Les Cahiers du GERAD ISSN: 0711–2440

High-Order Spatial Direct and Cross-Statistics for Categorical Attributes

D.F. Machuca-Mory
R. Dimitrakopoulos

G–2013–62

September 2013

High-Order Spatial Direct and Cross-Statistics for Categorical Attributes

David F. Machuca-Mory
Roussos Dimitrakopoulos*

COSMO – Stochastic Mine Planning Laboratory
Department of Mining and Materials Engineering
McGill University
Montreal (Quebec) Canada, H3A 2A7

david.machucamory@mcgill.ca
roussos.dimitrakopoulos@mcgill.ca

* *and GERAD*

September 2013

Les Cahiers du GERAD
G–2013–62

Abstract: The characterization of the spatial continuity of categorical variables, such as geological units, is a longstanding subject in geostatistics. Indicator covariances and variograms are used to measure spatial relationships of categorical data between pairs of points. Alternatively, transition probabilities, or transiograms, have been proposed to measure the probability of transition from one category to another as a function of distance. Recently, high-order moments and cumulants built from them have been proposed as measures of complex non-linear spatial relationships for arrangements of multiple points in 3D space. This paper extends the spatial high order statistics, originally conceived for continuous data, to the analysis of categorical spatial datasets. In addition the concept of two-point conditional transition probabilities is expanded to multiple point conditioned probabilities. The algorithm for high-order statistics, HOSC, has been updated to allow for the proposed high-order indicator spatial statistics. A third extension developed is the inference of indicator cross-cumulants and transiograms for two different categories. The experimental spatial indicator cumulants and transition probabilities for different scattered datasets are compared with those obtained from corresponding exhaustive datasets and training images. These comparisons show that significant information about the high-order and multiple point spatial continuity of categorical variables can be extracted from scattered samples. These results open an avenue for the development of simulation algorithms that rely more on data and less on training images.

Key Words: Spatial indicator cumulants, multiple point transition probabilities, categorical variables, training image.

1 Introduction

The modeling of the spatial structure of categorical attributes is a common task in various disciplines related to the earth and environmental sciences and engineering. In mining, for instance, the modelling of geological units is often required to define domains that differentiate populations of grades and, consequently, it is critical for the mineral resources evaluation. Geological units and rock types are examples of mutually exclusive categorical attributes, i.e. they can take only one category or state at each location. The indicator formalism has been proposed as the basis of a non-parametrical approach to deal with distributions of categorical attributes. Under this formalism, a binary coding is applied to data depending if a certain category is observed or not at the datum location (Goovaerts 1997). The variograms and covariances applied on these indicator transforms measure the 2-point spatial continuity of the corresponding attributes. These statistics are used in sequential indicator simulation to generate stochastic models of the spatial distributions of categorical attributes (Alabert 1987). Transition probabilities, or transiograms, were proposed as a more interpretable alternative to indicator covariances and variograms (Carle and Fogg 1996). Being 2-point statistics, indicator variograms and covariances fail to characterize the complex patterns that geological categorical variables often exhibit (Journel 2005). Consequently, traditional simulation algorithms based on these 2-point statistics are not able to reproduce complex spatial features, but result in patchy and disconnected patterns.

The popularity of simulation techniques based on 2-point statistics was justified by the difficulty to extract high-order statistics from very sparse data. The use of training images as a source of multiple-point statistics was proposed in the early nineties by Guardiano and Srivastava (1992). In this first implementation, the training image was scanned to obtain the multiple-point statistics required to infer the conditional probability distribution function (*cpdf*) at each unsampled node. But it was not until improved algorithms and data structures for storing large number of multiple-point statistics, such as the search tree, were incorporated that simulation based on multiple-point statistics became practical (Strebelle 2000). Presently, algorithms such as SNESIM (Strebelle 2002) and its various upgrades are used for geological modeling, particularly in the petroleum industry, where hard data is usually scarce and indirect information is available to assist the construction of training images. Contrarily, in the mining industry, hard data is often abundant. Accordingly, the inference of high-order spatial statistics, such as moments and cumulants from dense datasets is viable (Dimitrakopoulos, Mustapha and Gloaguen 2010). High-order statistics are used for approximating non-Gaussian conditional distributions that can be used in the simulation of continuous spatial attributes (Mustapha and Dimitrakopoulos 2010a). However, a training image is still needed to complement the information provided by hard data.

This paper explores the inference of high-order spatial moments, cumulants and multiple-point transition probabilities for categorical attributes in hard data. Spatial direct and cross indicator cumulants are presented as measures of high-order spatial continuity for one or more categorical attributes. Multiple-point transition probabilities express the conditional probability of having a particular category at one point given that the categories at various surrounding points are known. The same spatial high-order indicator moments that are used to build the direct and cross indicator cumulants are used to build the multiple-point transition probabilities. The application of these statistics to continuous and categorical data is shown with the help of 2-D and 3-D datasets. The resulting cumulant and probability maps are compared with those obtained from corresponding geological models and exhaustive training images. The implementation of these high-order statistics are intended to be the first step towards simulation methods that rely more on the multiple-point spatial continuity informed by hard data besides relying only on training images.

2 High-order Spatial Indicator Statistics

As in traditional geostatistics, high-order and multiple point indicator statistics are based on the indicator transformation. If $z(\boldsymbol{u})$ represents a categorical attribute, and s_k is a category or state among a finite number K of states, the indicator transform $i(\boldsymbol{u}; s_k)$ is one if the state s_k is present at $\boldsymbol{u}$, and zero otherwise (Goovaerts

1997). This is commonly expressed as:

$$i(\boldsymbol{u}; s_k) = \begin{cases} 1 & \text{if } z(\boldsymbol{u}) = s_k \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

In this paper, the categories s_k, $k \in \{1, \ldots, K\}$ are regarded as mutually exclusive, i.e. only one category s_k can be present at a location $\boldsymbol{u}$.

In geostatistics, the uncertainty related to the spatial distribution of an attribute is modeled by the joint probabilistic distribution of multiple random variables $Z(\boldsymbol{u}_\alpha)$ defined at N locations $\boldsymbol{u}_\alpha$, $\alpha = 1, \ldots, N$. The joint random variables form a random field and their corresponding indicator transforms are denoted as $I(\boldsymbol{u}_\alpha; s_k)$, with $\alpha = 1, \ldots, N$ and $k \in \{1, \ldots, K\}$.

Both spatial indicator spatial cumulants and multiple-point transition probabilities are derived from high-order moments, which can be used to describe a distribution. Consider a stationary categorical random field $Z(\boldsymbol{u})$ in a spatial domain V, and a moving template formed by a tail at point $\boldsymbol{u}_0 \in V$ plus n heads at the end of vectors $\boldsymbol{h}_1, \ldots, \boldsymbol{h}_n$, such as $\boldsymbol{u}_1 = \boldsymbol{u}_0 + \boldsymbol{h}_1, \ldots, \boldsymbol{u}_n = \boldsymbol{u}_0 + \boldsymbol{h}_n \in V$. Also, for the sake of generality, consider that, for each point $\boldsymbol{u}_\alpha$, $\alpha = 0, \ldots, n$ in the template, the indicator transform is defined for a particular category s_{k_α}, with $k \in \{1, \ldots, K\}$. The moment of order $\omega = j_0 + j_1 + \ldots + j_n$ for the indicator transforms at the $n+1$ template endpoints is expressed as

$$\begin{aligned} \mu_{0\ldots n; j_0 \ldots j_n} &= E\left[I^{j_0}(\boldsymbol{u}_0; s_{k_0}) \cdot I^{j_1}(\boldsymbol{u}_1; s_{k_1}) \cdot \ldots \cdot I^{j_n}(\boldsymbol{u}_n; s_{k_n})\right] \\ &= E\left[I(\boldsymbol{u}_0; s_{k_0}) \cdot I(\boldsymbol{u}_1; s_{k_1}) \cdot \ldots \cdot I(\boldsymbol{u}u_n; s_{k_n})\right] \end{aligned} \tag{2}$$

Thus, when dealing with indicator transforms the order of a moment is equal to its number of points. Due to this the terms order and points are used interchangeably in this paper.

There is a direct equivalence between the high-order indicator moment and the joint probability distribution function (*pdf*) of Z:

$$\begin{aligned} E\left[I(\boldsymbol{u}_0; s_{k_0}) \cdot I(\boldsymbol{u}_1; s_{k_1}) \cdot \ldots \cdot I(\boldsymbol{u}u_n; s_{k_n})\right] &= \Pr\left[Z(\boldsymbol{u}_0) = s_{k_0} \wedge Z(\boldsymbol{u}_1) = s_{k_1} \wedge \ldots \wedge Z(\boldsymbol{u}_n) = s_{k_n}\right] \\ &= f_Z(\boldsymbol{u}_0, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_n; s_{k_0}, s_{k_1}, \ldots, s_{k_n}) = p_{s_{k_0}, \ldots s_{k_n}}(\boldsymbol{h}_1, \ldots, \boldsymbol{h}_n), \end{aligned} \tag{3}$$

where $p_{s_{k_0}, \ldots s_{k_n}}(\boldsymbol{h}_1, \ldots, \boldsymbol{h}_n)$ is a joint probability that depends on vectors $\boldsymbol{h}_1, \ldots, \boldsymbol{h}_n$. When $s_{k_\alpha} = s_k$, $\forall \alpha = 0, \ldots, n$, the moment is referred to as an indicator direct moment of order $n+1$, or $(n+1)$-point indicator direct moment. Otherwise, it is an indicator cross-moment of order $n+1$ or $(n+1)$-point indicator cross-moment. If two or more points in the template are coincident, the order of the indicator direct moment is reduced; for instance,

$$E\left[I(\boldsymbol{u}_0; s_k) \cdot I(\boldsymbol{u}_1; s_k) \cdot \ldots \cdot I(\boldsymbol{u}_{n-1}; s_k) \cdot I(\boldsymbol{u}_n; s_k)\right] = E\left[I(\boldsymbol{u}_0; s_k) \cdot I(\boldsymbol{u}_1; s_k) \cdot I(\boldsymbol{u}_n; s_k)\right], \quad \text{if } \boldsymbol{u}_1 = \boldsymbol{u}_2 = \cdots = \boldsymbol{u}_{n-1}.$$

When two or more points in the template are coincident, the indicator-cross moment is zero, since two or more different categories cannot be present in the same point:

$$E\left[I(\boldsymbol{u}_0; s_{k_0}) \cdot I(\boldsymbol{u}_1; s_{k_1}) \cdot \ldots \cdot I(\boldsymbol{u}_{n-1}; s_{k_{n-1}}) \cdot I(\boldsymbol{u}_n; s_{k_n})\right] = 0, \quad \text{if } \boldsymbol{u}_\alpha = \boldsymbol{u}_\beta \text{ and } s_{k_\alpha} \neq s_{k_\beta} \text{ for } \alpha, \beta = 0, \ldots, n.$$

In a spatial context, high-order moments can be inferred directly from data when it is abundant and assuming that the joint probabilistic distribution within a region, or domain, is stationary. The next expression allows the experimental calculation of the high-order spatial moments from scattered samples:

$$\begin{aligned} &\hat{E}\left[I(\boldsymbol{u}_0; s_{k_0}) \cdot I(\boldsymbol{u}_1; s_{k_1}) \cdot \ldots \cdot I(\boldsymbol{u}_n; s_{k_n})\right] \\ &= \frac{1}{N_{\boldsymbol{h}_1, \ldots, \boldsymbol{h}_n}} \sum_{\alpha=1}^{N_{\boldsymbol{h}_1, \ldots, \boldsymbol{h}_n}} i(\boldsymbol{u}_\alpha; s_{k_0}) \cdot i(\boldsymbol{u}_\alpha + \boldsymbol{h}_1; s_{k_1}) \cdot \ldots \cdot i(\boldsymbol{u}_\alpha + \boldsymbol{h}_n; s_{k_n}) \end{aligned}. \tag{4}$$

Where $N_{\boldsymbol{h}_1,\ldots,\boldsymbol{h}_n}$ is the number of replicates that can be found for the template defined by the point $\boldsymbol{u}_\alpha$ and the head points of vectors $\boldsymbol{h}_1,\ldots,\boldsymbol{h}_n$. As for traditional variogram inference (Deutsch and Journel 1998), angular and distance tolerances can be allowed to deal with irregularly spaced data. Wider tolerances will lead to a more robust inference of the multiple-point moments, but at the price of masking features that describe actual patterns in the spatial distribution of the category.

2.1 Indicator spatial cumulants

Spatial cumulants are built as non-linear combinations of spatial moments (Mustapha and Dimitrakopoulos 2010b). The general relation between moments and cumulants (C) of any order is given by (Smith 1995)

$$\begin{aligned} C_I\left(Z_0,\ldots Z_{n-1},Z_n\right) = \sum_{a_0=0}^{1}\cdots\sum_{a_{n-1}=0}^{1}\sum_{a_n=0}^{1}\binom{1}{a_0}\cdots\binom{1}{a_{n-1}}\binom{1}{a_n} C_I\left(Z_0^{1-a_0},\ldots Z_n^{1-a_{n-1}},Z_n^{1-a_n}\right) \\ \times E\left[I^{a_0}(\boldsymbol{u}_0;s_0)\cdot\ldots\cdot I^{a_n}(\boldsymbol{u}_n;s_n)\right] \end{aligned}. \tag{5}$$

Note that $C_I\left(Z_0^{a_0},\ldots,Z_n^{a_n}\right) = C_I\left(Z_0,\ldots,Z_n\right)$, $\forall a_j \geq 1$. This implies that the order and the number of points for indicator cumulants is the same. Spatial cumulants can be expressed in terms of joint probabilities. Assuming stationarity and an ergodic random field the indicator cumulants up to the 4-order is given next:

$$\begin{aligned} C_I(\boldsymbol{u}_0;s_{k_0}) &= p_{s_{k_0}} \\ C_I(\boldsymbol{h}_1;s_{k_0},s_{k_1}) &= p_{s_{k_0}s_{k_1}}(\boldsymbol{h}_1) - p_{s_{k_0}}\cdot p_{s_{k_1}} \\ C_I(\boldsymbol{h}_1,\boldsymbol{h}_2;s_{k_0},s_{k_1},s_{k_2}) &= p_{s_{k_0}s_{k_1}s_{k_2}}(\boldsymbol{h}_1,\boldsymbol{h}_2) - p_{s_{k_0}}\cdot p_{s_{k_1}s_{k_2}}(\boldsymbol{h}_2-\boldsymbol{h}_1) - p_{s_{k_1}}\cdot p_{s_{k_0}s_{k_2}}(\boldsymbol{h}_2) \\ &\quad - p_{s_{k_2}}\cdot p_{s_{k_0}s_{k_1}}(\boldsymbol{h}_1) + 2p_{s_{k_0}}\cdot p_{s_{k_1}}\cdot p_{s_{k_2}} \\ C_I(\boldsymbol{h}_1,\boldsymbol{h}_2,\boldsymbol{h}_3;s_{k_0},s_{k_1},s_{k_2},s_{k_3}) &= p_{s_{k_0}s_{k_1}s_{k_2}s_{k_3}}(\boldsymbol{h}_1,\boldsymbol{h}_2,\boldsymbol{h}_3) - p_{s_{k_0}s_{k_1}}(\boldsymbol{h}_1)\cdot p_{s_{k_2}s_{k_3}}(\boldsymbol{h}_3-\boldsymbol{h}_2) \\ &\quad - p_{s_{k_0}s_{k_2}}(\boldsymbol{h}_2)\cdot p_{s_{k_1}s_{k_3}}(\boldsymbol{h}_3-\boldsymbol{h}_1) - p_{s_{k_0}s_{k_3}}(\boldsymbol{h}_3)\cdot p_{s_{k_1}s_{k_2}}(\boldsymbol{h}_2-\boldsymbol{h}_1) \\ &\quad - p_{s_{k_0}}\cdot p_{s_{k_1}s_{k_2}s_{k_3}}(\boldsymbol{h}_2-\boldsymbol{h}_1,\boldsymbol{h}_3-\boldsymbol{h}_2) - p_{s_{k_1}}\cdot p_{s_{k_0}s_{k_2}s_{k_3}}(\boldsymbol{h}_2,\boldsymbol{h}_3) \\ &\quad - p_{s_{k_2}}\cdot p_{s_{k_0}s_{k_1}s_{k_3}}(\boldsymbol{h}_1,\boldsymbol{h}_3) - p_{s_{k_3}}\cdot p_{s_{k_0}s_{k_1}s_{k_2}}(\boldsymbol{h}_1,\boldsymbol{h}_2) \\ &\quad + 2p_{s_{k_0}}\cdot p_{s_{k_1}}\cdot p_{s_{k_2}s_{k_3}}(\boldsymbol{h}_3-\boldsymbol{h}_2) + 2p_{s_{k_0}}\cdot p_{s_{k_2}}\cdot p_{s_{k_1}s_{k_3}}(\boldsymbol{h}_3-\boldsymbol{h}_1) \\ &\quad + 2p_{s_{k_0}}\cdot p_{s_{k_3}}\cdot p_{s_{k_1}s_{k_2}}(\boldsymbol{h}_2-\boldsymbol{h}_1) + 2p_{s_{k_1}}\cdot p_{s_{k_2}}\cdot p_{s_{k_0}s_{k_3}}(\boldsymbol{h}_3) \\ &\quad + 2p_{s_{k_1}}\cdot p_{s_{k_3}}\cdot p_{s_{k_0}s_{k_3}}(\boldsymbol{h}_2) + 2p_{s_{k_2}}\cdot p_{s_{k_3}}\cdot p_{s_{k_0}s_{k_3}}(\boldsymbol{h}_1) \\ &\quad - 6p_{s_{k_0}}\cdot p_{s_{k_1}}\cdot p_{s_{k_2}}\cdot p_{s_{k_3}} \end{aligned} \tag{6}$$

If $s_{k_0}=s_{k_1}=s_{k_2}=s_{k_3}$ the expressions above are direct indicator cumulants, otherwise, they are indicator cross-cumulants. Figure 1 shows a small example of direct third order indicator cumulants for continuous and categorical variables. The exhaustive dataset used for this example consists of five square areas with value 1 on a zero-value field (see Figure 1(a)). A three point template with lag vectors $\boldsymbol{h}_X$ and $\boldsymbol{h}_Y$ parallel to the coordinate axes was used to calculate the cumulants. Figure 1(b) shows the resulting 3$^{\text{rd}}$-order direct indicator cumulant map for category 1, whereas Figure 1(c) shows a similar map for category 0. In this case the two categories $s=1$ and $s'=0$ are complementary, i.e. $\Pr[z(\boldsymbol{u})=s'] = 1-\Pr[z(\boldsymbol{u})=s]$. The high value areas in these graphs are the result of lag vectors $\boldsymbol{h}_X$ and $\boldsymbol{h}_Y$ that permit collecting large number of 3-point replicates of categories s and s' for their corresponding indicator direct cumulants. Note that the high and low value spots in Figure 1(b) and (c), respectively, are spaced by the same distance units as the exterior category 1 squares in Figure 1(a). The impact of the central category 1 square in the cumulant maps at the right is minimal due to the particular configuration of the template.

When two categories are complementary, the corresponding 3$^{\text{rd}}$-order indicator direct cumulants are the negative of each other:

$$C_I(\boldsymbol{h}_1,\boldsymbol{h}_2;s,s,s) = -C_I(\boldsymbol{h}_1,\boldsymbol{h}_2;s',s',s') \text{ if } \Pr\left[Z(\boldsymbol{u})=s\right] + \Pr\left[Z(\boldsymbol{u})=s'\right] = 1. \tag{7}$$

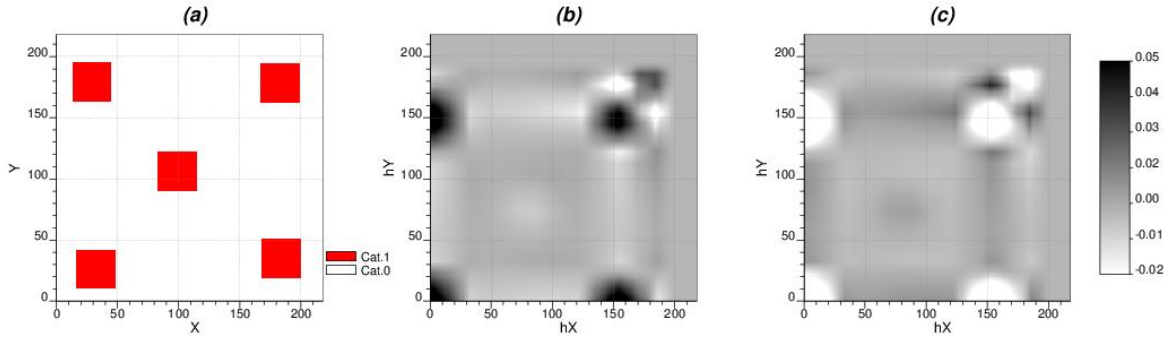

Figure 1: (a) a simple binary dataset, 3rd-order indicator direct cumulants for (b) category 1 and (c) category 0.

Also, the next equivalences can be established between 3rd-order indicator direct and cross-cumulants for complementary categories:

$$\begin{aligned} C_I(\boldsymbol{h}_1, \boldsymbol{h}_2; s, s', s) &= C_I(\boldsymbol{h}_1, h_2; s, s, s') = C_I(\boldsymbol{h}_1, \boldsymbol{h}_2; s', s', s'), \\ C_I(\boldsymbol{h}_1, \boldsymbol{h}_2; s, s', s') &= C_I(\boldsymbol{h}_1, \boldsymbol{h}_2; s, s, s). \end{aligned} \tag{8}$$

The demonstration of equivalences (7) and (8) and others between higher order indicator direct and cross-cumulants is not presented in this paper for the sake of brevity. Nevertheless, the top row of Figure 2 shows these equivalences for the template formed by lag vectors $\boldsymbol{h}_X$ and $\boldsymbol{h}_Y$ parallel to the coordinate axis, whereas the bottom row shows the indicator cross-cumulants using the $[\boldsymbol{h}_X, \boldsymbol{h}_Y]$ template after a 45° clockwise rotation, that is $[\boldsymbol{h}_{X'(Az.135°)}, \boldsymbol{h}_{Y'(Az.45°)}]$. The left column of Figure 2 shows the indicator cross-cumulant maps when the heads of lag vectors $\boldsymbol{h}_X$ and $\boldsymbol{h}_{X'}$ correspond to category 0 (Figure 2(a) and (d), respectively). The central column corresponds to indicator cross-cumulants when only the head of lag vectors $\boldsymbol{h}_Y$ (Figure 2(a)) and $\boldsymbol{h}_{Y'}$ (Figure 2(b)) correspond to category 0. In the right column of this figure, all the lag vector heads correspond to category 0. In all cases, the tail point corresponds to category 1. The indicator cross-cumulant maps for the rotated template shows the interaction between the central and exterior category 1 squares in Figure 1.

2.2 Multiple-point transition probabilities

The inference of 2-point transition probabilities, or transiograms, directly from categorical scattered data was proposed by Li (2006, 2007). This idea is expanded to multiple-point transition probabilities. The 1-point probability is the proportion of the category and it is defined by:

$$p_{s_{k_0}}(\boldsymbol{u}_0) = \Pr\left[Z(\boldsymbol{u}_0) = s_{k_0}\right] = E\left[I(\boldsymbol{u}_0; s_{k_0})\right]. \tag{9}$$

The 2-point transition probability is the conditional probability of having the category s_{k_0} in point $\boldsymbol{u}_0$, the tail, given that category s_{k_1} is present at the head of vector $\boldsymbol{h}_1$. This conditional probability can be obtained by the quotient of the 2-point indicator moment $E\left[I(\boldsymbol{u}_0; s_{k_0}) \cdot I(\boldsymbol{u}_1; s_{k_1})\right]$ with the 1-point indicator moment $E\left[I(\boldsymbol{u}_1; s_{k_1})\right]$ (Carle and Fogg 1996):

$$\begin{aligned} t_{s_{k_0}/s_{k_1}}(\boldsymbol{h}_1) &= \Pr\left[Z(\boldsymbol{u}_0) = s_{k_0} | Z(\boldsymbol{u}_1) = s_{k_1}\right] \\ &= \frac{\Pr\left[Z(\boldsymbol{u}_0) = s_{k_0} \wedge Z(\boldsymbol{u}_1) = s_{k_1}\right]}{\Pr\left[Z(\boldsymbol{u}_1) = s_{k_1}\right]} = \frac{E\left[I(\boldsymbol{u}_0; s_{k_0}) \cdot I(\boldsymbol{u}_1; s_{k_1})\right]}{E\left[I(\boldsymbol{u}_1; s_{k_1})\right]} \end{aligned} \tag{10}$$

As for indicator cumulants, if $s_{k_0} = s_{k_1}$ the expression above is a 2-point direct transition probability, or a cross- transition probability, if not. The multiple-point transition probability is built as the conditional probability of having a category s_{k_0} at $\boldsymbol{u}_0$ given that the head points of the vectors in the template are in the same or different categories:

$$t_{s_{k_0}/s_{k_1}\ldots s_{k_n}}(\boldsymbol{h}_1, \ldots, \boldsymbol{h}_n) = \Pr\left[Z(\boldsymbol{u}_0) = s_{k_0} | Z(\boldsymbol{u}_1) = s_{k_1} \wedge \ldots \wedge Z(\boldsymbol{u}_n) = s_{k_n}\right].$$

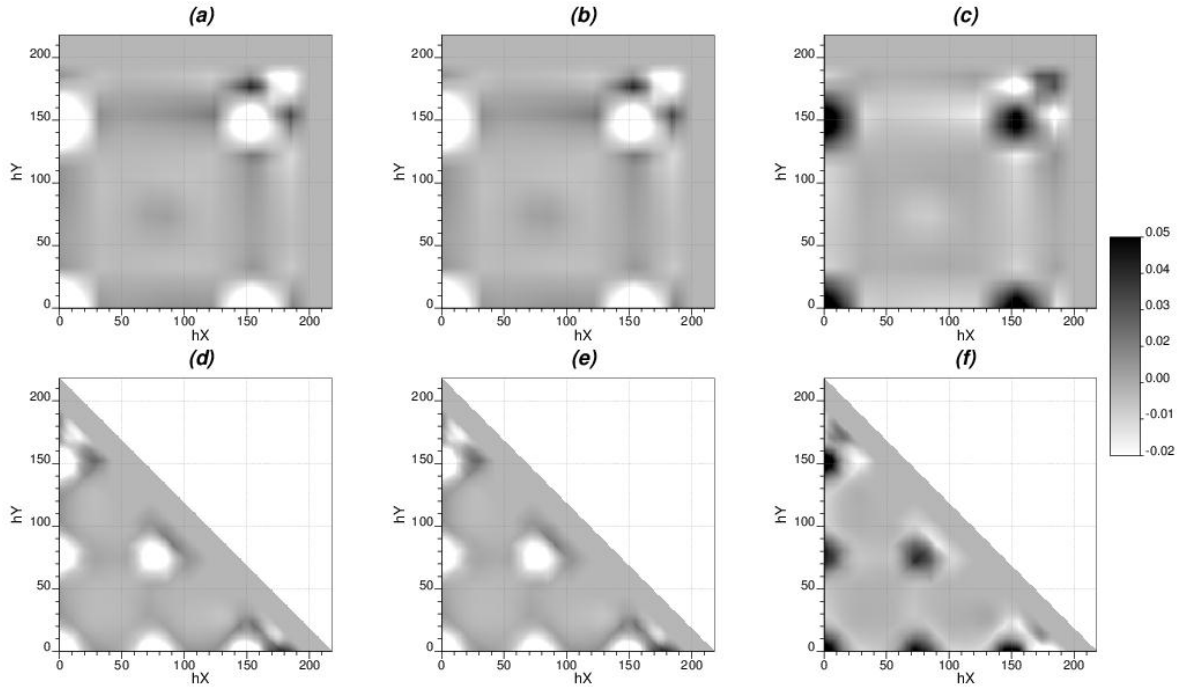

Figure 2: Top row, (a), (b) and (c) indicator cross-cumulant maps for a L shape template with lag vectors parallel to the coordinate axes. Bottom row, (d), (e) and (f) indicator cross-cumulant maps for a 45° rotated L shape template.

This conditional probability can be built by the quotient of the indicator moments $E[I(\boldsymbol{u}_0; s_{k_0}) \cdot \ldots \cdot I(\boldsymbol{u}_n; s_{k_n})]$ of order $n+1$ and $E\left[I(\boldsymbol{u}_1; s_{k_1}) \cdot \ldots \cdot I(\boldsymbol{u}_n; s_{k_n})\right]$ $E_{1\ldots n; s_{k_1}\ldots s_{k_n}}$ of order n:

$$\begin{aligned} t_{s_{k_0}/s_{k_1}\ldots s_{k_n}}(\boldsymbol{h}_1, \ldots, \boldsymbol{h}_n) &= \frac{\Pr\left[Z(\boldsymbol{u}_0) = s_{k_0} \wedge Z(\boldsymbol{u}_0 + \boldsymbol{h}_1) = s_{k_1} \wedge \ldots \wedge Z(\boldsymbol{u}_0 + \boldsymbol{h}_n) = s_{k_n}\right]}{\Pr\left[Z(\boldsymbol{u}_0 + \boldsymbol{h}_1) = s_{k_1} \wedge \ldots \wedge Z(\boldsymbol{u}_0 + \boldsymbol{h}_n) = s_{k_n}\right]} \\ &= \frac{E\left[I(\boldsymbol{u}_0; s_{k_0}) \cdot I(\boldsymbol{u}_1; s_{k_1}) \cdot \ldots \cdot I(\boldsymbol{u}_n; s_{k_n})\right]}{E\left[I(\boldsymbol{u}_1; s_{k_1}) \cdot \ldots \cdot I(\boldsymbol{u}_n; s_{k_n})\right]} \end{aligned} \tag{11}$$

As before, if $s_{k_0} = s_{k_1} = \ldots = s_{k_n}$, the expression above is a direct transition probability of order $n+1$, otherwise, it is a cross transition probability of the same order. Multiple-point transition probabilities can be inferred only if the denominator of the expression above is greater than zero, this is $E[I(\boldsymbol{u}_1; s_{k_1}) \cdot \ldots \cdot I(\boldsymbol{u}_n; s_{k_n})] > 0$, but they are defined as zero if the numerator, $E[I(\boldsymbol{u}_0; s_{k_0}) \cdot I(\boldsymbol{u}_1; s_{k_1}) \cdot \ldots \cdot I(\boldsymbol{u}_n; s_{k_n})]$, is also zero.

To further elucidate the basic properties of multiple point direct and cross transition probabilities let us analyse the 3-point case:

$$\begin{aligned} t_{s_{k_0}/s_{k_1}s_{k_2}}(\boldsymbol{h}_1, \boldsymbol{h}_2) = \Pr\left[Z(\boldsymbol{u}_0) = s_{k_0} | Z(\boldsymbol{u}_1) = s_{k_1} \wedge Z(\boldsymbol{u}_2) = s_{k_2}\right] &= \frac{p_{s_{k_0}s_{k_1}s_{k_2}}(\boldsymbol{h}_1, \boldsymbol{h}_2)}{p_{s_{k_1}s_{k_2}}(\boldsymbol{h}_2 - \boldsymbol{h}_1)} \\ &= \frac{E\left[I(\boldsymbol{u}_0; s_{k_0}) \cdot I(\boldsymbol{u}_1; s_{k_1}) \cdot I(\boldsymbol{u}_2; s_{k_2})\right]}{E\left[I(\boldsymbol{u}_1; s_{k_1}) \cdot I(\boldsymbol{u}_2; s_{k_2})\right]} \end{aligned} \tag{12}$$

The direct transition probability, that is for $s_{k_0} = s_{k_1} = s_{k_2}$, is 1 when $\boldsymbol{u}_0 = \boldsymbol{u}_1 = \boldsymbol{u}_2$, unless the numerator of expression (12) is zero.

Figure 3(a) shows a 3-point direct transition probability map of category $s_{k_0} = 1$ obtained from the same dataset presented in Figure 1(a). The pattern of the high probability areas reflects the arrangement of the category 1 squares in the dataset. The zero probability areas in this figure indicate that there are no instances

when the heads, $\boldsymbol{u}_X$ and $\boldsymbol{u}_Y$, of the corresponding $\boldsymbol{h}_X$ and $\boldsymbol{h}_Y$ vectors are in category 1 for a tail $\boldsymbol{u}_0$ that also is in such category. Figure 3(b) shows the cross-transition probability map of $t_{s_{k_0}=1/s_{k_X}=1,s_{k_Y}=0}(\boldsymbol{h}_X, \boldsymbol{h}_Y) = \Pr[Z(u_0) = 1 | Z(u_X) = 1 \wedge Z(u_Y) = 0]$. If $\boldsymbol{u}_X = \boldsymbol{u}_0$, the value of the transition probability is 1, since

$$t_{1/1,0}(0, \boldsymbol{h}_Y) = \frac{\Pr[Z(\boldsymbol{u}_0) = 1 \wedge Z(\boldsymbol{u}_0) = 1 \wedge Z(\boldsymbol{u}_Y) = 0]}{\Pr[Z(\boldsymbol{u}_0) = 1 \wedge Z(\boldsymbol{u}_Y) = 0]} = 1,$$

unless $E[I(\boldsymbol{u}_0; 1) \cdot I(\boldsymbol{u}_X; 1) \cdot I(\boldsymbol{u}_Y; 0)] = 0$. Whereas, if $\boldsymbol{u}_Y = \boldsymbol{u}_0$, the transition probability is zero. The two high probability bands parallel to the Y axis have the same separation as the horizontal distance between the exterior category 1 squares in Figure 1(a). These bands are created by replicates with both: the tail and the $\boldsymbol{u}_X$ head are in category one, whereas the $\boldsymbol{u}_Y$ head is in category zero. The bands become thinner for $\boldsymbol{h}_Y$ distances where the $\boldsymbol{u}_Y$ head is also in category one. A similar behavior is observed now parallel to the X axis when the head point in the complementary category is the end of vector $\boldsymbol{h}_X$, as in Figure 3(c), which shows the cross transition probability map for $t_{s_{k_0}=1/s_{k_X}=0,s_{k_Y}=1}(\boldsymbol{h}_X, \boldsymbol{h}_Y) = \Pr[Z(\boldsymbol{u}_0) = 1 | Z(\boldsymbol{u}_X) = 0 \wedge Z(\boldsymbol{u}_Y) = 1]$. Figure 3(d) presents the cross transition probability map when both head points are in the complementary category, this is $t_{s_{k_0}=1/s_{k_X}=0,s_{k_Y}=0}(\boldsymbol{h}_X, \boldsymbol{h}_Y) = \Pr[Z(\boldsymbol{u}_0) = 1 | Z(\boldsymbol{u}_X) = 0 \wedge Z(\boldsymbol{u}_Y) = 0]$. In this case the transition probability values along both template axes are zero. Note that in Figure 3(d), the maximum cross-transition probability happens for lag distances, in both $\boldsymbol{h}_X$ and $\boldsymbol{h}_Y$, for which no more category 1 areas can be found. Figure 3 (e), (f), (g), and (h), at the bottom, show similar direct and cross-transition probabilities than those in the top row of Figure 3, but they were inferred using the 45° rotated template. The features in these 3-point transition probability maps show the impact of the central category 1 square of Figure 1(a).

When comparing the 3$^{\text{rd}}$-order cumulant maps of Figure 2 with the 3-point transition probability maps of Figure 3, it is clear that the later show sharper features. This can be explained by considering that the indicator cumulants contain more information from single and joint probabilities than the transition probabilities.

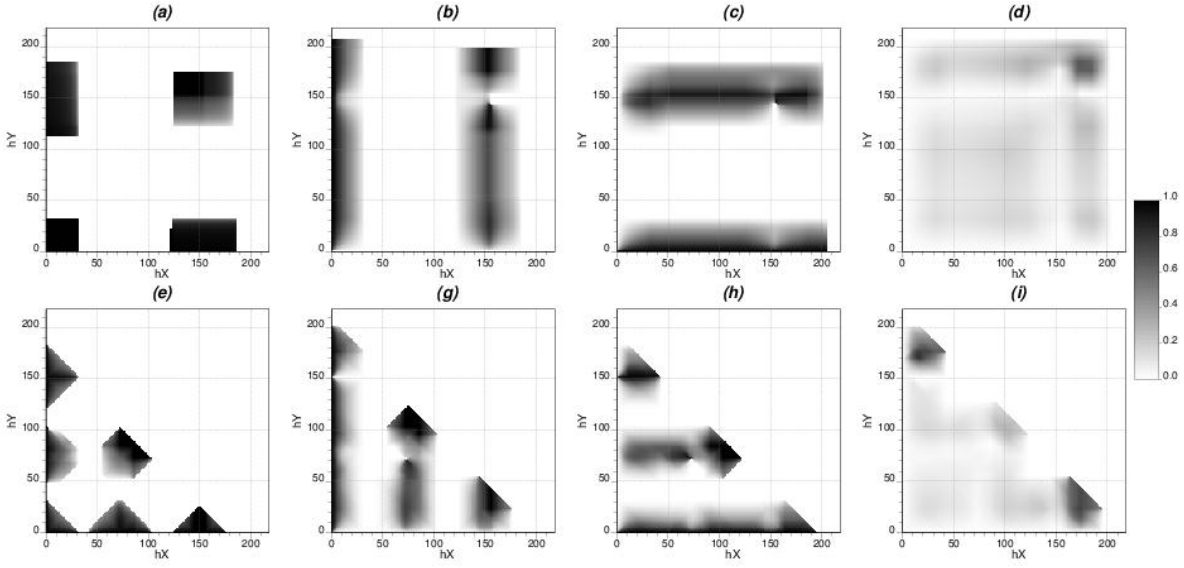

Figure 3: 3-point direct, (a) and (e), and cross, (b) to (d) and (f) to (h) transition probability maps for a L-shape template (top row) and 45° rotated L-shape template (bottom row).

3 Implementation and Case Studies

The prototype program, HOSC+, implements the calculation of indicator direct and cross-cumulants and transition probabilities from regularly and irregularly spaced data. In its current implementation, HOSC+ can deal with up to two different categories at the same time in the inference of cross high-order statistics.

HOSC+ was used to generate the direct cumulant and transition probability maps in the 2-D example and the direct and cross transition probability maps and volumes in the two 3-D cases study that are shown next.

3.1 A two-dimensional case

A horizontal 100 × 100 pixel slice of the Stanford V Reservoir Data Set (Mao and Journel, 1999) was selected for this 2-D example. The original continuous dataset has been transformed to a categorical image by applying a cut-off that divides the high values in the channels from the low values in the background. Figure 4(a) shows the exhaustive image of the selected slice. 380 samples are taken from this image on a pseudo regular 5 × 5 pixel grid. Figure 4(b) shows the location of the scattered samples. An L-shape template with lag vectors parallel to the X and Y coordinate axes was used for obtaining the 3^{rd}-order indicator cumulants and transition probabilities. For the 4^{th} order statistics a third lag vector parallel to the Z axis was considered. Incremental lags of 1 distance units were used for the exhaustive image, and 10 distance units for the scattered samples. Distance tolerances equal to half of the lag size were allowed, as well as angular tolerances of ±5°.

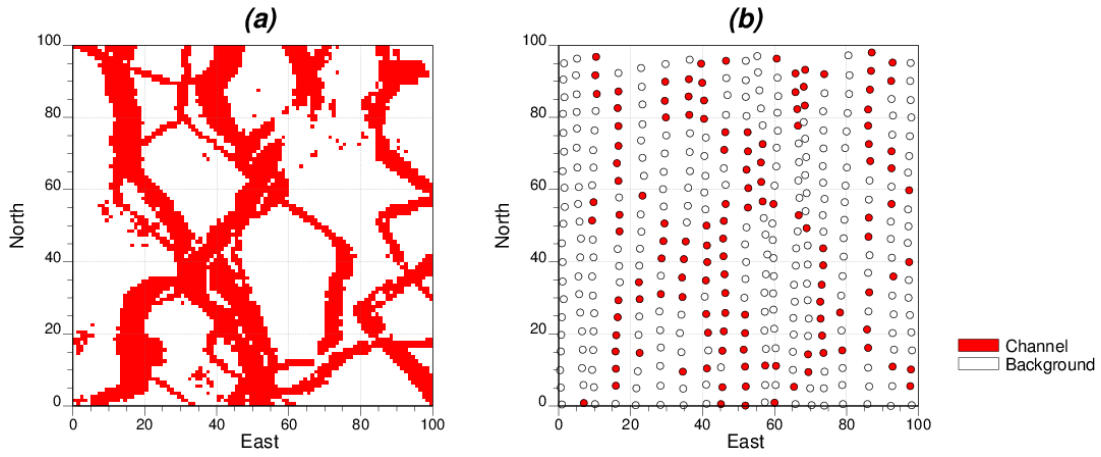

Figure 4: (a) exhaustive 2-D data set, (b) scattered samples taken from it.

The cumulants and transition probabilities maps shown next, as well as in the second example, were produced by interpolating the inferred values of these statistics. This facilitates their visualization and interpretation. Figure 5(a) and (b) show the indicator direct cumulant maps obtained from the exhaustive data set and the samples, respectively. Some common features can be identified between the two maps, such the elongated zero-value contours. These features reflect the geometry of the high value channels. The areas above zero result when the templates capture a large proportion of high values in the heads, while the tail value is below the cut-off. Figure 6(a) and (b) show the $X-Y$, $X-Z$ and $Y-Z$ slices of the 4^{th}-order indicator direct cumulant volumes for the exhaustive and scattered dataset, respectively. As for the 3^{rd}-order indicator cumulants in Figure 5, some similar features related to the geometry of the channels can be distinguished for the 4^{th}-order, particularly for the shorter lag distances.

Figure 7(a) and (b) show the three-point direct transition probabilities for the exhaustive and scattered datasets, respectively. These figures represent the conditional probability of the tail value being within the channels given that the two head values in the template also are in channels. The low probability banding stretched parallel to the North-South axe for $\boldsymbol{h}_X$ lags with lengths between 5 and 30 distance units reflect the short scale transition from channels to background. The low probability band for very large $\boldsymbol{h}_X$ lag vectors indicate that very few channels are separated horizontally by such distances. For medium $\boldsymbol{h}_X$ lags, both graphs show increased conditional probabilities, but those obtained from the exhaustive data set (Figure 7(a)) are higher than the obtained from the scattered samples (Figure 7(b)).

Figure 8 shows the 4-point indicator transition probability maps created using a template with three lag vectors with azimuths of 0°, 90° and 270°, respectively. These graphs express the conditional probability

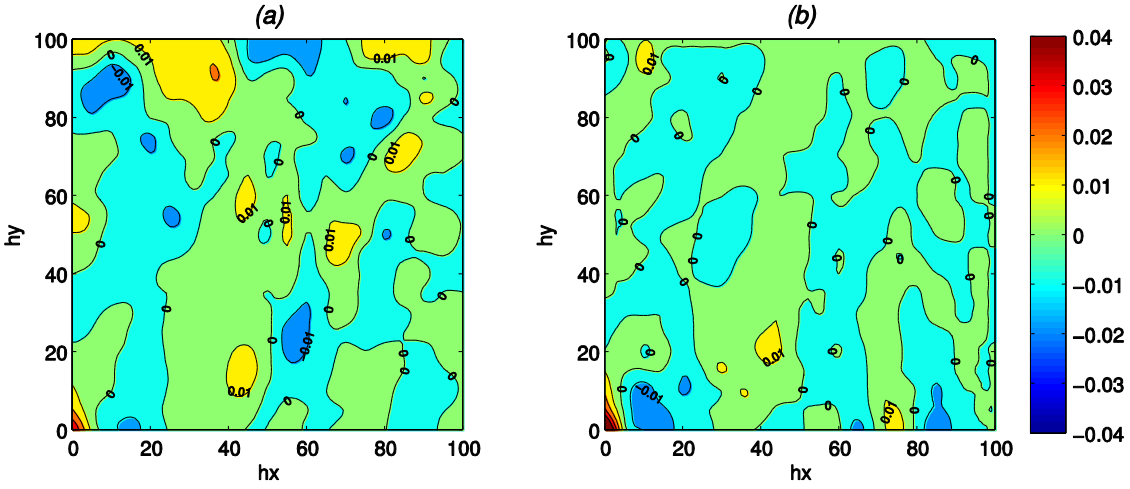

Figure 5: 3rd-order indicator direct cumulants for the exhaustive data (a) and for the scattered samples (b).

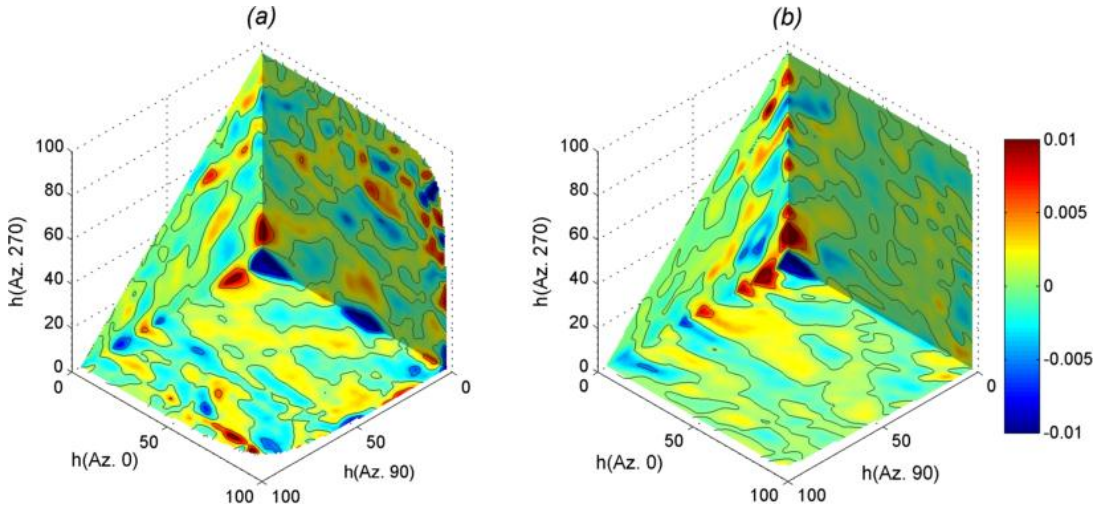

Figure 6: 4th-order indicator direct cumulants for the exhaustive data (a) and for the scattered samples (b).

of being inside the channel category when the samples at the North, East and West also are at different distances. The high probability bands parallel to the zero azimuth vectors reflect the geometry of the channels. These look less continuous for the exhaustive dataset (see Figure 8(a)) than for the scattered samples (see Figure 8(a)) since the former contains much more information about the short scale features of the channels.

3.2 3-D case: A structurally-controlled gold deposit

The dataset used for this three-dimensional case comes from the Apensu Gold Deposit in Ghana (Jones et al., 2011). The primary geological structure in this deposit is a north-east striking fault and its area of influence. Two families of subsidiary structures are present at eastwards of the fault zone. Figure 9(a) shows a view of the main fault zone and the most important of the families of subsidiary structures. Figure 9(b) shows the traces of 376 drill holes that jointly contain 10253 samples of 4.5m length. These samples are coded according to the geological structures they intercept. The holes were drilled sub-perpendicular to the main fault. Between Figure 9(a) and 9(b) there is a representation of the template geometry that was used for the inference of the 3 and 4-points transition probabilities. The first axis of this template, $\boldsymbol{h}_1$, is parallel to the

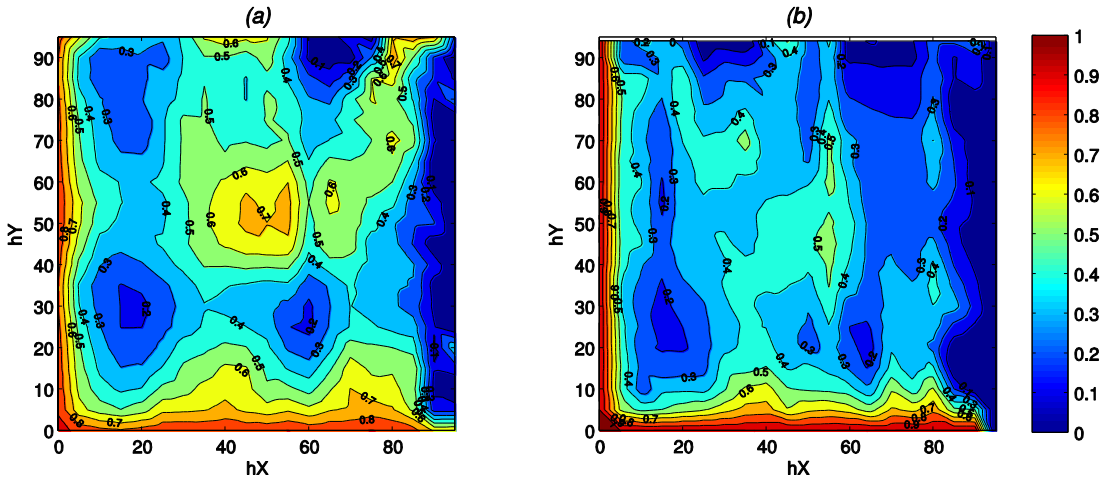

Figure 7: 3-Point transition probabilities for the exhaustive data (a) and the scattered samples (b).

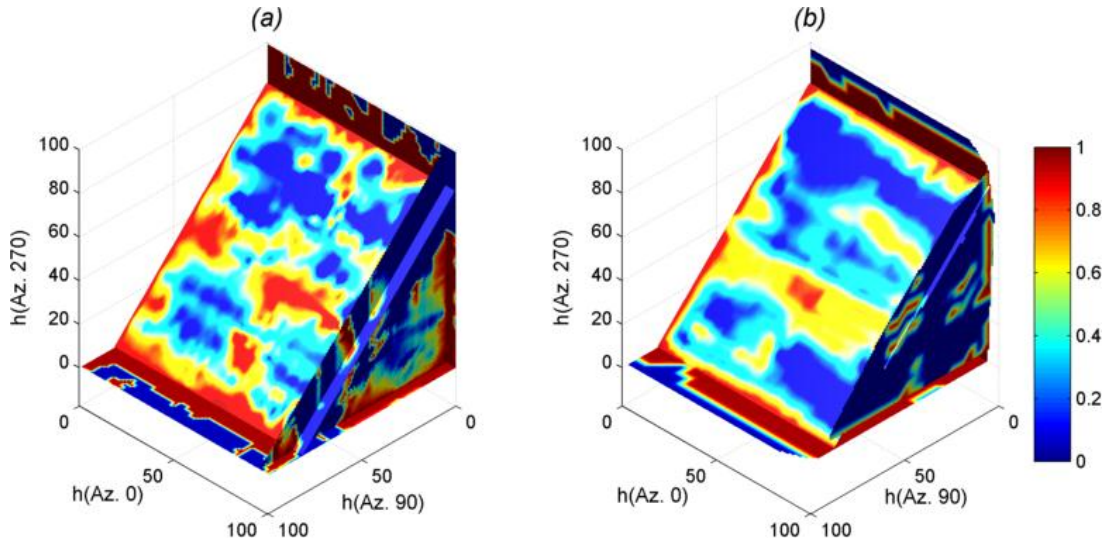

Figure 8: 4-point direct transition probability maps for (a) the exhaustive image, (b) the sample dataset.

average down-the-hole direction. The second axis, $\boldsymbol{h}_2$, has an azimuth of 35° and 0° dip, which correspond to the main fault's strike. The third axis of the template, $\boldsymbol{h}_3$, is perpendicular to the plane formed by the first two axes. This template geometry corresponds to the directions along where most conditioning sample replicates can be found.

Figure 10 presents various 3-point direct and cross-transition probabilities maps obtained from the drill hole sampling dataset using 3-point subsets of the template described above. The categories considered for the transition probabilities are the fault zone (F) and a family subsidiary structures (S1). The direct transition probabilities corresponding to the subsidiary structures in Figure 10(a) reflect the banding and cyclicity of this geological unit. The direct transition probabilities maps in Figure 10(d) for the fault zone category quickly fall to zero along the down-the hole direction but are preserved in the parallel direction to the fault structure and also along the 3$^{\text{rd}}$ template axes. This behavior conforms to the geometry of the fault zone. The two of the 6 possible cross transition probabilities maps that are shown in Figure 10 show the conditional probabilities of being in the category S1 given the head points of the template are at categories F and S1.

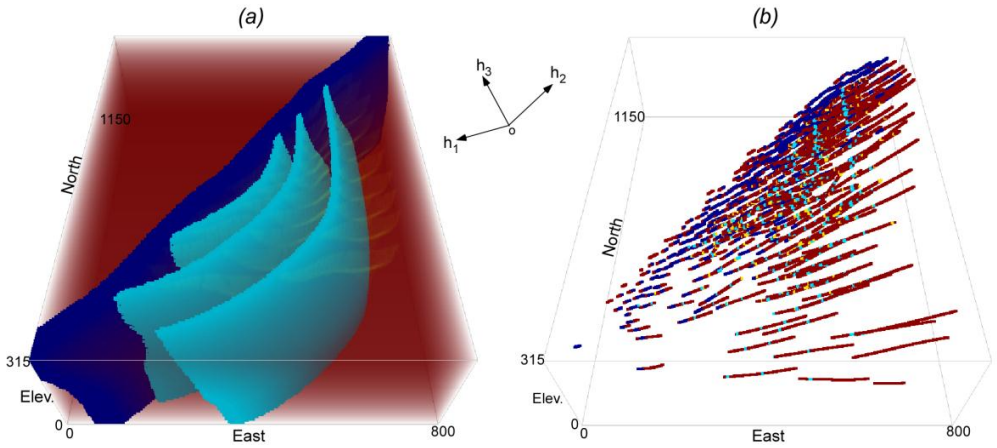

Figure 9: (a) geological model showing the thrust fault zone (F) and one family of subsidiary structures (S1). (b) Drill-hole traces coded by their intersections with the geological units.

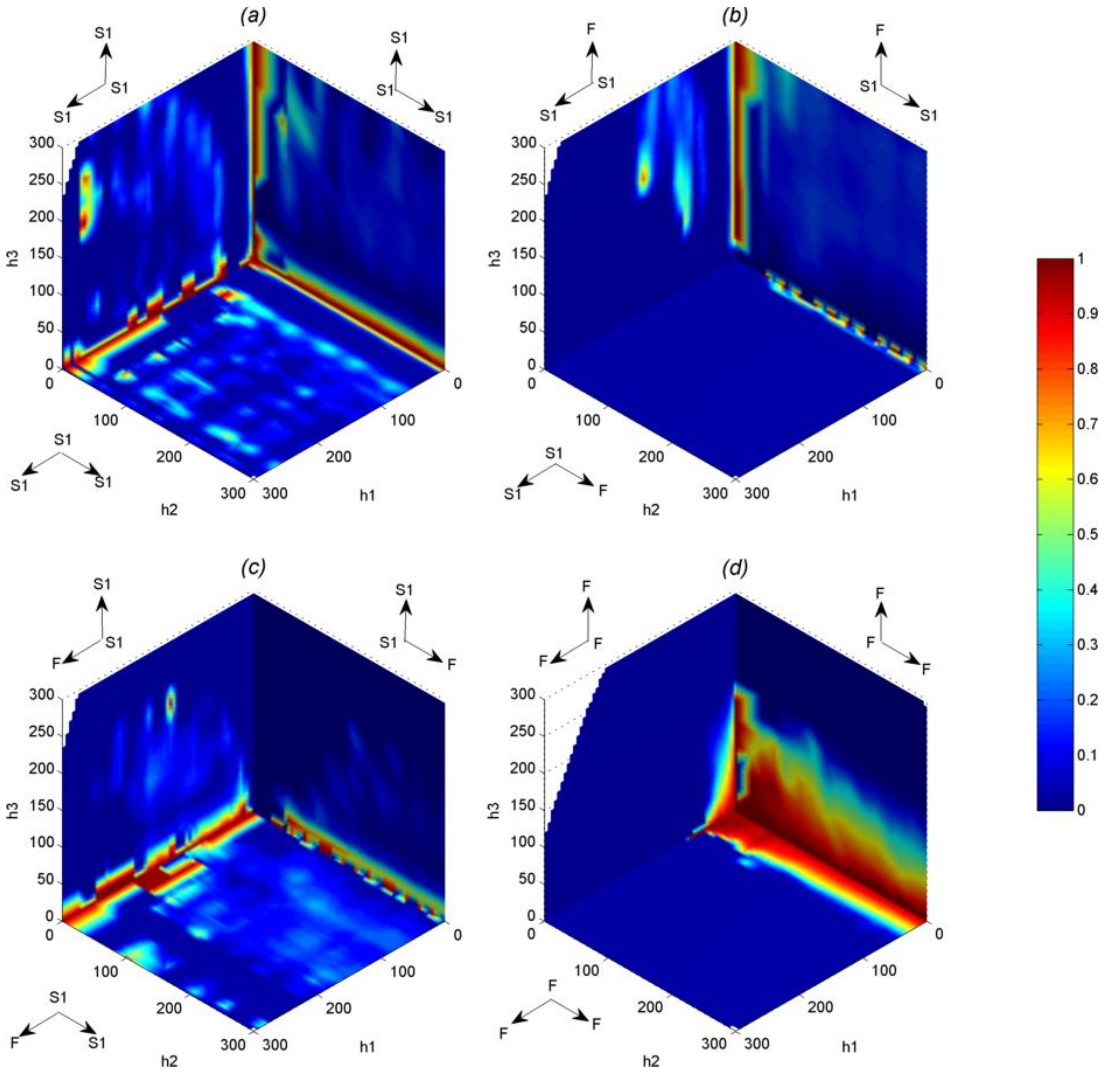

Figure 10: 3-point (a), (d) direct and cross (b), (c) transition probabilities maps for the Fault (F) and Structure Family 1 (S1) categories.

Figure 11 presents isosurfaces corresponding to the 0.25 conditional probability of being in the subsidiary structure when the 3 conditioning samples fall either in the subsidiary structure or in the fault zone. In Figure 11(a) all the conditioning samples fall in the subsidiary structures. The elongated isosurfaces parallel to $\boldsymbol{h}_2$ shown in this figure are produced by the interaction of samples within category S1. The separations between the three strings of isosurfaces correspond approximately to the separations between individual structures in the family S1. In Figure 11(b), all points but the head of $\boldsymbol{h}_3$ vector fall in the subsidiary structure. Two high probability bands related to the subsidiary structures are still present in this case, but the probability of finding this particular data arrangement fades with the distance to the fault. In Figure 11(c), the head of $\boldsymbol{h}_1$ vector fall in the fault zone, whereas all other heads fall in category S1. In Figure 11(d), the conditional probability surfaces are few and smaller since the instances when all heads in the template fall within the fault zone are restricted to the proximity of the fault.

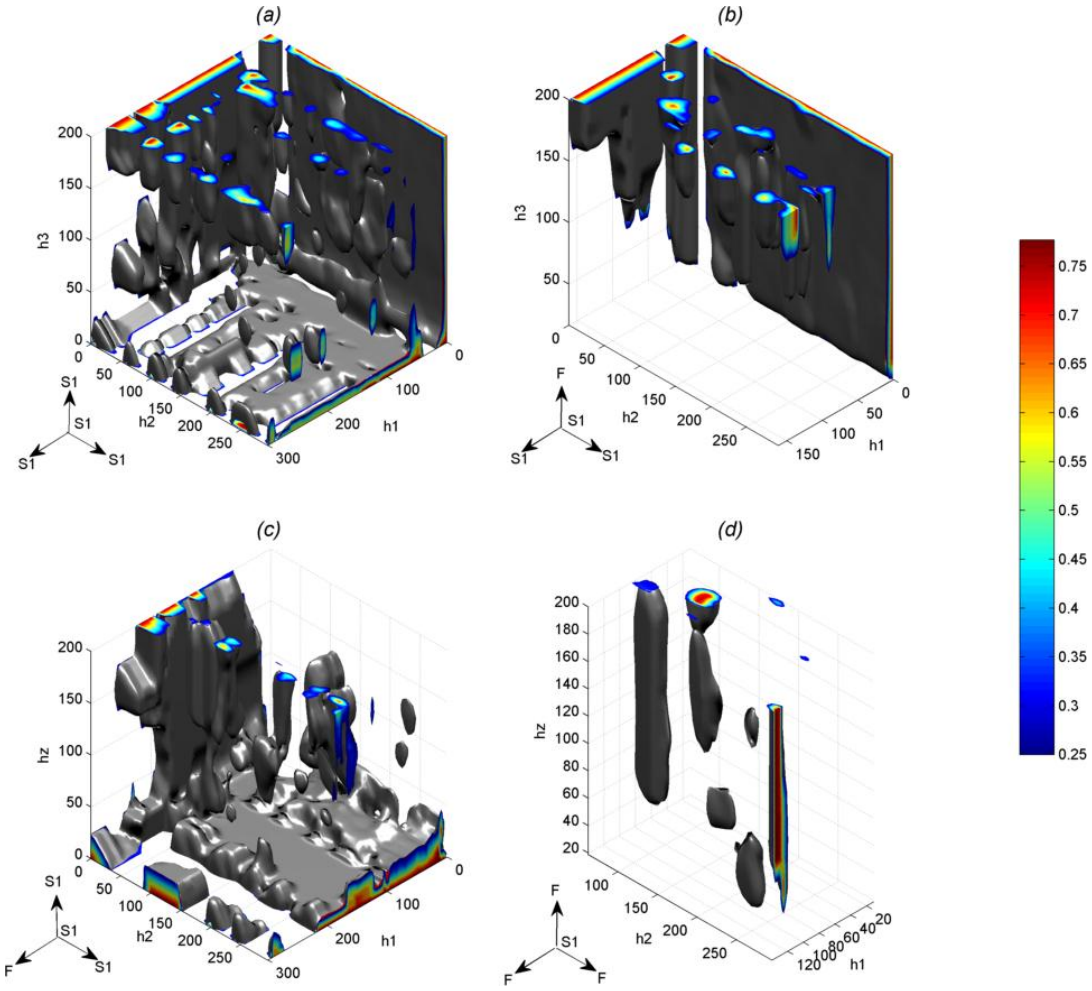

Figure 11: Direct, (a) and cross (b), (c) and (d) 4-point transition probabilities P25 isosurfaces for the Fault (F) and Structure Family 1 (S1) categories.

3.3 3-D case: A kimberlitic diamond pipe

The data for the last case study comes from the Fox kimberlitic diamond pipe, located in the Ekati property, Northwest Territories, Canada. This data consist of a geological 3D model of the pipe (see Figure 12(a)) and 3610 composited drill hole samples (see Figure 12(b)). The multiple rock types were grouped in four main geological units: crater, diatreme, xenoliths and host rock (cyan, yellow, red and blue, respectively in Figure 12).

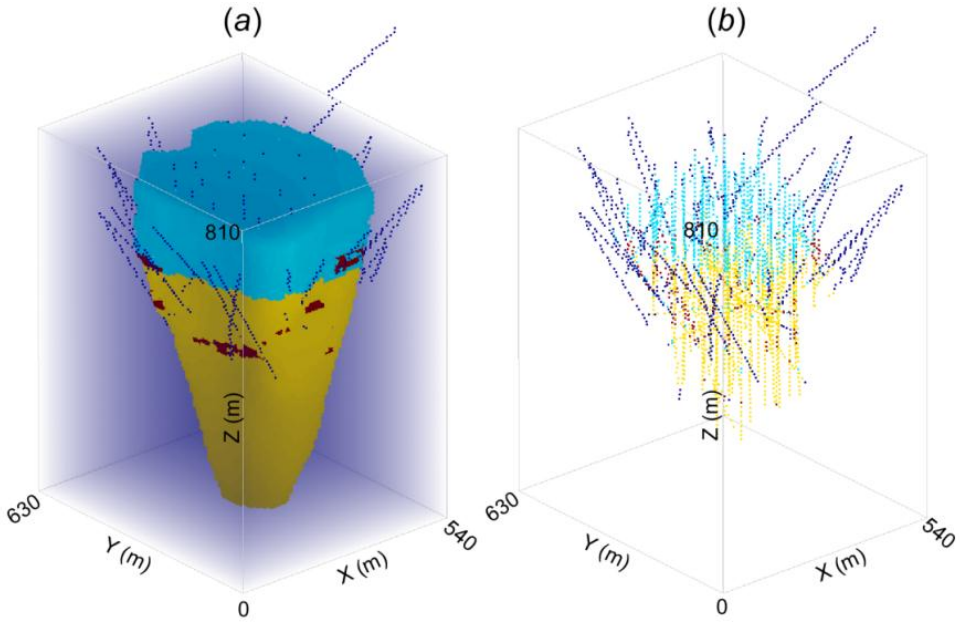

Figure 12: (a) Geological model of the Fox kimberlitic diamond pipe. (b) drill hole traces in the same pipe. Cyan: crater, yellow: diatreme, red: xenoliths, blue: host rock.

The 4-point template used for obtaining the direct transition probabilities is formed by the tail and heads of three lag vectors, $\boldsymbol{h}_X$, $\boldsymbol{h}_Y$ and $\boldsymbol{h}_{-Z}$, parallel to the axes X and Y, and vertical downwards, respectively. Figure 13 presents the 25% conditional probability isosurfaces obtained from the 4-point multiple point direct and cross-transition probabilities between diatreme and the other three geological units. Figure 13(a) expresses the conditional probability of being inside the diatreme given that the three heads also are within this geological unit. The resulting isosurface reflects the geometry of the diatreme boundaries. Figure 13(b) shows the conditional probability of being inside the diatreme given the downwards point of the template is also in diatreme, whereas the eastward and northward points fall in the host rock. The shape of the resulting 25% probability isosurface is related to the geometry of the contact between the diatreme and the host rock. Figure 13(c) corresponds to the 25% probability of being in diatreme conditioned to the downwards sample also being in diatreme while the others are in the crater. This isosurface is thin and fades beyond 40m depth, which corresponds to a smooth and unique crater-diatreme contact in the geological model. Figure 13(d) represents the 25% probability of transitioning from diatreme to a xenolith in the downward direction. The irregular shape of this isosurface reflects the irregular pattern of the xenoliths within the diatreme.

Figure 14 shows equivalent conditional probability isosurfaces to those in Figure 13, but they correspond to the drill hole samples. The shape of the sample transition probabilities isosurfaces is much less continuous than those obtained from the geological model. This happens because hard data is incomplete and also because the geological units in the samples show more spatial variability than in the geological model. For instance, Figure 14(c) indicates that there is more than 25% of transitioning horizontally from diatreme to crater at deeper points than indicated by the equivalent transition probabilities obtained from the geological model in Figure 13(c). This is because the scattered dataset actually contains samples coded as crater much deeper than the crater-diatreme contact surface in the geological model.

4 Discussion and Conclusions

Experimental indicator cumulants and transition probabilities are able to characterize complex spatial relationships from indicator transforms of continuous or categorical scattered data. As it can be observed from the case studies, much of the multiple-point spatial structure of geological units can be characterized directly

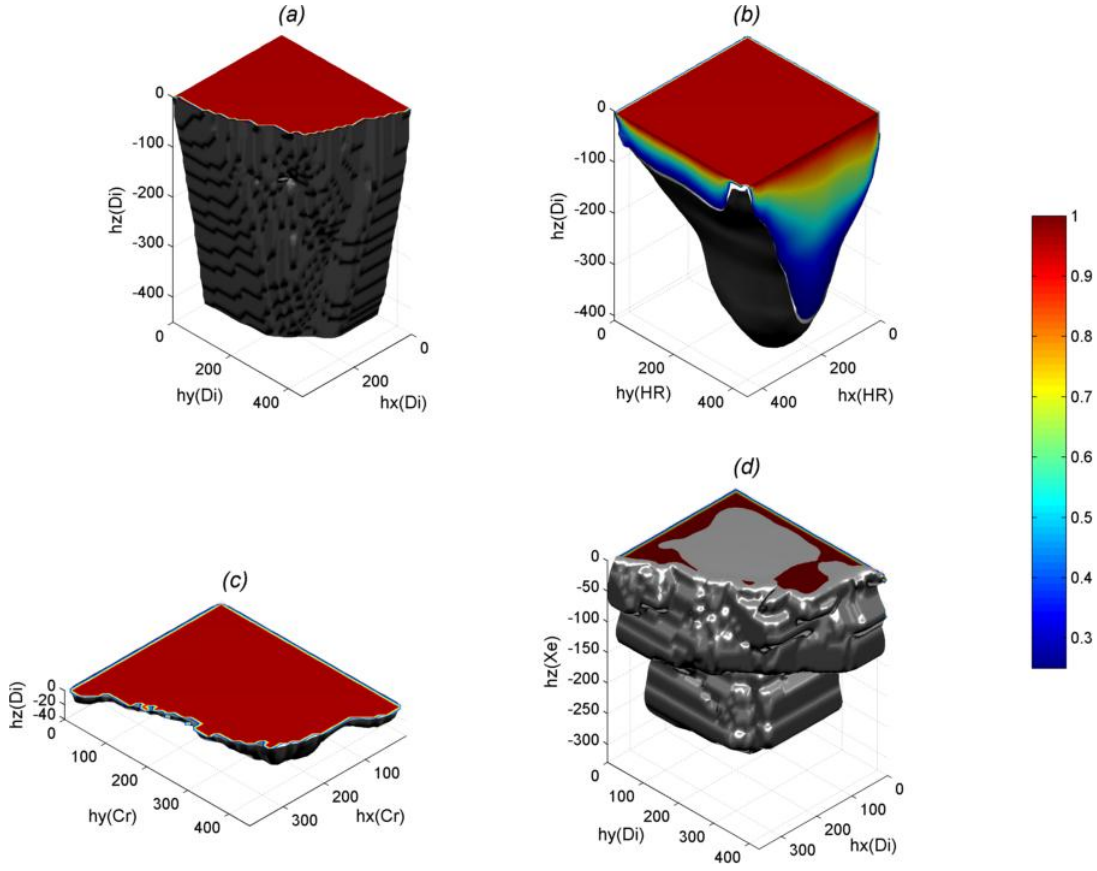

Figure 13: Isosurfaces obtained from the geological model representing the 25% direct and cross-transition probabilities within diatreme (a), diatreme and host rock (b), diatreme and crate (c) and diatreme and xenoliths (d).

from the information provided by scattered samples. This opens an avenue for the development of stochastic geological modeling methods based on actual hard and soft data rather than in preconceived training images.

The high-order indicator statistics maps obtained from only hard data often look either discontinuous or as if lacking detail. This is due mainly to the incompleteness of samples and the higher variability of its categorical values when compared with those of training images. Including soft data may result in more robust indicator high-order statistics.

Indicator cumulants carry information of multiple low- and high-order univariate and joint distributions. However, they are not straightforward to interpret. Transition probabilities are simpler and easier to interpret, but the main advantage of transition probabilities over indicator cumulants is that the former can be used directly to build the conditional distribution for a given arrangement of conditioning data and categories. These conditional probabilities can be further used for simulating geological categories.

A very important limitation for both cross indicator cumulants and cross transition probabilities is the exponential growth of the complexity of these statistics as the number of different categories considered increases. So far, the current implementation of HOSC+ considers only up to two different categories. The use of these statistics in a simulation algorithm will have to take into account this limitation. A way to overcome this may include the use of efficient data structures, such as search trees.

Future work includes the incorporation of transition probabilities obtained from scattered data in conditional simulation of categorical attributes. Another pending task is the implementation of HOSC+ as a SGeMS plug-in.

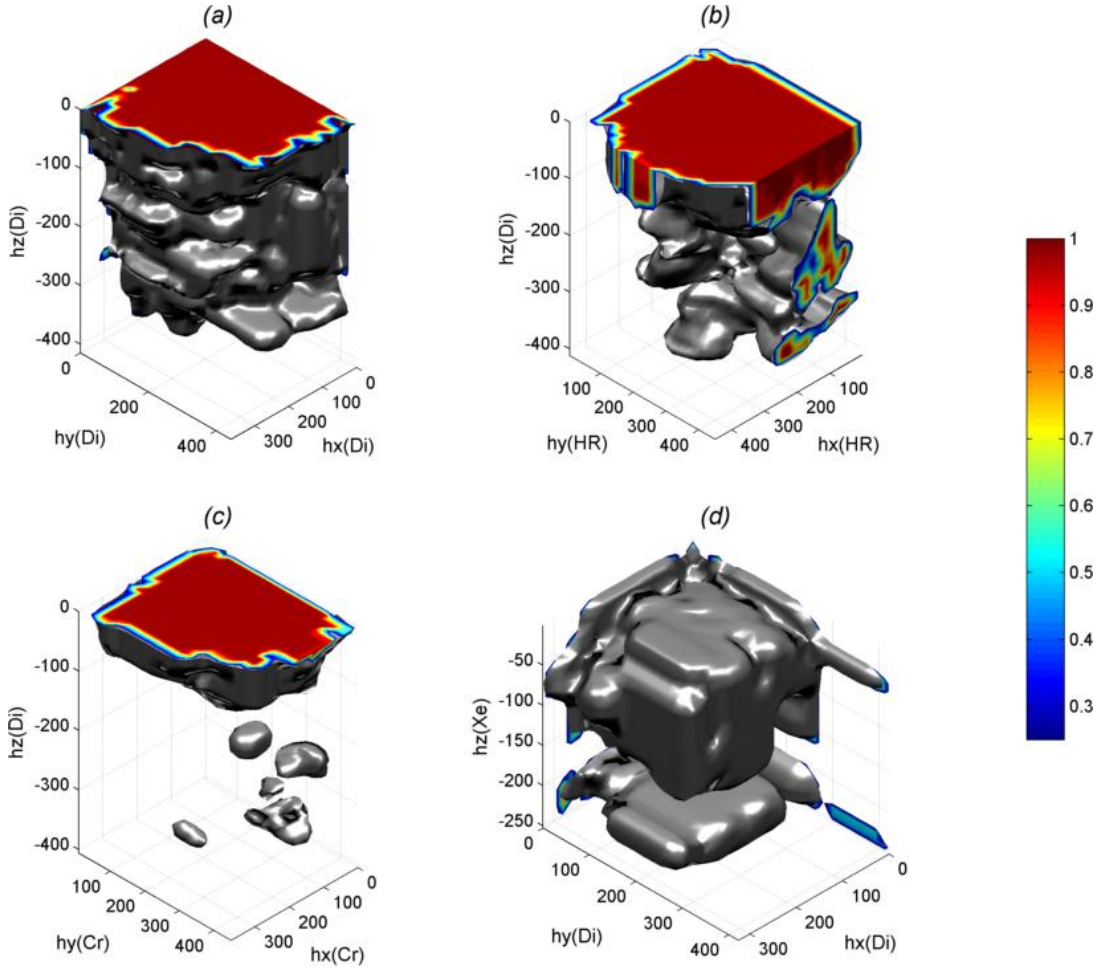

Figure 14: Isosurfaces obtained from the drill hole samples representing the 25% direct and cross-transition probabilities within diatreme (a), diatreme and host rock (b), diatreme and crate (c) and diatreme and xenoliths (d).

References

Alabert, F.G. (1987) Stochastic Imaging of Spatial Distributions Using Hard and Soft Information. Department of Applied Earth Sciences, Stanford University, Stanford, USA.

Carle, S.F. & G.E. Fogg (1996) Transition probability-based indicator geostatistics. Mathematical Geology, 28, 453–476.

Deutsch, C. & A. Journel (1998) GSLIB: Geostatistical software library and user's guide. New York: Oxford University Press.

Dimitrakopoulos, R., H. Mustapha & E. Gloaguen (2010) High-order Statistics of Spatial Random Fields: Exploring Spatial Cumulants for Modeling Complex Non-Gaussian and Non-linear Phenomena. Mathematical Geosciences, 42, 65-99.

Goovaerts, P. (1997) Geostatistics for natural resources evaluation. New York: Oxford University Press.

Guardiano, F. & R.M. Srivastava (1992) Multivariate geostatistics: Beyond bivariate moments. In Geostatistics-Troia, ed. A.M. Soares, 133–144. Dordrecht: Kluwer.

Jones, P., I. Douglas & A. Jewbali (2011) Modelling geological uncertainty in mining using multiple-point statistics. World Gold 2011, 3rd International Conference, Montréal, Canada.

Journel, A.G. (2005) Beyond Covariance: The Advent of Multiple-Point Geostatistics. In Geostatistics Banff 2004, O. Leuangthong & C.V. Deutsch (eds.), Quantitative Geology and Geostatistics Series, 225–233. Springer Netherlands.

Li, W. (2006) Transiogram: A spatial relationship measure for categorical data. International Journal of Geographical Information Science, 20, 693–699.

Li, W. (2007) Transiograms for Characterizing Spatial Variability of Soil Classes. Soil Sci. Soc. Am. J., 71, 881–893.

Mao, S. & A. Journel (1999) Generation of a reference petrophysical and seismic 3D data set: the Stanford V reservoir. Annual Meeting Report. Stanford, USA: Stanford Center for Reservoir Forecasting.

Mustapha, H. & R. Dimitrakopoulos (2010a) High-order Stochastic Simulation of Complex Spatially Distributed Natural Phenomena. Mathematical Geosciences, 42, 457–485.

Mustapha, H. & R. Dimitrakopoulos (2010b) A new approach for geological pattern recognition using high-order spatial cumulants. Computers & Geosciences, 36, 313–334.

Smith, P.J. (1995) A Recursive Formulation of the Old Problem of Obtaining Moments from Cumulants and Vice Versa. The American Statistician, 49, 217–218.

Strebelle, S. (2002) Conditional Simulation of Complex Geological Structures Using Multiple-Point Statistics. Mathematical Geology, 34, 1–21.

Strebelle, S. (2000) Sequential simulation drawing structures from training images. Ph.D. Thesis, Department of Geological and Environmental Sciences, Stanford University, Stanford, USA.