Monte Carlo statistics. Our algorithm has some similarities with Markov chain models (e.g., Sinhval and Sinhval, 1992). In Markov chains, stratigraphic sequences are simulated using probabilities that a series of lithologies follow each other in a predictable pattern. In our simulation algorithm, the predictable patterns are captured in terms of geology-related rules and constraints attached to the entities of a generic integration framework. The stochastic information is supplied in the form of probability density functions (pdfs) and correlation-coefficients.

The Monte Carlo method is a procedure that involves sampling based on probabilities to approximate the solution of mathematical or physical problems in a statistical way. Monte Carlo statistics are used for a variety of different problems. In geoscientific applications, the method is used, e.g., for reserves estimations and for prospect evaluations. In cases where the stochastic vectors are correlated, the simulation is not straightforward because the variables cannot be drawn independently.

Manuscript received by the Editor January 16, 1995; revised manuscript received September 12, 1995.
*de Groot-Bril Earth Sciences B.V., Boulevard 1945—nr.24, 7511 AE Enschede, The Netherlands.
‡TNO Institute of Applied Geoscience. Schoemakerstraat 97, P.O. Box 6012, 2600 JA Delft, The Netherlands.

The conventional way, assuming multivariate normal distributions, is to transform the correlated system into an independent system where the transformed stochastic vectors can be simulated. Inverse transformation then yields the correlated stochastic vectors. A major disadvantage of this approach, is however, vectors are simulated rather than variables. The simulation of entire vectors implies that the procedure is difficult to combine with rules and constraints. In other words, additional geological knowledge cannot be incorporated into the simulation. Instead, our algorithm simulates correlated variables one at the time. Therefore, we are able to determine at any time, whether a variable must be simulated and which constraints should be satisfied (de Groot et al., 1993). For example, it is possible to simulate a package consisting of sand and shale layers. The sand and shale layers have different probabilities to be selected, hence we can generate sand-prone and shale-prone packages. First a thickness is simulated for the package. Subsequently the package is filled with sands and shales with varying thicknesses and varying physical properties such as sonic values, density values, and user-defined properties. Each variable (sand thickness, sonic, density, etc.) can be correlated with any other variable; e.g., sonic and density distributions are often correlated negatively, thicknesses might be correlated to simulate pinch-outs, etc. Each simulated value honors the correlations and is evaluated against the constraints before it is accepted.

In this article, we describe how this algorithm is used to simulate realistic wells; i.e., 1-D stratigraphic profiles with attached physical properties but without spatial information. The mathematics is explained in the Appendix. The simulated sonic and density log responses of these wells can be used to synthesize seismic traces for geophysical reservoir characterization purposes, (de Groot, 1995b).

SIMULATING WELLS

The aim of the simulation is to generate a set of wells that are representative of the variations in stratigraphy and physical properties of the target interval. When used for reservoir characterization, the acoustic properties of these simulated wells are used to synthesize seismic traces that are subsequently analyzed or inverted back to the original well properties. For this reason, it is important that the physical properties of the wells are related directly to meaningful geological entities. The wells are therefore constructed from acoustic-stratigraphic entities that relate physical well properties to geology. The acoustic-stratigraphic entities are defined in a generic integration framework.

INTEGRATION FRAMEWORK

The generic integration framework, defines acoustic-stratigraphic entities at three scale levels. An example is given in Table 1. We will refer to these entities as: units, subunits, and lithologies. To facilitate data management and lithological and stratigraphic addressing, each entity in the framework is given a user-specified code. Acoustic entities in simulated wells can be recognized by these codes. If a code occurs more than once in a particular well, an occurrence number is added by the system. This provides a unique identification for each entity in the wells. To handle hydrocarbon properties in the system, a rock type is assigned to each of the framework lithologies. We distinguish seals, waste, and reservoir rock types. Seals are used by the simulation algorithm to simulate gross hydrocarbon columns. Reservoir rocks can have a fluid content. Separate acoustic probability density functions can be specified for each fluid-fill. Waste zones are nonsealing, noneconomic lithologies used to calculate net pay zones.

Table 1. Integration framework defining the acoustic-stratigraphic entities of the target interval.

Unit	Facies	Lithology	Type	Code
Carbonate A	Carbonate	Carbonate	Waste	crba.crb.crb
		Shale	Waste	crba.crb.shl
Carbonate B	Massive	Anhydrite	Waste	crbb.msv.anh
		Carbonate	Waste	crbb.msv.crb
	Anhydrite	Anhydrite	Waste	crbb.anh.anh
		Carbonate	Waste	crbb.anh.crb
		Shale	Waste	crbb.anh.shl
	Alternating	Carbonate	Waste	crbb.alt.crb
		Shale	Waste	crbb.alt.shl
Seal	Seal	Seal	Seal	seal.seal.seal
Reservoir	Massive Type 3	Type 3 Sand	Reservoir	res.mt3.t3s
		Silt/Shale	Waste	res.mt3.slt
	Massive Type 2	Type 2 Sand	Reservoir	res.mt2.t2s
		Silt/Shale	Waste	res.mt2.slt
	Massive SiltShale	Silt/Shale	Reservoir	res.msl.slt
		Type 2 or 3 Sand	Waste	res.msl.snd
	Laminated	Type 2 or 3 Sand	Reservoir	res.lam.snd
		Silt/Shale	Waste	res.lam.slt
Palaeozoic	Marine	Shale	Waste	pal.mar.shl

The simulation algorithm

Input to the simulation algorithm is a combination of stochastic information and geological knowledge. It is considered important that simulated models are realistic representations of the subsurface. This implies that the simulation must be controlled by geological reasoning and that unrealistic stochastic realizations can be redrawn. In our simulation algorithm, the predictable patterns are captured in terms of geology-related rules and constraints attached to the entities of the defined integration framework. The rules will be explained later.

Two types of constraints are used by the algorithm: simulation constraints and hard constraints. Simulation constraints are a special kind of geology-related rules. They determine the probability of a framework entity to be used in the construction of a well. These will be explained in more detail later. Hard constraints are constraints set on the upper and lower boundary of probability density functions (pdfs). Stochastic realizations are evaluated against these boundaries. If the hard constraints are not satisfied, a variable can either be drawn again or accepted.

To deal with the uncertainty, stochastic input is supplied in the form of pdfs and correlation coefficients. Pdfs and correlation coefficients are, in practice, determined from factual well data. For this purpose our software system (named geoProbe; predecessor of a commercial system named dgB-GDI) offers a well data analysis module. The information derived from factual data may be modified in the simulation to reflect geological probabilities of areas not penetrated by the drill-bit.

The simulation algorithm requires the following input to be specified:

1) Pdfs for each of the physical properties: thickness of geological entity, sonic and density at the top of each lithology.
2) Hard constraints on the upper and lower boundary of the pdfs.
3) Correlation coefficients between pairs of stochastic variables.
4) Geology-related rules. The following rules have been implemented in the geoProbe system: xor, sum, iterate, relative. The last three rules have two versions: one in which the smaller scale entities are selected in a random order and one in which they are selected in the order in which they have been defined in the framework.

These rules are best explained with an example. We will describe the simulation of the Massive Type 3 sub-unit of the framework of Table 1 as follows.

Attaching the xor rule to the Massive Type 3 sub-unit means that Type 3 Sand and Silt/Shale are mutually exclusive. The thickness of the Massive Type 3 sub-unit is simulated from the defined pdf. The thickness of the selected lithology is made equal to the simulated Massive Type 3 sub-unit thickness. The xor rule supports an optional parameter to indicate the probability for a smaller scale entity to be selected. For example xor 40/60 denotes that Type 3 Sand has a 40% chance to be selected against a 60% chance for Silt/Shale.

Attaching the sum rule to the Massive Type 3 sub-unit means that the sub-unit is constructed from one Type 3 Sand and one Silt/Shale lithology. The thickness of the Massive Type 3 sub-unit is the sum of the simulated thicknesses of Type 3 Sand and Silt/Shale.

Attaching the iterate rule to the Massive Type 3 sub-unit means that the sub-unit is constructed from as many Type 3 Sands and Silt/Shale lithologies as are required to fill the simulated sub-unit thickness. First a thickness for the sub-unit is simulated from the defined pdf. Subsequently, lithologies Type 3 Sand and Silt/Shale are selected and thicknesses for these are simulated from their respective pdfs. This process is continued until the sum of the lithology thicknesses exceeds the simulated sub-unit thickness. The thickness of the last selected lithology is adjusted accordingly.

Attaching the relative rule to the Massive Type 3 sub-unit means that the sub-unit contains one Type 3 Sand and one Silt/Shale lithology while the thickness-ratio is maintained in the final realization. The thicknesses of the simulated lithologies are adjusted (stretched or squeezed) to fit the simulated thickness of the Massive Type 3 sub-unit. In other words, the relative thickness (or thickness-ratio) is kept constant.

5) Simulation constraints. The following rules have been implemented in the geoProbe system: presence, generation, occurrence.

These constraints are used in combination with the geology-related rules explained above. To illustrate these constraints we will again describe the simulation of the Massive Type 3 sub-unit (Table 1).

The presence constraint operates on the full simulation data set. It denotes that the entity can be present in a percentage of the simulated wells only. A 60% presence attached to the Massive Type 3 sub-unit means that only 60% of the simulated wells comprize the Massive Type 3 sub-unit.

The generation constraint indicates the probability of an entity to be selected. An 80% generation attached to Type 3 Sand and a 20% generation attached to Silt/Shale, in combination with the iterate rule attached to the Massive Type 3 sub-unit indicates that the sub-unit will be sand-prone.

The occurrence constraint operates on a well-by-well basis. It denotes the number of occurrences of the entity per well. An occurrence of 2 attached to Type 3 Sand, in combination with the iterate rule attached to the Massive Type 3 sub-unit indicates that only 2 Type 3 Sand lithologies can be selected to fill the sub-unit.

Additional input to the simulation algorithm can be specified optionally in the form of:

1) Pdfs for sonic and density at the bottom of each lithology to simulate linear trends as a function of depth over an interval.
2) Pdfs for gross hydrocarbon column lengths.
3) Pdfs for hydrocarbon filled sonic and density variables at the top of each reservoir lithology.
4) Pdfs for hydrocarbon filled sonic and density variables at the bottom of each reservoir lithology.
5) Pdfs for user-defined variables.

A combination of rules, constraints, and correlations is used to control the simulation. Various stratigraphic settings can be simulated in this way. An example of what can be done with these input specifications is presented here.

The statistical description of the simulation algorithm is given in the Appendix. A theorem is given for drawing a value for a variable that is correlated to another variable already drawn, e.g., drawing a sonic given a density value. This theorem requires a full correlation matrix, which is, in general, not supplied by the user. Therefore a second theorem provides a method for filling missing elements of the correlation matrix. An example is given to illustrate how the correlation matrix is filled and how a set of values for correlated variables is drawn.

Example

The simulation algorithm is applied to a simulated oil field. To make the simulation as realistic as possible, the field is modeled after an existing oil field in the Middle East. The trap is a structural dome with a parasitic structure on the limb of a monocline. The reservoir is an Upper Carboniferous to Lower Permian fluvial/fan assemblage sitting unconformably on Silurian marine shales. The reservoir formation is in turn covered conformably by Upper Permian carbonates. The field can be considered a labyrinth of interconnecting and isolated reservoir bodies. Oil production rate is primarily a function of reservoir development. Considerable volumes of oil can be produced from relatively thin (10–25′) sandstone intervals. The reservoir formation deposits are predominantly floodplain and playa lake deposits with reduced sand/shale ratios. There are no laterally correlatable horizons within the reservoir formation, but the top reservoir can be mapped on seismic data. Figure 1 shows a hypothetical cross-section through the crest of the structure.

An integration framework was established for the simulated field based on the major structural elements and sedimentology. Integration sub-unit subdivisions and lithology typing were derived from analysis of well data and formation analysis logs of the real field. The framework for this example is defined as follows (Table 1):

1) There are four main units. Each unit has one or several sub-units (geological or seismic). Each sub-unit has one or several lithologies. The fourth column shows the user-defined codes that enable the user to identify and manipulate data items at three scale levels.
2) The four main units occur sequentially as shown in the framework. Sub-units within the carbonate units occur sequentially while sub-units within the reservoir can vary. The reservoir unit has been divided into four sub-units: Massive Type 3, Massive Type 2, Massive Silt-Shale, and Laminated. This subdivision is based on common grouping of certain lithologies observed in the Formation Analysis logs and corresponds to the genetic units of the formation. In individual wells, the sub-units order may vary. Sub-unit could be completely absent or present multiple times.
3) Each sub-unit is assigned several lithologies. The lithological composition of an interval determines a particular sub-unit type. Type 3 sandstones, for example, can only occur in Massive Type 3 sub-unit or Laminated sub-unit while Type 2 sandstones will only be found in Laminated or Massive Type 2. Lithologies could occur in any order, repeat themselves, or be completely absent.

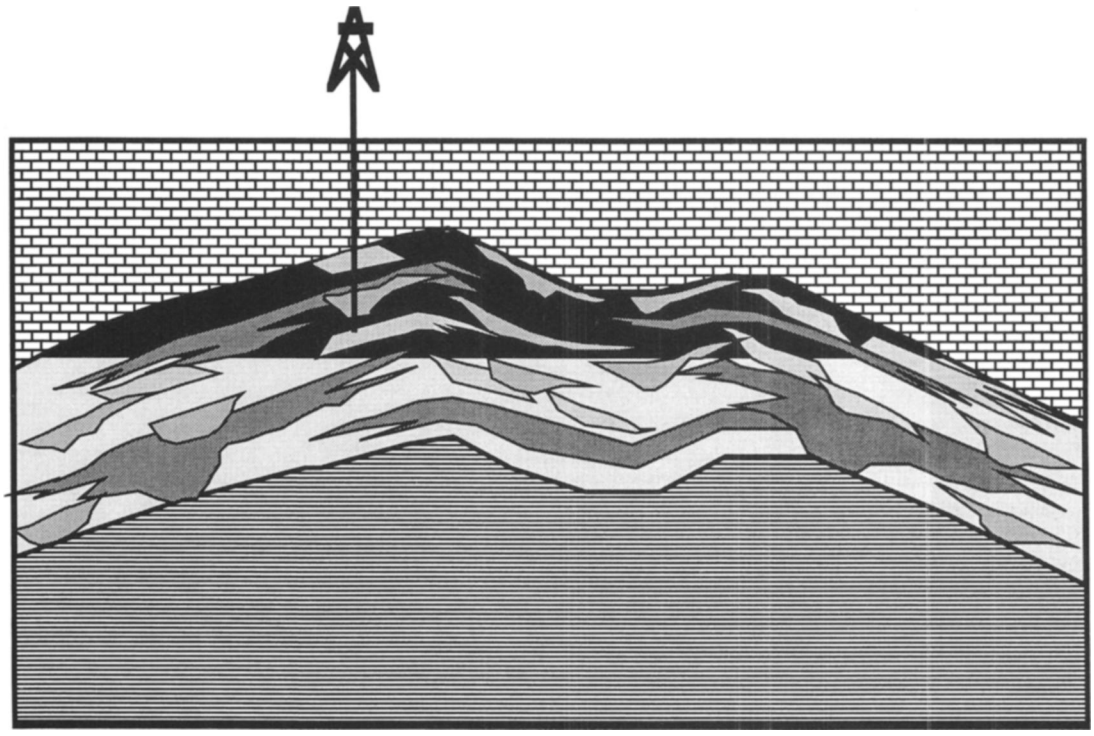

FIG. 1. Hypothetical cross-section through the simulated oil-field. Within this fluvial-deltaic setting our algorithm will simulate one-dimensional stratigraphic profiles with attached physical properties but without spatial information.

4) Oil-bearing sands occur only in the reservoir unit. Essentially, all sand lithologies, regardless of which sub-unit they belong to, are considered to be producing if they occur within the oil column. The fine grained nonproducing sandstones, silts, and shales were grouped into one lithotype (Silt/Shale) for the purposes of this study and are considered waste zones.
5) There is only one hydrocarbon column, which is attached to the overlying seal at the base of the carbonate unit.

The geoProbe system recognizes hydrocarbon columns by the seal to which the column has been attached. If the lithology of the layer directly overlying the reservoir can vary, as in this case, geoProbe has a problem. For this reason a virtual seal has been introduced immediately above the reservoir unit. In the simulations this seal is given a constant thickness of 0.1 foot. Based on factual well data and knowledge about the geological setting, a decision was made to feed the simulation algorithm with the following information:

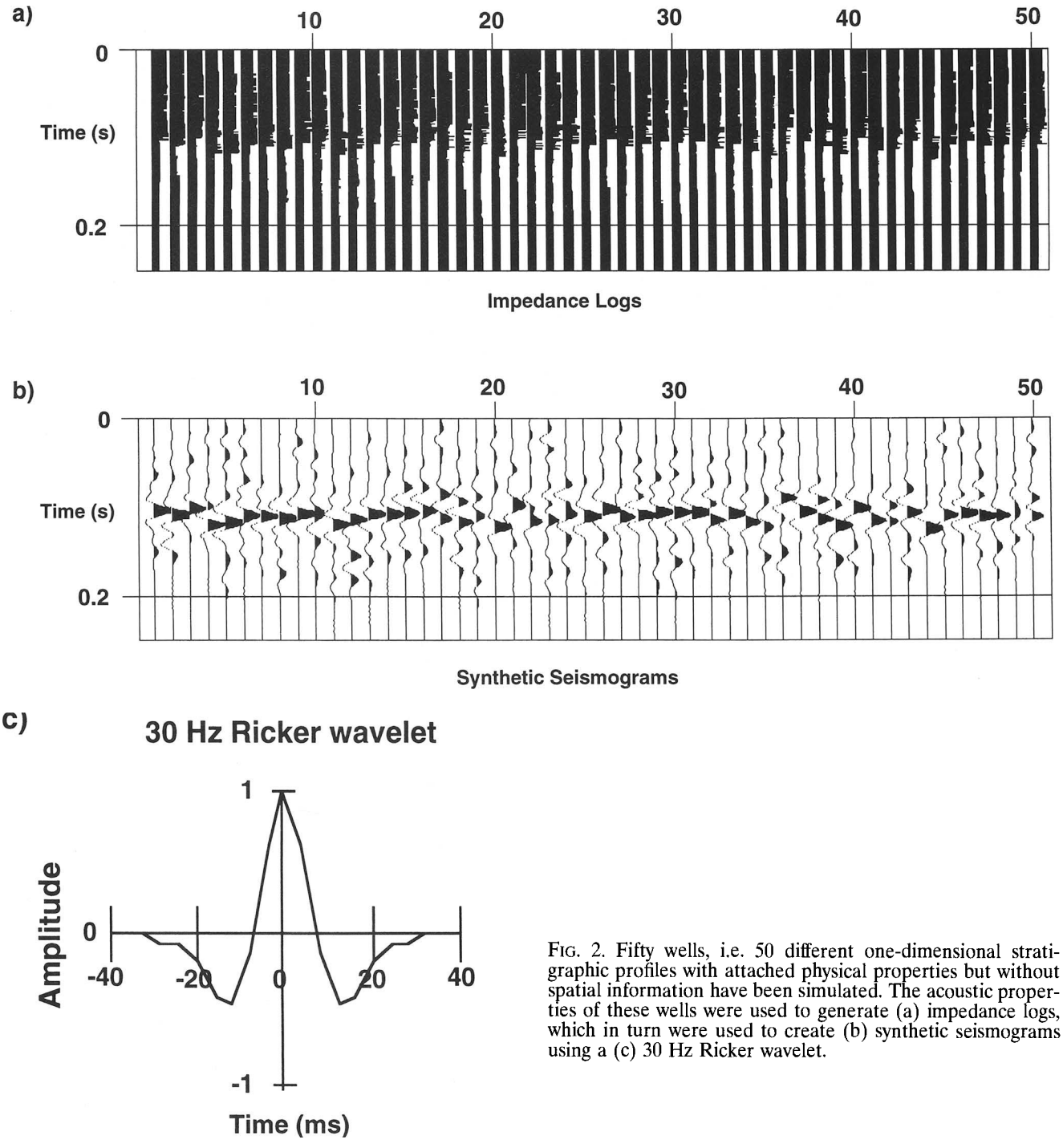

FIG. 2. Fifty wells, i.e. 50 different one-dimensional stratigraphic profiles with attached physical properties but without spatial information have been simulated. The acoustic properties of these wells were used to generate (a) impedance logs, which in turn were used to create (b) synthetic seismograms using a (c) 30 Hz Ricker wavelet.

1) The random iterate rule for Carbonate B unit, Reservoir unit and each of the facies.
2) The sum sequential rule for Carbonate A, Seal and Palaeozoic units.
3) Pdfs for thicknesses of all framework entities, for gross oil column thickness, for sonic and density of each lithology, for sonic and density of oil filled reservoir lithologies.
4) Correlations for sonic and density and for sonic of oil-filled and sonic of brine-filled reservoir lithologies to reflect that porefill is independent of rock properties.
5) A hard constraint for the minimum and maximum oil column length and minimum and maximum thickness of the Reservoir unit. Values were redrawn until these constraints were met.
6) Generation constraints for the reservoir lithologies to control the sand/silt ratio's of the various facies entities.

The simulation algorithm was used to simulate 50 wells, i.e., 50 different 1-D stratigraphic profiles with attached physical properties but without spatial information. Each of the wells is fully described in terms of integration framework entities. The acoustic properties of the wells were used to create reflectivity logs. These were converted into synthetic seismograms by depth-time conversion, anti-alias filtering to 2 ms and convolution with a 30 Hz Ricker wavelet. The acoustic logs with corresponding synthetic seismograms are shown in Figure 2. Please note the vertical scale differences between impedance logs and synthetic seismic response. When using these data for reservoir characterization, the seismic responses are analyzed, or inverted back, to the underlying well properties (de Groot, 1995b).

CONCLUSIONS

It has been shown that wells, i.e. 1-D stratigraphic profiles with attached physical properties but without spatial information can be realistically simulated by combining Monte Carlo statistics with geology-related rules, correlations and constraints. This combination is possible because we are able to draw correlated variables one by one. Our algorithm operates on an integration framework that defines acoustic-stratigraphic entities of the target interval at three scale levels. Simulated wells are constructed from framework entities and attached physical properties. The simulated data set can be used subsequently to investigate relations between seismic response and well properties for seismic lateral prediction purposes, (de Groot, 1995b).

ACKNOWLEDGMENTS

Tom Kavli of SINTEF is thanked for his insight into the design of the neural networks and his cooperation throughout all phases of this project. We also thank Subhashis Mallick and an annonymous reviewer for their comments on earlier versions of this manuscript. We are indebted to the sponsors of the Probe consortium (TNO Institute of Applied Geoscience, SINTEF, Saudi Arabian Oil Company, BEB Erdöl und Erdgas GMBH., IBM Nederland N.V., High Tech Automation B.V., The European Union under Thermie contract no. OG/178/92 NL and the Norwegian Research Council NTNF) for supporting this work and to allow publication of this paper.

REFERENCES

Barnes, C., and Tarantola, A., 1993, Probabilistic description of geological information: 55th Conference, Eur. Assn. Expl. Geophys., Expanded Abstracts, paper B048.
de Groot, P. F. M., Campbell, A. E., and Kavli, T., 1993, Seismic reservoir characterisation using artificial neural networks and stochastic modeling techniques: 55th Conf. Eur. Assn. Expl. Geophys., Stavanger, Expanded Abstracts, paper B047.
de Groot, P. F. M., 1995a, Total space inversion–Concept and experiments in model space: 57th Conf. Eur. Assn. Expl. Geophys., Expanded Abstracts.
——— 1995b, Seismic reservoir characterisation employing factual and simulated wells: PhD. thesis. Delft University of Technology.
Kerner, C., and Harris, P. E., 1994, Scattering attenuation in sediments modeled by ARMA processes-Calidation of simple Q models: Geophysics, **59**, 1813–1826.
Mardia, K. V., Kent, J. T., and Bibby, J. M., 1979, Multivariate analysis: Academic Press Inc.
Sherrif, R. E. (ed.), 1992, Reservoir geophysics: Investigations in geophysics, No. 7: Soc. Expl. Geophys.
Sinvhal, A., and Sinvhal, H., 1992, Seismic modeling and pattern recognition in oil exploration: Kluwer Academic Publishers.
Walden, A. T., 1993, Simulation of realistic synthetic reflection sequences: Geophys. Prosp., **41**, 313–321.
Walden, A. T., and Hosken, J. W. J., 1986, The nature of the non-Gaussianity of primary reflection coefficients and its significance for deconvolution: Geophys. Prosp., **34**, 1038–1066.

APPENDIX

MONTE CARLO STATISTICS; SIMULATING CORRELATED MULTIVARIATE STOCHASTIC VARIABLES

The following mathematical description is used in a simulation algorithm aimed at simulating wells, i.e., 1D-stratigraphic profiles with attached physical properties. In the algorithm, wells are constructed from so-called integration framework entities. These entities are grouped at three different scale levels. It is considered important that geological knowledge controls the selection of framework entities and that unrealistic realizations of variables can be redrawn. This implies that wells must be constructed one-by-one, entity-by-entity, and variable-by-variable.

Variables in a computer are simulated using a (pseudo-) random number generator. When random variables are correlated, it is not simple, however, to simulate random draws using such a (pseudo-) random number generator. This is especially true when the variables must be drawn one-by-one, as in our application. The realizations of previously drawn variables will in that case influence the realization of the variable to be drawn. For example, let us assume that a positive correlation exists between the thicknesses of two layers. When a small thickness is drawn for the first layer, then also for the second layer a small thickness must be drawn. In the case of normally distributed random variables, it is possible to draw the variables consecutively from the marginal distributions. Each time a variable is to be drawn, its marginal distribution must first be updated for the variables already drawn to which it is correlated.

In the following discussion $\mathbf{X}$ is a stochastic vector. In our algorithm, $\mathbf{X}$ comprizes all stochastic variables required for the

simulation. A component of $\mathbf{X}$ is denoted by X_i. Examples of components are sonic, density, thickness, and user-defined variables attached to framework entities. Each component X_i is assumed to be normally distributed with expectation μ_i and variance σ_i^2, symbolically written as: $X_i \sim N(\mu_i, \sigma_i^2)$. The vector of expectation will be denoted $\boldsymbol{\mu}$. The components are assumed to be correlated. The covariance between components i and j is indicated by σ_{ij}. Note, that the covariance between component i and itself, σ_{ii} equals σ_i^2. The matrix of covariances will be denoted as Σ. When the covariance σ_{ij} is normalized with the standard deviations σ_i and σ_j, we obtain the correlation coefficient ρ_{ij}, symbolically written as: $\rho_{ij} = \sigma_{ij}/(\sigma_i\sigma_j)$. The matrix of correlation coefficients will be denoted by C. Sets of components can be grouped into subvectors of $\mathbf{X}$ denoted by $\mathbf{X}^{(i)}$. An example of a subvector $\mathbf{X}^{(i)}$ is that part of stochastic vector $\mathbf{X}$ comprising correlated thicknesses of a set of layers. The theorems given hereafter apply to the general case of drawing entire subvectors. However, for design reasons, the variables are drawn one-by-one in the final implementation of the algorithm. In other words the subvector $\mathbf{X}^{(i)}$ to be drawn has only one component. This is illustrated by the example at the end of this Appendix.

We require two theorems for our algorithm to work. Theorem A-1 is used for updating the expectation and covariance matrix of a variable to be drawn, given some already drawn correlated variables (Mardia, 1979). This theorem requires the covariance matrix to be specified completely. In general, the user will not be in a position to specify all coefficients. Therefore, the unspecified correlation coefficients must be approximated first. This is accomplished with Theorem A-2 (A. M. H. Meeuwissen and R. H. Cooke, personal communication).

In the following discussion, first the two theorems are given followed by an illustration of their use with an example.

Theorem A-1

First we introduce some notation. Let $\mathbf{X}$ be an n-dimensional stochastic vector that is partitioned as follows:

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} \end{pmatrix}, \tag{A-1}$$

with expectation $E[\mathbf{X}]$ equal to $\boldsymbol{\mu}$:

$$\boldsymbol{\mu} = E[\mathbf{X}] = \begin{pmatrix} \boldsymbol{\mu}^{(1)} \\ \boldsymbol{\mu}^{(2)} \end{pmatrix}, \tag{A-2}$$

and a positive definite covariance matrix $Cov(\mathbf{X})$ given by

$$\Sigma = Cov(\mathbf{X}) = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}. \tag{A-3}$$

Suppose $\mathbf{X}$ is multivariate normally distributed with expectation $\boldsymbol{\mu}$ and covariance matrix Σ, which can be symbolically written as

$$\mathbf{X} \sim MVN(\boldsymbol{\mu}, \Sigma). \tag{A-4}$$

Here, $\sim$ denotes "is distributed as" and MVN indicates multivariate normally distributed. Then the conditional distribution of $\mathbf{X}^{(1)}$ given a realization $\mathbf{x}^{(2)}$ of $\mathbf{X}^{(2)}$ is multivariate normally distributed with expectation

$$\hat{\boldsymbol{\mu}}^{(1)} = \boldsymbol{\mu}^{(1)} + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)}), \tag{A-5}$$

where $\hat{\boldsymbol{\mu}}^{(1)}$ is the updated expectation. The updated covariance matrix $\hat{\Sigma}_{11}$ is given by

$$\hat{\Sigma}_{11} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \tag{A-6}$$

Theorem A-2

Suppose X_1, X_2, and X_3 are correlated random variables that satisfy

$$E[X_1|X_2 = x_2] \quad \text{is linear in } x_2, \tag{A-7}$$

and

$$E[X_1|X_3 = x_3] \quad \text{is linear in } x_3. \tag{A-8}$$

Then, given the correlation coefficients ρ_{12} between the pairs X_1 and X_2 and ρ_{13} between X_1 and X_3, the correlation coefficient ρ_{23} is given by

$$\rho_{23} = \rho_{12}\rho_{13}. \tag{A-9}$$

The conditions in the theorem imply, say for X_1, X_2, that given a realization x_2 of variable X_2, the expectation of X_1 shifts linearly towards x_2. For normal distributions this is always satisfied, as can be seen from theorem A-1, equation (A-5).

Although this theorem applies to three variables with one missing correlation coefficient only, we are also going to use it, without strict theoretical justification, for more than three variables where several correlation coefficients may be missing. We must note here, that, for more then three correlated variables, the positive definiteness of the covariance matrix may be violated by this procedure. In practice, we have seen this happen only in rare cases.

The following example illustrates the use of these theorems. Suppose the correlation matrix has been specified for five variables as follows:

$$C = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & * & 0.8 & * \\ 0 & * & 1 & * & 0.6 \\ 0 & 0.8 & * & 1 & 0.4 \\ 0 & * & 0.6 & 0.4 & 1 \end{bmatrix}. \tag{A-10}$$

In this particular example, ρ_{24}, ρ_{35}, and ρ_{45} are known coefficients and ρ_{34}, ρ_{25}, and ρ_{23} are unknown, which is indicated in the matrix by the * symbol. Using equation (A-9), we can determine two of the unspecified correlation coefficients.

$$\rho_{34} = \rho_{35}\rho_{54} = 0.24, \tag{A-11}$$

and

$$\rho_{25} = \rho_{24}\rho_{45} = 0.32. \tag{A-12}$$

However, ρ_{23} cannot be determined by the combination of two of the given correlation coefficients. In a second step, we can approximate ρ_{23} using the correlation coefficients determined previously:

$$\rho_{23} = \rho_{24}\rho_{43}, \tag{A-13}$$

which can be expanded using equation (A-11) to

$$\rho_{23} = \rho_{24}\rho_{35}\rho_{54} = 0.192. \tag{A-14}$$

Note, that we could also have used

$$\rho_{23} = \rho_{25}\rho_{53} = \rho_{24}\rho_{45}\rho_{53}. \quad \text{(A-15)}$$

In this particular case, the same value for ρ_{23} will be obtained for equations (A-14) and (A-15). In general, however the approximation is not unique. If several combinations are possible, when the number of initially specified correlation coefficients differs, then a selection is made from the combinations with the least number of initial coefficients. From these we arbitrarily choose one of the possible combinations. Thus, if in a different example, ρ_{23}, ρ_{34}, ρ_{35}, ρ_{45}, would have been specified, then we could obtain ρ_{25}, either from

$$\rho_{25} = \rho_{23}\rho_{35}, \quad \text{(A-16)}$$

or from

$$\rho_{25} = \rho_{23}\rho_{34}\rho_{45}. \quad \text{(A-17)}$$

The former expression is favored because it contains less specified correlation coefficients. With respect to the approximate nature of the procedure, we emphasize that after multiplying correlation coefficients, the resulting number comes closer and closer to zero. Therefore, the effect of the resulting approximation of the correlation coefficient decreases rapidly. Hence, we argue that making an error in the approximation has little effect when many terms are involved.

After application of the above procedure, the correlation matrix of equation (A-10) can be approximated by

$$\tilde{C} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0.192 & 0.8 & 0.32 \\ 0 & 0.192 & 1 & 0.24 & 0.6 \\ 0 & 0.8 & 0.24 & 1 & 0.4 \\ 0 & 0.32 & 0.6 & 0.4 & 1 \end{bmatrix}. \quad \text{(A-18)}$$

We can now draw samples for all variables. Suppose we would like to draw them in the order X_3, X_5, X_1, X_2, X_4. When selecting X_3, no other variable has been drawn, so we can simply draw it from its marginal probability density function $X_3 \sim N(\mu_3, \sigma_3^2)$. Now X_5 must be drawn, conditioned on the x_3 value. Using theorem A-1, we find

$$\hat{\mu}_5 = \mu_5 + \sigma_{35}(\sigma_3^2)^{-1}(x_3 - \mu_3), \quad \text{(A-19)}$$

and

$$\hat{\sigma}_5^2 = \sigma_5^2 - \sigma_{35}(\sigma_3^2)^{-1}\sigma_{53}, \quad \text{(A-20)}$$

where

$$\sigma_{35} = \rho_{35}\sigma_3\sigma_5, \quad \text{(A-21)}$$

is the covariance between X_3 and X_5. Now X_5 can be drawn from $N(\hat{\mu}_5, \hat{\sigma}_5^2)$.

Now X_1 is to be drawn. Since it is independent of X_2, X_3, X_4, and X_5 it can be drawn from its marginal distribution $N(\mu_1, \sigma_1^2)$. Finally, for X_2 and X_4 we use

$$\hat{\mu}_2 = \mu_2 + [\sigma_{23}\sigma_{25}]\begin{bmatrix} \sigma_3^2 & \sigma_{35} \\ \sigma_{35} & \sigma_5^2 \end{bmatrix}^{-1}\begin{bmatrix} x_3 - \mu_3 \\ x_5 - \mu_5 \end{bmatrix}, \quad \text{(A-22)}$$

$$\hat{\sigma}_2^2 = \sigma_2^2 - [\sigma_{23}\sigma_{25}]\begin{bmatrix} \sigma_3^2 & \sigma_{35} \\ \sigma_{35} & \sigma_5^2 \end{bmatrix}^{-1}\begin{bmatrix} \sigma_{23} \\ \sigma_{25} \end{bmatrix}, \quad \text{(A-23)}$$

and

$$\hat{\mu}_4 = \mu_4 + [\sigma_{24}\sigma_{34}\sigma_{54}]\begin{bmatrix} \sigma_2^2 & \sigma_{23} & \sigma_{25} \\ \sigma_{23} & \sigma_3^2 & \sigma_{35} \\ \sigma_{25} & \sigma_{35} & \sigma_5^2 \end{bmatrix}^{-1}\begin{bmatrix} x_2 - \mu_2 \\ x_3 - \mu_3 \\ x_5 - \mu_5 \end{bmatrix}, \quad \text{(A-24)}$$

$$\hat{\sigma}_4^2 = \sigma_4^2 - [\sigma_{24}\sigma_{34}\sigma_{54}]\begin{bmatrix} \sigma_2^2 & \sigma_{23} & \sigma_{25} \\ \sigma_{23} & \sigma_3^2 & \sigma_{35} \\ \sigma_{25} & \sigma_{35} & \sigma_5^2 \end{bmatrix}^{-1}\begin{bmatrix} \sigma_{24} \\ \sigma_{34} \\ \sigma_{54} \end{bmatrix}, \quad \text{(A-25)}$$

respectively. This allows us to draw the variables one by one in any order. Also, we can redraw any one of the variables when needed, and condition on the latest drawn value for each of the correlated variables.