Computers & Geosciences 70 (2014) 190–205

Contents lists available at ScienceDirect

Computers & Geosciences

journal homepage: www.elsevier.com/locate/cageo

Verifying the high-order consistency of training images with data for multiple-point geostatistics

Cristian Pérez [a,b,*], Gregoire Mariethoz [c,d], Julián M. Ortiz [a,b]

[a] ALGES Laboratory – Advanced Mining Technology Centre (AMTC), Universidad de Chile, Santiago, Chile
[b] Department of Mining Engineering, Universidad de Chile, Santiago, Chile
[c] School of Civil and Environmental Engineering, The University of New South Wales, Sydney, Australia
[d] National Centre for Groundwater Research and Training, Australia

ARTICLE INFO

Article history:
Received 6 November 2013
Received in revised form
30 May 2014
Accepted 4 June 2014
Available online 13 June 2014

Keywords:
Multiple-point statistics
Direct sampling
Sensitivity analysis
Training image
Inference
Selection

ABSTRACT

Parameter inference is a key aspect of spatial modeling. A major appeal of variograms is that they allow inferring the spatial structure solely based on conditioning data. This is very convenient when the modeler does not have a ready-made geological interpretation. To date, such an easy and automated interpretation is not available in the context of most multiple-point geostatistics applications. Because training images are generally conceptual models, their preparation is often based on subjective criteria of the modeling expert. As a consequence, selection of an appropriate training image is one of the main issues one must face when using multiple-point simulation. This paper addresses the development of a geostatistical tool that addresses two separate problems. It allows (1) ranking training images according to their *relative* compatibility to the data, and (2) obtaining an *absolute* measure quantifying the consistency between training image and data in terms of spatial structure. For both, two alternative implementations are developed. The first one computes the frequency of each pattern in each training image. This method is statistically sound but computationally demanding. The second implementation obtains similar results at a lesser computational cost using a direct sampling approach. The applicability of the methodologies is successfully evaluated in two synthetic 2D examples and one real 3D mining example at the Escondida Norte deposit.

1. Introduction

Geological models are often built using deterministic techniques, meaning that their construction relies on the knowledge and experience of a specialist that assigns geological attribute values to a given volume. This practice is however not satisfying because it does not allow quantifying geological uncertainty (Refsgaard et al., 2012). With increasing frequency in recent years, geostatistical simulation has been used to construct stochastic models. Most of these techniques rely on the available sparse data (such as for example boreholes) to infer spatial continuity. This is accomplished using statistics based on relations between pairs of points, as for example variograms or correlograms (Caers, 2005; Deutsch and Journel, 1992; Isaaks and Srivastava, 1990; Journel and Huijbregts, 1978; Kitanidis, 1997). These conventional 2-point simulation methods are relatively simple and often show appropriate results. However, they present certain inherent limitations. Among those:

– Curvilinear or complex structures can be poorly represented by Gaussian simulations and may require higher order statistics (Journel and Zhang, 2006).
– Variograms do not inform about contacts or contact geometries between different categories (e.g. Carle and Fogg, 1996).
– The reliance of variogram-based Geostatistics on the maximum entropy, multi-Gaussian distribution to model all statistics beyond the two-point statistics results in maximum disconnectivity of extremes and the reproduction of only linear spatial features (e.g. Boisvert et al., 2007; Zinn and Harvey, 2003).

As an alternative, multiple point simulation algorithms, have recently become an important point of focus, with a wealth of different methods developed in the last decade (Arpat and Caers, 2007; Gloaguen and Dimitrakopoulos, 2009; Guardiano and Srivastava, 1993; Honarkhah and Caers, 2010; Mariethoz et al., 2010; Parra and Ortiz, 2011; Straubhaar et al., 2011; Strebelle, 2002; Tahmasebi et al., 2012; Zhang et al., 2006). These algorithms

* Corresponding author at: Department of Mining Engineering, Universidad de Chile, Avenida Tupper 2069, Santiago, Chile. Tel./fax: +56 222772556.
E-mail address: criperez@ing.uchile.cl (C. Pérez).

http://dx.doi.org/10.1016/j.cageo.2014.06.001
0098-3004/

model spatial continuity using higher order statistics, and therefore do not use variographic models to impose a spatial structure. As the amount of higher-order events contained in scattered data is usually insufficient, multiple point simulation algorithms require inferring the statistics of spatial patterns from a training image. Because training images are conceptual models, their preparation is often based on subjective criteria of the modeling expert. As a consequence, verifying the consistency of the training image with data is one of the main issues a modeling professional must face when using multiple-point simulation. It is even more important than variogram modeling for classical geostatistical simulation, as it controls higher-order, as well as second order spatial relations.

1.1. Background on training image selection

Two different types of data can be distinguished to serve as a base for the selection of training images. The first type is indirect state data such as flow and transport, which are typically integrated through inverse methods. The problem consists in evaluating the compatibility of a training image with dynamic outputs (for example time series of contaminant output). Approaches to select training images on this basis have been proposed in the framework of distance-based approaches (Suzuki and Caers, 2008). Another avenue in this direction is to consider each candidate training image as a prior model and to weight the different priors using a Bayesian mixture model (Khodabakhshi and Jafarour, 2013). A drawback of such approaches is that in general the selection of training images based on state data requires expensive forward model runs.

The other type of data, which is the specific focus of this paper, is static data. As for variograms, it consists in quantifying the consistency of a training image based on spatial statistics derived on both training image and data. To date, significant developments are lacking regarding objective criteria for verifying high-order training image consistency with the available scattered data. In this paper we focus on the consistency problem, and we leave aside some related issues as for example training image scaling issues (e.g. Ortiz et al., 2007).

One of the first approaches proposed for training image selection based on static data was initially proposed by Ortiz and Deutsch (2004). It consists in comparing the cumulative distribution of runs of the training image with the cumulative distribution of runs observed in 1D wells. Boisvert et al. (2007) proposed another method based on the comparison of multiple point histograms for vertical one-dimensional patterns. The training image and the conditioning data are scanned using a search template and the resulting statistics are compared.

A different approach is suggested by Eskandaridalvand (2008), who proposes a spiral search method. It loops over all conditioning nodes, and for each of them, over all training nodes. If the node in the training image has the same value as the conditioning node, the method loops over the close conditioning nodes from the nearest to the farthest. The values are compared to the nodes in the training image that show the same spatial configuration relative to the central node. If both values have the same relationship, i.e. if both increase or decrease in the same manner, a counter for compatible nodes increases. The method allows obtaining a distribution of compatible training nodes for each conditioning node and a unique distribution of maximum compatible nodes. These distributions can be used for deriving a measure of consistency between the training image and the conditioning information.

As an alternative to the methods mentioned above, we note that spatial cumulants are promising because they offer a parametric description of the high-order spatial statistics (Dimitrakopoulos et al., 2010). However, to this day cumulants have not been used in the context of training image selection.

This paper addresses the development of a geostatistical tool that provides two measurable criteria for selecting training images based on their consistency with given data. The first method, applicable in cases where more than one conceptual model is available, allows ranking training images according to their compatibility with a data set in terms of low and high order spatial structures. This represents a *relative* compatibility measurement. With this relative measurement, even a somewhat incompatible image can potentially be top ranked if all other available images are even less compatible. To overcome this limitation, an *absolute* compatibility measurement method is developed, which computes the probability of finding patterns in a given training image. These metrics are able to compare conditioning data with pattern statistics of training images across a range of statistical orders. It is important to remark that if modeling takes place in a deformed space, the data to be used should be considered after deformation. Three examples are addressed: the first one is a simple synthetic 2D example, in the second one our method is used to identify non-stationarity in a large synthetic 3D alluvial model, and the third one demonstrates the applicability on a real 3D mining example.

2. Methodology

The purpose of the proposed algorithm is to generate a ranking of several training images, according to their compatibility with conditioning data. In essence, the algorithm is given a conditioning data set and a series of training images. The method works by defining conditioning data events, which are patterns of spatially distributed data values, and computing their frequency of occurrence in the different training images. The training images that have a higher frequency of data events are deemed more coherent with the data. The result is a ranking of the training images according to their data consistency. The overall algorithm can be divided in a number of steps whose implementation is described below in details. We first present an algorithm that is statistically straightforward but computationally inefficient. In a second step, we present an equivalent alternative that yields similar results with a much lesser computational burden.

2.1. Conditioning data event definition

The first step of the method is to migrate the scattered data to a regular grid. The grid structure permits to accelerate future spatial searches of data events, therefore reducing computation times. To avoid scale issues, the nodes spacing in the user defined

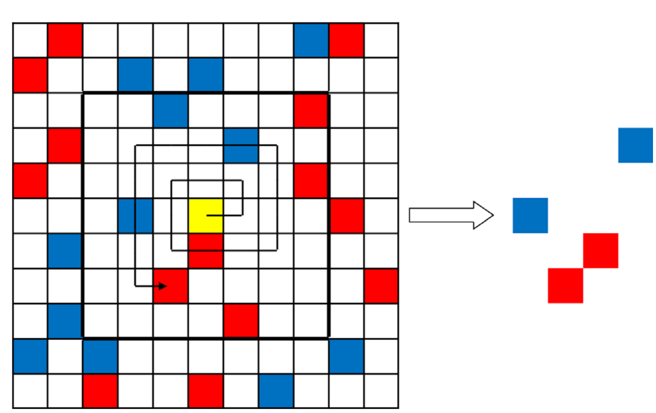

Fig. 1. Spiral search (left) and defined conditioning event (right).