Short Note on Models of Coregionalization

Oy Leuangthong (oy@ualberta.ca)
Department of Civil & Environmental Engineering, University of Alberta

Abstract

This note provides a synopsis of models of coregionalization that are common in geostatistical practice. Theoretical development is shown for most analytical models and brief discussions are given for heuristic models developed from simplifying multivariate statistical techniques.

Introduction

In practice, we are often interested in modelling more than one property and/or secondary data are often available as additional information for the modelling of a primary variable. Multivariate geostatistical techniques must be applied to capitalize on the available information. A longstanding challenge in multivariate geostatistics is the inference of a statistical model of coregionalization that permits simultaneous consideration of multiple variables.

Models such as the linear model of coregionalization (LMC) and the Markov Model are common in practice. Other analytical and heuristic models exist that are also useful in certain circumstances. This note provides an overview of some analytical and heuristic models of coregionalization for multivariate geostatistics.

Notation. To avoid confusion, the following mathematical notation is adopted. Random variables are defined with a subscript index to specify the data type, that is,

$$Z_i(\mathbf{u}_\alpha), i = 1, \ldots, P; \alpha = 1, \ldots, N$$

where

$$\begin{aligned} \alpha &= \text{index that specifies location in domain } \mathcal{A}, \\ N &= \text{total number of locations in domain } \mathcal{A} \\ i &= \text{index that specifies the data type,} \\ P &= \text{total number of different data types,} \end{aligned}$$

A particular outcome of the random variable is defined with a lower case z replacing the upper case Z. Furthermore, where reference is made to some arbitrary location in the domain $\mathcal{A}$, the subscript α will be dropped for simplicity. The number of data locations is n where $n < N$.

The first and second order moments of the RF are denoted as:

1. Mean:

$$E\{Z_i(\mathbf{u})\} = \mu_i(\mathbf{u})$$

2. Covariance:

$$Cov\{Z_i(\mathbf{u}) \cdot Z_j(\mathbf{u}+\mathbf{h})\} = C_{ij}(\mathbf{u}, \mathbf{u}+\mathbf{h}),\ i,j = 1,\ldots,P$$

For geostatistical inference, a decision of stationarity is required. For practical purposes, second order stationarity is often sufficient, that is,

$$\begin{aligned} E\{Z_i(\mathbf{u})\} &= \mu_i, \forall \mathbf{u} \in \mathcal{A} \\ Cov\{Z_i(\mathbf{u}) \cdot Z_j(\mathbf{u}+\mathbf{h})\} &= C_{ij}(\mathbf{h}), \forall i,j,\mathbf{h} \text{ and } \mathbf{u} \in \mathcal{A} \end{aligned}$$

Analytical Models

Consider a multivariate data set consisting of P types of data. Explicit characterization of the spatial relationship between the P variables requires a matrix of stationary covariance functions $C_{ij}(\mathbf{h}), i,j = 1,\ldots,P$:

$$C(\mathbf{h}) = \begin{bmatrix} C_{11}(\mathbf{h}) & \cdots & C_{1P}(\mathbf{h}) \\ \vdots & \ddots & \vdots \\ C_{P1}(\mathbf{h}) & \cdots & C_{PP}(\mathbf{h}) \end{bmatrix}, \forall \mathbf{h}$$

This covariance matrix is often assumed to be symmetric (i.e. $C_{ij}(\mathbf{h}) = C_{ji}(\mathbf{h})$). To ensure that all variances are non-negative, the covariance matrix must be positive semi-definite, that is all leading principal minor determinants of order k must be non-negative, $k = 1,\ldots,P$:

$$\det C(\mathbf{h}) = \sum_{i=1}^{P} (-1)^{i+j} \det U_{ij}(\mathbf{h}) \geq 0$$

where $\det U_{ij}(\mathbf{h})$ is a minor determinant of order $(P-1)$ of the $P \times P$ covariance matrix C, with the indices i and j denoting the row and column of C removed in order to form the $(P-1) \times (P-1)$ matrix [5]. For the simple case of a second order covariance matrix, the positive semi-definite constraint requires that

$$\begin{aligned} C_{ii}(\mathbf{h}) &\geq 0 \\ C_{jj}(\mathbf{h}) &\geq 0 \\ C_{ii}(\mathbf{h})C_{jj}(\mathbf{h}) &\geq C_{ij}(\mathbf{h})C_{ji}(\mathbf{h}),\ \forall i,j,\mathbf{h} \end{aligned}$$

Linear Model of Coregionalization

Consider P stationary random functions, $\mathbf{Z} = \{Z_1,\ldots,Z_P\}$. Suppose that each random function $Z_i, i = 1,\ldots,P$ can be expressed as a linear combination of K independent second-order stationary random functions, $Y_k, k = 1,\ldots,K$, each with zero mean and covariance function $C_k(\mathbf{h})$:

$$Z_i(\mathbf{u}) = \sum_{k=1}^{K} a_{ik} Y_k(\mathbf{u}) + \mu_i \tag{1}$$

where

$$\begin{aligned} E\{Z_i(\mathbf{u})\} &= \mu_i \\ E\{Y_k(\mathbf{u})\} &= 0 \\ C\{Y_k(\mathbf{u}), Y_{k'}(\mathbf{u}+\mathbf{h})\} &= C_k(\mathbf{h}), \text{ if } k = k' \\ &= 0, \text{ otherwise} \end{aligned} \tag{2}$$

Note that the RFs $Y_k, k = 1, \ldots, K$ are underlying and unknown (latent variables in statistical jargon). If the RFs $Y_k, k = 1, \ldots, K$ are grouped by those RFs Y_k with the same direct covariances $C_k(\mathbf{h})$, then Equation 1 can be written as:

$$Z_i(\mathbf{u}) = \sum_{l=0}^{L} \sum_{k=1}^{n_l} a_{ik}^l Y_k^l(\mathbf{u}) + \mu_i \tag{3}$$

with

$$\begin{aligned} C\{Y_k^l(\mathbf{u}), Y_{k'}^{l'}(\mathbf{u}+\mathbf{h})\} &= C^l(\mathbf{h}), \text{ if } k = k' \text{ and } l = l' \\ &= 0, \text{ otherwise} \end{aligned} \tag{4}$$

where $L+1$ is the number of groups with different direct covariances, and n_l is the number of RFs with the same covariance in group l. Based on Equation 3, the cross covariance of two RVs $Z_i(\mathbf{u})$ and $Z_j(\mathbf{u}+\mathbf{h})$ is

$$\begin{aligned} C_{ij}(\mathbf{h}) &= E\left\{ \left(\sum_{l=0}^{L} \sum_{k=1}^{n_l} a_{ik}^l Y_k^l(\mathbf{u}) \right) \left(\sum_{l'=0}^{L} \sum_{k'=1}^{n_{l'}} a_{jk'}^{l'} Y_{k'}^{l'}(\mathbf{u}+\mathbf{h}) \right) \right\} \\ &= \sum_{l=0}^{L} \sum_{l'=0}^{L} \sum_{k=1}^{n_l} \sum_{k'=1}^{n_{l'}} a_{ik}^l a_{jk'}^{l'} C\{Y_k^l(\mathbf{u}) Y_{k'}^{l'}(\mathbf{u}+\mathbf{h})\} \end{aligned} \tag{5}$$

Using the covariance in Equation 4 simplifies Equation 5 to

$$\begin{aligned} C_{ij}(\mathbf{h}) &= \sum_{l=0}^{L} \sum_{k=1}^{n_l} a_{ik}^l a_{jk}^l C\{Y_k^l(\mathbf{u}) Y_k^l(\mathbf{u}+\mathbf{h})\} \\ &= \sum_{l=0}^{L} \sum_{k=1}^{n_l} a_{ik}^l a_{jk}^l C^l(\mathbf{h}) \end{aligned} \tag{6}$$

From Equation 6, the sill of the l^{th} covariance structure, $C^l(\mathbf{h})$, is given by $\sum_{k=1}^{n_l} a_{ik}^l a_{jk}^l$. Defining $b_{ij}^l, i, j = 1, \ldots, P$ such that

$$b_{ij}^l = \sum_{k=1}^{n_l} a_{ik}^l a_{jk}^l$$

simplifies Equation 6 to

$$C_{ij}(\mathbf{h}) = \sum_{l=0}^{L} b_{ij}^l C^l(\mathbf{h}) \tag{7}$$

It only remains to determine $C^l(\mathbf{h}), l = 0, \ldots, L$ and the $(L+1) \cdot P^2$ parameters b_{ij}^l so that covariances are jointly positive definite. If the covariance models $C^l(\mathbf{h}), l = 0, \ldots, L$ are chosen to be known positive semi-definite models, this amounts to requiring that the $L+1$ matrices of b_{ij}^l coefficients are also positive semi-definite [2, 4, 6, 8]. For 2 variables, this constraint requires that

$$b_{ii}^l \cdot b_{jj}^l \geq b_{ij}^l \cdot b_{ji}^l, \forall i, j, l$$

In practice, this requires that for P variables, a total of $P(P+1)/2$ licit variograms are required to be modelled simultaneously to ensure positive definiteness. Consequences of a non-positive semi-definite covariance matrix are singular kriging systems and negative estimation errors.

Markov Models

Two models exist under this heading: Markov Model I and Markov Model II. The former is the more common Markov assumption used in most collocated co-kriging applications, while the latter is a variation of the original model for cases where the volume support of the secondary data is much larger than that of the primary data.

Markov Model I. Modelling direct and cross-variograms is a complex and tedious task. A Markov-type model of coregionalization simplifies this process. Consider two standard jointly Gaussian RVs, $Z_i(\mathbf{u})$ and $Z_j(\mathbf{u}), i \neq j$, which are the primary and secondary variable, respectively. The Markov-type assumption states that collocated hard data will screen the influence of other hard data that is further away [1, 15], i.e.

$$E\{Z_j(\mathbf{u}) | Z_i(\mathbf{u}) = z, Z_i(\mathbf{u}+\mathbf{h}) = z'\} = E\{Z_j(\mathbf{u}) | Z_i(\mathbf{u}) = z\} \tag{8}$$

Derivation of the Markov cross covariance model is based on determining the covariance of $Z_j(\mathbf{u})$ given $Z_i(\mathbf{u}) = z$ and $Z_i(\mathbf{u}+\mathbf{h}) = z'$, where $f_{\mathbf{h}}(z, z')$ is the bivariate pdf of $Z_i(\mathbf{u})$ and $Z_i(\mathbf{u}+\mathbf{h})$:

$$
\begin{aligned}
C_{ij}(\mathbf{h}) &= E\{Z_j(\mathbf{u}) \cdot Z_i(\mathbf{u}+\mathbf{h})\} \\
&= \int\int E\{Z_j(\mathbf{u}) \cdot Z_i(\mathbf{u}+\mathbf{h}) | Z_i(\mathbf{u}) = z, Z_i(\mathbf{u}+\mathbf{h}) = z'\} f_{\mathbf{h}}(z,z') dz dz' \\
&= \int\int z' E\{Z_j(\mathbf{u}) | Z_i(\mathbf{u}) = z, Z_i(\mathbf{u}+\mathbf{h}) = z'\} f_{\mathbf{h}}(z,z') dz dz' \\
&= \int\int z' E\{Z_j(\mathbf{u}) | Z_i(\mathbf{u}) = z) f_{\mathbf{h}}(z,z') dz dz' \text{ based on the Markov assumption}
\end{aligned}
$$

Since the two RVs are jointly Gaussian, the regression of Z_j on Z_i is

$$
E\{Z_j(\mathbf{u}) | Z_i(\mathbf{u}) = z) \quad = \quad \rho_{ij}(\mathbf{0}) \cdot z
$$

where $\rho_{ij}(\mathbf{0})$ is the correlation between $Z_i(\mathbf{u})$ and $Z_j(\mathbf{u})$ (i.e. collocated). This result gives the Markov cross covariance model:

$$
\begin{aligned}
C_{ij}(\mathbf{h}) &= \rho_{ij}(\mathbf{0}) \cdot \int\int z' z f_{\mathbf{h}}(z,z') dz dz' \\
&= \rho_{ij}(\mathbf{0}) \cdot C_{ii}(\mathbf{h}), \forall \mathbf{h} \qquad (9)
\end{aligned}
$$

For standardized variables, that is, random variables with unit variance, Equation 9 becomes

$$
\rho_{ij}(\mathbf{h}) = \rho_{ij}(\mathbf{0}) \cdot \rho_{ii}(\mathbf{h}), \forall \mathbf{h} \qquad (10)
$$

The Markov model only requires that the covariance function of the primary variable be modelled. The cross covariance between the primary and secondary variable is approximated using the relation in Equations 9 or 10. Use of only the collocated secondary data means that the covariance function of the secondary data is not required [1, 15].

One situation in which the Markov approximation is a poor assumption is the integration of data of significantly different volume supports. For example, suppose there is seismic and core data available at location $\mathbf{u}$. The small scale data from the core sample cannot screen the seismic data that informs a much larger volume although both data are centered at the same location.

Markov Model II. For the case when the secondary variable $Z_j(\mathbf{u})$ is defined on a much larger support than the primary variable $Z_j(\mathbf{u})$, Journel introduced a variation of the Markov assumption referred to as Markov Model II [7]. Simply stated, Markov Model II assumes that collocated *secondary* data will screen the influence of other *secondary* data that is further away [7], i.e.

$$
E\{Z_i(\mathbf{u}) | Z_j(\mathbf{u}) = z, Z_j(\mathbf{u}+\mathbf{h}) = z'\} = E\{Z_i(\mathbf{u}) | Z_j(\mathbf{u}) = z\} \qquad (11)
$$

The corresponding cross covariance model is

$$C_{ij}(\mathbf{h}) = C_{ij}(\mathbf{0}) \cdot C_{jj}(\mathbf{h}), \forall \mathbf{h} \tag{12}$$

Derivation of 12 follows from the derivation of Markov Model I. Unlike the more popular Markov Model I which requires only modelling of the covariance model of the primary variable, $C_{ii}(\mathbf{h})$ to define the cross covariance, Markov Model II requires that the covariance model of the secondary variable, $C_{jj}(\mathbf{h})$ be modelled. The resulting cross-covariance is obtained via relation 12. Further, the Markov Model II defines the primary covariance as a function of the secondary covariance and a residual covariance, $C_R(\mathbf{h})$ [7, 12]:

$$C_{ii}(\mathbf{h}) = C_{ij}^2(\mathbf{0}) \cdot C_{jj}(\mathbf{h}) + (1 - C_{ij}^2(\mathbf{0})) \cdot C_R(\mathbf{h}), \forall \mathbf{h} \tag{13}$$

Since Markov Model II requires modelling of the secondary covariance, from which subsequent definition of the primary and cross covariance is possible, the resulting model of coregionalization must be checked to ensure the model is positive semi-definite.

Markov-Bayes

Use of the LMC and the Markov models of coregionalization in traditional Gaussian algorithms only allows for transfer of linear, homoscedastic correlation. The Markov-Bayes model aims to account for the entire conditional distribution, not only the parameters that define the conditional distribution but also the *shape* of the distribution at each location $\mathbf{u}$ [9, 16]. This model was developed for the purpose of improving non-parametric geostatistics, specifically indicator simulation.

Suppose Z_i is the sparsely sampled primary variable, and Z_j is the densely sampled secondary variable. The primary data are considered "hard" data and are coded as indicators:

$$i(\mathbf{u}; z) = \begin{cases} 1, & \text{if } z_i(\mathbf{u}) \leq z \\ 0, & \text{otherwise} \end{cases}$$

where $z_i(\mathbf{u})$ is the primary data value at location $\mathbf{u}$. By convention the indicator random variable is denoted $I(\mathbf{u}; z)$ with outcome $i(\mathbf{u}; z)$, which is not to be confused with the variable index "i" for the primary variable.

Secondary data $Z_j(\mathbf{u})$ are used to define a "local prior" distribution of $Z_i(\mathbf{u})$. Secondary data are coded as probabilities or Y data:

$$y(\mathbf{u}; z) = prob\{Z_i(\mathbf{u}) \leq z | \text{related information}\}$$

where $y(\mathbf{u}; z) \in [0, 1]$ and $\neq F_i(z)$. For locations where hard data exists (i.e. $z_i(\mathbf{u})$ is known), the local prior cdf becomes

$$y(\mathbf{u}; z) = \begin{cases} 1, & \text{for all } z \leq z_i(\mathbf{u}) \\ 0, & \text{for all } z > z_i(\mathbf{u}) \end{cases}$$

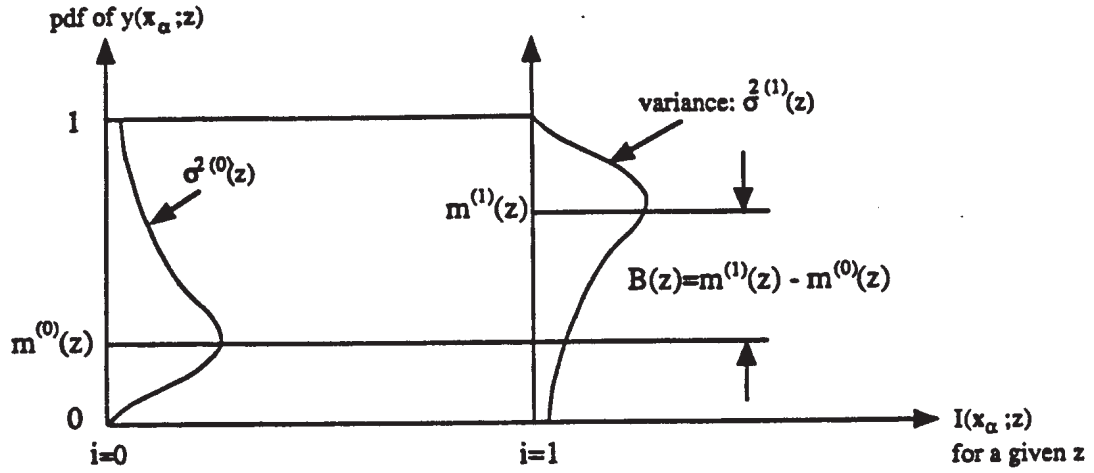

Figure 1: Graphical Interpretation of Calibration Parameter $B(z)$ of Soft Data in Markov-Bayes Updating. Source: Journel and Zhu, 1990

Secondary information can also be coded as a constraint interval or a continuous interval based on calibration to a bivariate distribution representing correlation between primary and secondary data [9, 16].

This model requires (1) a Markov-type assumption to simplify modelling of cross covariance models between $I(\mathbf{u};z)$ and $Y(\mathbf{u};z)$, and (2) use of Bayes theorem to update local prior distributions and obtain posterior conditional distributions, given that direct and cross covariances are known.

Based on the Markov approximation, the direct and cross covariances are calibrated by $B(z)$ (see Figure 1) [9, 16]:

$$C_{IY}(\mathbf{h};z) = B(z) \cdot C_I(\mathbf{h};z) \ \forall h$$

$$C_Y(\mathbf{h};z) = \begin{cases} B^2(z) \cdot C_I(\mathbf{h};z), & \forall \mathbf{h} > 0 \\ |B(z)| \cdot C_I(\mathbf{h};z), & \mathbf{h} = 0 \end{cases}$$

where

$$\begin{aligned} B(z) &= m^1(z) - m^0(z) \\ m^1(z) &= E\{Y(\mathbf{u};z)|I(\mathbf{u};z) = 1\} \\ m^0(z) &= E\{Y(\mathbf{u};z)|I(\mathbf{u};z) = 0\} \end{aligned}$$

Zhu and Journel (1990,1993) interpret the parameters $m^1(z)$ and $m^0(z)$ as a measure of accuracy of the local prior distributions of $y(\mathbf{u};z)$ in predicting $Z_i(\mathbf{u}) \leq z$ and $Z_i(\mathbf{u}) >$

z, respectively. The value of $B(z)$ is then indicative of the accuracy of inference using soft data [9, 16]. A calibration of $B(z) = 1$ is considered the best in terms of accuracy since this means that the primary and secondary RV are perfectly spatially correlated, i.e. $C_Y(\mathbf{h}; z) = C_I(\mathbf{h}; z)$ and $C_{IY}(\mathbf{h}; z) = C_I(\mathbf{h}; z) \forall \mathbf{h}$. Conversely, $B(z) = -1$ is interpreted as perfect "error" where the event $Z_i(\mathbf{u}; z) \leq z$ is actually assigned the probability of $Z_i(\mathbf{u}; z) > z$. The worst case occurs when $B(z) = 0$ which indicates that soft information $Y(\mathbf{u}; z)$ does not help in predicting the value of the indicator $I(\mathbf{u}; z)$.

Once the calibration parameter $B(z)$ is established, the model of coregionalization is fully defined. The Markov approximation along with Bayesian updating requires only the direct covariance of the primary variable or hard data be modelled. As this model is dependent on a Markov assumption, it is also a poor approximation when conditioning to data of significantly different supports.

Empirical Models

Contrary to the previous models of coregionalization, those discussed in this section are not easily defined analytically. The models that fall within this group are those that result from transformation techniques applied to simplify multivariate (geo)statistics.

Principal Component Analysis. The objective of this multivariate transformation is to reduce the dimensionality of the data by identifying new variables that are linear combinations of the original variables. It produces new variables that are uncorrelated with each other. Reducing the dimension of the problem relieves the effort required to infer a model of coregionalization. Orthogonality of the new variables is assumed to extend to spatial coordinates. In practice, this assumption is verified by visual inspection of the cross-covariance models between pairs of variables [13, 14].

Goovaerts explored the relation between the covariance of the principal components and the covariance of the actual regionalized variables in greater detail for the cases of intrinsic correlation, and for two and three basic structures with a nugget effect [3].

Stepwise Conditional Transform. This technique is a multivariate transformation technique that produces independent multiGaussian variables at $\mathbf{h} = 0$, with the added possibility of independent simulation [10, 11]. Secondary variables are conditionally transformed to standard normal distributions based on the probability classes of previously transformed primary variables. The result is a set of transformed secondary variables that are a combination of multiple variables; the corresponding covariance model for these variables reflects the mixture of both primary and secondary variables, i.e. the direct and cross covariance models of the transformed variables are a function of the direct and cross covariances at all *lower* levels [10].

Leuangthong and Deutsch explored the model of coregionalization that results from applying this transformation, with specific results given for the intrinsic correlation scenario [10].

Remarks

Regardless of the model that is adopted, the matrix of covariances must be checked to ensure that the system is indeed positive semi-definite so as to guarantee non-negative kriging variances. Once the model is verified, we have all the information required in order to solve the kriging equations for conventional multivariate geostatistics.

References

[1] A. S. Almeida and A. G. Journel. Joint simulation of multiple variables with a Markov-type coregionalization model. *Math Geology*, 26:565–588, 1994.

[2] C. Deutsch. Geostatistical methods for modelling earth sciences data. Unpublished MIN E 612 Course Notes, University of Alberta, Alberta, Canada, 1999.

[3] P. Goovaerts. Spatial orthogonality of the principal components computed from coregionalized variables. *Math Geology*, 25:281–302, 1993.

[4] P. Goovaerts. *Geostatistics for Natural Resources Evaluation*. Oxford University Press, New York, 1997.

[5] J. Harris and H. Stocker. *Handbook of Mathematics and Computational Science*. Springer-Verlag, New York, 1998.

[6] E. H. Isaaks and R. M. Srivastava. *An Introduction to Applied Geostatistics*. Oxford University Press, New York, 1989.

[7] A. Journel. Markov models for cross-covariances. *Mathematical Geology*, 31:955–964, 1999.

[8] A. G. Journel and C. J. Huijbregts. *Mining Geostatistics*. Academic Press, New York, 1978.

[9] A. G. Journel and H. Zhu. Integrating soft seismic data: Markov-Bayes updating, an alternative to cokriging and traditional regression. In *Report 3, Stanford Center for Reservoir Forecasting*, Stanford, CA, May 1990.

[10] O. Leuangthong and C. Deutsch. Stepwise conditional transformation for simulation of multiple variables. *Mathematical Geology*. Submitted 2001.

[11] M. Rosenblatt. Remarks on a multivariate transformation. *Annals of Mathematical Statistics*, 23(3):470–472, 1952.

[12] L. Shmaryan and A. Journel. Two markov models and their applications. *Mathematical Geology*, 31:965–988, 1999.

[13] V. Suro-Pérez. Indicator kriging based on principal component analysis. Master's thesis, Stanford University, Stanford, CA, 1988.

[14] V. Suro-Pérez. Indicator principal component kriging: the multivariate case. In A. Soares, editor, *Geostatistics-Troia*, volume 1, pages 441–454. Kluwer, 1993.

[15] W. Xu, T. T. Tran, R. M. Srivastava, and A. G. Journel. Integrating seismic data in reservoir modeling: the collocated cokriging alternative. In *67th Annual Technical Conference and Exhibition*, pages 833–842, Washington, DC, October 1992. Society of Petroleum Engineers. SPE paper # 24742.

[16] H. Zhu and A. G. Journel. Formatting and integrating soft data: Stochastic imaging via the Markov-Bayes algorithm. In A. Soares, editor, *Geostatistics Troia 1992*, volume 1, pages 1–12. Kluwer, 1993.