Paper presented at the VIII Geostatistical Congress, Santiago, Chili
December 1-5, 2008. To appear in the proceedings.

COLLOCATED CO-SIMULATION USING PROBABILITY AGGREGATION

G. MARIETHOZ[1], PH. RENARD[1], R. FROIDEVAUX[2].
[1] CHYN, University of Neuchâtel, rue Emile Argand 11, CH - 2009 Neuchâtel, Switzerland
[2] FSS Consultants, 9, rue Boissonnas, CH - 1227 Geneva, Switzerland

ABSTRACT

In this paper, we propose a new cosimulation algorithm for simulating a primary variable using one or several secondary attributes known exhaustively on the domain.

At each node of the grid to be simulated, two conditional distribution functions are inferred. The first one comes from the available conditioning data of the main attribute using, for instance, a multi-Gaussian framework. The second one is the distribution function of the main attribute conditioned to the co-located value of the secondary attribute. Using a conjunction of probabilities approach, these two distribution functions are then combined into a single one from which an outcome is drawn.

The paper discusses the probabilistic model underlying this approach and illustrates its performances using both synthetic and actual field data. The synthetic examples cover a range of different types of joint probability distributions inspired from relations between geophysical parameters and hydraulic parameters. The field example uses remote sensing images.

INTRODUCTION

Exhaustive maps of secondary information are often available and can guide the simulation of a primary variable by using an appropriate coregionalization model. For instance, geophysical surveys provide exhaustive maps of electric resistivity, which is related to hydraulic conductivity, the attribute of main interest.

Several approaches have been proposed to address this problem: full co-kriging and co-located cokriging which assume a linear model of coregionalization (Journel 1999; Rivoirard 2001) and the so-called cloud transform technique (Bashore et al. 1994; Kolbjørnsen and Abrahamsen 2004) which uses a non-

parametric description of the bivariate distribution between the main and secondary attributes thus allowing to accommodate situations in which the marginal distributions are multimodal and the relationship between the two attributes are non-linear or heteroscedastic. This joint distribution can be inferred from data by interpolation in the probability space (e.g. kernel smoothing, Epanechnikov 1969; Kolbjørnsen and Abrahamsen 2004), or deduced from known physical laws or from interpretation.
In all the above methods a single local conditional distribution function of the main attribute is estimated directly. We propose a new approach, in which two separate distribution functions are inferred locally. The first one is estimated considering the available conditioning data only. The second one is extracted from the bivariate distribution model using the collocated value of the secondary attribute. These two distributions are then combined into a single one using the concept of probability conjunction (Tarantola 2005), which can be seen as a particular case of the theory of Bordley (1982) used in management science for aggregating expert's opinions. Note that similar ideas were used by Ortiz and Deutsch (2004) to combine indicator kriging probabilities with multiple-points statistics.

SIMULATION BY PROBABILITY AGGREGATION

Outline of the Method

Denote:

$Z(\mathbf{u})$: the attribute of main interest.
$S(\mathbf{u})$: the co-located attribute.
$z(\mathbf{u}_i)$, $i=[1 \dots N]$: available conditioning data for the main attribute.
$f(z,s)ds,dz = \text{Prob}\{z<Z\leq z+dz \ , \ s<S\leq s+ds\}$: the joint probability density function.

Given this joint probability model, the marginal probability distribution function $f(z)$ is given by the integral on the real line:

$$f(z) = \frac{1}{\eta}\int_R f(z,s)ds\,, \tag{1}$$

where η is a normalizing factor.
In all generality at location $\mathbf{u}$, the distribution function of the main attribute $Z(\mathbf{u})$ conditional to the neighboring data is given by :

$$F^1(\mathbf{u};z) = \text{Prob}\left\{Z(\mathbf{u}) \leq z \mid z(\mathbf{u}_1),...,z(\mathbf{u}_N)\right\} \tag{2}$$

This ccdf can be estimated using any suitable geostatistical method (e.g. multiGaussian kriging, indicator kriging).

At the same location **u**, the distribution function of Z(**u**) conditional to the co-located attribute s(**u**) is:

$$F^2(\mathbf{u};z) = \text{Prob}\{Z(\mathbf{u}) \le z \mid S(\mathbf{u}) = s(\mathbf{u})\} \tag{3}$$

The issue is therefore to combine (aggregate) $F^1(\mathbf{u};z)$ and $F^2(\mathbf{u};z)$ into a single ccdf $F(\mathbf{u};z)$ which would be an approximation of:

$$\text{Prob}\{Z(\mathbf{u}) \le z \mid z(\mathbf{u}_1),...,z(\mathbf{u}_N), S(\mathbf{u}) = s(\mathbf{u})\} \tag{4}$$

Once this ccdf is available, the simulation proceeds as usual in sequential simulation: an outcome is drawn by Monte-Carlo from $F(\mathbf{u};z)$ and treated as conditional data thereafter.

Probability Conjunction

Due to the sequential character of pixel-based simulations methods, ccdfs defined in (2) and (3) are not independent, because $F^1(\mathbf{u};z)$ is based on previously simulated nodes that already integrated information from the joint distribution.

Management science provides methods for aggregating expert's opinions in a Bayesian framework while dealing with data interaction. Bordley (1982) demonstrates that aggregating n probability density functions (pdfs) $f^k = F^{k\prime}$, $k = [1 \dots n]$, while accounting for data interaction, can be accomplished by:

$$f(z) = \left(\prod_{k=1}^{n} \left(f^k(z) \right)^{w_k} \right) \left(f^0(z) \right)^{\left(1 - \sum_{k=1}^{n} w_k\right)}, \tag{5}$$

each probability having the weight w_k, that can be seen as a way of quantifying redundancy or as a confidence factor. f^0 is the prior density function which is, in our case, the marginal pdf $f(z)$ defined in (1) (i.e. the homogeneous state of information).

Equation (5) is closely related to *tau* and *nu* models (Journel 2002; Krishnan 2005; Polyakova and Journel 2007). These models use a parallel approach that formulates the problem in terms of odd functions (Bordley 1982). In the same spirit, Tarantola (2005) defines the conjunction of probability densities by the operation

$$f(\mathbf{u};z) = f^1(\mathbf{u};z) \wedge f^2(\mathbf{u};z) = \frac{1}{\eta} \frac{f^1(\mathbf{u};z)\, f^2(\mathbf{u};z)}{f(z)}, \tag{6}$$

which is a particular case of (5) with two probabilities being aggregated and identical confidence factors equal to 1.

Remarks

1. Although in our case all the weight w_k are equal and set to 1, such does not necessarily be the case: they could be used to assign relative weight to an expert-provided bivariate distribution model or to account for possible non-stationary model uncertainty. Setting a weight to 0 would produce the uniform distribution, resulting in the corresponding source of information having no influence.
2. The proposed method can easily account for more than one secondary attribute.

Step-by-Step Algorithm

The proposed algorithm performs the following steps:

- Compute the marginal cdf of the primary attribute from the joint probability model.
- Assign conditioning data to nearest grid nodes.
- Define a suitable path through the grid nodes.
- At each node:
    - Collect conditioning information.
    - Estimate $f^1(\mathbf{u};z)$ using an appropriate method.
    - Extract $f^2(\mathbf{u};z)$ from the bivariate model.
    - Estimate $f(\mathbf{u};z)$ by conjunction of probabilities.
    - Draw an outcome $z'(\mathbf{u})$ from $F(\mathbf{u};z)$ and add it to the data set.

SYNTHETIC EXAMPLES

The probability aggregation algorithm has been tested on synthetic data sets, with multiGaussian kriging used for estimating $f^1(\mathbf{u};z)$. For each example, a bivariate density function is known. A reference field for the primary variable is obtained by applying a normal score transform to a Gaussian simulation. The secondary variable is constructed from the (fully known) primary variable by drawing for each node **u** a value in

$$\text{Prob}\left\{S(\mathbf{u}) \le s \mid Z(\mathbf{u}) = z_1\right\} . \quad (7)$$

Then, for each example, the primary variable is sampled at 50 random locations. This dataset is used as conditioning data. The simulation grid size is 50x50 cells for all synthetic examples, and 100 realizations are computed.
Simulations are then compared to the reference which is known exhaustively, in order to evaluate the performances of the method. The comparison criteria are the reproduction of the histogram and variogram, the errors compared to the known reality and the visual aspect of the simulations.

Figure 1 illustrates the method with a primary variable (fig. 1.a) that has an exponential variogram model with a range of 5 and a sill of 6.3. The resulting simulations are shown in fig. 1.b and 1.c. Locations of samples data are marked by circles. Inside the circles are stars whose color indicates the value of the conditioning data.
A noisy secondary variable (fig. 1.d) results from the custom crescent-shaped joint PDF (fig. 1.e), but it still contains enough information to guide the simulations, where features of the reference are present at locations where no data are available (for example the dark channel that runs through the field from left to right).
The joint distribution (fig. 1.f), the reference histogram (fig 1.g) and variogram (fig 1.h) are well reproduced (solid line represents the reference, dotted lines the simulations and circles the samples data). There is no systematic bias, as shown by the histogram of errors of the simulated variable that is centered on 0 (fig. 1.i).
The next examples are illustrated in the same manner. Figure 2 uses a relationship that presents a low correlation coefficient (0.5) and Figure 3 uses an almost perfect dependency (correlation coefficient of 0.99), with pure nugget effect for the primary variable.
For all synthetic cases, the method allows an overall good reproduction of histograms, variograms and joint distributions. Moreover, the features of the primary variable field are generally well reproduced.

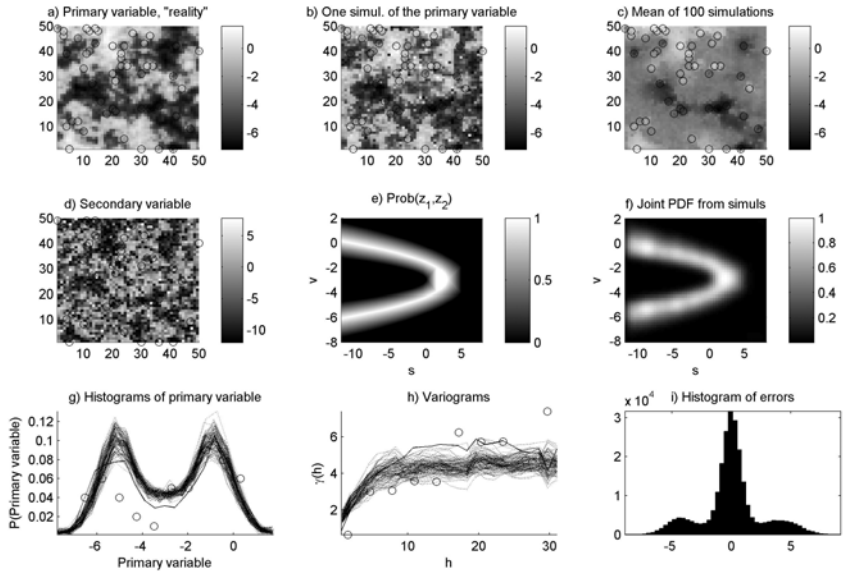

Figure 1: Synthetic example using a custom non-bijective joint PDF. The primary variable has a spherical variogram model (sill=5.3, range=12, adjusted on the 50 sample data).

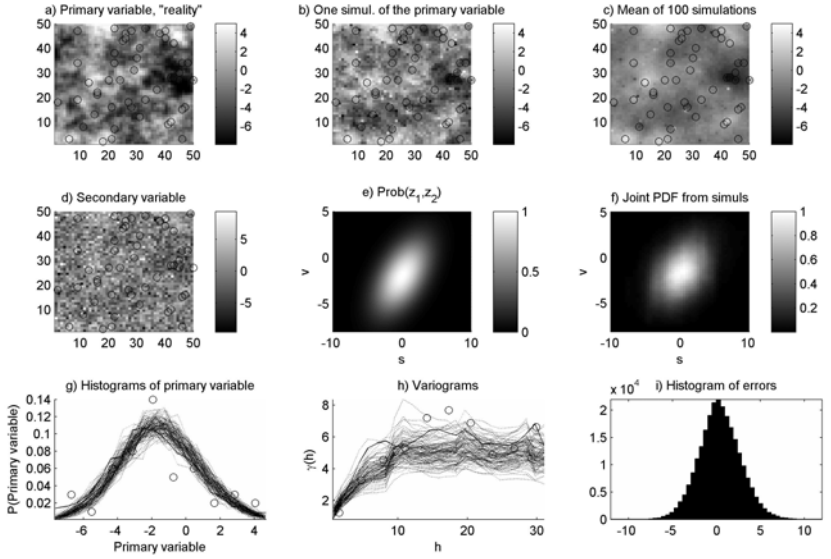

Figure 2: Synthetic example using a loosely correlated joint PDF *(ρ=0.5)*. The primary variable has an exponential variogram model (sill=6.3, range=5, adjusted on the 50 sample data).

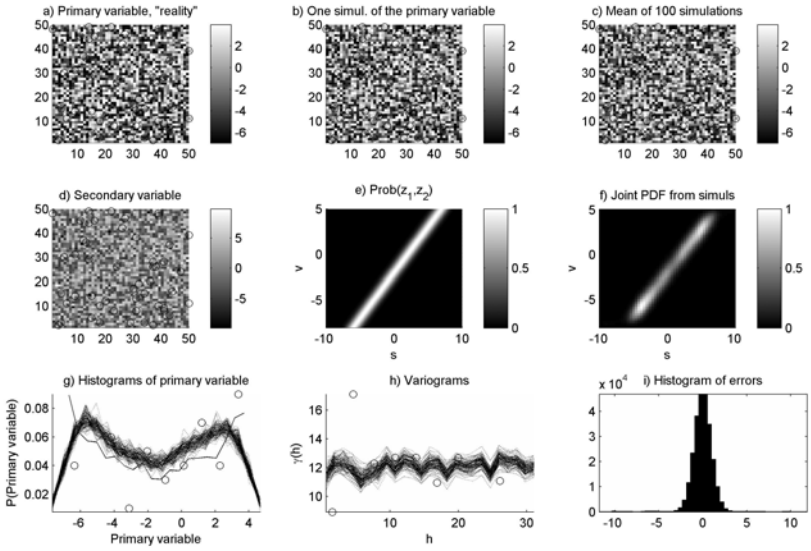

Figure 3: Synthetic example using a joint PDF with *ρ=0.99*. The primary variable is pure nugget effect (sill=12, adjusted on the 50 sample data).

FIELD EXAMPLE

In order to test further the method, a real case data set has been used. It is based on two Landsat 7 satellite images of the same area corresponding to two different wavelengths. One image is considered to be the primary variable while the other is the secondary variable. Each image highlights different features of the land surface and the joint relationship is thus complex.

As for synthetic examples, multiGaussian kriging is used for estimating $f'(\mathbf{u};z)$. Again, the first image is used as a reference, sampled at 100 random locations, while the second is the auxiliary variable. The size of the simulation grid is 181x201. This dataset is used to build the joint distribution by kernel smoothing. Figure 1 presents the results of the 100 simulations.

Results show a good match with the reference primary variable field, even if the joint PDF is inaccurate because it was constructed using only 100 points, which is not enough to capture all features of the real field (as shown by the discrepancy between the reference variogram and the variogram of the data).

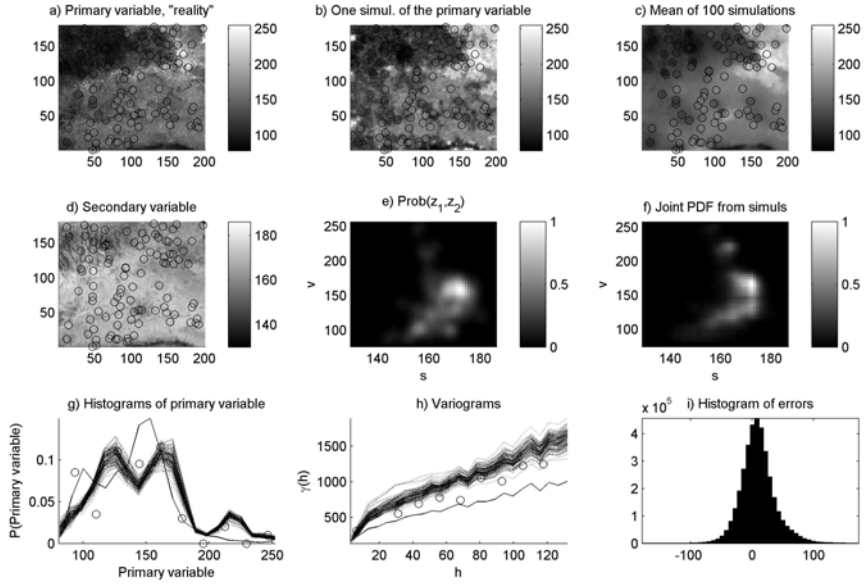

Figure 1: Real case example. The primary variable has an exponential variogram model (sill=1110, range=30, adjusted on the 100 sample data).

DISCUSSION AND CONCLUSION

We presented a new method for conditional co-simulations using a secondary variable and accounting for a non-parametric joint distribution.

The main advantage of the proposed method, as compared to traditional co-simulation methods, is that the algorithm can be applied when the relationship

between variables is too complex for a linear model to hold and in particular when this relation is not bijective. The method is based on the concept of probability aggregation and is implemented in the framework of Sequential Gaussian Simulation. It could be extended to any sequential, pixel-based simulation technique that uses local ccdfs. Moreover, it is not limited to continuous variables.

A point to discuss is that the respective weights of the spatial correlation model and the collocated information have been set arbitrarily to one and remain identical during the simulation process when aggregating these probabilities. This may be questionable due to the complex interaction between the two sources of information during the simulation process. However, the numerical examples indicate that the method provides satisfactory results.

Adjusting the weights could be a powerful way of parametrizing and extending the method. Without entering deeply into the details, the important aspects to bear in mind are first that the sum of the weights (equal to 2 here) allows to reinforce (or not) the information provided separately by each source of information, and second that the individual weights are related to the confidence that is associated with a given source of information (Bordley, 1982). Ongoing work aims at calibrating the weights to aggregate variables having different information contents.

At this stage of the work, spatial cross-correlations between primary and secondary variables were neglected, but integrating them in the simulation algorithm would not change drastically the theoretical background.

Finally, this paper shows that the concept of probability conjunction or aggregation, originating from management science, is a precious tool for integrating secondary variables in geostatistical simulations. It allows combining information originating from diverse sources in a straightforward implementation.

ACKNOWLEDGEMENTS

Funding for this work was provided by the Swiss National Science Foundation (contract PP002-1065557). We want to thank Denis Allard (INRA) for his constructive comments. We also thank Albert Tarantola (Institut de Physique du Globe de Paris) for enlightening lessons and François Bertone (BCEOM Engineering) for initiating this project by giving us a real-case problem.

REFERENCES

Bashore W, U.Araktingi, Levy M, Schweller U (1994) Importance of a Geological framework for Reservoir modelling and subsequent Fluid-Flow Predictions. In: Chambers JMYaRL (ed) AAPG Computer application in geology: 159-175

Bordley RE (1982) A multiplicative formula for aggregating probability assessments. Management Science 28: 1137-1148

Epanechnikov VA (1969) Nonparametric estimation of a multidimensional probability density. Theoretical Probability Applications: 153-158

Journel A (1999) Markov Models for Cross-Covariances. Mathematical Geology 31: 955-964

Journel AG (2002) Combining Knowledge From Diverse Sources: An Alternative to Traditional Data Independence Hypotheses. Mathematical Geology 34: 573-596

Kolbjørnsen O, Abrahamsen P (2004) Theory of the cloud transform for applications. In: Deutsch OLaCV (ed) Geostatistics Banff 2004. Kluwer Academic Publisher: 45-54

Krishnan S (2005) Combining diverse and partially redundant information in the earth sciences. Ph.D, Stanford University

Ortiz JM, Deutsch CV (2004) Indicator Simulation Accounting for Multiple-Point Statistics. Mathematical Geology 36: 545-565

Polyakova E, Journel A (2007) The Nu Expression for Probabilistic Data Integration Mathematical Geology 39: 715-733

Rivoirard J (2001) Which models for collocated cokriging. Mathematical Geology 33: 117-131

Tarantola A (2005) Inverse Problem Theory and Methods for Parameter estimation. Society for Industrial and Applied Mathematics, Philadelphia