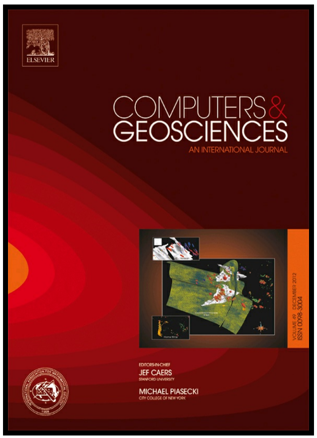
spMC: an R-package for 3D lithological reconstructions based on spatial Markov chains

Luca Sartore, Paolo Fabbri, Carlo Gaetan



www.elsevier.com/locate/cageo

PII: S0098-3004(16)30147-9
DOI: http://dx.doi.org/10.1016/j.cageo.2016.06.001
Reference: CAGEO3771

To appear in: *Computers and Geosciences*

Received date: 8 October 2015
Revised date: 17 March 2016
Accepted date: 2 June 2016

Cite this article as: Luca Sartore, Paolo Fabbri and Carlo Gaetan, spMC: an R package for 3D lithological reconstructions based on spatial Markov chains *Computers and Geosciences,* http://dx.doi.org/10.1016/j.cageo.2016.06.001

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

Luca Sartore[a,b], Paolo Fabbri[c], Carlo Gaetan[b]

[a]*National Institute of Statistical Science, 19 T.W. Alexander Drive, P.O. Box 14006, Research Triangle Park, NC 27709-4006, U.S.A.*
[b]*Dipartimento di Scienze Ambientali, Informatica e Statistica, Università "Ca' Foscari" di Venezia, Campus Scientifico, Via Torino 155, I-30172 Mestre-Venezia, Italy*
[c]*Dipartimento di Geoscienze, Università di Padova, via Gradenigo 6, 35131 Padova, Italy*

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

Email addresses: lsartore@niss.org (Luca Sartore), paolo.fabbri@unipd.it (Paolo Fabbri), gaetan@unive.it (Carlo Gaetan)

considering the example data set of the case study included in the package.

Keywords: Categorical data, Transition probabilities, Transiogram modeling, Indicator CoKriging, Bayesian entropy, 3D lithological conditional simulation/prediction

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 modelling tool for 3-D applications and also embedded in some commercial groundwater modelling 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 litholological 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 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

$$t_{ij}(\mathbf{h}) = \Pr(Z(\mathbf{s}+\mathbf{h}) = j | Z(\mathbf{s}) = i), \ \forall i, j = 1, \ldots, K,$$

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

$$\mathbf{R}_{\mathbf{e}_k} = \operatorname{diag}(\boldsymbol{\ell}_{\mathbf{e}_k})^{-1} \left[\mathbf{F}_{\mathbf{e}_k} - \mathbf{I}\right],$$

or for the reversibility of the chain as

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

where $\boldsymbol{\ell}_{\mathbf{e}_k}$ 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 $\overline{L}_{i,\mathbf{e}_k}$ 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 $r_{ii,\mathbf{e}_k} = -1/\overline{L}_{i,\mathbf{e}_k}$, while the other rates are calculated as $r_{ij,\mathbf{e}_k} = f^*_{ij,\mathbf{e}_k}/\overline{L}_{i,\mathbf{e}_k}$, 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):

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

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:

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

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.

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

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 $\mathbf{W}_l$; such weights are calculated by solving the following system of linear equations:

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

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:

$$\Pr\left(Z(\mathbf{s}_0) = z_i \,\middle|\, \bigcap_{l=1}^{n} Z(\mathbf{s}_l) = z(\mathbf{s}_l) \right) \approx \Pr\left(Z(\mathbf{s}_0) = z_i \,\middle|\, \bigcap_{l=1}^{m} Z(\mathbf{s}_l) = z_{k_l} \right) \propto$$

$$\propto t_{k_1 i}(\mathbf{s}_0 - \mathbf{s}_1) \prod_{l=2}^{m} t_{ik_l}(\mathbf{s}_0 - \mathbf{s}_l).$$

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 successively considered by Allard et al. (2011) to justify the usage of the following approximation:

$$\Pr\left(Z(\mathbf{s}_0) = z_i \left| \bigcap_{l=1}^{n} Z(\mathbf{s}_l) = z(\mathbf{s}_l) \right.\right) \approx \frac{p_i \prod_{l=1}^{n} t_{ik_l}(\mathbf{s}_0 - \mathbf{s}_l)}{\sum_{i=1}^{K} p_i \prod_{l=1}^{n} t_{ik_l}(\mathbf{s}_0 - \mathbf{s}_l)}.$$

3. spMC features

The spMC package is basically a collection of functions not implemented in other software, which can be grouped according to their purposes as summarized in Table 1. Since the package was designed for intensive geostatistical computations, part of the code deals with parallel computing via the OpenMP constructs (OpenMP Architecture Review Board, 2008). For example, the `setCores()` function permits the user to choose the number of CPU cores that will be used by the other functions of the spMC package.

Some of the functions implement descriptive geostatistical tools, which are useful for a better understanding of the process and essential for the parameter estimation of the model.

Graphical tools were developed to help the user to choose the model. These tools are often used for initial evaluations on the input data. From a visual inspection of these graphics, it is possible to analyze the distribution of the stratum thicknesses/lengths along a given direction.

Once the transition rates have been estimated with the chosen model fitting algorithm, it is possible to calculate the theoretical transition probabilities for a set of multidimensional lags. This transition probabilities are

Table 1: Most important user functions in the spMC package.

Tasks and functions	Techniques implemented in the spMC package
Descriptive geostatistical tools	
`which_lines`	Points classification through directional lines
`getlen`	Estimation of stratum lengths for embedded chains
`density.lengths`	Empirical densities of stratum lengths
`mlen`	Mean length estimation for embedded chains
Estimations of continuous lag models	
`transiogram`	Empirical transition probabilities estimation
`pemt`	Multi-directional transiograms estimation
`embed_MC`	Transition probabilities estimation for embedded chains
`tpfit`	One-dimensional model parameters estimation
`multi_tpfit`	Multidimensional model parameters estimation
Categorical spatial random field simulation and prediction	
`sim`	Random field simulations and predictions
`quench`	Quenching algorithm for simulation adjustments
Graphical tools	
`plot.transiogram`	Plot one-dimensional transiograms
`mixplot`	Plot multiple one-dimensional transiograms
`contour.pemt`	Display contours with multi-directional transiograms
`image.pemt`	Images with multi-directional transiograms
`image.multi_tpfit`	Images with multidimensional transiograms
`boxplot.lengths`	Boxplot of stratum lengths
`hist.lengths`	Histograms of stratum lengths
High performance computational tools	
`setCores`	Set the number of CPU cores for HPC

used in spMC package for simulation of the lithological categories, while predictions are by-products of the function `sim()`.

3.1. Descriptive tools

Most of the descriptive tools of the spMC package are based on graphical analyses, with a subset adopted for inferential purposes. In fact, the study of stratum thicknesses/lengths is relevant for guiding the decision of which computational method to adopt for estimating the mean lengths. The analysis of the empirical distribution of stratum lengths is mainly based on the evaluation of quartiles and extreme values through the basic technique of the boxplot diagrams, which is implemented in the function `boxplot.lengths()`. Another technique is available for the empirical estimation of the stratum lengths distribution, which is performed by the function `density.lengths()`, and it is based on the kernel-smoothing approach.

Further descriptive tools are the analyses of empirical, multi-directional and theoretical transiograms. However, the descriptive analysis of the transiograms can be performed only after an accurate inferential analysis. For example, the function `mixplot()` is used to check for probabilistic anisotropies by comparing one-dimensional empirical transiograms along several direction. Similar analyses can be performed also for multidimensional models, e.g. when the function `contour.pemt()` is applied to an object resulting from the function `pemt()`.

3.2. Inferential tools

The implementation of the one-dimensional experimental transiogram computation is based on two subsequent steps. In primis, a selection of points which belong to specific directional-lines is common to all transiogram estimation methods. This technique is implemented in the function `which_lines()`, which classifies observation coordinates along a chosen direction. After this, the estimation of the empirical transiogram is performed by counting the transitions among categories along the classified lines. The absolute transition frequencies are then normalized to obtain the transition probabilities as relative frequencies. Both directional classification and transition probability estimation are performed by the usage of the function `transiogram()`, which also computes the standard errors by assuming the asymptotic normality of the estimates. These standard errors are then used by the function `plot.transiogram()` to produce confidence intervals by the inversion of the Wald type interval for the log odds (Stone, 1996; Brown et al., 2001).

One-dimensional theoretical transiograms are computed differently, because they require the estimation of the model parameters for computing the transition probabilities. In practice, the function `tpfit()` allows the selection from three different rate estimation techniques through a specific argument:

- the mean lengths method (`method = "ml"`), which is based on the estimation of mean lengths and the transition probabilities of the embedded Markov chains by the functions `mlen()` and `embed_MC()` respectively. The resulting quantities are used to estimate the parameters;
- the maximum entropy algorithm (`method = "me"`), which is iterative and requires few iterations to converge;
- the iterated least squares technique (`method = "ils"`), which was de-

veloped for reducing the discrepancies between the experimental transiogram and the theoretical model by relaxing the mathematical constraints on the parameters.

Multidimensional transiogram estimation can be viewed as an extestion of the one-dimensional methods. The function `multi_tpfit()` allows for the parameter estimation along multiple orthogonal axes. These parameters will be ellipsoidally interpolated for the calculation of transition rates along non-orthogonal directions. As for the one-dimensional models, the three estimation techniques previously exposed are chosen by a specific argument of the functions `multi_tpfit()`.

Multi-directional transiograms are computed either with ellipsoidal interpolation or without. The function `pemt()` allows for the computation of theoretical transition probabilities for any chosen direction without ellipsoidal interpolation.

3.3. Simulations and predictions tools

Three different techniques were considered to approximate the conditional probability in (3). The function `sim()` allows the selection of the method for simulation, in particular:

- the Kriging methods are implemented for the indicator Kriging and indicator CoKriging. The Kriging approach is usually adopted for prediction, but it is used in the spMC package mainly for sequential simulations. In addition, it is possible to adjust the simulations by performing the quenching algorithm implemented in the function `quench()`;

- a fixed and random path algorithms are available, and they can be selected by logical argument `fixed`. By default a random path algorithm is performed, because its results are more consistent with reality;

- the maximum entropy approach, which was proposed by Allard et al. (2011) for avoiding the entropy optimization. It performs an aggregation of transition probabilities to approximate the optimal solution. This particular setting reduces the computations with respect to the Bogaert's proposal (2002).

Furthermore, these three methods produce also predictions by combining the transition probabilities calculated through the theoretical model in (1), where the transition rates in the matrix $\mathbf{R_h}$ are calculated as in (2). In doing so, a considerable computational efficiency is achieved for computing an approximation of the distribution at each point in the simulation grid.

4. Case study

The package includes the 3D data-set ACM, related to a sediment deposit of about 300 m in longitude (X direction), 500 m in latitude (Y direction) and 400 m in depth (Z direction), located in Scorzé area (Venetian plain, NE Italy) (Figure 1), consisting of a collection of eleven simplified lithostratigrafical borehole data. The lithologies of these boreholes were simplified in three different cases. In the first categorical data set (MAT5) the local lithology was simplified in five lithologies (Clay, Sand, Mix of Sand and Clay, Gravel, Mix Sand and Gravel), in the second one (MAT3) in three lithologies (Clay, Sand, Gravel) and finally the third one the lithostratigraphy was simplified

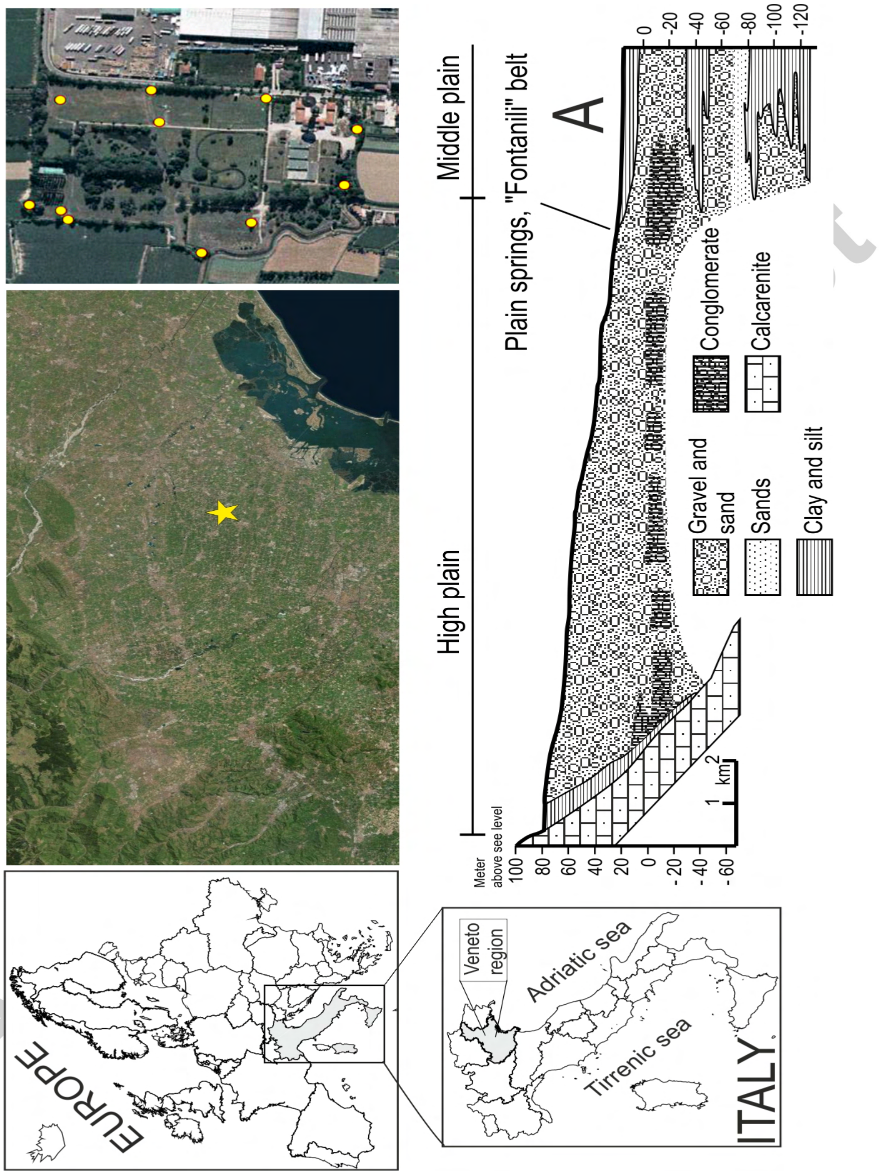


Figure 1: Geographical location of the borehole data.

in only two permeability categories (TRUE, FALSE). Geologically Venetian plain can be roughly divided in “high”, “low” plain. The high plain is essentially of fluvial origin, but also glacial and fluvioglacial origin near the pre-Alps. This area is principally composed of gravel, particularly the sediments are made by very permeable gravel and pebbly materials. Transition between the high and low plain, of about 2-5 kilometers wide, is represented by the “fontanili” belt. In this zone the gravels decrease in thickness splitting them into sub-horizontal gravelly layers separated by silty and/or clayey beds, sometimes interbedded with clay layers. The low plain starts where the gravel layers move to sand until the Adriatic coast. Low plain presents a subsoil composed essentially by silt and clay layers interposed with sandy layers. In this part the gravels are absent, with some exceptions found, at considerable depths (e.g. up to 300 meters in depth)(Carraro et al., 2013; Fabbri et al., 2011). In the high plain an undifferentiated aquifer is present, where water table is at maximum depth, this aquifer Southeastern becomes a multi-layered confined or semi-confined aquifer system directly connected with the unconfined. The water table outcrops in the most depressed zones originating the typical plain springs called “fontanili”, where the water table, being very shallow, intersects the topographic surface (Vorlicek et al., 2004; Fabbri and Piccinini, 2013). This discharge band of the unconfined aquifer can be from 2 to 10 kilometers wide, draining the unconfined aquifer and representing the source of some important Venetian river. Hydrogeologically ACM data set concerns the area southern of the “fontanili” belt in area of essentially gravelly multi-layered confined or semi-confined aquifer system.

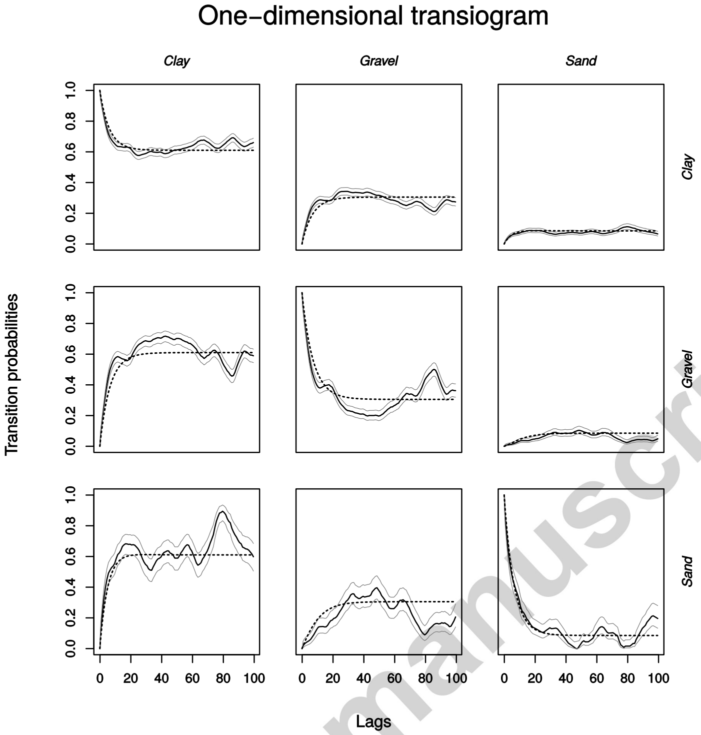


Figure 2: Empirical (full black line) and theoretical (dashed line) transiogram along Z direction. They are calculated with the MAT3 variable. The light-grey lines correspond to the upper and lower confidence bounds for 99% coverage probability.

4.1. One-dimensional lags model

The empirical transiogram exposed in Figure 2 is computed with 100 lags of 1 meter by considering all couples of points along Z direction within a maximum distance of 100 meters. The light-grey lines corresponds to the upper and lower confidence bounds calculated with 99% coverage probability. From a graphical inspection of the transiogram, it is possible to establish if the process is stationary. In fact, the empirical transition probabilities should approximately converge to the relative frequency of the observed materials as the lag-length tends to infinity (see theoretical transiogram by looking at each column in Figure 2). For this reason, the transition probabilities (by columns) corresponding to the farthest distances are respectively close to 0.62, 0.30 and 0.08 for Clay, Gravel and Sand.

By comparing two or more transiograms drawn for different directions, one can check if there is directional dependence on the data (especially if these are located on a regular sample grid). The process is anisotropic if the transition probabilities are dependent on the directions. In most cases, this aspect is more obvious when the distances between points along different directions are measured at different scales. For example, the distance between points along Z direction can be measured in meters, while it is expressed in kilometers along X and/or Y direction. However, a more quantitative method for inspecting this issue makes use of multidirectional transiograms and is useful when relatively abundant data are available in all three dimensions.

Multidirectional transiograms are based on theoretical transition probabilities calculated from the estimates of transition rates per multiple chosen directions. This method exploits the implementation of the `tpfit_ml()` function, which is computationally faster than the `tpfit_me()` function. Once the transition probabilities are calculated for specific lags, they can be organized and represented on few graphics as in the left column of Figure 3.

4.2. Multidimensional lags model

Multidimensional models are required to calculate transition probabilities in multidimensional spaces. In fact, even if it is possible to estimate for any direction the transition rates, and hence the corresponding probabilities, it is not computationally feasible to deal with one-dimensional models along

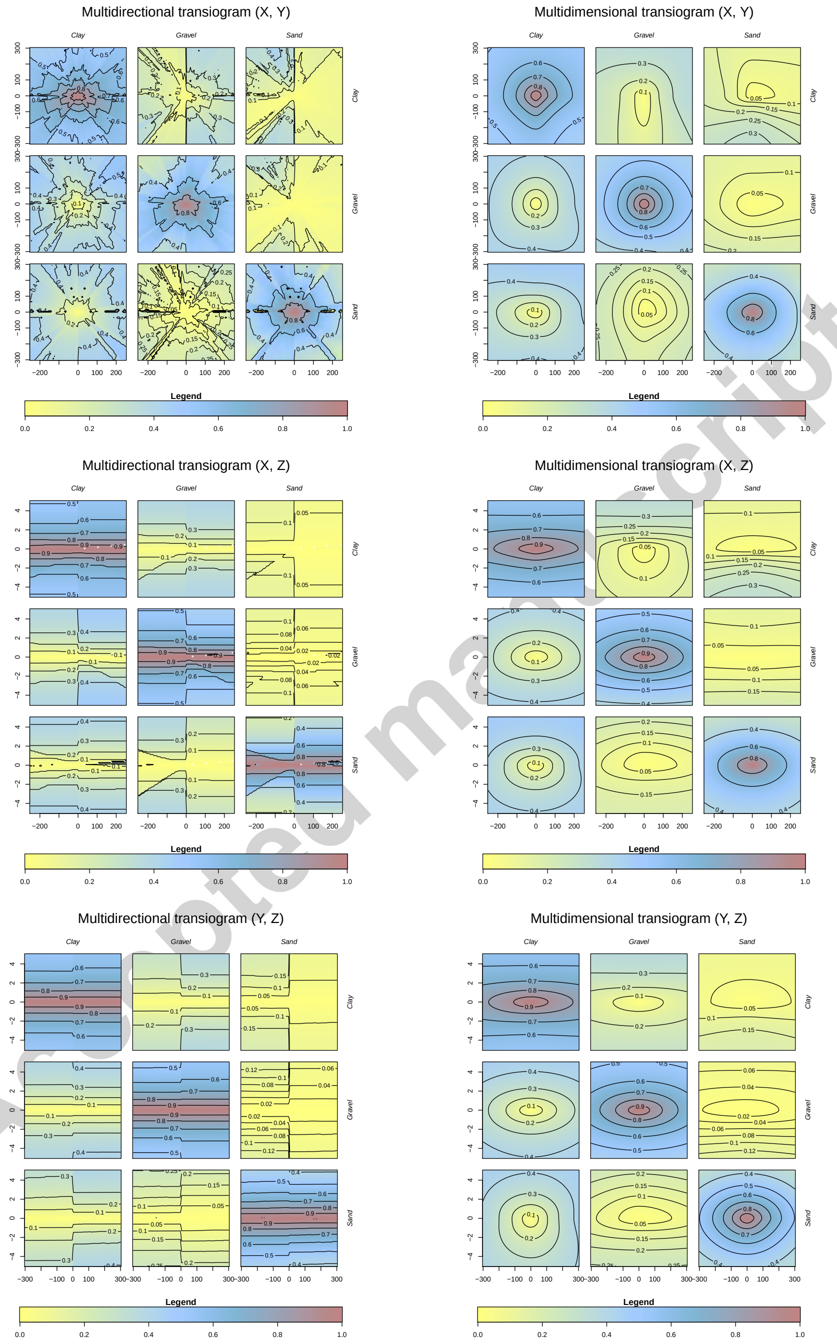


Figure 3: Multidirectional transiograms, and multidimensional transiograms derived from the interpolation of the theoretical model in (1).

multiple directions. Multidimensional model interpolate the transition rates along the main axis to obtain a suitable approximation. In so doing, the resulting transition probabilities are more regular, as shown in Figure 3. Since the evaluation of these probabilities is computationally more efficient, it is preferable to adopt theoretical probabilities calculated with interpolated rates, especially when the number of points in the simulation grid is large.

The transition probabilities shown in the right column of graphics in Figure 3 share some common patterns with those exposed on the left column. This tool is used to study the probabilistic anisotropy along several directions, the juxtaposition of categories, and the variations of the transition probabilities with respect to both the direction and the distance from the center of each representation.

4.3. Spatial simulations and predictions

From a geological viewpoint, spatial simulations and predictions are necessary tools for lithological reconstruction and mapping. However, these statistical techniques can be computationally intensive, and therefore, exploitation of HPC techniques can be advantageous.

The main computational issues in classical geostatistics are related to the inversion of a variance-covariance matrix to obtain Kriging predictions for a large number of points in the simulation grid. In this context, both indicator Kriging and CoKriging must solve a system of simultaneous equations where the only few k-nearest neighbors are used instead of the whole observations. Similarly, the method proposed by Allard et al. (2011) can also use a reduced conditional probability for better computational achievements (even when parallel computing is not performed). In the following, a value of $k = 12$ was considered, which is the default value of the function `sim()`. The choice of k is subjective, because, at the best of our knowledge, no selection methods for k have been developed for lithological data yet.

To show the computational advantages of the implemented algorithms, a regular simulation grid is constructed within the sample space. It consists in 21 × 21 × 21 simulation points, which cover a volume of 293m × 477m × 400m. Spatial simulations and predictions were performed with a 16-core AMD64 CPU at 2.4 GHz. Simulations were repeated by using 1, 8 and 16-cores. In particular, Kriging algorithms were executed by considering 32-nearest neighbors and path algorithms with a search radius of 200 meters. The efficient maximum entropy method was performed by considering the transition probabilities among all points (as in the original formulation) and also with 32-nearest neighbors.

Table 2: Execution time in seconds.

	IK	ICK	FP	RP	MCS	MCSKNN
Serial (1 core)	7.301	7.963	12.554	13.216	97.882	3.886
Parallel (8 cores)	2.738	3.352	12.553	13.212	21.307	1.408
Parallel (16 cores)	2.445	3.233	12.557	13.216	16.948	0.967

From Table 2, which reports the elapsed execution time for each algorithm, one can perceive a drastic time reduction with respect to sequential computing. Indicator Kriging (IK) and CoKriging (ICK) are similar, even if indicator Kriging performs faster because it processes less information than CoKriging. Path algorithms are sequential. They are not affected by the use of multiple processors. However, the fixed path algorithm (FP) perform faster then the random path algorithm (RP), because the sequence of points to predict is already known and it does not require extra calculations. The efficient maximum entropy categorical simulations (MCS) are the slowest, while they become the fastest when the conditional probability is calculated with the k-nearest neighbors (MCSKNN).

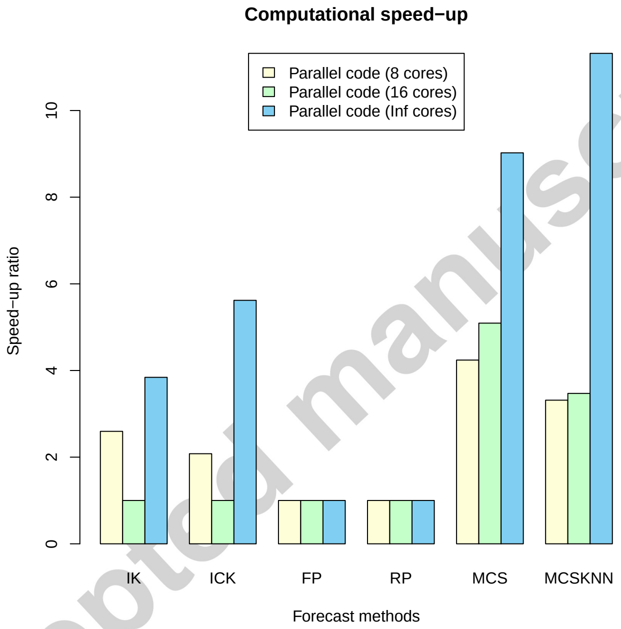


Figure 4: Computational efficiency of the simulation and prediction methods.

After looking at the Figure 4, it is possible to establish which algorithm has a strong impact on high performance computing and scalability (Kumar and Gupta, 1994). In fact, the computational efficiency (measured as speed-up time) is calculated through the ratio of the execution time between serial code and parallel. The maximum speed-up with infinity cores can be approximately computed as the product of the sequential execution time and the fraction of code which is not parallelizable. Figure 4 shows that Kriging methods improve their efficiency when HPC techniques are used. However, the most substantial improvements obtained by parallel computations are shown for the efficient maximum entropy methods.

5. Conclusions

In comparison with other software used for predicting and simulating lithological categories, spMC is based on a theoretical framework which focuses on transition probabilities rather than covariances/variograms or multipoint geostatistics. The spMC package is able to produce results more efficiently by high performance computational techniques, and it can be used on several platforms (Linux, Windows and Mac). It is the unique open-source software which implements several estimation procedures of transition probabilities, and the more advanced simulation-prediction techniques based on maximum entropy by geostatistical transition probabilities. Currently, the Gslib library (Deutsch and Journel, 1998) and SGeMS (Remy et al., 2009) are the most known free-source softwares for lithological simulation/prediction based on variogram via Kriging/CoKriging. T-PROGS (Carle, 1999) is based on transition probabilities and Kriging/CoKriging, which is also available as a stand-alone or as an add-on in GMS groundwater model. Mainly, spMC supports parallel computing, and hence its results are produced more efficiently and several lithological categories can be more readily supported. The results of spMC package can be visualized into R through other packages, or exported from R and used in other software. For example, they can be exported in ASC format and imported in GIS software or can be used in groundwater modeling. They can be exported in CSV format and used to draw probabilistic maps in open-source software like ParaView (Squillacote, 2007) or for the visualization per each category of the occupancy volumes (see, for example, the probability map of Figure 5 for Sand category).

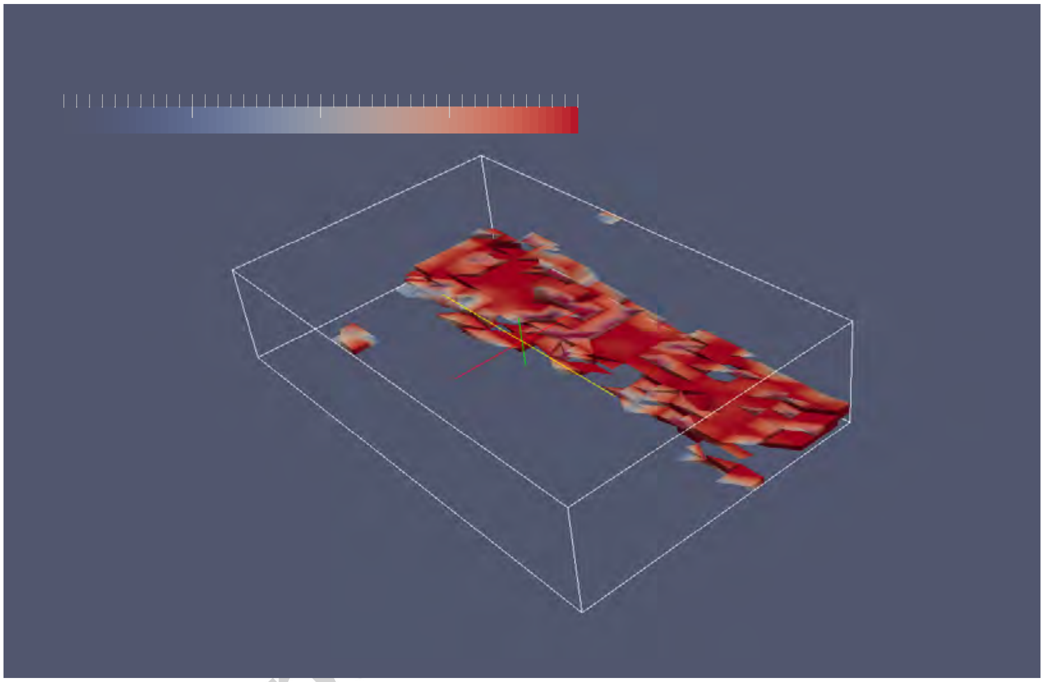


Figure 5: Random-path results, obtained for Sand category, as displayed by Paraview software.

The development of the spMC package will continue. In the future, we plan to include non-parametric estimates of transiograms by means of kernel methods (Allard et al., 2011) and other probabilistic aggregations (Allard et al., 2012). Additional validation functions will be also included to allow for the comparison of simulation/prediction probabilities and actual categorical variables.

Acknowledgements

Critical reading of the manuscript and comments by professor Graham E. Fogg and one anonymous reviewer are gratefully acknowledged. The authors would also like to thank Kelly Toppin for proof-reading the article. This research was supported by grant "GEO-RISKS – Geological, morphological and hydrological processes: monitoring, modeling and impact in the Northeastern Italy" funded by the University of Padua (Italy).

References

Allard, D., Communian, A., Renard, P., 2012. Probability aggregation methods in geoscience. Mathematical Geosciences 44, 545–581.

Allard, D., D'Or, D., Froidevaux, R., 2011. An efficient maximum entropy approach for categorical variable prediction. European Journal of Soil Science 62, 381–393.

Aquaveo, L.L.C., 2015. Groundwater Modeling System Version 10.1, build date, December 14, 2015. UT, USA.

Bogaert, P., 2002. Spatial prediction of categorical variables: the Bayesian maximum entropy approach. Stochastic Environmental Research and Risk Assessment 16, 425–448.

Brown, L.D., Cai, T.T., DasGupta, A., 2001. Interval estimation for a binomial proportion. Statistical Science 16, 101–133.

Carle, S.F., 1999. T-PROGS: Transition probability geostatistical software.

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

Carle, S.F., Fogg, G.E., 1997. Modeling spatial variability with one and multidimensional continuous-lag Markov chains. Mathematical Geology 29, 891–918.

Carle, S.F., Labolle, E.M., Weissmann, G.S., Van Brocklin, D., Fogg, G.E., 1998. Conditional simulation of hydrofacies architecture: a transition probability/Markov approach, in: Fraser, G.S., Davis, J.M. (Eds.), SEPM Hydrogeologic Models of Sedimentary Aquifers, Concepts in Hydrogeology and Environmental Geology No. 1. Tulsa, Oklahoma, pp. 147–170.

Carraro, A., Fabbri, P., Giaretta, A., Peruzzo, L., Tateo, F., Tellini, F., 2013. Arsenic anomalies in shallow Venetian plain (Northeast Italy) groundwater. Environmental Earth Sciences 70, 3067–3084.

Christakos, G., 1990. A Bayesian/maximum-entropy view to the spatial estimation problem. Mathematical Geology 22, 763–777.

Deutsch, C., Journel, A.G., 1998. GSLIB. Geostatistical Software Library and User's Guide. Oxford University Press.

Fabbri, P., Gaetan, C., Zangheri, P., 2011. Transfer function-noise modelling of an aquifer system in NE Italy. Hydrological Processes 25, 194–206.

Fabbri, P., Piccinini, L., 2013. Assessing transmissivity from specific xapacity in an alluvial aquifer in the middle Venetian plain (NE, Italy). Water Science & Technology 67, 2000–2008.

Goodman, L.A., 1968. The analysis of cross-classified data: independence, quasi-independence, and interaction in contingency table with and without missing entries. Journal of the American Statistical Association 63, 1091–1131.

Kumar, V.P., Gupta, A., 1994. Analyzing scalability of parallel algorithms and architectures. Journal of Parallel and Distributed Computing 22, 379–391.

Li, W., 2007. A fixed-path Markov chain algorithm for conditional simulation of discrete spatial variables. Mathematical Geology 39, 159–176.

Li, W., Zhang, C., 2007. A random-path Markov chain algorithm for simulating categorical soil variables from random point samples. Soil Science Society of America Journal 71, 656–668.

OpenMP Architecture Review Board, 2008. OpenMP Application Program Interface Version 3.0.

R Core Team, 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Remy, N., Boucher, A., Wu, J., 2009. Applied geostatistics with SGeMS. Cambridge University Press.

Ritzi, R.W., 2000. Behavior of indicator variograms and transition probabilities in relation to the variance in lengths of hydrofacies. Water resources research 36, 3375–3381.

Sartore, L., 2013. spMC: Modelling Spatial Random Fields with Continuous Lag Markov Chains. R Journal 5.

Squillacote, A., 2007. The ParaView Guide. Kitware.

Stone, C.J., 1996. A course in probability and statistics. Duxbury Press Belmont.

Vorlicek, P.A., Antonelli, R., Fabbri, P., Rausch, R., 2004. Quantitative hydrogeological studies of Treviso alluvial plain (north east of Italy). Quarterly Journal of Engineering Geology and Hydrogeology 37, 23–29.

Weissmann, G.S., Carle, S.F., Fogg, G.E., 1999. Three-dimensional hydrofacies modeling based on soil surveys and transition probability geostatistics. Water Resources Research 35, 1761–1770.

Weissmann, G.S., Fogg, G.E., 1999. Multi-scale alluvial fan heterogeneity modeled with transition probability geostatistics in a sequence stratigraphic framework. Journal of Hydrology 226, 48–65.