



ELSEVIER

Contents lists available at ScienceDirect

## Computers &amp; Geosciences

journal homepage: [www.elsevier.com/locate/cageo](http://www.elsevier.com/locate/cageo)

## Case study

## spMC: an R-package for 3D lithological reconstructions based on spatial Markov chains

Luca Sartore<sup>a,b,\*</sup>, Paolo Fabbri<sup>c</sup>, Carlo Gaetan<sup>b</sup><sup>a</sup> National Institute of Statistical Science, 19 T.W. Alexander Drive, P.O. Box 14006, Research Triangle Park, NC 27709-4006, USA<sup>b</sup> Dipartimento di Scienze Ambientali, Informatica e Statistica, Università "Ca' Foscari" di Venezia, Campus Scientifico, via Torino 155, I-30172 Mestre-Venezia, Italy<sup>c</sup> Dipartimento di Geoscienze, Università di Padova, via Gradenigo 6, 35131 Padova, Italy

## ARTICLE INFO

## Article history:

Received 8 October 2015

Received in revised form

17 March 2016

Accepted 2 June 2016

Available online 7 June 2016

## Keywords:

Categorical data

Transition probabilities

Transiogram modeling

Indicator Cokriging

Bayesian entropy

3D lithological conditional simulation/

prediction

## ABSTRACT

The paper presents the spatial Markov Chains (spMC) R-package and a case study of subsoil simulation/prediction located in a plain site of Northeastern Italy. spMC is a quite complete collection of advanced methods for data inspection, besides spMC implements Markov Chain models to estimate experimental transition probabilities of categorical lithological data. Furthermore, simulation methods based on most known prediction methods (as indicator Kriging and CoKriging) were implemented in spMC package. Moreover, other more advanced methods are available for simulations, e.g. path methods and Bayesian procedures, that exploit the maximum entropy. Since the spMC package was developed for intensive geostatistical computations, part of the code is implemented for parallel computations via the OpenMP constructs. A final analysis of this computational efficiency compares the simulation/prediction algorithms by using different numbers of CPU cores, and considering the example data set of the case study included in the package.

© 2016 Elsevier Ltd. All rights reserved.

## 1. Introduction

The paper aims to introduce the spMC package (Sartore, 2013) which is an extension package for the R software (R Core Team, 2016). Its main purpose is to provide recent tools for the analysis, simulation and prediction of lithological data under the methodological framework of the spatial Markov chains. The first software implementation of lithological simulation and prediction for spatial Markov chains, stemming from the seminal work of Carle and Fogg (1996, 1997); Carle et al. (1998); Weissmann et al. (1999), and Weissmann and Fogg (1999), was the geostatistical software T-PROGS (Carle, 1999). This software is a well-established stochastic modeling tool for 3-D applications and also embedded in some commercial groundwater modeling software (e.g. GMS, (Aquaveo, 2015)). In T-PROGS transition probabilities are estimated for describing the stratigraphical characteristics of the geological data. Then simulations are performed through CoKriging and simulated annealing methods. The spMC package in its present version is a complete collection of advanced methods for data inspection, statistical estimation of parameter models, and lithological simulation and prediction. It includes common tools for predicting and simulating lithofacies at pixel level which are

typically used like sequential indicator simulation (SISIM, (Deutsch and Journel, 1998)) as well as the more recent advances (Li, 2007; Allard et al., 2011). We think there are three features of spMC that can be of value in the geostatistical community. First, it is an extension package of an increasingly used software like R. Second, a particular strength of the package is the exploitation of high performance computational (HPC) techniques, such as parallel computing, by allowing to deal better with a large number of categories. Finally, we can find the implementation of the more recent advances in simulation of lithological data. In the next section we briefly recall the methodological framework. In Section 3 we illustrate the main features of spMC by examining a case study (Section 4). Concluding remarks are addressed in Section 5.

## 2. Background on spatial Markov chain in geostatistics

The spMC package provides several functions to deal with categorical spatial data and continuous lag Markov chain, where the lag is the difference between two spatial positions. Traditionally, a Markov chain is described by a probabilistic temporal model for one-dimensional discrete lags, i.e. the model quantifies the probability to observe any specific state in the future given the knowledge of the current state. The extension of this concept arises by the definition of a Markov process involving continuous

\* Corresponding author.

multidimensional lags in a  $d$  dimensional space.

We consider the stationary transition probability between two states (or categories),  $i$  and  $j$ , in two locations,  $\mathbf{s}$  and  $\mathbf{s} + \mathbf{h}$ , namely  $t_{ij}(\mathbf{h}) = \Pr(Z(\mathbf{s} + \mathbf{h}) = j | Z(\mathbf{s}) = i)$ ,  $\forall i, j = 1, \dots, K$ ,

where  $K$  is the total number of states that the random variable  $Z$  can assume as outcome and  $\mathbf{h}$  is a multidimensional lag of dimension. In continuous-lag formulation of a Markov chain model (Carle and Fogg, 1997) the transition probability  $t_{ij}(\mathbf{h})$  is the element in the  $i$ -th row and in the  $j$ -th column of the matrix  $\mathbf{T}(\mathbf{h})$  such that

$$\mathbf{T}(\mathbf{h}) = \exp(\|\mathbf{h}\| \mathbf{R}_{\mathbf{h}}). \tag{1}$$

The transition rate matrix  $\mathbf{R}_{\mathbf{h}}$  depends on the direction given by the lag  $\mathbf{h}$ .

Carle and Fogg (1997) introduced an approximation of the rate matrix  $\mathbf{R}_{\mathbf{h}}$  by the ellipsoidal interpolation which makes the rate matrix for the direction of  $\mathbf{h}$  dependent on the rate matrices  $\mathbf{R}_{\mathbf{e}_k}$  estimated for the main axial directions. The vector  $\mathbf{e}_k$  indicates the standard basis vector of dimension  $d$ , whose  $k$ -th component is one and the others are zero. In particular, the matrix  $\mathbf{R}_{\mathbf{e}_k}$  can be computed as

$$\mathbf{R}_{\mathbf{e}_k} = \text{diag}(\ell_{\mathbf{e}_k})^{-1} [\mathbf{F}_{\mathbf{e}_k} - \mathbf{I}],$$

or for the reversibility of the chain as

$$\mathbf{R}_{-\mathbf{e}_k} = \text{diag}(\mathbf{p}) \mathbf{R}_{\mathbf{e}_k}^T \text{diag}(\mathbf{p})^{-1},$$

where  $\ell_{\mathbf{e}_k}$  is the mean vector of the stratum thicknesses/lengths along the direction  $\mathbf{e}_k$ , the matrix  $\mathbf{F}_{\mathbf{e}_k}$  denotes the transition probabilities for consecutive blocks made of adjacent points with the same category,  $\mathbf{I}$  is the identity matrix, and  $\mathbf{p}$  is the vector of relative frequencies corresponding to the estimate of the stationary distribution.

The rate  $r_{ij,\mathbf{h}}$  in the  $i$ -th row and  $j$ -th column of the matrix  $\mathbf{R}_{\mathbf{h}}$  is then calculated as

$$|r_{ij,\mathbf{h}}| = \sqrt{\sum_{k=1}^d \left( \frac{h_k}{\|\mathbf{h}\|} r_{ij,\mathbf{e}_k} \right)^2}, \tag{2}$$

where  $r_{ij,\mathbf{h}}$  is non-positive when  $i=j$ , otherwise it is non-negative;  $d$  represents the dimension of the lag  $\mathbf{h}$  (and hence the number of coordinates of  $\mathbf{s}$ ), and  $r_{ij,\mathbf{e}_k}$  denotes the components in the  $i$ -th row and  $j$ -th column of the matrix  $\mathbf{R}_{\mathbf{e}_k}$ .

From a statistical viewpoint, two problems arise. The former is related to how to estimate the components  $r_{ij,\mathbf{h}}$ , while the latter is associated to the formulation of the conditional probability used for simulations and predictions.

sPMC provides a variety of estimation methods. We implemented the mean length method and the maximum entropy method suggested in Carle and Fogg (1997) and Carle (1999). These methods are both based on the mean lengths  $\bar{L}_{i,\mathbf{e}_k}$  and the transition probabilities of embedded occurrences  $f_{ij,\mathbf{e}_k}^*$ , which are the components of the matrix  $\mathbf{F}_{\mathbf{e}_k}$ . The autotransition rates are derived by  $r_{ii,\mathbf{e}_k} = -1/\bar{L}_{i,\mathbf{e}_k}$ , while the other rates are calculated as  $r_{ij,\mathbf{e}_k} = f_{ij,\mathbf{e}_k}^*/\bar{L}_{i,\mathbf{e}_k}$ , i.e. for any  $i \neq j$ . The mean lengths are usually computed by means of the average of the observed stratum thicknesses/lengths, while the transition probabilities of embedded occurrences are estimated as the average of the relative transition frequencies, or through an iterative procedure based on the entropy (Goodman, 1968).

A maximum likelihood method is implemented in which we consider the stratum thicknesses/lengths distributed as log-normal random variables (Ritzi, 2000). There also exist robust alternatives for estimating the mean lengths which are based on the trimmed median and the trimmed average.

Finally, we have considered a least squares approach in which we minimize the sum of the squared discrepancies between the empirical transition probabilities and theoretical probabilities given by the model (1). Such minimization is performed under the constraints (Carle and Fogg, 1997):

$$\sum_{j=1}^K r_{ij,\mathbf{h}} = 0, \quad \forall i = 1, \dots, K \quad \text{and} \\ \sum_{i=1}^K p_i r_{ij,\mathbf{h}} = 0, \quad \forall j = 1, \dots, K,$$

where  $p_i$  denotes the  $i$ -th component of the vector  $\mathbf{p}$ .

In order to perform lithological simulations and predictions, an approximation of the following conditional probability must be considered:

$$\Pr\left(Z(\mathbf{s}_0) = j \mid \bigcap_{l=1}^n Z(\mathbf{s}_l) = z(\mathbf{s}_l)\right), \quad \forall j = 1, \dots, K, \tag{3}$$

where  $\mathbf{s}_0$  denotes a simulation or prediction location,  $\mathbf{s}_l$  represents the  $l$ -th spatial position which corresponds to the  $l$ -th observation, and  $z(\mathbf{s}_l)$  indicates the observed value of the random variable  $Z(\mathbf{s}_l)$ . The approximation proposed by Carle and Fogg (1996) is based on indicator Kriging and CoKriging methods, which are then adjusted by a quenching procedure based on the simulated annealing method. Other approximations are based on path methods (Li, 2007; Li and Zhang, 2007), while those that are based on the Bayesian entropy perspective (Christakos, 1990) were considered by Bogaert (2002) and modified by Allard et al. (2011).

The Kriging approximations are calculated through a linear combination of weights, i.e.

$$\Pr\left(Z(\mathbf{s}_0) = j \mid \bigcap_{l=1}^n Z(\mathbf{s}_l) = z(\mathbf{s}_l)\right) \approx \sum_{l=1}^n \sum_{i=1}^K w_{ij,l} c_{il},$$

where

$$c_{il} = \begin{cases} 1 & \text{if } z(\mathbf{s}_l) = i, \\ 0 & \text{otherwise,} \end{cases}$$

and the weight  $w_{ij,l}$  is the component in the  $i$ -th row and  $j$ -th column of the matrix  $\mathbf{W}_j$ ; such weights are calculated by solving the following system of linear equations:

$$\begin{bmatrix} \mathbf{T}(\mathbf{s}_1 - \mathbf{s}_1) & \dots & \mathbf{T}(\mathbf{s}_n - \mathbf{s}_1) \\ \vdots & \ddots & \vdots \\ \mathbf{T}(\mathbf{s}_1 - \mathbf{s}_n) & \dots & \mathbf{T}(\mathbf{s}_n - \mathbf{s}_n) \end{bmatrix} \begin{bmatrix} \mathbf{W}_1 \\ \vdots \\ \mathbf{W}_n \end{bmatrix} = \begin{bmatrix} \mathbf{T}(\mathbf{s}_0 - \mathbf{s}_1) \\ \vdots \\ \mathbf{T}(\mathbf{s}_0 - \mathbf{s}_n) \end{bmatrix}.$$

This system of equations, which can also lead to the CoKriging equations, is singular. However, it can be solved through the constraints proposed by Carle and Fogg (1996).

In order to obviate axiomatic problems arising from the Kriging approximation, the path methods (Li, 2007; Li and Zhang, 2007) considered the following approximation under the assumption of conditional independence:

$$\Pr\left(Z(\mathbf{s}_0) = z_i \mid \bigcap_{l=1}^n Z(\mathbf{s}_l) = z(\mathbf{s}_l)\right) \approx \Pr\left(Z(\mathbf{s}_0) = z_i \mid \bigcap_{l=1}^m Z(\mathbf{s}_l) = z_{k_l}\right) \\ \propto t_{k_i}(\mathbf{s}_0 - \mathbf{s}_1) \prod_{l=2}^m t_{i k_l}(\mathbf{s}_0 - \mathbf{s}_l).$$

These methods are characterized by following a fixed or random path of unknown points, which are predicted or simulated by conditioning on the of the previous prediction point.

Other approximations were proposed in order to improve the Kriging deficiencies. In particular, Bogaert (2002) introduced a Bayesian procedure exploiting the maximum entropy, which was

Download English Version:

<https://daneshyari.com/en/article/506993>

Download Persian Version:

<https://daneshyari.com/article/506993>

[Daneshyari.com](https://daneshyari.com)