Bernstein copula-based spatial stochastic simulation of petrophysical properties using seismic attributes as secondary variable

Martín A. Díaz-Viera[1], Arturo Erdely[2], Tatiana Kerdan[3], Raúl del-Valle-García[4] and Francisco Mendoza-Torres[5]

Abstract A novel Bernstein copula-based spatial stochastic co-simulation (BCSCS) method for petrophysical properties using seismic attributes as a secondary variable is presented. The method is fully non-parametric and it has the ad-vantages of not requiring linear dependence between variables. The method-ology is illustrated in a case study from a marine reservoir in the Gulf of Mexico and the results are compared with sequential Gaussian co-simulation method.

Introduction

Modeling the spatial distribution of petrophysical properties in the framework of reservoir characterization is a crucial and difficult task due to the lack of data and hence the degree of uncertainty associated with it. For this reason, in recent years a stochastic simulation approach for the spatial distribution of petrophysical properties has been adopted.

Seismic attributes have been extensively used as secondary variables in static reservoir modeling for petrophysical property prediction but usually assuming linear dependence and Gaussian distribution (Parra & Emery, 2013).

Quite recently, copulas have become popular for being a flexible means of representing dependency relationships in the financial sector and applications are already emerging in the field of geostatistics (Bardossy & Li, 2008) (Kazianka & Pilz, 2010).

[1] Instituto Mexicano del Petróleo, Eje Central Lázaro Cárdenas Norte 152, CP 07730, Ciudad de México, México, mdiazv@imp.mx,

[2] Facultad de Estudios Superiores Acatlán, UNAM, México, arturo.erdely@comunidad.unam.mx

[3] Instituto Mexicano del Petróleo, Eje Central Lázaro Cárdenas Norte 152, CP 07730, Ciudad de México, México, tkerdan@imp.mx,

[4] Instituto Mexicano del Petróleo, Eje Central Lázaro Cárdenas Norte 152, CP 07730, Ciudad de México, México, rvalleg@imp.mx,

[5] Posgrado de Ciencias de la Tierra, UNAM, México, mentofran@gmail.com

A geostatistical simulation method, based on Bernstein copula approach as a tool to represent the underlying dependence structure between petrophysical properties and seismic attributes, is proposed. The procedure basically consists of applying the simulated annealing method with a joint probability distribution model estimated by a Bernstein copula in a completely non-parametric fashion (Hernández-Maldonado, Díaz-Viera, & Erdely, 2012).
The method has the advantages of not requiring linear dependence or a specific type of distribution. The application of the methodology is illustrated in a case study where the results are compared with sequential Gaussian co-simulation method.

Methodology

As stated in the introduction, the main goal of this work is to show the application of a Bernstein copula-based spatial co-simulation method for petrophysical property predictions using seismic attributes as secondary variables and its comparison with the classical sequential Gaussian co-simulation method. In what follows a brief description of both methods and a general workflow outline are presented.

Sequential Gaussian co-simulation (SGCS)

The sequential Gaussian co-simulation (SGCS) method is very well established in the geostatistics literature, so here we will just mention the details of its application. Usually this method is applied with a linear model of coregionalization (Chiles & Delfiner, 1999) which is mostly unnatural, forced, very complicated and difficult to establish. The method assumes the existence of very strong linear dependence between primary and secondary variables, which is its main assumption and at the same time its main drawback. Here we choose to use an alternative variant, the Markov Model, given in (Chiles & Delfiner, 1999, p. 305) and implemented in SGeMS (Remy, Boucher, & Wu, 2009).

Bernstein copula-based spatial stochastic co-simulation (BCSCS)

A Bernstein copula-based spatial stochastic co-simulation (BCSCS) method has been previously presented in a series of papers (Hernández-Maldonado, Díaz-Viera, & Erdely, 2012), (Hernández-Maldonado, Díaz-Viera, & Erdely, 2014) and has been mainly applied in one dimension for petrophysical properties at well-log scale.

The method basically consists of establishing a dependence model between a primary and a secondary variable, and then use this model in conjunction with the spatial dependence structure (variogram) of the primary variable to predict the first one using the second one as a conditioning variable. This can be done in a global optimization framework using simulated annealing method, but other methods, such as genetic algorithms, could also be applied.

The modern way to analyze dependencies is by copula approach (Joe, 1997). Copula approach assumes neither a predetermined nor a priori type of dependency, but from the data one tries to establish the best model that represents the existing dependence on them.

In particular, here it is preferred to use a completely non-parametric approach to modeling the dependence by using Bernstein copulas, which gives name to the method. However other approaches, parametric (Díaz-Viera & Casar-González, 2005) and semi-parametric (Erdely & Diaz-Viera, 2010), are also possible. The Bernstein copulas introduced by (Sancetta & Satchell, 2004) are nothing more than an approximation of the sample copula by Bernstein polynomials. Its main shortcoming is the curse of dimensionality, as it quickly becomes computationally prohibitive for more than two dimensions. Alternatives have been proposed using vine copulas (Erdely & Diaz-Viera, 2016).

In summary, the algorithm consists of two stages:

1. A dependence model, using a Bernstein copula, is established from which a number of sample values are generated (see Appendix A).
2. A stochastic spatial simulation is performed using a simulated annealing method with a variogram model and a bivariate distribution functions as objective functions (Deutsch & Cockerham, 1994), (Deutsch & Journel, 1998).

Additional details about the mathematical formulation of the method and its computational implementation can be found on (Hernández-Maldonado, Díaz-Viera, & Erdely, 2012) and (Hernández-Maldonado, Díaz-Viera, & Erdely, 2014)

Workflow outline

One of the biggest challenges in these applications is to simultaneously handle multiple scales. Here, we have two scales: a well-log scale and a seismic scale. But sometimes, due to the amount of data, an additional intermediate scale is required, since the well-log scale data is a very large dataset and upscaled well-logs may have from a statistical point of view not enough data. Log data usually have a sampling interval in the range of 10-25 centimeters, while seismic data are in the range of several meters. So it is necessary to perform some upscaling process to make well-logs compatible with seismic data. For the upscaling process there is no single recipe because it is largely dependent on the data. Here we will use the median as upscaling procedure.

The general workflow is as follows: 1- univariate data analysis, 2- bivariate dependence analysis, 3- variography analysis and 4- simulations.

Case study

Data used in the case study are from a marine reservoir in the Gulf of Mexico. The reservoir is siliciclastic and it is formed mainly by alternating sequences of sands and shales.

Data description

The data consist of a total porosity well-log from a well and seismic attribute (P-impedance) obtained in a vertical (inline) section. The well-log has a sample interval of 0.1 m. The section has a length of 412.5 m and covers an interval of 336.4 m in depth and was chosen so that the well was located in the middle of it (see Figure 1). Seismic grid is made of 33 intervals of 12.5 m in X direction and 60 intervals of 5.5 m in depth direction.

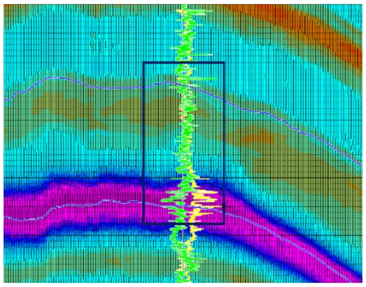

Figure 1. Vertical (inline) section with P-impedance as a result of seismic inversion. The color scale represents impedance values. In the middle of the section two logs are plotted along a well: in yellow P-impedance and in green total porosity.

At well-log scale, the P-wave impedance log is obtained from the product of P-wave velocity and density logs. At the seismic scale, a seismic inversion method was used, based on the “LP Sparse Spike” approach by (Li, 2001). Proper care was taken in incorporating low frequency impedance trend so that the impedance from the log and the impedance from the seismic section are equivalent around the well.

The impedance in general depends on the type of rock and its petrophysical properties as well as the containing fluid types and their saturations. It is very common in reservoir geophysics to take advantage of dependency relationships between petrophysical properties (for instance, total porosity and P-impedance) at well-log scale to predict the former ones (total porosity) using seismic attributes (P-impedance) at the seismic scale.

In particular, in this work the total porosity was considered as primary variable (variable to predict) and P wave impedance as secondary variable (conditioning variable). As mentioned in the previous section three scales are considered: a well-log, a seismic and an additional intermediate “one-meter” scale. Hereafter the following notation will be used:

- PhiT_T and Ip_T for total porosity and P-impedance from original well-logs (well-log scale)
- PhiT_T_1m and Ip_T_1m for total porosity and P-impedance from original well-logs subsampled every meter (one-meter scale)
- PhiT_T_median and Ip_T_median for total porosity and P-impedance from original well-logs upscaled using median upscaling procedure (seismic scale)
- Ip_inline for P-impedance from the vertical inline section (seismic scale)
- Ip_inline_U for P-impedance from the vertical inline section restricted to the corresponding well coordinates, i.e., only along the well trajectory (seismic scale)

Univariate data analysis

In Figure 2 are shown histograms and boxplots for PhiT and Ip at the three scales, and in Table 1 and Table 2 a summary of corresponding basic univariate statistics.

Table 1 Statistics summary of original and one-meter upscaled well logs.

Statistics	PhiT_T	Ip_T	PhiT_T_1m	Ip_T_1m
n	4059	4059	337	337
Minimum	0.030	4802.22	0.057	4802.22
1st. Quartile	0.218	6163.18	0.230	6064.66
Median	0.261	6717.99	0.273	6481.54
Mean	0.257	6906.39	0.266	6740.08
3rd. Quartile	0.299	7270.29	0.304	7099.52
Maximum	0.571	11812.36	0.556	11013.43
Variance	0.004	1264603	0.003	1133016

Table 2 Statistics summary of median upscaled well logs and Ip at seismic scale.

Statistics	PhiT_T_med	Ip_T_med	Ip_inline_U	Ip_inline
n	60	60	60	1980
Minimum	0.110	5730.23	5940.26	5940.16
1st. Quartile	0.239	6074.77	6080.53	6083.72
Median	0.273	6426.24	6350.74	6340.04
Mean	0.267	6619.13	6623.50	6602.87
3rd. Quartile	0.296	6967.11	6896.91	6893.98
Maximum	0.340	10430.77	8726.18	8773.20
Variance	0.002	630041	577250	485570

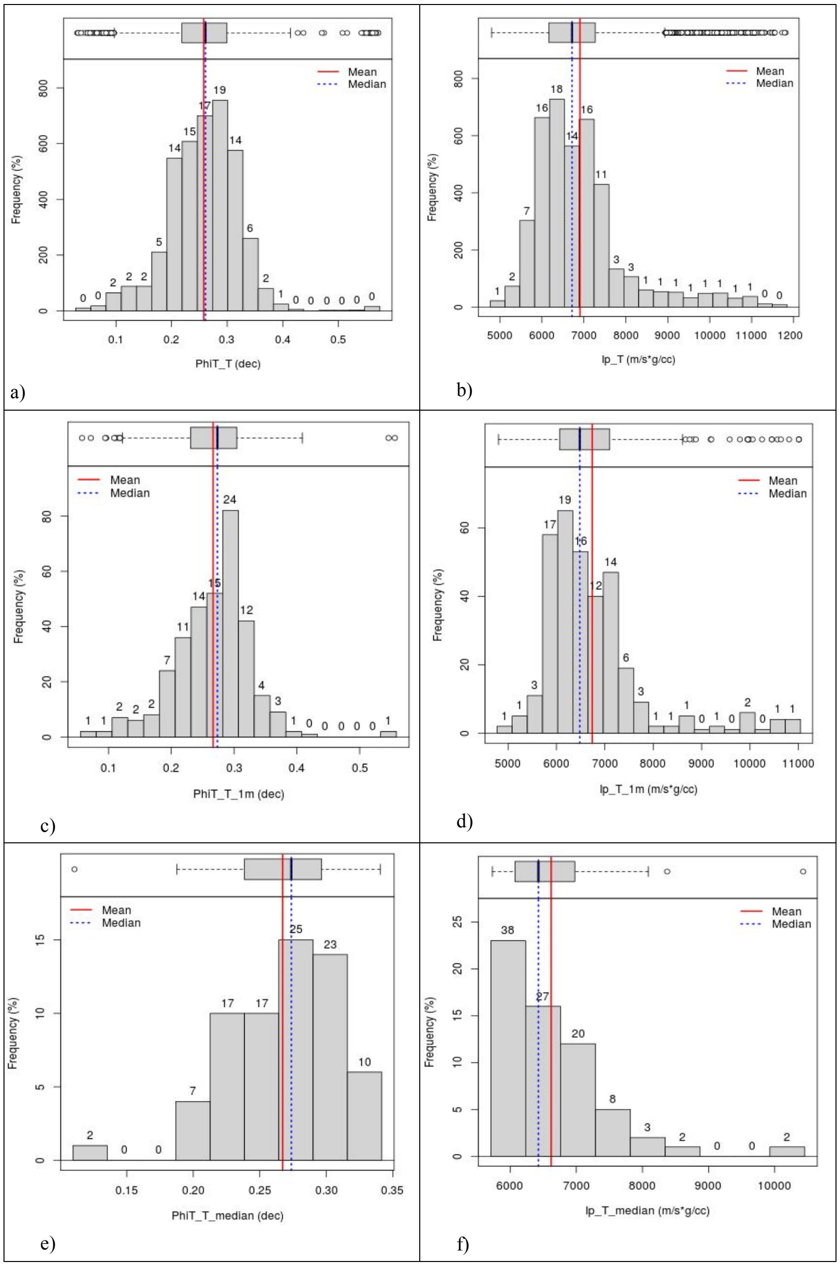

Figure 2. Histograms and boxplots for PhiT and Ip at well-log scale (a, b), at one-meter scale (c, d) and at seismic scale (e, f), respectively.

Note that median and mean are pretty close for one-meter scale and seismic scales, while Ip_inline_U and Ip_inline have very consistent statistics.

Bivariate dependence analysis

In Figure 3 are given the scatterplots with marginal histograms and boxplots for PhiT vs Ip a) at well-log, b) at one-meter and c) at seismic scale, respectively, while

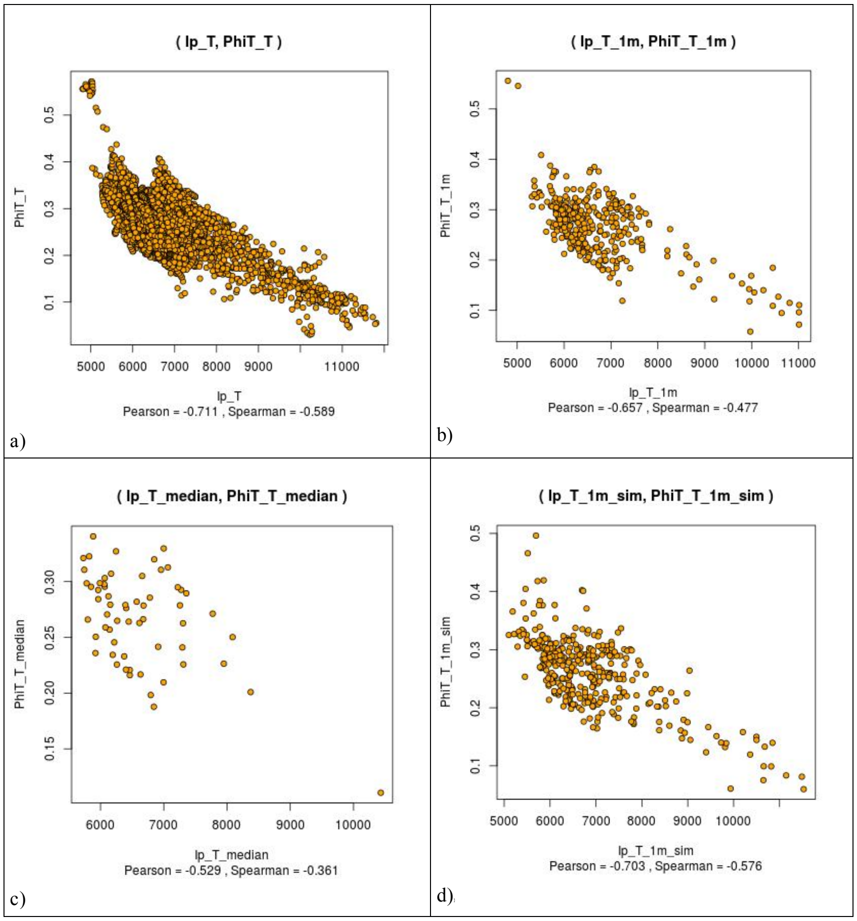

Figure 3. Scatterplots with marginal histograms and boxplots for Ip vs. PhiT, a) at well-log scale, b) at one-meter scale, c) at seismic scale, d) a non-conditional bivariate simulation with a Bernstein copula at one-meter scale.

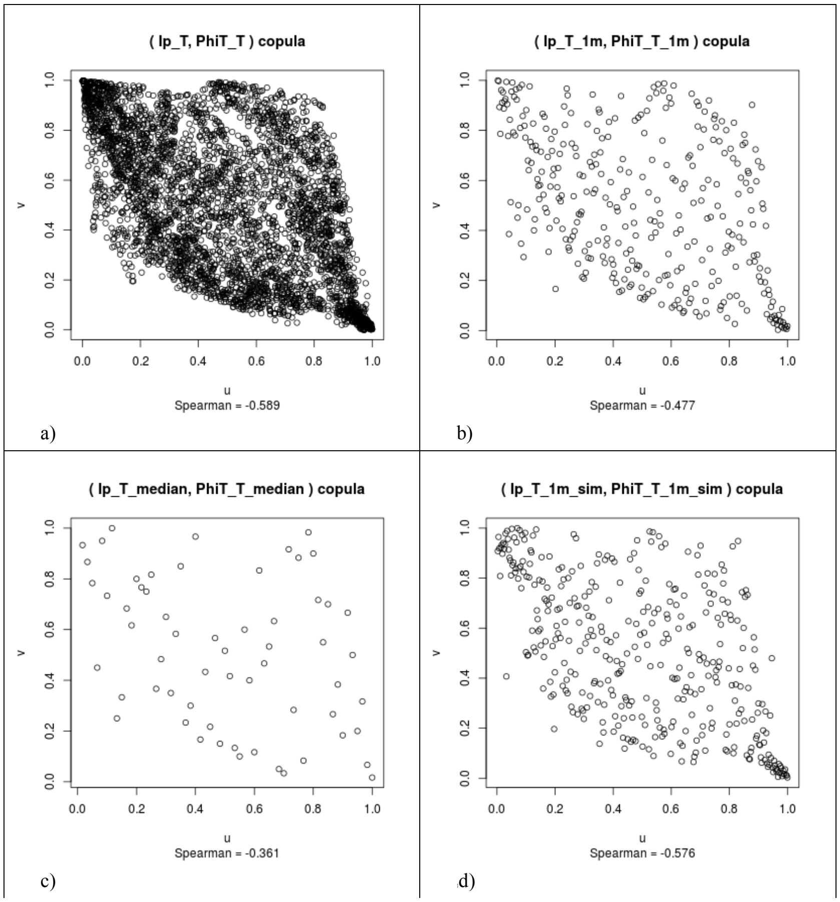

Figure 4. Pseudo-observations (sample copulas) scatterplots of Ip-PhiT data ranks, rescaled to [0,1] a) at well-log scale, b) at one-meter scale, c) at seismic scale, and d) a non-conditional bivariate simulation with a Bernstein copula at one-meter scale.

Table 3 Summary of correlation coefficients for Ip vs PhiT at well-log, one-meter and seismic scales, and for a non-conditional bivariate Ip-PhiT simulation with a Bernstein copula at one-meter scale (BCS_1m).

Correlation Coefficients	well-log scale	one-meter scale	Seismic scale	BCS_1m
Spearman	-0.589	-0.477	-0.361	-0.576
Pearson	-0.711	-0.657	-0.529	-0.703

in the Figure 3 d) the corresponding scatterplot for a non-conditional Ip-PhiT simulation using a Bernstein copula at one-meter scale. In Figure 4 are given pseudo-observations (sample copula) scatterplots for Ip-PhiT, a) at well-log scale, b) at one-meter scale, c) at seismic scale, and d) for a non-conditional bivariate simulation with a Bernstein copula at one-meter scale. In

Table 3 a summary of corresponding correlation (Spearman and Pearson) coefficients is given. It can be observed that the dependence is weakened with the increasing of the scale.

Variography analysis

In the Figure 5 are shown estimated variograms (a, b) and best fit variogram models (c, d) for PhiT and Ip at seismic scale in depth direction. As is evident in Figure 5 b) the sample variogram of Ip_inline_U shows a typical behavior related with the presence of trend, which means that at least the intrinsic hypothesis is not satisfied. Then, the trend, which in this case was of second order, was estimated and removed, resulting a new variable Ip_inline_U_r2 without trend. The same previous procedure was applied to Ip_Inline and a resulting detrended variable was named Ip_inline_r2. Note, in Figure 5 d) the variogram was obtained after removing trend from Ip_inline_U. While in the Figure 6 are displayed estimated variograms and best fit variogram models for impedance at seismic scale in a) X and b) depth directions, respectively, after removing trend from Ip_inline.

Because of lack of data for total porosity in the X direction the same variogram structure of the impedance in this direction is adopted, considering that they show almost the same structure in the depth direction (see Figure 5). A variogram model for porosity at seismic scale is proposed so that the total variance of PhiT_T median is preserved. Which basically it is to consider a correlation range equal to the impedance variogram in the X direction (see Figure 6). For both simulation methods the following variogram model for porosity is used: model=spherical, nugget= 0.0002, structure contribution=0.0016, ranges: maximum=160, medium=50, minimum=1, angles: x=90, y=0, z=0.

SGCS simulations

A sequential Gaussian co-simulation (SGCS) with Markov Model variant (MM1), implemented in SGeMS (Remy, Boucher, & Wu, 2009) was performed with the following parameters: primary variable: PhiT_T_median, secondary variable: Ip_Inline_r2, grid: 33x60x1 (the same as Ip_Inline_r2), Kriging type: Simple Kriging (SK), max conditioning data: 12, correlation coefficient= -0.657, search ellipsoid: 160, 50, 1, variogram model of primary variable=spherical, nugget=

0.0002, structure contribution=0.0016, ranges: maximum=160, medium=50, minimum=1, angles: x=90, y=0, z=0.
The resulting simulation is named PhiT_SGC and its map in the vertical (inline) section is given in the Figure 8 a).

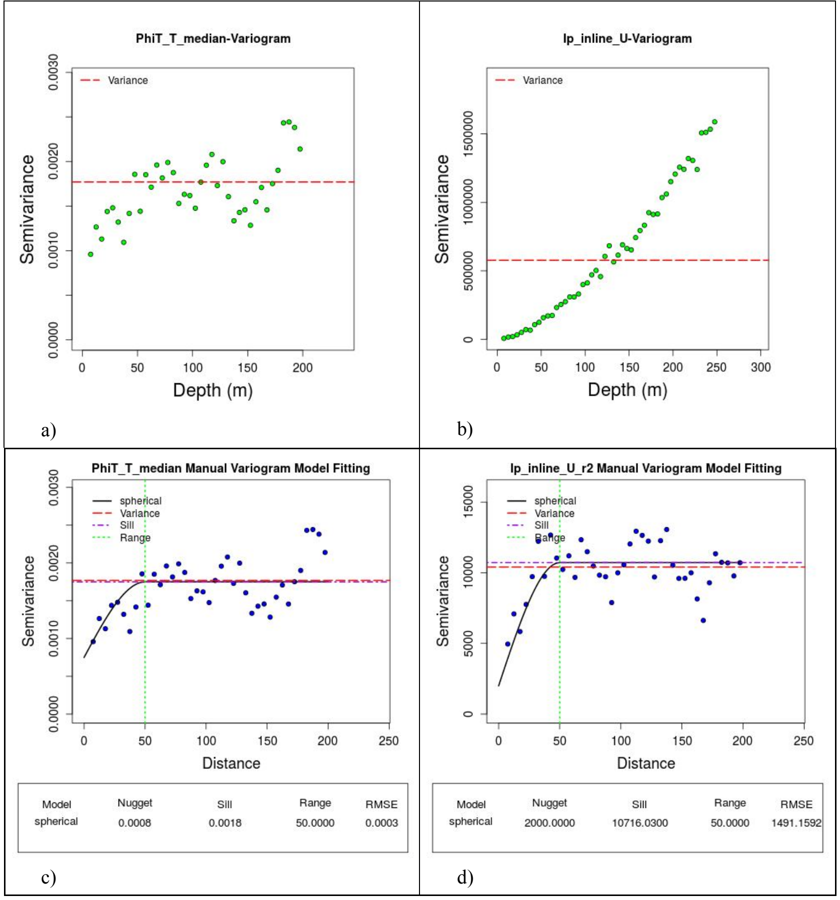

Figure 5. PhiT and Ip estimated variograms (a, b) and best fit variogram models (c, d) at seismic scale in depth direction. Note, in d) variogram after removing trend from Ip.

BCSCS simulations

A Bernstein copula-based spatial stochastic co-simulation (BCSCS) was performed using the procedure explained before. First, a dependence model, using a Bernstein copula at one-meter scale (see Figure 4 b) was obtained from which 40,000 conditional bivariate simulations (BCsim40000_cond) conditioning by secondary variable were generated (see Figure 7). Then, as the simulated annealing program was used a modified version of SASIM from GSLIB (Deutsch & Journel, GSLIB: Geostatistical software library and user's guide, 1998) with the following parameters: primary variable: PhiT_T_median, secondary variable: Ip_Inline, grid: 33x60x1 (the same as Ip_Inline), objective function: variogram and bivariate distribution function, paired data: 40,000 conditional bivariate simulations using a Bernstein copula (BCsim40000_cond), number of primary thresholds=10, number of secondary thresholds=10, number of variogram lags: 40, variogram model of primary variable= spherical, nugget= 0.0002, structure contribution=0.0016, ranges: maximum=160, medium=50, minimum=1, angles: x=90, y=0, z=0. A map of the resulting simulation is given in the Figure 8 b).

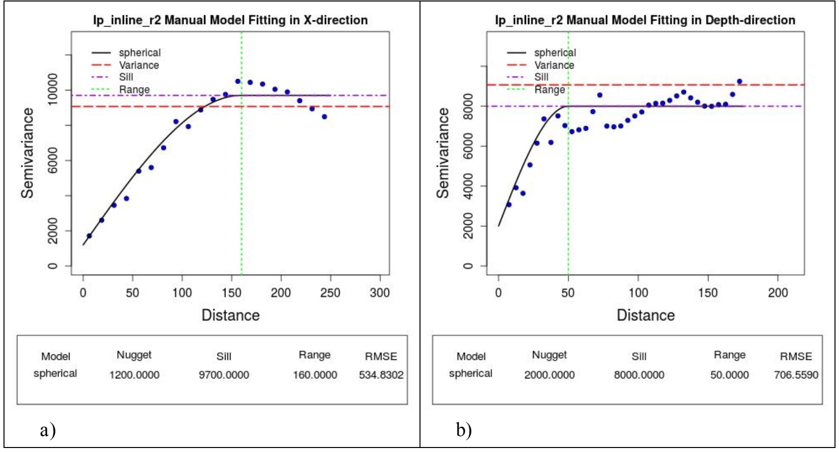

Figure 6. Estimated variograms and best fit variogram models for Ip at seismic scale after removing trend in a) X and b) depth directions, respectively.

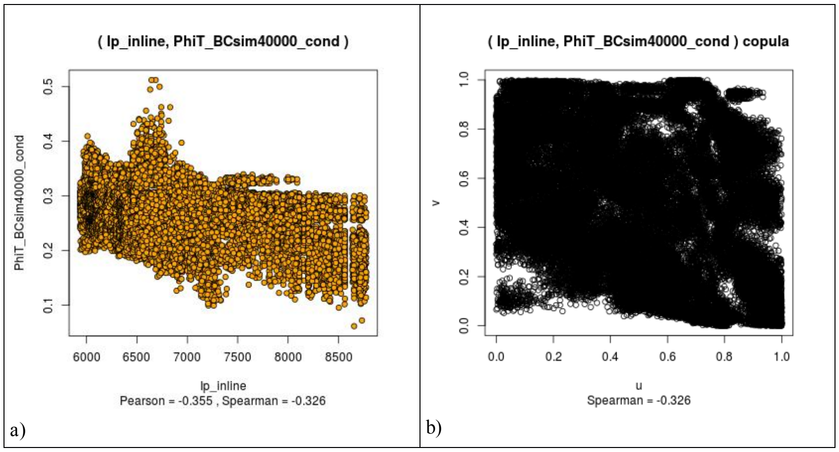

Figure 7. a) Scatterplot with marginal histograms and boxplots and .b) pseudo-observations (sample copulas) scatterplot for 40,000 conditional bivariate simulations using a Bernstein copula at one-meter scale.

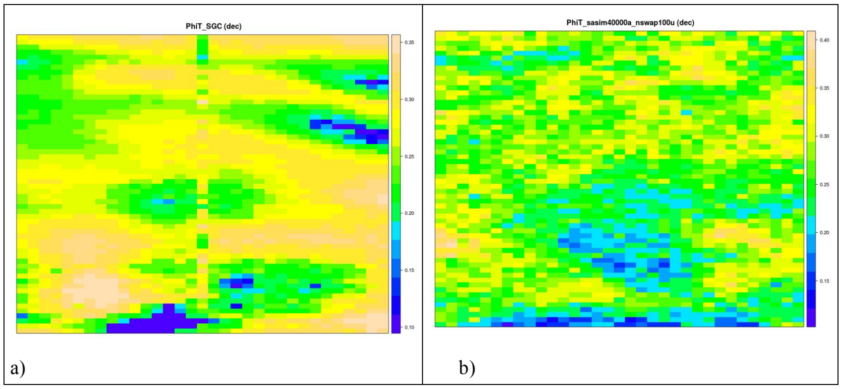

Figure 8. Maps for a) a PhiT sequential Gaussian co-simulation, and b) a PhiT Bernstein copula-based co-simulation in the vertical (inline) section.

SGCS vs. BCSCS, a comparative summary

In comparison with the SGCS method the BCSCS method better reproduce the statistics in terms of variance and extreme values (see Figure 9), and both methods reproduce quite well the spatial structure (see Figure 11), but the sequential Gaussian co-simulation shows spurious correlation dependence, which does not exist in the original data, highlighted in red color in Figure 10. This is the main reason of the difference between Figures 8a and 8b.

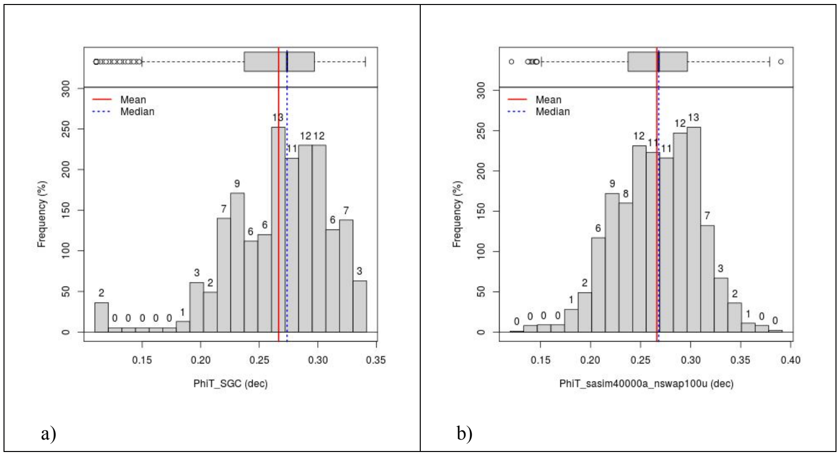

Figure 9. PhiT histograms and boxplots for a) a sequential Gaussian co-simulation and b) a Bernstein copula-based co-simulation, respectively.

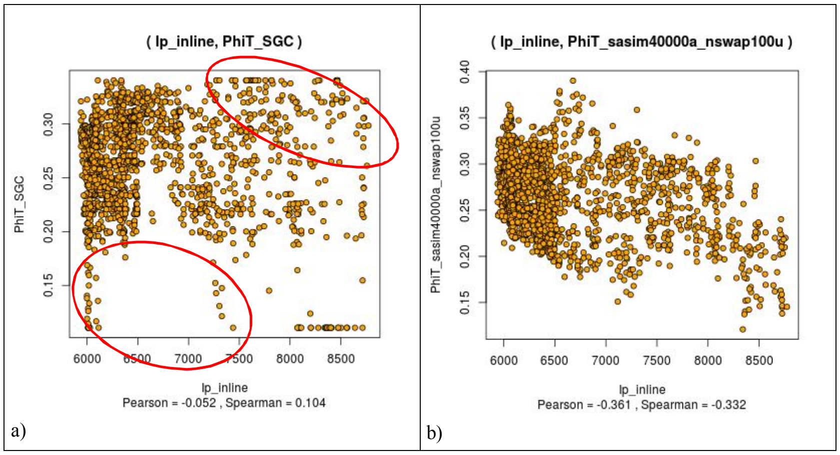

Figure 10. Ip vs. PhiT scatterplots with marginal histograms and boxplots for a) a sequential Gaussian co-simulation and b) a Bernstein copula-based co-simulation, respectively. Simulated values with spurious dependence, which does not exist in the original data, are highlighted in red color.

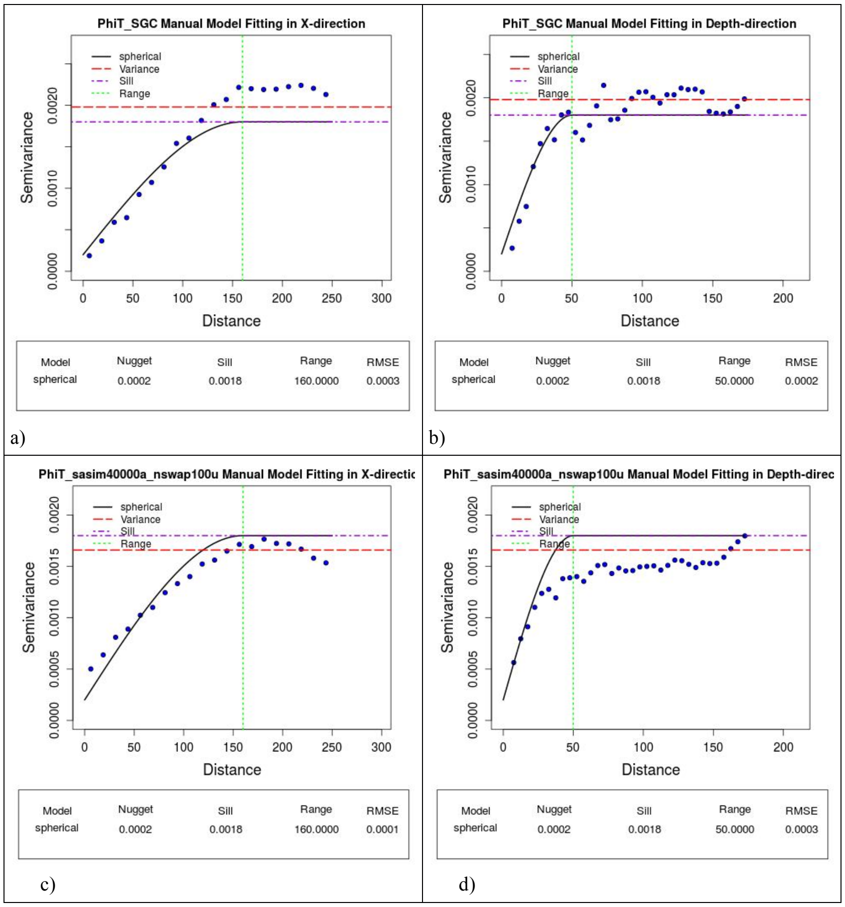

Figure 11. Estimated variograms and best fit variogram models in X and depth directions for a sequential Gaussian co-simulation (a, b) and a Bernstein copula-based co-simulation (c, d), respectively.

Final remarks and future work

A Bernstein copula-based spatial stochastic co-simulation (BCSCS) method presented in this paper possess several advantages over the classical sequential Gaussian co-simulation (SGCS). Firstly of all, it does not require of a strong linear dependence between variables, on the contrary it can capture and reproduce the existing dependence between them. The method is non-parametric which means that it does not need a specific distribution function. Since the Bernstein copula is based on the sample distribution function it may reproduce the variability and the extreme values. Another advantage is that there is no need to make back transformations, which are potentially biased, since copulas are invariant under strictly increasing transformations.

Instead of using a single seismic attribute, it could be used the best combination of them depending on the primary (explanatory) variable by applying standard multivariate statistical procedures such as principal component and factorial analysis. Another option would be using a multivariate copula with three or more variables to directly exploit their dependencies.

This work can be easily extended to 3D problems but it depends on the computing power available. Although in this work the aim was to show the performance of the simulation method, a simpler and efficient alternative, perhaps more convenient for 3D large problems, could be the median regression approach already shown in previous works (Erdely & Diaz-Viera, 2010) and (Hernández-Maldonado, Díaz-Viera, & Erdely, 2012) in 1D.

Acknowledgments

The present work was supported by the IMP project D.61037 “Interpretación Sísmica Cuantitativa Guiada por Litofacies para la Caracterización de Yacimientos”.

Bibiography

Bardossy, A., & Li, J. (2008). Geostatistical interpolation using copulas. *Water Resources Research, 44*, 1-15.

Chiles, J. P., & Delfiner, P. (1999). *Geostatistics: modeling spatial uncertainty.* Wiley series in probability and statistics. Applied probability and statistics section.

Deutsch, C. V., & Cockerham, P. W. (1994). Geostatistical Modeling of Permeability with Annealing Cosimulation (ACS). *SPE 69th Annual Technical Conference and Exhibition* (pp. 523-532). New Orleans, U.S.A.: Society of Petroleum Engineers.

Deutsch, C. V., & Journel, A. G. (1998). *GSLIB: Geostatistical software library and user’s guide* (Second Edition ed.). New York-Oxford: Oxford University Press.

Díaz-Viera, M., & Casar-González, R. (2005). Stochastic simulation of complex dependency patterns of petrophysical properties using t-copulas. *Proceedings of IAMG’05: GIS and Spatial Analysis*, *2*, pp. 749-755.

Erdely, A., & Diaz-Viera, M. (2010). Nonparametric and semiparametric bivariate modeling of petrophysical porosity-permeability dependence from well-log data. In P. Jaworski, F. Durante, W. Härdle, & T. Rychlik (Eds.), *Copula Theory and Its Applications. Lecture Notes in Statistics 198* (pp. 267-278). Berlin Heidelberg: Springer-Verlag.

Erdely, A., & Diaz-Viera, M. (2016). A vine and gluing copula model for permeability stochastic simulation. In J. R. Bozeman, T. Oliveira, & C. H. Skiadas (Eds.), *Stochastic and Data Analysis Methods and Applications in Statistics and Demography* (pp. 199-207).

Hernández-Maldonado, V., Díaz-Viera, M., & Erdely, A. (2012). A joint stochastic simulation method using the Bernstein copula as a flexible tool for modeling nonlinear dependence structures between petrophysical properties. *Journal of Petroleum Science and Engineering, 92-93*, 112-123.

Hernández-Maldonado, V., Díaz-Viera, M., & Erdely, A. (2014). A multivariate Bernstein copula model for permeability stochastic simulation. *Geofisica Internacional, 53*(2), 163-181.

Joe, H. (1997). *Multivariate models and dependence concepts.* London: Chapman & Hall.

Kazianka, H., & Pilz, J. (2010). Copula-based geostatistical modeling of continuous and discrete data including covariates. *Stochastic Environmental Research and Risk Assessment, 24*(5), 661-673.

Li, Q. (2001). LP Sparse Spike Impedance Inversion. Hampson-Russell Software Services Ltd. - CSEG.

Nelsen, R. B. (2006). *An Introduction to Copulas (Springer Series in Statistics)* (2nd Edition ed.). New York: Springer.

Parra, J., & Emery, X. (2013). Geostatistics applied to cross-well reflection seismic for imaging carbonate aquifers. *Journal of Applied Geophysics, 92*, 68-75.

Remy, N., Boucher, A., & Wu, J. (2009). *Applied Geostatistics with SGeMS: a user's guide.* New York: Cambridge University Press.

Sancetta, A., & Satchell, S. (2004). The Bernstein copula and its applications to modeling and approximations of multivariate distributions. *Econometric Theory, 20*(3), 535-562.

Sklar, A. (1959). Fonctions the repartition à n dimensions et leurs marges. *Publ. Inst. Statist, 8*, 229–331.

Appendix A: Copula-based approach for dependence modeling

A theorem by (Sklar, 1959) proved that there exists a functional relationship between the joint probability distribution function of a random vector and its univariate marginal distribution functions. In the bivariate case, for example, if (X, Y) is a random vector with joint probability distribution $F_{XY}(x, y) = P(X \leq x, Y \leq y)$

then the marginal distribution functions of X and Y are $F_X(x) = P(X \leq x) = F_{XY}(x, \infty)$ and $F_Y(y) = P(Y \leq y) = F_{XY}(\infty, y)$, respectively, but in the marginalization of F_{XY} some information is lost since with the only knowledge of the marginal distributions F_X and F_Y is not generally possible to specify F_{XY} because the marginals only explain the probabilistic individual behavior of the random variables they represent. *Sklar's Theorem* proves that there exists a function $C_{XY}: [0,1]^2 \rightarrow [0,1]$ such that

$$F_{XY}(x, y) = C_{XY}(F_X(x), F_Y(y))$$

C_{XY} is called *copula function* associated to (X, Y) and contains information about the dependence relationship between X and Y, independently from their marginal probabilistic behavior. C_{XY} is uniquely determined on $Ran\ F_X \times Ran\ F_Y$, and therefore if F_X and F_Y are continuous then C_{XY} is unique on $[0,1]^2$. Among several properties of copula functions, see (Nelsen, 2006), we have the following:

- $C(u, 0) = 0 = C(0, v)$
- $C(u, 1) = u, \quad C(1, v) = v$
- $C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0 \quad \text{if } u_1 \leq u_2, v_1 \leq v_2$
- C is uniformly continuous on its domain $[0,1]^2$.
- The horizontal, vertical, and diagonal sections of a copula C are all nondecreasing and uniformly continuous on $[0,1]$.
- $W(u, v) \leq C(u, v) \leq M(u, v)$ where $W(u, v) = \max(u + v - 1, 0)$ and $M(u, v) = \min(u, v)$ are also copulas known as the lower and upper Fréchet-Hoeffding bounds.
- A convex linear combination of copula functions is also a copula function.
- If X and Y are continuous random variables with copula C_{XY}, and if α and β are strictly increasing functions on $Ran\ X$ and $Ran\ Y$, respectively, then $C_{\alpha(X)\beta(Y)} = C_{XY}$. Thus C_{XY} is invariant under strictly increasing transformations of X and Y.

Copula functions are a useful tool to build joint probability models in a more flexible way since we may choose separately the univariate models for the random variables of interest and the copula function that better represents the dependence among them, in each case in a parametric or non-parametric way. In the case of a multivariate normal model, for example, all the marginal distributions have to be normally distributed, with no tail dependence at all, and with finite second moments for the correlations to be well defined. In fact, the multivariate normal model is a particular case when the underlying copula is Gaussian and all the univariate marginals are normally distributed.

In case F_X and F_Y are continuous, by elementary probability we know that $U = F_X(X)$ and $V = F_Y(Y)$ are continuous Uniform$(0,1)$ random variables, and the underlying copula C for the random vector (U, V) is the same copula corresponding to

(X,Y), and by Sklar's Theorem we have that the joint probability distribution function for (U,V) is equal to $F_{UV}(u,v) = C\big(F_U(u), F_V(v)\big) = C(u,v)$. Therefore, in case F_X and F_Y are known and F_{XY} is unknown, if $\{(x_1,y_1),\ldots,(x_n,y_n)\}$ is an observed random sample from (X,Y), the set $\{(u_k,v_k) = \big(F_X(x_k),F_Y(y_k)\big) : k = 1,\ldots,n\}$ would be an observed random sample from (U,V) with the same underlying copula C as (X,Y), and since $C = F_{UV}$ we may use the (u_k,v_k) values (called copula observations) to estimate C as a joint empirical distribution:

$$\hat{C}(u,v) = \frac{1}{n}\sum_{k=1}^{n} 1_{\{u_k \le u\,,\, v_k \le v\}}$$

Strictly speaking, the estimation $\hat{C}$ is not a copula since it is discontinuous and copulas are always continuous. If F_X, F_Y, and F_{XY} are all unknown (the usual case) we estimate F_X and F_Y by univariate empirical distribution functions:

$$\widehat{F_X}(x) = \frac{1}{n}\sum_{k=1}^{n} 1_{\{x_k \le x\}} \qquad \widehat{F_Y}(y) = \frac{1}{n}\sum_{k=1}^{n} 1_{\{y_k \le y\}}$$

Now the set of pairs $\{(u_k,v_k) = \left(\widehat{F_X}(x_k),\widehat{F_Y}(y_k)\right) : k = 1,\ldots,n\}$ is referred to as *copula pseudo-observations*. It is straightforward to verify that $\widehat{F_X}(x_k) = \frac{1}{n}\,rank(x_k)$ and $\widehat{F_Y}(y_k) = \frac{1}{n}\,rank(y_k)$. In this case the concept of *empirical copula*, see Nelsen (2006), is defined as the following function $C_n\colon I_n^2 \to [0,1]$, where $I_n = \{\frac{i}{n} : i = 0,\ldots,n\}$, given by:

$$C_n\left(\frac{i}{n},\frac{j}{n}\right) = \frac{1}{n}\sum_{k=1}^{n} 1_{\{rank(x_k)\le i\,,\;\, rank(y_k)\le j\}}$$

Again, C_n is not a copula but it is an estimation of the underlying copula C on the grid I_n^2 that may be extended to a copula on $[0,1]^2$ by means of, for example, Bernstein polynomials, as proposed and studied in (Sancetta & Satchell, 2004), which leads to what is known as a *Bernstein copula* non-parametric estimation $\tilde{C}\colon [0,1]^2 \to [0,1]$ given by:

$$\tilde{C}(u,v) = \sum_{i=0}^{n}\sum_{j=0}^{n} C_n\left(\frac{i}{n},\frac{j}{n}\right)\binom{n}{i}u^i(1-u)^{n-i}\binom{n}{j}v^j(1-v)^{n-j}$$

As summarized in (Erdely & Diaz-Viera, 2010) in order to simulate replications from the random vector (X,Y) with the dependence structure inferred from the observed data $\{(x_1,y_1),\ldots,(x_n,y_n)\}$ we have the following:

Algorithm 1

1. Generate two independent and continuous Uniform(0,1) random variates u and t.
2. Set $v = c_u^{-1}(t)$ where $c_u(v) = \frac{\partial\tilde{C}(u,v)}{\partial u}$.

3. The desired pair is $(x, y) = \left(\widetilde{Q_n}(u), \widetilde{R_n}(v)\right)$ where $\widetilde{Q_n}$ and $\widetilde{R_n}$ are empirical quantile functions for X and Y, respectively.

For a value x in the range of the random variable X and a given $0 < \alpha < 1$ let $y = \varphi_\alpha(x)$ denote the solution to the equation $P(Y \leq y|X = x) = \alpha$. Then the graph of $y = \varphi_\alpha(x)$ is the *α-quantile regression curve* of Y conditional on $X = x$. In (Nelsen, 2006) is proven that:

$$P(Y \leq y|X = x) = c_u(v)|_{u=F_X(x),v=F_Y(y)}$$

This result leads to the following algorithm to obtain the *α-quantile regression curve* of Y conditional on $X = x$:

Algorithm 2

1. Set $c_u(v) = \alpha$.
2. Solve for v the regression curve, say $v = g_\alpha(u)$.
3. Replace u by $\widetilde{Q_n^{-1}}(x)$ and v by $\widetilde{R_n^{-1}}(y)$.
4. Solve for y the regression curve, say $y = \varphi_\alpha(x)$.