# Modeling the Volume Dependent Distribution of Categorical Variables

Zhou Lan and Clayton V. Deutsch

Centre for Computational Geostatistics (CCG)
Department of Civil and Environmental Engineering
University of Alberta

*A categorical variable describes the objects in a population which can be divided into a group of categories. The distribution of lithofacies, an example of categorical variable, is one of the most important factors in reservoir modeling. Many critical petrophysical properties in reservoir analysis and estimation, such as porosity and permeability, are highly correlated with facies type. A correct and precise modeling and estimation on the spatial distribution pattern of various facies categories over the area of interest will be significantly helpful for our modeling and estimation on the reservoir characteristics and performance. In the practice of facies modeling and analysis, we need to deal with data at different support, either point data or scaled up block data. The scaled up facies proportion will take different values and have different distribution along with the changes in block scale. In this paper, we will start from a facies category training image in CCG training image library to analyze the distribution at different scales. The mean, variance and variography of the scaled up facies proportion, as well as the shapes of the distribution, will be observed and analyzed at different scale. Analytical fitting of the marginal and joint distribution of multivariate indicator variables will be derived and its application multiscale indicator mapping will be discussed.*

## Introduction

A categorical variable describes the objects in a population which can be divided into a group of categories. One important example of categorical variable in mining, petroleum industry and geological study is the lithofacies. Suppose in a three-dimensional space $\Omega$, there exist $K$ facies categories $S_1, S_2, ..., S_K$. Each single point $\mathbf{u}_\alpha$ in the space is corresponding to an exact facies category, that is, a set of indicator variables $I(\mathbf{u}_\alpha, k)$ ($k = 1, 2, ..., K$), such that $I(\mathbf{u}_\alpha, k) = 1$ when the facies category at $\mathbf{u}_\alpha$ is $S_k$ and $I(\mathbf{u}_\alpha, k) = 0$ otherwise. Scaling up the facies categories over a neighborhood $v_\alpha$ of location $\mathbf{u}_\alpha$, the proportion of category $S_k$ is obtained by:

$$p_k = p_v(\mathbf{u}, k) = \frac{1}{v}\int_v I(\mathbf{u}, k)dv, \; k = 1, 2, ..., K.$$

The values and distributions of proportions $p_v(\mathbf{u}, k)$ are volume-dependent, that is, at different scales of the neighborhood $v_\alpha$, we will get different values and distribution. Figure 1 gives a brief illustration on the volume-dependent distribution of a categorical variable.

In this paper, we will take lithofacies as an example, start from a facies category training image in CCG training image library and discover how distribution of a categorical variable changes with the volume. In part I of this paper, we give a brief review on the previous research in this area. In part II, we give a brief description about the data. In part III, we observe and describe the changes in distributions of facies proportion in three aspects: i) shape and continuity, ii) mean and variance, and iii) variography. In part IV, we apply a series of parametric distribution to fit the marginal as well as the multivariate facies proportion model and compare the results with the training image data. In part V, an application of the analytical fitting in multiscale facies indicator mapping will be discussed and Part VI will give the related sample case study.

**Scaling Laws**

For many years, researches have been carried out by many scientists and geostatisticians about volume dependent distribution on facies categories and proportions, particularly with effort to discover the scaling laws, governing the changes in mean, variance, covariance and variograms of facies categories and facies proportions based on a series of volumetric supports, some of the important works include: (1) Journel and Huijbregts (1978) develop a series of theoretical concepts and theorems which are widely applied and analyzing scaling laws of categorical variables in geostatistical study. (2) Isaaks and Srivastava (1989) gave further discussion the scaling laws based a practical case study. (3) Deutsch and Frykman (1999) gave a full discussion on semivariogram modeling at different volumetric support as well as sequential simulation based on multiscale data. (4) Deutsch, Tran and Xie (2001) also gave a detail discussions in direct sequential simulation based with multiscale well, seismic and production data. The following concepts and theories are of particular importance in understanding the volume dependent distribution of categorical variables:

Given two different volumetric supports $v$ and $V$, three important concepts: dispersion variance $\sigma^2(v,V)$, average variogram $\bar{\gamma}(v,V)$ and mean covariance $\bar{C}(v,V)$ are defined as (Journel and Huijbregts, 1978):

$$\sigma^2(v,V) = E[m_v - m_V]^2$$

$$\bar{\gamma}(v,V) = \frac{1}{Vv}\int_{V(\mathbf{u})}\int_{v(\mathbf{u}')}\gamma(y-y')dy'dy$$

$$\bar{C}(v,V) = \frac{1}{Vv}\int_{V(\mathbf{u})}\int_{v(\mathbf{u}')}C(y-y')dy'dy$$

where $m_v, m_V$ are means at the support of scale $v$ and $V$ respectively. The average variogram and mean covariance are in fact the mean values of, respectively, the point variograms $\gamma(\mathbf{h})$ and covariance $C(\mathbf{h})$, where one extremity of the distance vector $\mathbf{h}$ describes the domain of $V(\mathbf{u})$ and the other extremity independently describes the domain $v(\mathbf{u}')$. The follow relationship is easy to reach (Journel and Huijbregts, 1978):

$$\sigma^2(v,\Omega) = \sigma^2(V,\Omega) + \sigma^2(v,V) \quad (\ v \subset V \subset \Omega\ )$$

$$\sigma^2(v,V) = \bar{C}(v,v) - \bar{C}(V,V) = \bar{\gamma}(V,V) - \bar{\gamma}(v,v)$$

$\sigma^2(\bullet,\bullet) = 0$ and $\overline{\gamma}(\Omega,\Omega) = \sigma^2(\bullet,\Omega)$. The symbol "$\bullet$" here is used to denote the point data. All these results provide the essential ideals in understanding changes in variances along with the volumes based a certain covariance and variogram structure.

For scaled up variable $Z_v$, volumetric support variogram $\gamma_v(\mathbf{h})$ is defined as:

$$\gamma_v(\mathbf{h}) = E\left\{[Z_v(\mathbf{u}) - Z_v(\mathbf{u}+\mathbf{h})]^2\right\} \text{ where } Z_v(\mathbf{u}) = \frac{1}{v}\int_{v(\mathbf{u})} Z(y)dy$$

Journel and Huijbregts (1978) showed that scaled up semivariogram could be expressed as:

$$2\gamma_v(\mathbf{h}) = 2\overline{\gamma}[v(\mathbf{u}), v(\mathbf{u}+\mathbf{h})] - \overline{\gamma}[v(\mathbf{u}), v(\mathbf{u})] - \overline{\gamma}[v(\mathbf{u}+\mathbf{h}), v(\mathbf{u}+\mathbf{h})]$$

Based on stationarity assumption, we have

$$\overline{\gamma}[v(\mathbf{u}), v(\mathbf{u})] = \overline{\gamma}[v(\mathbf{u}+\mathbf{h}), v(\mathbf{u}+\mathbf{h})] = \overline{\gamma}(v, v)$$

and thus

$$\gamma_v(\mathbf{h}) = \overline{\gamma}[v(\mathbf{u}), v(\mathbf{u}+\mathbf{h})] - \overline{\gamma}(v, v)$$

For a large distance $\mathbf{h}$ compared with the size of block $v$, the value of $\overline{\gamma}[v(\mathbf{u}), v(\mathbf{u}+\mathbf{h})]$ will be much closed to the point variogram $\gamma(\mathbf{h})$. In this case, $\gamma_v(\mathbf{h}) \simeq \gamma(\mathbf{h}) - \overline{\gamma}(v, v)$. The changes in variograms can therefore be predicted along with the changes in volumetric support.

As discussed by C.V. Deutsch and P.Frykman (1999), The fitted variogram model at arbitrary scale $v$ is defined as:

$$\gamma_v(\mathbf{h}) = C_v^0 + \sum_{i=1}^{nst} C_v^i \Gamma^i(\mathbf{h})$$

where $\Gamma^i(\mathbf{h})$ represents $i^{th}$ nested structure and $nst$ the total number of nested structures. Also $C_v^0$ denotes the nugget effect and $C_v^i$ the variance contribution the $i^{th}$ nested structure. The sum of variance contribution equals the dispersion variance, that is:

$$\sigma^2(v, \Omega) = C_v^0 + \sum_{i=1}^{nst} C_v^i$$

with $\Omega$ the volume of entire space of interest. The range of the volumetric supported variogram at a larger volume $V$ increases as the increase in volume size ($|V| - |v|$) in each particular direction, that is:

$$a_V = a_v + (|V| - |v|)$$

Depending on the shape of large volume $V$, the range may increase in some directions and stay the same in other directions. The purely random component, the nugget effect, decreases with an inverse relationship of the volume, that is:

$$C_V^0 = C_v^0 \cdot \frac{|v|}{|V|}$$

The changes (decreases) variance contribution of each nested structure along with supporting volumes are determined by the average variogram $\bar{\Gamma}$ calculated from the nested structure $\Gamma^i$, that is:

$$C_V^i = C_v^i \cdot \frac{1-\bar{\Gamma}(V,V,\mathbf{a}^i)}{1-\bar{\Gamma}(v,v,\mathbf{a}^i)}$$

**An Example**

The data for this paper is a facies category training image over a three-dimensional space denoted as 256×256×128 in terms of x×y×z coordinates, or equivalently denoted as East×North×Depth. We have six facies categories, $S_0, S_1, S_2, S_3, S_4$ and $S_5$ of interest. Figure 2 gives the slice maps along the planes x=50, y=50 and z=50. Figure 3 shows a group of 3-dimensional pictures about spatial distribution of the facies categories over the area. Figure 4 gives the maps of the vertical facies proportions over the horizontal area. In this training image, $S_0, S_1$ and $S_2$ are the three most important categories and their proportions sum to over 90% in most of the area. The proportion of category $S_3$ is 0 and categories $S_4$ and $S_5$ take only very small part.

In order to obtain a clear picture about the distribution of facies proportion over different scales, we divided the entire space into blocks of equal scale using a series of sizes: 2×2×2, 4×4×4, 8×8×8, 16×16×16, 32×32×32 and 64×64×64. The proportion of each facies category was calculated for each block.

Taking as an example the facies category $S_0$, Figure 5 gives the histogram of proportion $p_0$ over different scales. At scale of $v = l_1 \times l_2 \times l_3$, the facies proportions take some values among $\left\{0, \frac{1}{v}, \frac{2}{v}, ..., 1\right\}$, At a small scale, the facies proportions take a group of discrete values. As the scale increases, more and more continuous distributions occur. Also, at a small scale, an obvious bimodal distribution of facies proportion $p_0$ is observed. As the scale increases, there exists a trend of converging to a symmetric unimodal distribution. In Figure 6, we have the histogram of $p_1$ and a similar trend is shown.

The cumulative distribution (CDF) curves give more clear pictures about changes in distribution along with scales. Figure 7 shows that at scales 2×2×2 and 4×4×4, step-shape CDF curves are obtained. As the scale increases, the curves become more and more continuous.

Now take a further look at the cdf of proportion $p_k$ for facies categories $S_k$: Let $p$ be its prior global proportion. The value and distribution of $p_k$ depend on the scale of supporting volume $v$. For $v = 0$, $p_k = I(\mathbf{u}, k)$, which is either 1 or 0, and $P[p_k = 0] = 1 - p$, $P[p_k = 1] = p$. The cumulative distribution function (CDF):

$$F_k(x) = P[p_k \le x] = \begin{cases} 1-p & \forall 0 \le x < 1 \\ p & \text{for } x \ge 1 \end{cases}$$

For $v = \infty$, we have $p_k = p$, and the CDF:

$$F_k(x) = P[p_k \le x] = \begin{cases} 0 & \forall x < p \\ 1 & \forall x \ge p \end{cases}$$

Figure 8 gives an illustration on shapes of CDF's at each of these extreme cases. As the scale increases between 0 and $\infty$, the CDF curves shift between the above extreme cases, Figure 9 shows the changes.

Taking as an example the facies category $S_0$, the mean and standard deviation (S.d.) of proportion $p_0$ are tabulated below:

Table 1 Mean and standard deviation of $p_0$ at different scales

| Scale | 2×2×2 | 4×4×4 | 8×8×8 | 16×16×16 | 32×32×32 | 64×64×64 |
|---|---|---|---|---|---|---|
| Mean | 0.6648 | 0.6648 | 0.6648 | 0.6648 | 0.6648 | 0.6648 |
| S.d. | 0.4514 | 0.4232 | 0.3898 | 0.3232 | 0.2419 | 0.1372 |

The mean remains unchanged over different scales and the standard deviation decreases as the scale increases. Given a certain scale $v$, where $n$ individual points are located in each block. The proportion $p_v(\mathbf{u}_\alpha, k)$ of facies category $S_k$ can be defined as:

$$p_v(\mathbf{u}_\alpha, k) = \frac{1}{n}\sum_{i=1}^{n} I(\mathbf{u}_\alpha^i, k) \quad k = 1, 2, ..., K$$

Suppose the entire space of interest is divided into $m$ blocks of the same size, then mean of the proportion $p_k$ can be obtained by :

$$E[p_v(k)] = \frac{1}{m}\sum_{\alpha=1}^{m}\frac{1}{n}\sum_{i=1}^{n} I(\mathbf{u}_\alpha^i, k) = \sum_{\alpha=1}^{m}\sum_{i=1}^{n}\frac{1}{m}\frac{1}{n} I(\mathbf{u}_\alpha^i, k)$$

$$= \sum_{i=1}^{N}\frac{1}{N} I(\mathbf{u}^i, k) = \frac{1}{N}\sum_{i=1}^{N} I(\mathbf{u}^i, k) = \mu_k$$

The mean of facies proportion over the entire area of interest is independent of the size of the scale and is equal to the global mean $\mu_k$. The variance depends on the scale (as described above):

$$\sigma^2(V, \Omega) = \sigma^2(v, \Omega) - \sigma^2(v, V) \text{ where } v \subset V \subset \Omega.$$

Furthermore, based on the average variogram $\bar{\gamma}(v, V)$ and covariance $\bar{C}(v, V)$ we have:

$$\sigma^2(v, V) = \bar{C}(v, v) - \bar{C}(V, V) = \bar{\gamma}(V, V) - \bar{\gamma}(v, v)$$

or equivalently, the variance of the scaled up variable:

$$\sigma^2(v, \Omega) = \bar{\gamma}(\Omega, \Omega) - \bar{\gamma}(v, v) = \sigma^2(\bullet, \Omega) - \bar{\gamma}(v, v)$$

Where $\sigma^2(\bullet, \Omega)$ is the variance of the point data over the entire volume ($\Omega$) of our interest.

Based on the block $v$ of a certain scale, the semivariogram of facies proportions is defined as:

$$\gamma_v(\mathbf{h},k) = E\left\{[p_v(\mathbf{u},k) - p_v(\mathbf{u}+\mathbf{h},k)]^2\right\}$$

Figures 10 and 11 give the semivariograms of facies proportion $p_0$ and $p_1$ in x-direction (in red), y-direction (in green) and z-direction (in blue) at various scales. As the scale increases, the sill of indicator semivariogram in each direction is decreasing and the plots are flattening, suggesting a trend of getting more continuous as the scale increase. The changes in ranges are not obvious in our sample case. The changes in semivariograms of indicators and scaled up facies proportions are governed by the rules described above.

**Univariate Multiscale Distributions**

An Analytical description about the distribution of facies proportion is desirable and it enables more precise modeling and prediction. In this part of the paper, several parametric probability distributions and their fittings are tested.

Consider the multinomial distribution. Suppose there exists a prior facies proportion for $\tilde{p}_k$ for category $S_k$. The variable $x_k = \sum_{i=1}^{n} I(\mathbf{u}_\alpha^i, k)$ , the number of points where $S_k$ occur within a certain block $v(\mathbf{u}_\alpha)$ satisfies the following condition: $\sum_{k=1}^{K} x_k = n$, where $n$ the entire number of gridding nodes in the block and $\sum_{k=1}^{K} \tilde{p}_k = 1$. This suggests a multinomial distribution of variables $X_k$, $k = 1, 2, ..., K$, that is

$$p[X_1 = x_1, ..., X_K = x_K] = \frac{n!}{x_1!x_2!...x_K!}\tilde{p}_1^{x_1}\tilde{p}_2^{x_2}\cdots\tilde{p}_K^{x_K}$$

One important assumption for binomial distribution is that: indicator variable $I(\mathbf{u}_\alpha, k)$ and $I(\mathbf{u}_\beta, k)$ are independent from each other for any different locations $\mathbf{u}_\alpha$ and $\mathbf{u}_\beta$. However, this is not true in many geological data. See an illustration of semivariogram in Figure 13, those points with distance less than range $a$ positively correlated and those with variogram values above sill are negative correlated. Only those with variogram value at sill might be independent. And the variograms at different direction may have different sills and ranges. The histograms from simulated multinomial realizations are shown in Figure 12. Here the parameter $n$ takes the volume of the scale, i.e., at scale of $2\times2\times2$ we assign $n = 8$. Obviously, they are different from our sample data.

Consider the beta distribution. The beta distribution is defined for a random variable $X$ within a close interval [0,1], which has the probability density functions (pdf):

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$$

where the gamma function is defined as $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$ and $\Gamma(z+1) = z\Gamma(z)$. The shapes of the CDF curves are determined by the parameters $\alpha$ and $\beta$. Based on the known expected values (global mean $\tilde{p}_k$) and variances $Var_v(p_k)$, the parameters $\alpha$ and $\beta$ are determined as:

$$\alpha = \tilde{p}_k[\frac{\tilde{p}_k(1-\tilde{p}_k)}{Var_v(p_k)} - 1] \text{ and } \beta = (1-\tilde{p}_k)[\frac{\tilde{p}_k(1-\tilde{p}_k)}{Var_v(p_k)} - 1].$$

Furthermore, let $\Omega$ be the entire space of interest, and $D_k^2()$ be the dispersion variances of proportion $p_k$ based on certain supports, we have:

$$\frac{\tilde{p}_k(1-\tilde{p}_k)}{Var_v(p_k)} - 1 = \frac{D_k^2(\bullet,\Omega)}{D_k^2(v,\Omega)} - 1 = \frac{D_k^2(\bullet,\Omega)}{D_k^2(\bullet,\Omega) - D_k^2(\bullet,v)} - 1 = \frac{D_k^2(\bullet,v)}{D_k^2(\bullet,\Omega) - D_k^2(\bullet,v)} = \frac{1}{\theta - 1}$$

Where $\theta = \frac{D_k^2(\bullet,\Omega)}{D_k^2(\bullet,v)} = \frac{\tilde{p}_k(1-\tilde{p}_k)}{D_k^2(\bullet,v)}$. That is: $\alpha = \frac{\tilde{p}_k}{\theta - 1}$ and $\beta = \frac{1-\tilde{p}_k}{\theta - 1}$. The expected value and variances are calculated as follows:

$$E_v[p_k] = \frac{\alpha}{\alpha+\beta}, \quad Var_v[p_k] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Figure 14 gives comparison between the sample distributions (left) and simulated beta distributions (right) for $p_0$ at various scales while Figure 15 gives the overlapped curves of sample CDF and beta simulated CDF for $p_0$ and $p_1$. Some observations:

- The beta simulated realizations give good reproductions of the marginal sample distributions at most of the scales and the facies categories.
- The Frequencies of extreme proportion values (around 0 and 1) are partially over-estimated, making the simulated CDF curves over-smooth at both ends.
- When $Var_v(p_k) \to 0$, both $\alpha, \beta \to \infty$, the distribution goes to normal. Figure 16 gives the simulated realizations for $p_0$ at $\alpha$ and $\beta$ values based on some very variance. We see that it is a normal distribution with mean $\tilde{p}_0$ =0.6648
- Some other training images were tested and similar results were obtained for categories with prior global proportions values not very close to 0 or 1. For those categories with extreme prior proportions, e.g., greater than 0.99 or less than 0.01, the beta simulated realization did not reproduced the sample distributions. Figure 17, gives some examples.

If we are to proceed with multiscale facies modeling we require a multivariate distribution and not simply a univariate distribution.

**Dirichlet Distribution for Multivariate Multiscale Distributions**

Generally, for $K$ facies categories $S_0, S_1, \ldots, S_{K-1}$ in the area of interest, the proportions ($p_0, p_1, \ldots, p_{K-1}$) will fall on a hyperplane determined by $p_0 + p_1 + \ldots + p_{K-1} = 1$.

In the previous discussion, we realize that a beta distribution will be a workable parametric distribution in fitting the marginal distribution for each facies proportion. A generalized form of beta distribution, the Dirichlet Distribution, is therefore considered to fit the joint distribution of $p_0, p_1, \ldots, p_{K-1}$

A Dirichlet distribution (Johnson and Kotz, 2000), is defined for $n$ random variables $x_1, x_2, \ldots, x_n$ that have the joint probability density function (joint-pdf) given as:

$$f(\mathbf{x};\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{i=1}^{n}\alpha_i)}{\prod_{i=1}^{n}\Gamma(\alpha_i)}\prod_{i=1}^{n}x_i^{\alpha_i-1}$$

where $x_1, x_2, \ldots, x_n \in [0,1]$ and $\sum_{i=1}^{n} x_i = 1$. And $\alpha_1, \alpha_2, \ldots, \alpha_n$ are shape parameters. The expected value and variance of each variable are given as:

$$E[x_i] = \frac{\alpha_i}{\sum_{i=1}^{n}\alpha_i}, \text{ and } Var[x_i] = \frac{\alpha_i[\sum_{j=1}^{n}\alpha_j - \alpha_i]}{[\sum_{i=1}^{n}\alpha_i]^2[\sum_{i=1}^{n}\alpha_i + 1]}$$

It can be shown that the marginal distribution of $x_i$ follows a beta distribution with ($\alpha_i, \beta_i$).

Taking into consideration the constraint $\sum_{i=1}^{n} x_i = 1$, only $n-1$ variables are free and $x_n = 1 - \sum_{i=1}^{n-1} x_i$ . The the joint-pdf for Dirichlet distribution can then be expressed as:

$$f(x_1, x_2, \ldots, x_{n-1};\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{i=1}^{n}\alpha_i)}{\prod_{i=1}^{n}\Gamma(\alpha_i)}[\prod_{i=1}^{n-1}x_i^{\alpha_i-1}]\cdot(1-\sum_{i=1}^{n-1}x_i)^{\alpha_n-1}$$

Now come back to facies proportion $p_0, p_1, \ldots, p_{K-1}$, the joint pdf can then be fitted as:

$$f(p_0, p_1, \ldots, p_{K-1};\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{k=0}^{K-1}\alpha_k)}{\prod_{k=0}^{K-1}\Gamma(\alpha_k)}\prod_{k=0}^{K-1}p_k^{\alpha_k-1}$$

Or, taking into consideration the constraint $\sum_{k=0}^{K-1} p_k = 1$, only $K-1$ variables are free and $p_{K-1} = 1 - \sum_{k=0}^{K-2} p_k$ . The the joint-pdf for Dirichlet distribution can then be expressed as:

$$f(p_0, p_1, ..., p_{K-2}; \boldsymbol{\alpha}) = \frac{\Gamma(\sum_{k=0}^{K-1} \alpha_k)}{\prod_{k=0}^{K-1} \Gamma(\alpha_k)} [\prod_{k=0}^{K-2} p_k^{\alpha_k - 1}] \cdot [1 - \sum_{k=0}^{K-2} p_k]^{\alpha_{K-1}-1}$$

In case of only two facies categories, it is simplified to a standard beta distribution and parameters $\alpha$ and $\beta$ are uniquely determined by mean and variance of any one of the two facies proportions. In cases where more than two facies categories occur, the following conditions should all be satisfied:

$$E[x_i] = \frac{\alpha_i}{\sum_{i=1}^{n} \alpha_i}, \text{ and } Var[x_i] = \frac{\alpha_i [\sum_{j=1}^{n} \alpha_j - \alpha_i]}{[\sum_{i=1}^{n} \alpha_i]^2 [\sum_{i=1}^{n} \alpha_i + 1]}, \ i = 1, 2, ..., n$$

Here we have $n$ variables, $\alpha_1, \alpha_2, ..., \alpha_n$, to determined, satisfying $2n$ constraints. One possible way might be focusing on the expectation values and use only the variance of the most important category. From the mean constraints, we reach:

$$\alpha_i = E[x_i] \cdot \sum_{j=1}^{n} \alpha_j = E[x_i] \cdot \upsilon \text{ for all } i = 1, 2, ..., n$$

Here we denote the sum of $\alpha$'s as $\upsilon$. Substitute this into the constraint of $Var[x_d]$ where $x_d$ be the selected important category. Then we have:

$$\upsilon = \frac{E[x_d] \cdot (1 - E[x_d])}{Var[x_d]} \quad \text{and} \quad \alpha_i = E[x_i] \cdot \upsilon \quad \text{for all } i = 1, 2, ..., n.$$

A large number of realizations were simulated, using the variance for facies category $S_0$ when fitting the parameters. For facies proportion $p_0$, we got a simulated CDF very close to that as in beta simulation at each scale. Figure 18 gives some cross plots of joint CDF from the data versus the joint CDF from the Dirichlet simulated realizations for other facies categories and the following was observed in the simulation test:

- The shapes of the marginal distributions were approximately reproduced, particularly for the case of smaller scales.
- The mean of proportion of each facies category was reproduced.
- In the output distributions, the variances for facies proportion $p_0$ were reproduced at different scales of volumetric support, while the variances for other facies categories were over estimated to different extents. For $p_1$ and $p_2$, the estimated variances were close to the real levels, but for $p_4$ and $p_5$, the variances were strongly over-estimated and the shapes of distributions were changed particularly at a larger scale of support.
- In Figure 19, the joint cdf was approximately reproduced by Dirichlet distribution at small scale, but not at the larger scale.

**Ordinary Beta for Multivariate Multiscale Distributions**

One possible solution to problem of variances in Dirichlet distrbution lies in a generalized beta distribution introduced by Mauldon, 1959. Mauldon defined as follows an integral transformation ($\phi_\beta$) of $n$ random variables $x_1, x_2, ..., x_n$ with joint CDF $F(x_1, x_2, ..., x_n)$:

$$\phi_\beta = E\left[(t - \sum_{j=1}^{n} a_j x_j)^{-\beta}\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} (t - \sum_{j=1}^{n} a_j x_j)^{-\beta} dF(x_1, ..., x_n)$$

and defined $x_1, x_2, ..., x_n$ as forming a $n-$dimensional beta distribution when there exist parameters $c_{ij}$ and $\beta_i$ ($i = 1,2,...,r$) such that

$$\phi_\beta = \prod_{i=1}^{r} (t - \sum_{j=1}^{n} a_j c_{ij})^{-\beta_i} \text{ where } \beta = \sum_{i=1}^{r} \beta_i$$

The $c_{ij}$ parameters form a coordinate matrix. Mauldon showed that when the coordinate matrix is a unit matrix (with all $c_{ij}$ =1), and $x_1, x_2, ..., x_n$ fall within (0,1) and $x_1 + x_2 + ... + x_n = 1$, the joint pdf has the form:

$$f(x_1, ...x_n) = \frac{\Gamma(\beta)}{\prod \Gamma(\beta_i)} \prod_{j=1}^{n} x_j^{\beta_j - 1}$$

Mauldon called it basic beta distribution. Note it is in fact the Dirichlet distribution as we discussed above. Mauldon also showed that any $n-$dimensional beta distributions can be obtained by $\mathbf{y} = M\mathbf{x}$ from basic beta distributed variables $x_1, x_2, ..., x_n$ through matrix $M$. This is named as Ordinary Beta distribution by Mauldon.

The result from Mauldon is helpful in solving our problem. Let $p_1, p_2, ..., p_K$ be the $K$ facies proportion we are modeling, the joint distribution can be modeled by $\mathbf{p} = M\mathbf{x}$ where $\mathbf{x}$ forms a Dirichlet distribution with:

$$E[x_i] = \frac{\beta_i}{\beta} \text{ and } Var[x_i] = \frac{\beta_i(\beta - \beta_i)}{\beta^2(\beta + 1)}$$

and

$$E[\mathbf{p}] = M \cdot E[\mathbf{x}], \; COV[\mathbf{p}] = M \cdot COV[\mathbf{x}] \cdot M^T$$

where $COV[\mathbf{x}]$ denotes the covariance matrix of variable vector $\mathbf{x}$ and $M^T$ the transpose of matrix $M$. Solving the above equation systems we would be able find the matrix $M$ and parameters $\beta_i$ which make $p_1, ..., p_K$ honors population means, variances and covariances. Specifically, applying a diagonal matrix:

$$M = \begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{KK} \end{pmatrix}$$

And $\mathbf{x}=(x_1, x_2, ..., x_K)$ form a $K-$dimensional basic Beta (Dirichlet) distribution with parameters $\beta_i$ ($i=1,2,...,K$), we will reach the following system:

$$\begin{cases} E[p_i]=\dfrac{a_{ii}\beta_i}{\beta} \\ Var[p_i]=a_{ii}\cdot\dfrac{\beta_i(\beta-\beta_i)}{\beta^2(\beta+1)} & i=1,2,...,K \\ \beta=\beta_1+\beta_2+...+\beta_K \end{cases}$$

Or, equivalently:

$$\begin{cases} a_{ii}=\dfrac{\beta\cdot Var[p_i]+Var[p_i]+(E[p_i])^2}{E[p_i]} \\ \beta_i=\dfrac{\beta\cdot E[p_i]}{a_{ii}} & i=1,2,...,K \\ \beta=\beta_1+\beta_2+...+\beta_K \end{cases}$$

Solve this system for $a_{ii}$ and $\beta_i$ and thus reach distribution $\mathbf{p}=M\mathbf{x}$ which will honor both the means and the variances. One problem is that the values of $p_1, p_2, ..., p_K$ are not guaranteed to fall within [0,1] or sum to one. This problem can be solved by:

$$p_i^*=\frac{p_i}{\sum_{j=1}^{K}p_j} \quad i=1,2,...,K$$

Figure 20 gives cross plots of the data joint CDF versus the simulated joint CDF. Here we can see the joint CDF is pretty well reproduced at small scale and also reproduced at large scale. Several other training images were tested and the similar results were obtained. If a full matrix $M$ is adopted, the covariance can also be honored and a better fit can be expected. But that will require solving a very complicated equation system.

One major problem that might occurs in ordinary beta distribution fitting lies in the roots of $\beta$, $\beta_i$'s or $a_{ii}$. When dealing with system as discussed above, we will finally come to equations of $K^{th}$ order polynomials. When in any group of solution, values for $\beta$, $\beta_i$'s or $a_{ii}$ are not all positive, there will be risk when we draw the realization. It might also be possible that no real root exists and thus we can not go on. Fortunately, our observations suggest that all the non-real roots and most of the non-positive roots occur in those extreme situations where the expected values $E[p_k]$ for certain $k$'s in 1,2,…, $K$ are greater than 0.99 or less than 0.01. Note that for all $0\le p_k\le 1$, we have: $Var[p_k]=E[p_k^2]-(E[p_k])^2\le E[p_k]-(E[p_k])^2\le E[p_k]$. $10^9$ pairs of uniformly distributed random vectors ($\mathbf{u}$, $\mathbf{v}$), 5-dimentional or 4-dimentional, were drawn such that $0.01<\mathbf{u}(i)<0.99$ and $0.0005<\mathbf{v}(i)<\mathbf{u}(i)$, treated respectively as $E[\mathbf{p}]$ and $Var[\mathbf{p}]$ and did the test. Real roots occurred in all the cases and positive roots occurred in more than 97.5% of

the cases. In case all real roots were negative, we could slightly reduce the required variances and obtained positive roots. Based on this observation, when we build the conditional distribution, we can assign values 0.99 and 0.01 respectively, to $E[p_k]$ value when extreme value greater than 0.99 or less than 0.01 occurs. The result will be very close to the original one. And the problem of non-positive roots can be solved by slightly reducing the maximum of the target variances.

**Joint PDF for Ordinary Beta Distribution**

The parametric joint pdf for ordinary beta distribution can be derived applying Jacobian transformation rule. In the transformation $\mathbf{p} = M\mathbf{x}$, where $\mathbf{x}$ forms a Dirichlet distribution with joint pdf:

$$f_{\mathbf{x}}(x_1,...x_K) = \frac{\Gamma(\beta)}{\prod\Gamma(\beta_i)}\prod_{j=1}^{K} x_j^{\beta_j-1}$$

and $M$ an invertible matrix with inverse $M^{-1}=[b_{ij}]$, $i, j = 1,2,...,K$ we have

$$\mathbf{x} = M^{-1}\mathbf{p}$$

or equivalently $x_i(\mathbf{p}) = \sum_{j=1}^{K} b_{ij}p_j$ ($i, j = 1,2,...,K$). The Jacobian matrix for such transformation can be expressed as:

$$J = [s_{ij}], \text{ where } s_{ij} = \frac{\partial x_i(\mathbf{p})}{\partial p_j} = b_{ij} \quad (j = 1,2,...,K)$$

That is: $J = M^{-1}$. Denote the determinant of the Jacobian matrix $J$ as det($J$) and its absolute value as $|\det(J)|$, applying Jacobian transformation rule, we obtain joint pdf for $\mathbf{p}$ as:

$$f_{\mathbf{p}}(p_1,...p_K) = f_{\mathbf{x}}[x_1(\mathbf{p}),...x_K(\mathbf{p})]\cdot|\det(J)| = |\det(M^{-1})|\cdot\frac{\Gamma(\beta)}{\prod\Gamma(\beta_i)}\cdot\prod_{j=1}^{K}(\sum_{k=1}^{K} b_{jk}p_k)^{\beta_j-1}$$

In case we use diagonal matrix for $M$:

$$M = \begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{KK} \end{pmatrix}$$

where, as previously discussed, $a_{ii} > 0$ for all $i = 1,2,...,K$, the joint pdf $f_{\mathbf{p}}$ can be simplified as:

$$f_{\mathbf{p}}(p_1,...,p_K) = \frac{\Gamma(\beta)}{\prod_{i=1}^{K}[a_{ii}\cdot\Gamma(\beta_i)]}\cdot\prod_{j=1}^{K}\left(\frac{p_j}{a_{jj}}\right)^{\beta_j-1} = \frac{\Gamma(\beta)}{\prod_{i=1}^{K}[a_{ii}^{\beta_i}\cdot\Gamma(\beta_i)]}\cdot\prod_{j=1}^{K}\left(p_j\right)^{\beta_j-1}$$

and joint CDF is derived as:

$$F_{\mathbf{p}}(p_1,...,p_K)=\int_0^{p_1}...\int_0^{p_K} f_{\mathbf{p}}(t_1,...,t_K)dt_1...dt_K=\frac{\Gamma(\beta)}{\prod_{i=1}^{K}[a_{ii}^{\beta_i}\cdot\beta_i\cdot\Gamma(\beta_i)]}\cdot\prod_{j=1}^{K}\left(p_j\right)^{\beta_j}$$

**An Application**

Two important sources of data we usually deal with include well data and seismic data. The well data provide accurate measurements about the vertical distribution of lithofacies, as well as porosity, permeability and other critical properties in reservoir modeling. However, due to the high costs and various other reasons, the number of wells are usually too limit to provide enough information to support an accurate appraisal about a reservoir. In contrast, the seismic data are generally more abundant, though less precise. Usually the seismic data has less vertical resolution than the well data and approximately reflect lithofacies proportions and average porosity at a certain volumetric support. On the other hand, various other sources, such as historical production record, scientific research discovery, and so on, will also provide information at different scale of volumetric support about lithofacies and other geological properties. Integrating data of different scale has always been a critical issues.

Building 3-D realization of lithofacies codes, porosity at a sufficiently detailed resolution to provide a reliable basis for well planning, volumetric calculations and meaningful effective flow properties is a key problem in reservoir modeling (Deutsch, Srinivasan 1996). Two basic categories of algorithms are used in mapping the lithofacies, porosity and other variables: 1) Estimation algorithms, or, Interpolation algorithms, as named by A.G.Journel and W.Xu (1992), which yield a unique response, best in some sense. Kriging of different types form a critical family in this category. 2) Simulation, or stochastic imaging, [A.G.Journel and W.Xu (1992)], which provide multiple possible realizations of the variable of interest.

Many types of kriging algorithms can be used to map the lithofacies based on well data, seismic data and various other data of different scale support, such as kriging with external drift, kriging with local varying mean, block kriging and collocated cokriging. Simulation is closely related to the kriging algorithms. Usually, a certain type of kriging (or cokriging) approach is used to build the conditional distribution of a variable at a certain location based on the known data, and realization is then drawn from the conditional distribution.

Based on the discussion in part IV in this paper, we note that the multivariate distribution of facies proportions ( $p_1, p_2,..., p_K$ ) for $K$ facies categories ( $S_1, S_2,..., S_K$ ) can be estimated by an ordinary beta distribution determined by the mean and variance of each variable $p_k$ ( $k=1,2,...,K$ ). This result leads to another option of mapping the lithofacies categories and proportions based on the well data, seismic data and other data of different supporting scales. Suppose in a 3-dimensional space with volume $\Omega$, we have $n_1$ wells located at $\mathbf{u}_\alpha$, ($\alpha$ =1,2,…, $n_1$) and seismic data is available all over the area we are interested in and gives the block average at a support of $v_0$. Suppose as we want to build a 3-D realization of facies proportions $p_1, p_2,..., p_K$ for the $K$ facies categories ( $S_1, S_2,..., S_K$ ) over the area and at the volumetric support of $v$.

Applying the Ordinary Beta distribution, our problem is simplified to estimating the local mean and variance of facies proportion $p_k$ ($k = 1, 2, ..., K$) at each location ($\mathbf{u}$) based on the sample well data and seismic attributes. The following approaches can be used:

- ***Kriging with locally varying means***: Treat seismic attribute as the locally varying means. Perform simple kriging on the residual from the locally varying means to estimate $p_k(\mathbf{u})$ at each grid node $\mathbf{u}$ and treat it as means used to determine the conditional distribution.
- ***Collocated Cokriging:*** Treat the seismic data as a secondary variable and apply collocated cokriging algorithm to estimate $p_k(\mathbf{u})$ at each grid node $\mathbf{u}$ and treat it as means used to determine the conditional distribution.

Based on above estimated local means as well as the variances, local conditional joint distribution of facies proportions is built applying the ordinary beta distribution and realization is drawn from this distribution. Next, realization of facies categories is drawn from the conditional cdf. The simulated realization is then scaled up to reach the realization at the desired volumetric support.

In this example, a layer of depth 32 was extracted for the original training image. Figure 21 gives the facies maps of slice 50 at XZ, YZ directions and slice 16 at XY direction. Figure 22 gives the fitting of semivariogram models for the two horizontal directions: north-south and east-west (left), and the vertical direction (right). The dot curves show the experimental semivariogram and the dash curves give the variogram model. The red curves in the left figures represent the north-south direction and the blue curves represent the east-west direction. Suppose there is a well at each horizontal location determined by all pairs of coordinates (x, y) such that $\{x, y : x = 16i, y = 16j.\ i, j = 1, 2, ...16\}$, and at each well location, the vertical distribution of facies categories at all points separated by 4 unit were recorded. Thus we have a data of facies categories at each grid node of size $16 \times 16 \times 4$.

The original training image was scaled up at the support of $1 \times 1 \times 16$, that is at each of the $256 \times 256$ points in the horizontal plane facies proportions were obtained at the depth of every 16 units and the facies proportion was treated as the seismic derived facies proportion and assigned to each grid node with the same block. In this example, the facies proportions at scale $1 \times 1 \times 16$ was used as local varying mean and nonstationary simple kriging with local varying means was applied to obtained the local conditional means at each grid node of small size $1 \times 1 \times 1$. Ordinary beta distribution was then applied to build the conditional local distribution based on the conditional local means as well as the variance of each indicator variable. In practice, the `BlockSIS.for` program of CCG2005 was amended and ordinary beta distribution was inserted after kriging estimated local facies proportion was obtained. Indicator realization was then drawn at each of the small grid nodes (scale $1 \times 1 \times 1$) from its conditional local distribution. Here kriging option "2" was selected in `BlockSIS`. In the left columns of Figures 23 and 24, three realizations thus simulated are shown and compared with the results from the original `BlockSIS` Program (right columns). Here we can see these two results are very closed to each other, suggesting that cdf built directly from kriging estimated facies proportion (as is in `BlockSIS`) is honored by the ordinary beta distribution based on the means and variances.

## Conclusions

The distribution of scaled up proportions of categorical variables is complex. At a small scale, proportion of a facies category occurs as a set of discrete values. When the scale increases, a continuous distribution can be built. Means of the scaled up facies proportions are independent of the scales. The variance of facies proportion will decrease as the scale increases. Semivariograms will also decrease and flatten along with the increase in scale. The changes in variance and variograms can be predicted with the variogram and covariance models.

Both the marginal distribution of proportion for each facies category and the joint (multivariate) distribution of full set of facies categories depend on the means and variances of all facies categories. and therefore depend on the scale of the volumetric support. Beta distribution and Ordinary Beta distribution are workable in modeling the marginal and joint distribution, respectively, of the facies proportions based on the means and variance at different support.

## References


R.A.Behrens, M.K.Macleod, T.T.Tran, A.O.Alimi. *Incorporating Seismic Attribute Maps in 3D Reservoir Models*. SPE 36499, 1998

C. V. Deutsch. *Geostatistical Reservoir Modeling.* Oxford University Press, New York, 2002.

C. V. Deutsch. *A Sequential Indicator Simulation Program for Categorical Variables with Point and Block data: BlockSIS.* Centre for Computational Geostatistics (CCG) Annual Report 2005. p.402-1 – 402-22.

C. V. Deutsch and P.Frykman. *Geostatistical Scaling Laws Applied to Core and Log Data.* SPE 56822, 1999.

C. V. Deutsch and A. G. Journel. *GSLIB: Geostatistical Software Library and User's Guide.* Oxford University Press, New York, second Edition, 1998.

C. V. Deutsch and A. G. Journel. H.Kupfersberger. *Deriving Constraints on Small-Scale Variograms due to Variograms of Large-Scale Data* Mathematical Geology Vol 30, No.7, 1998 p.837-852.

C. V. Deutsch, T.T.Tran and Y.Xie. *Direct Geostatistical Simulation with Multiscale Well, Seismic and Production Data.* SPE 71323, 2001.

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

A. G. Journel and Ch. J. Huijbregts *Mining Geostatistics*. Academic Press, London, New York, San Francisco, 1978

S.Kotz, N.Balakrishnan, N.L.Johnson. *Continuous Multivariate Distributions.* John Willey & Sons, Inc. 2000.

J.G.Mauldon. *A Generalization of the Beta-Distribution.* The Annals of Mathematical Staitstics, Vol. 30, No. 2 1959. p.509-520.

W.Xu, T.T.Tran, R.M.Srivastava, A.G.Journel. *Integrating Seismic Data in Reservoir Modeling: The Collocated Cokriging Alternative*. SPE 24742, 1992

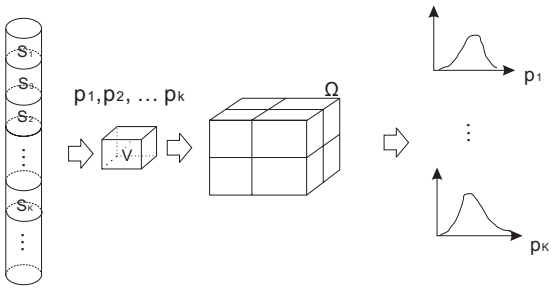


**Figure 1:** Multiscale Facies Modeling

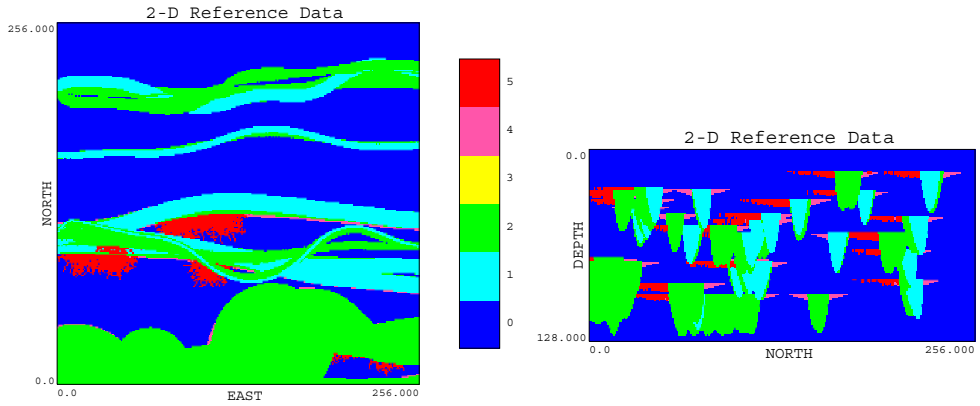


**Figure 2:** Slice Maps of Facies Categories

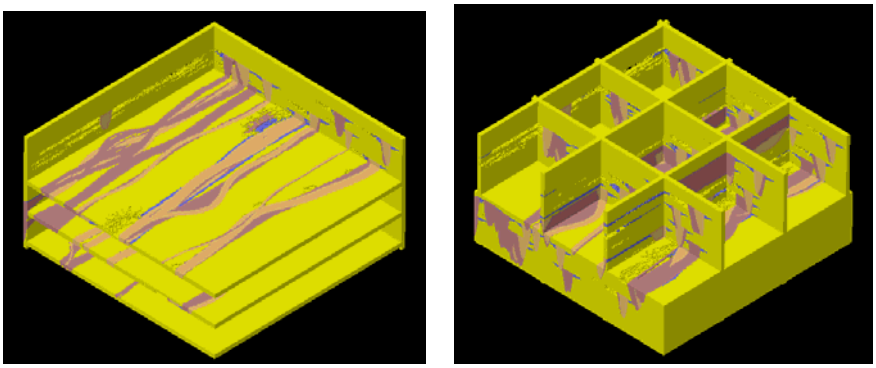

**Figure 3:** 3-Dimensional Pictures of Facies Categories

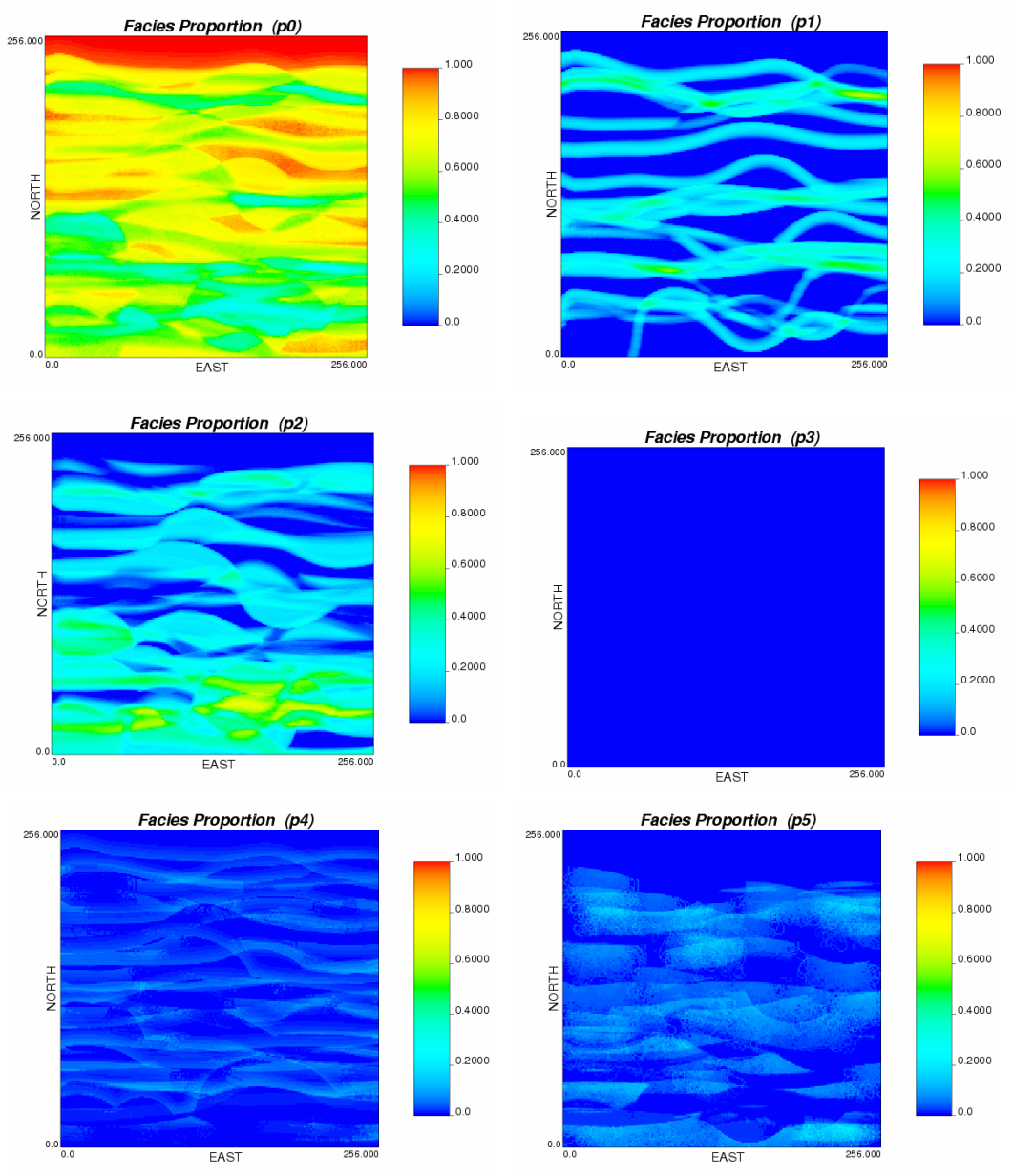


**Figure 4:** Facies Proportion Map

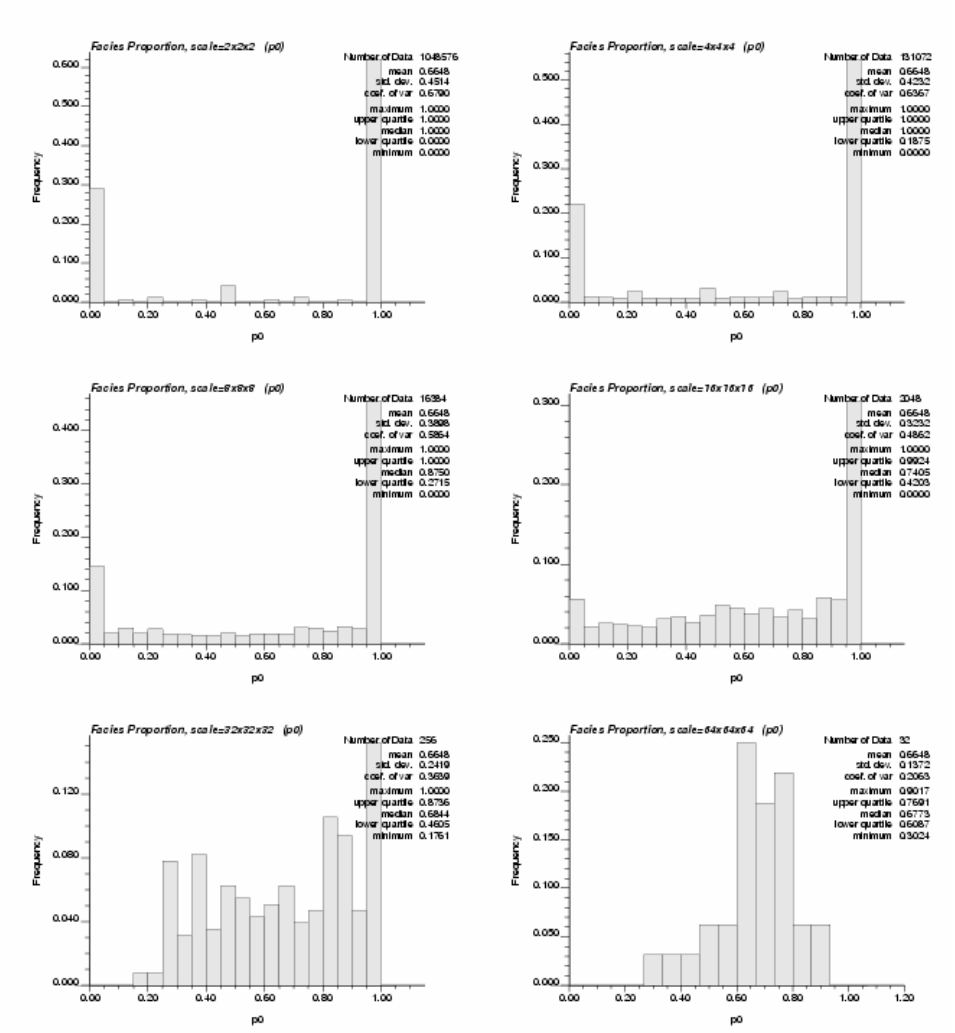


**Figure 5:** Facies Proportion of $S_0$ over Different Scales

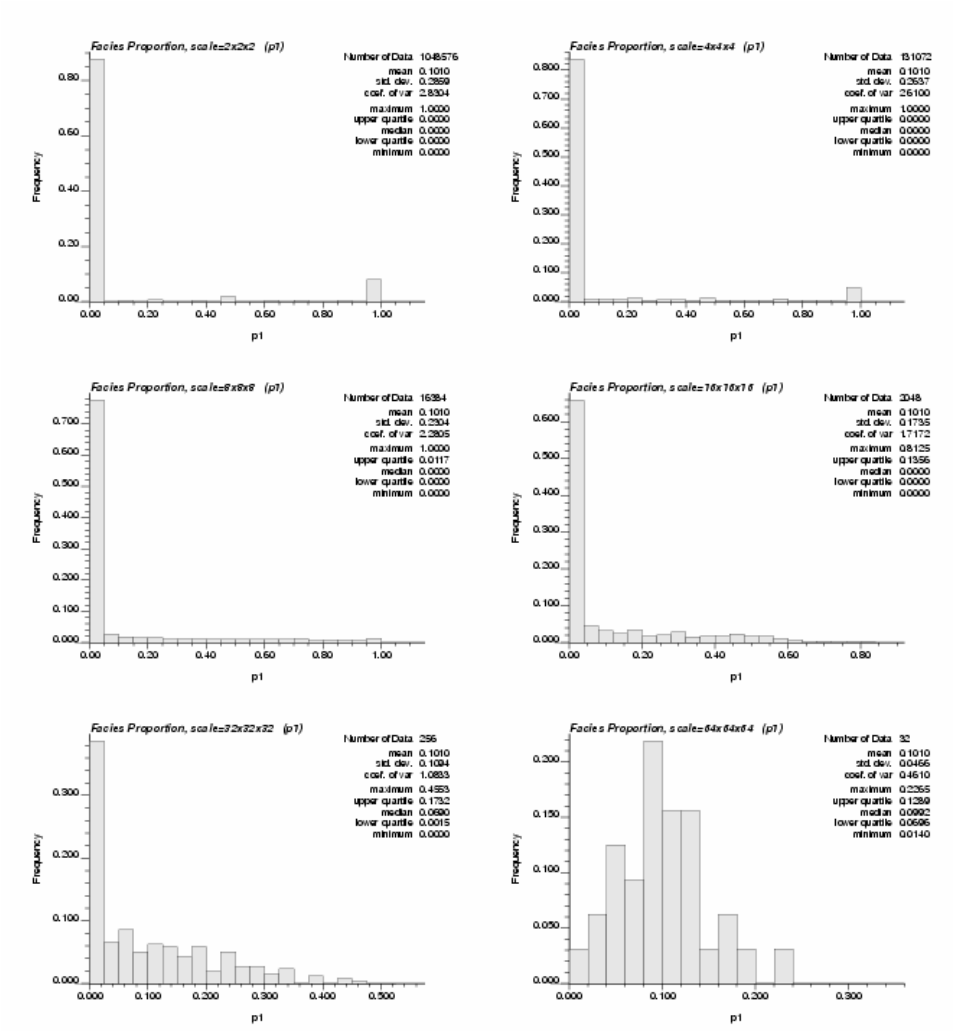


**Figure 6:** Facies Proportion of $S_1$ over Different Scales

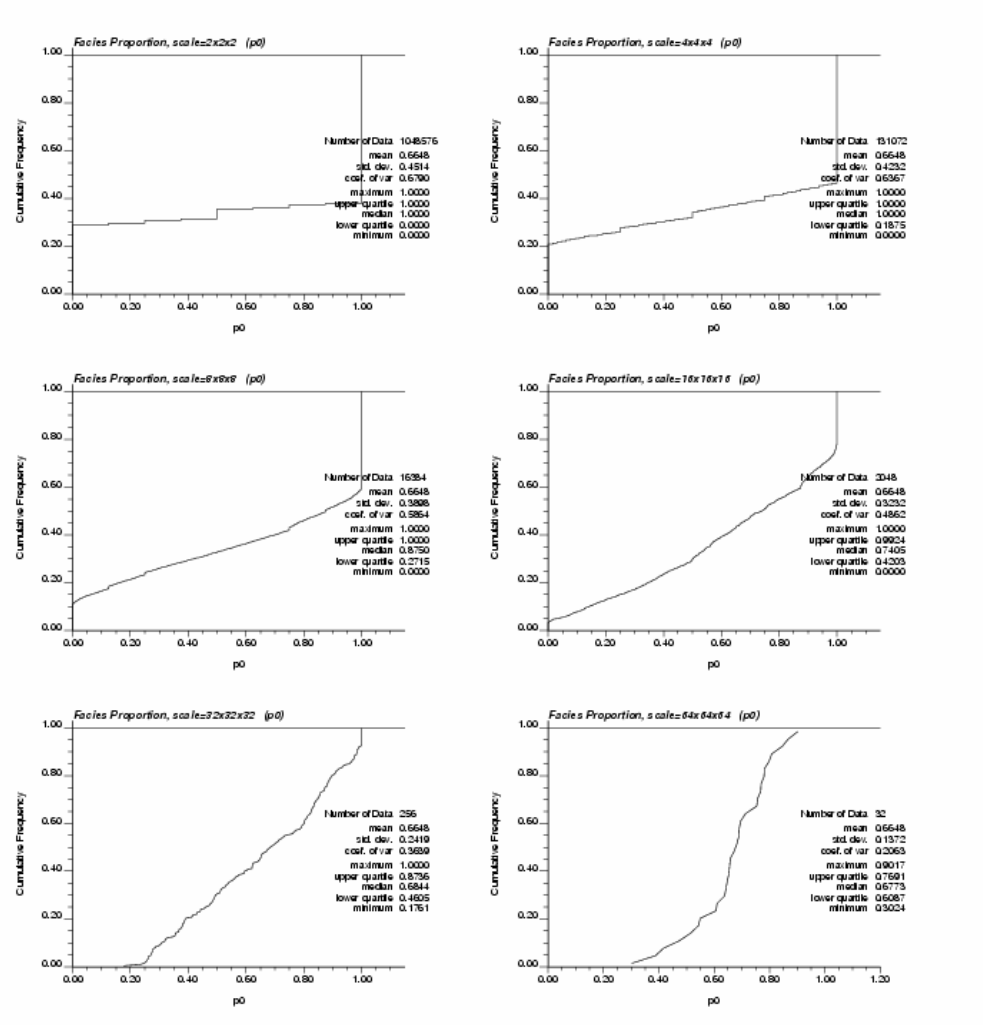


**Figure 7:** Cumulative Distribution of Proportion of $S_0$ over Different Scales

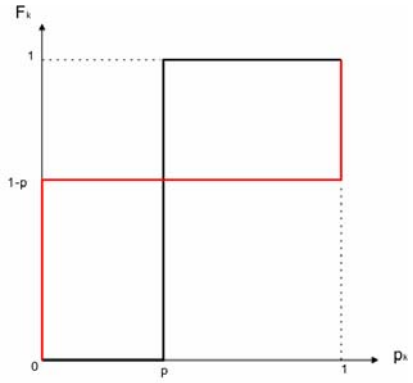


**Figure 8:** CDF of $p_k$ for cases $v = 0$ ( in red ) and $v = \infty$ ( in black )

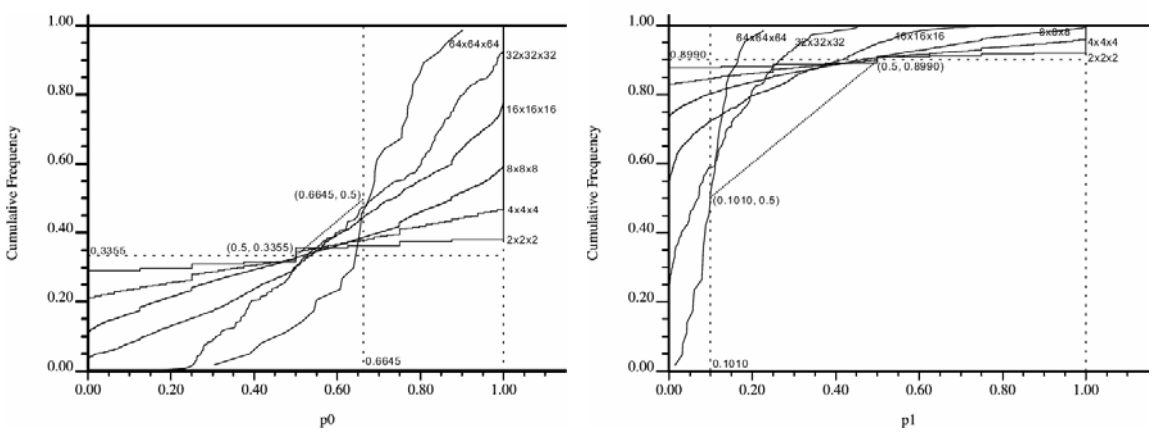


**Figure 9:**

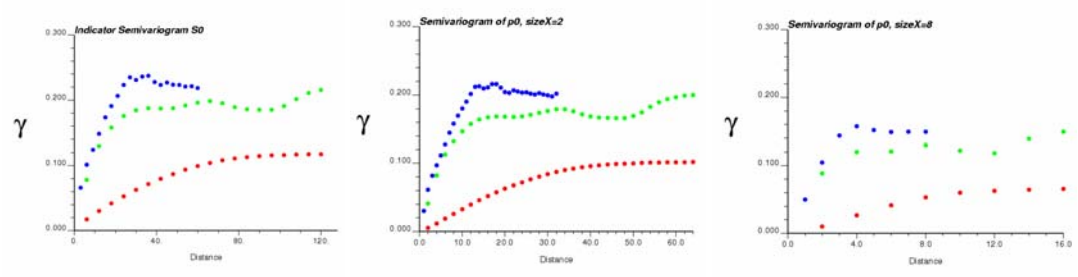


**Figure 10:** Semivariograms of Facies Proportion for $S_0$ at Various Scales. (X-direction: in red. Y-direction: in green. Z-direction: in blue. Unit lag distance: the distance between the centers of two adjacent blocks ).

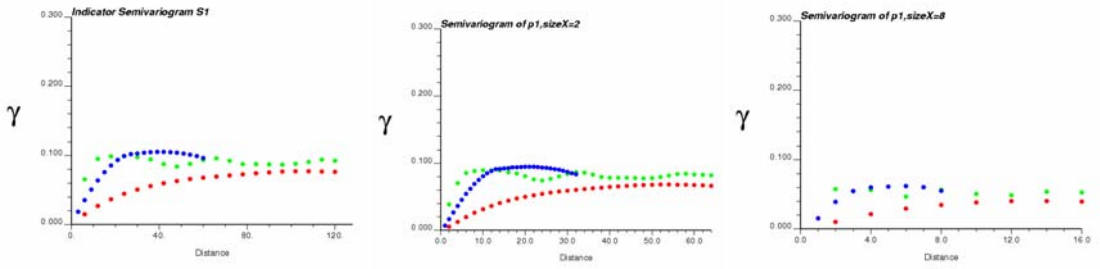


**Figure 11:** Semivariograms of Facies Proportion for $S_1$ at Various Scales. (X-direction: in red. Y-direction: in green. Z-direction: in blue. Unit lag distance: the distance between the centers of two adjacent blocks ).

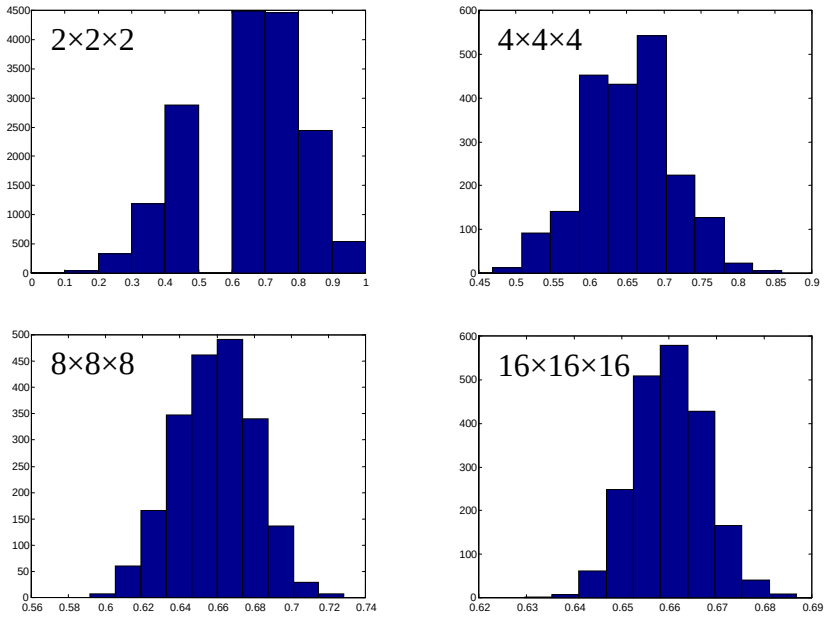


**Figure 12:** Binomial simulation ( $\tilde{p}_k$ =0.665 )

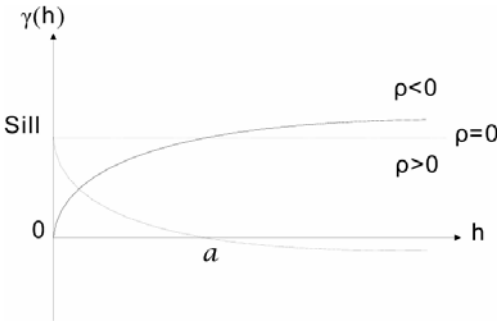


**Figure 13:** Semivariogram

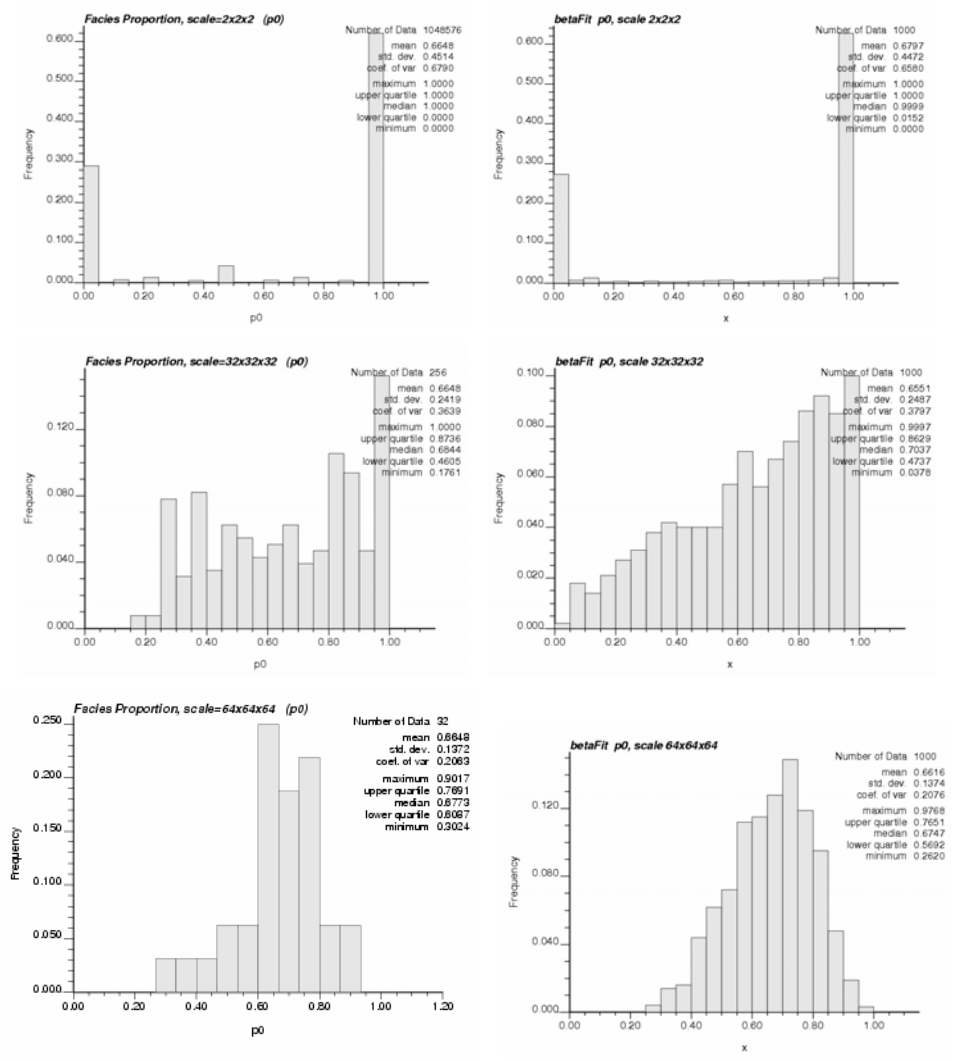


**Figure 14:** Beta simulated distributions. Histograms from data (left) compaired with the simulated histograms (right).

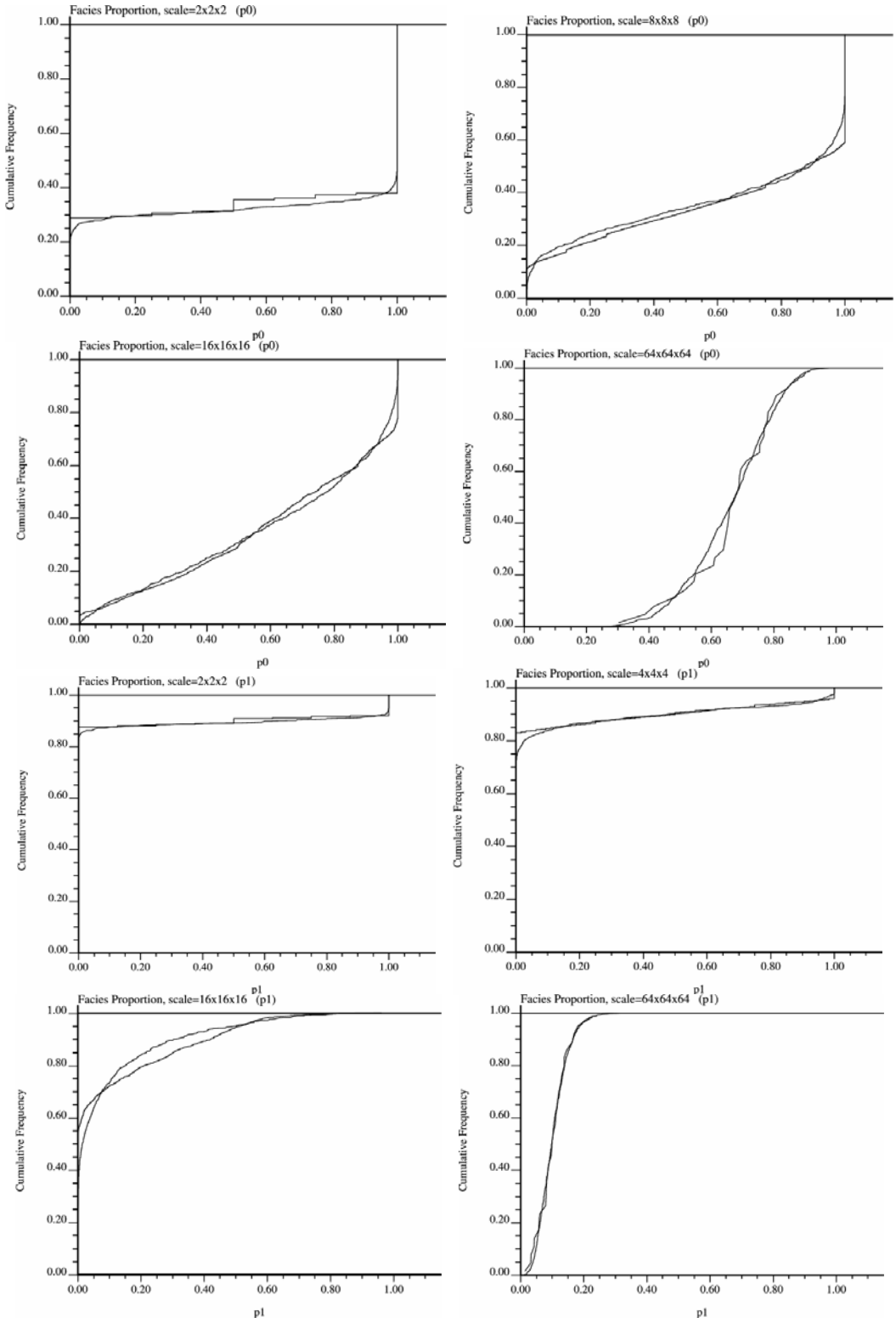


**Figure 15:** Beta simulated CDF fit

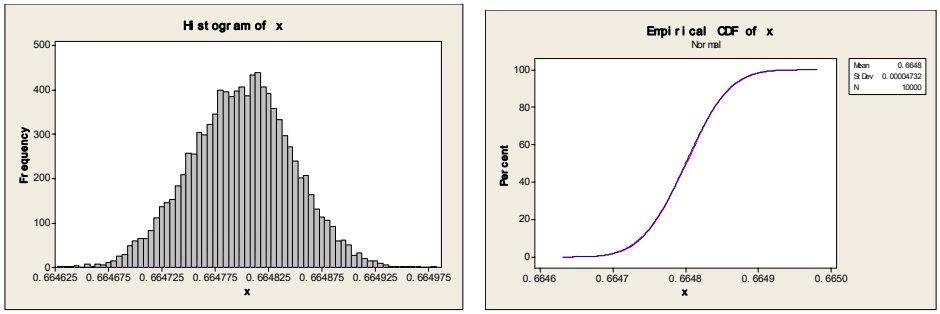


**Figure 16:** Beta simulated distribution for p0 at a very small variance

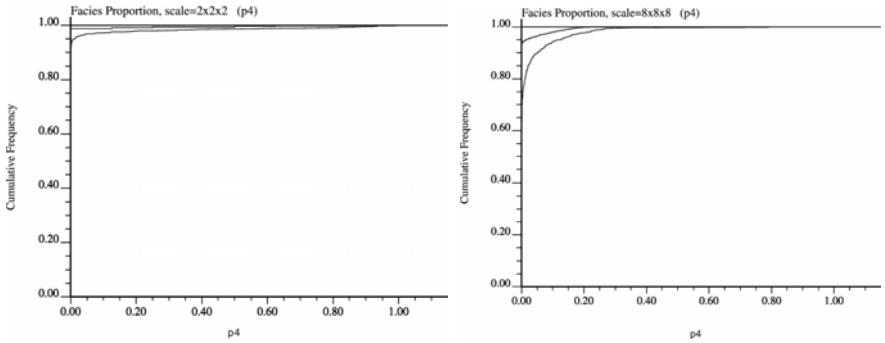


**Figure 17:** Figure 17 Beta simulated fits with prior global proportion 0.0047. The CDF's of simulated realizations lies below the CDF's of the data

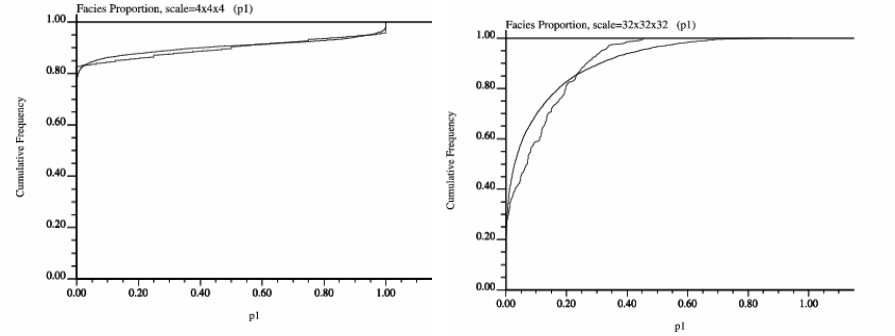


**Figure 18:** Dirichlet simulated realizations

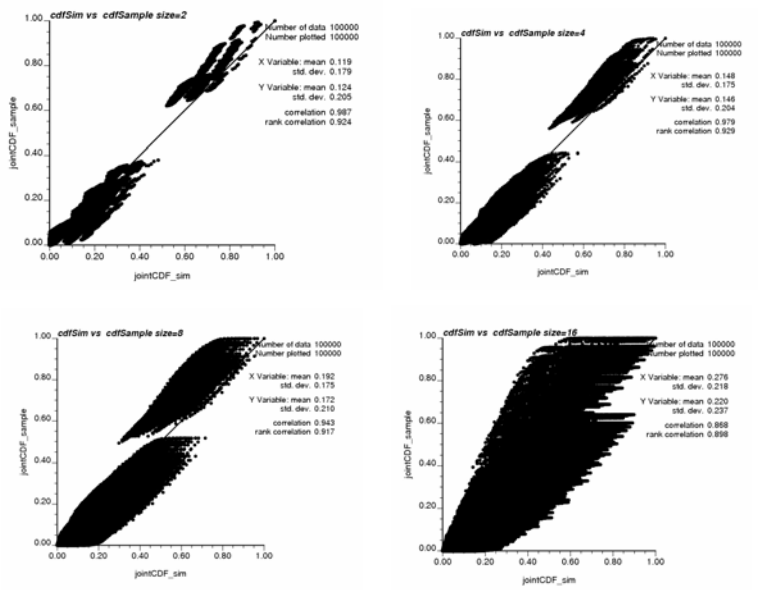


**Figure 19:** cross plots of sample jointCDF vs simulated jointCDF

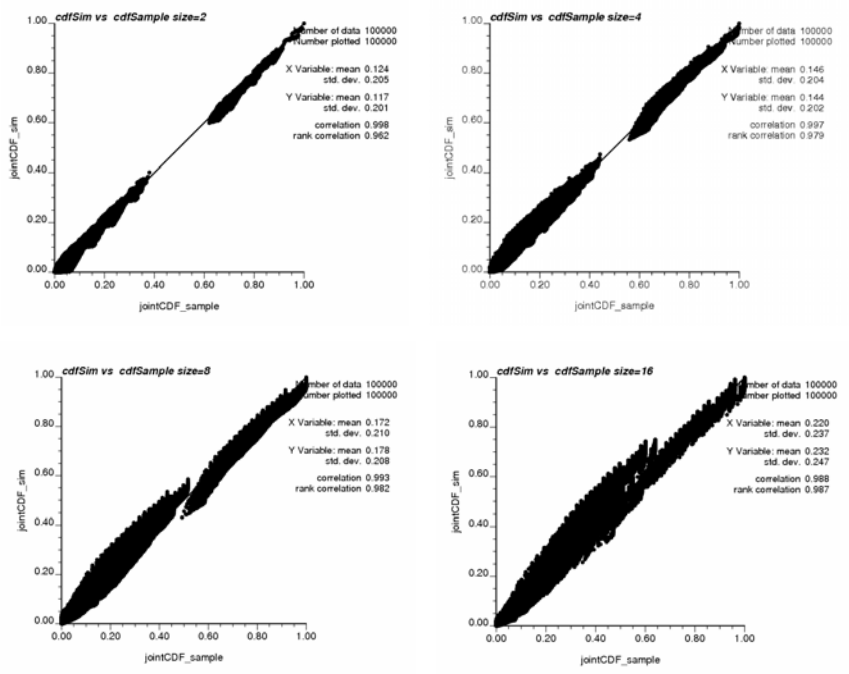


**Figure 20:** cross plots of sample jointCDF vs simulated jointCDF

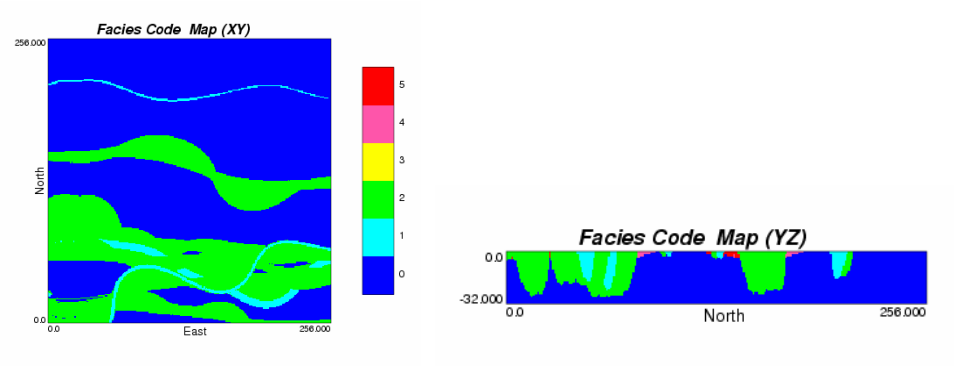


**Figure 21:** Facies maps XY: slice 16 XZ, YZ: slice 100

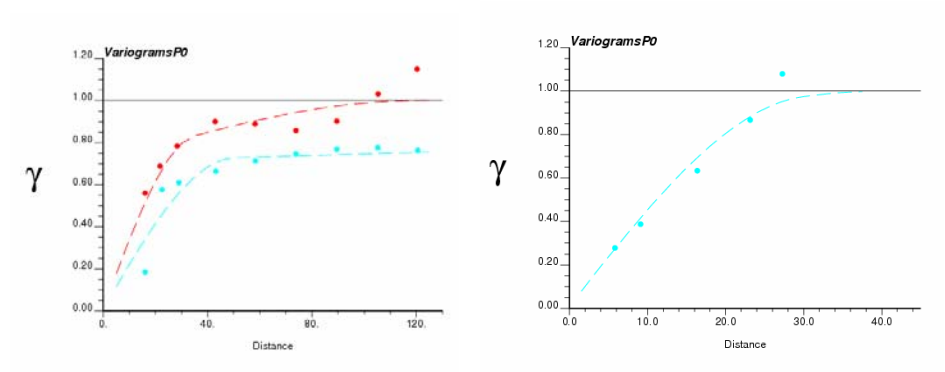


**Figure 22:** A variogram model fit

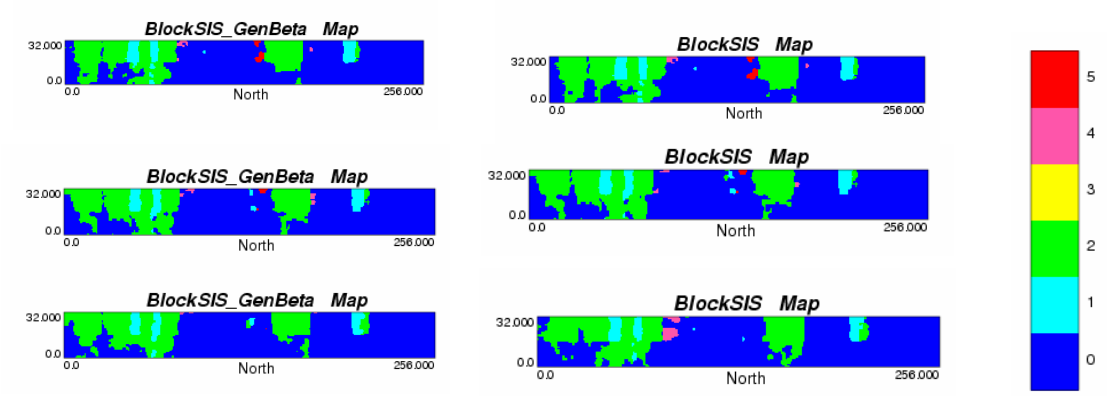


**Figure 23:** Realizations simulated from ordinary beta (left) and from `BlockSIS` (right)

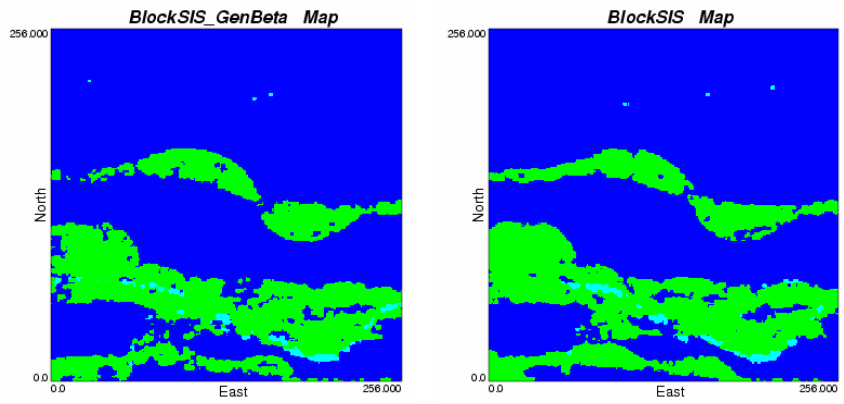


**Figure 24:** Realizations simulated from ordinary beta (left) and from `BlockSIS` (right)