Comments on "Generalized Coupled Markov Chain Model for Characterizing Categorical Variables in Soil Mapping"

In the paper of Park et al. (2007), the authors proposed a generalized coupled Markov chain (GCMC) model for soil type simulation based on the coupled Markov chain (CMC) model of Elfeki and Dekking (2001). The CMC model of Elfeki and Dekking (2001) extended the CMC model of Elfeki (1996) and was among the several multi-dimensional Markov chain models for lithofacies modeling published in geosciences before 2004. Li attempted to use the CMC model of Elfeki (1996) to simulate soil types and soil layer profiles during his postdoctoral research (Li, 1999). Unfortunately he did not obtain satisfactory results. Two major deficiencies were found: One was the diagonal pattern inclination and the second was the underestimation of small classes in simulated realizations (essentially in estimated local conditional probability distributions). It was found that these deficiencies were not gone in the CMC model of Elfeki and Dekking (2001), as demonstrated by the simulation cases in Elfeki and Dekking (2001), although they might be unapparent under some situations. The first deficiency may not be apparent when layers are naturally inclined, or thin/long and horizontally straight, and the second deficiency may not be apparent when different types have similar proportions. Both problems reach their extreme situations (i.e., diagonal inclination and disappearance of minor classes, respectively) in unconditional simulations. But when conditioning data were very dense (relative to layer/parcel sizes), both problems are not apparent. In addition, how to estimate the transition probability parameters from sample data was also a problem to solve, though it was a common issue in Markov chain spatial modeling. Considering that categorical spatial variable simulation in soil science and other fields of geosciences was important and that practical methods were rare, Li later made large efforts and spent many years to seek solutions for solving these problems. He first adhered to the basic idea of the CMC theory/model, but later had to propose a new theory and framework, that is, the Markov chain random field (MCRF) theory/model and the Markov chain geostatistical framework (Li, 2007).

Reading the paper of Park et al. (2007), we found there are some issues to clarify. For those issues that were already clarified by recent publications (e.g., Li, 2007; Li and Zhang, 2008; Zhang and Li, 2008), such as the problems mentioned above, we would like not to explain them in detail here. But there are some new issues occurring in Park et al. (2007). We think it is our obligation to clarify them to other readers in soil science.

We found that the GCMC model proposed by Park et al. (2007) is not much rational. Although the CMC model of Elfeki and Dekking (2001) is not practical in most real world cases due to some deficiencies in model and algorithm design, at least the one-dimensional Markov chain model they suggested is correct and interesting. The one-dimensional Markov chain model presented in Elfeki and Dekking (2001) was correctly written in Park et al. (2007, pp. 911, equation (4)) as

$$\Pr(Z_i = S_j \mid Z_{i-1} = S_l, Z_N = S_q) = \frac{p_{jq}^{N-i} p_{lj}}{p_{lq}^{N-i+1}}, \qquad \text{(EQ1)}$$

where Z is a random variable with a subscript denoting its location, S_q refers to a state q in the state space S, and p_{lq} represents a transition probability from state l to state q with a superscript denoting the number of spatial steps. This equation can be illustrated in **Fig. 1**, with the top arrow on each transition probability represents the direction of transition probability.

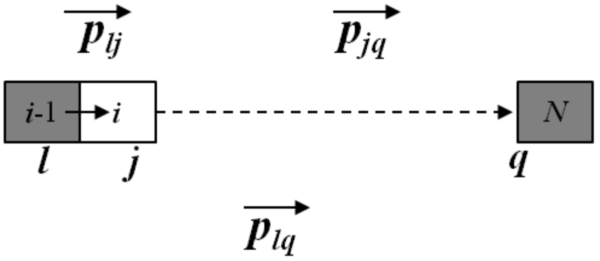
$\overrightarrow{p_{lj}}$
$\overrightarrow{p_{jq}}$
i-1
i
N
l
j
q
$\overrightarrow{p_{lq}}$

Fig. 1. Illustration of a one-dimensional Markov chain with conditioning to a future state

In **Fig. 1**, the states of solid cells *i*-1 and *N* are informed and the state of the empty cell *i* is uninformed. The Markov chain moves forward one step from state *l* at cell *i*-1 to state *j* at cell *i* with conditioning to a future state *q* at cell *N*. The state *j* at cell *i* is subject to estimation. All transition probabilities are forward. This is correct, because transition probabilities always have a head-tail direction. We think it should be generalized as

$$\Pr(Z_i = S_j \mid Z_M = S_p, Z_N = S_q) = \frac{p_{pj}^{|M-i|} p_{jq}^{|N-i|}}{p_{pq}^{|N-M|}}, \qquad \text{(EQ2)}$$

which can be illustrated as shown in Fig. 2, with the top arrow on each transition probability represents the direction of transition probability.

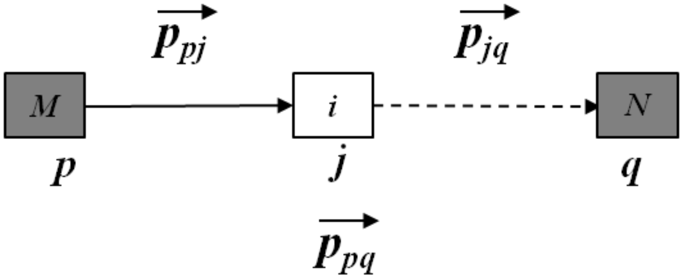
$\overrightarrow{p_{pj}}$
$\overrightarrow{p_{jq}}$
M
i
N
p
j
q
$\overrightarrow{p_{pq}}$

Fig. 2. Illustration of a generalized one-dimensional Markov chain with conditioning to a future state

In **Fig. 2**, the Markov chain moves multiple steps (or jump) from cell *M* to cell *i*, with conditioning to a future state at cell *N*. The state of cell *M* is *p* and that of cell *N* is *q*, both informed. The state *j* of cell *i* is uninformed, subject to estimation. All transition probability terms are forward. Such a generalization was done by Li and Zhang (2005) using a better way (i.e., using transiogram) so that grid point data could be used and the model was not limited by pixel size. The above EQ2 already takes directional asymmetry into account because Markov chains are directional and transition probabilities are asymmetric. In addition, transition probabilities also can be estimated uni-directionally.

However, what is surprising is that the authors did not generalize EQ1 to EQ2, but strangely added symbols "-" and "+" to their generalized one-dimensional Markov chain model (see equation (9) in Park et al., 2007, pp.912), making it become

$$\Pr(Z_i = S_j \mid Z_M = S_p, Z_N = S_q) = \frac{{}^{-}p_{pj}^{|M-i|}\,{}^{+}p_{jq}^{|N-i|}}{{}^{+}p_{pq}^{|N-M|}}. \qquad \text{(EQ3)}$$

They stated that "*Here we assume spatial asymmetry of the spatial indicator structures, and the negative and positive superscripts stand for the multistep transition probability in the negative or positive direction calculated from the TPMs for the negative or positive direction, respectively*" (Park et al., 2007, pp. 912).

The above EQ3 can be illustrated by a figure as shown in **Fig. 3**, with the top arrow on each transition probability represents the direction of transition probability.

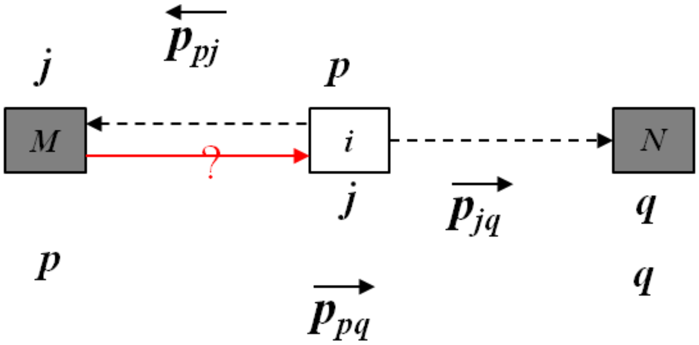

Fig. 3. Illustration of the generalized one-dimensional Markov chain model in Park *et al.* (2007) (called "*the generalized one-dimensional Markov chain transition conditional probability equation*")

In **Fig. 3**, the states of both cell M and cell i conflict for transition probability terms ${}^{-}p_{pj}^{|M-i|}$ and ${}^{+}p_{pq}^{|N-M|}$. This is not a correct one-dimensional Markov chain model, nor a correctly-derived equation. There is no way to derive their final GCMC model given in the following EQ4 by coupling two such one-dimensional wrong equations.

$$\Pr(Z_i = S_j \mid Z_M = S_p, Z_N = S_q, Z_o = S_f, Z_p = S_g) = \frac{{}^{-}p_{pj}^{\delta_w|M-i|}\,{}^{+}p_{jq}^{\delta_e|N-i|}\,{}^{-}p_{fj}^{\delta_n|o-i|}\,{}^{+}p_{jg}^{\delta_s|p-i|}}{\sum_{k}^{n} {}^{-}p_{pk}^{\delta_w|M-i|}\,{}^{+}p_{kq}^{\delta_e|N-i|}\,{}^{-}p_{fk}^{\delta_n|o-i|}\,{}^{+}p_{kg}^{\delta_s|p-i|}} \qquad \text{(EQ4)}$$

In addition, the CMC model is not much suitable for random path, as mentioned in Li (2007, pp. 323). The authors added a Dirac delta function on each transition probability term in their final GCMC model so that a random path might be used (Park et al., 2007, equation (11), pp. 912). It seems fancy but is not reasonable. Even if they did not add symbols "-" and "+", it would be improper because once data are absent in some directions the model is not a CMC model anymore. That was also one reason why Li and his colleagues did not use the CMC model or the model extended from the CMC model (Li and Zhang, 1996) for random path simulation. With the symbols "-" and "+", their GCMC model is just a wrong equation.

The authors used the way that geostatisticians used for estimating experimental variograms to do their simulation (Park et al., 2007, figure 4, pp.912). Such a simulation algorithm seems absurd. It is not surprising that they used three classes with similar proportions in their simulation case study, because under such a situation the small class underestimation problem of the CMC model, which was pointed out by us and also demonstrated by Elfeki and Dekking (2001) but the authors here seemed reluctant to display, became unclear in simulated realizations due to the absence of small class (note that a small or minor class here means a class with a proportion obviously less than the average). In addition, the authors did not provide any parameters (i.e., transition probability matrices) for the simulation.

While the authors admitted the pattern inclination problem and the small class underestimation problem that usually occur in simulated realizations of the CMC model, they made some confusing explanations. They stated that "*Therefore, unidirectional information transfer may predominate, and this can give rise to "artificial" lithology parcel inclination for certain transition probabilities, sampling intervals, and borehole spacings*" and "*Another problem that has persisted since the development of the CMC model is underestimation for undersampled or sparsely located indicators*" (Park et al., 2007, pp. 910). First, we found that parcel inclination was caused by the asymmetry of the simulation path with asymmetric neighborhoods used in the CMC model. Seeking for symmetric or quasi-symmetric simulation paths was exactly what Li and his colleagues did for solving the problem, as explained in Li and Zhang (2008). Second, we are afraid that "undersampled indicators" should be "underestimated" in simulated realizations for a correct model. “Undersampled” means that a class is sampled less than its due proportion compared to other classes. “Sparsely located” means a class occurs rarely in the space. So these two terms should have different meanings, at least for the situation under which the CMC model underestimates small classes in simulated realizations.

Since the small class underestimation issue and other related issues had been solved years ago (Li, 2007), there is no necessity to further argue on them anymore, especially on the misunderstandings of Park and his coauthors. However, some readers in soil science and other fields may not know these. To avoid further confusions, a clarification here may be still helpful.

REFERENCES

1. Elfeki, A.M. 1996. Stochastic characterization of geological heterogeneity and its impact on groundwater contaminant transport. Ph.D. diss. Delft University of Technology, Balkema Publisher, The Netherlands.
2. Elfeki, A.M., and F.M. Dekking. 2001. A Markov chain model for subsurface characterization: Ttheory and applications. Math. Geol. 33: 569-589.
3. Li, W. 1999. 2-D stochastic simulation of spatial distribution of soil layers and types using the coupled Markov-chain method. Postdoctoral Res. Rep. No. 1. Institute for Land and Water Management, K.U. Leuven. Leuven, Belgium.
4. Li, W. 2007. Markov chain random fields for estimation of categorical variables. Math. Geol. 39: 321-335.
5. Li, W. and C. Zhang. 2005. Application of transiograms to Markov chain modeling and spatial uncertainty assessment of land cover classes. GISci. Remote Sens. 42: 297-319.
6. Li, W., and C. Zhang. 2006. A generalized Markov chain approach for conditional simulation of categorical variables from grid samples. Trans. GIS 10(4): 651-669.
7. Li, W., and C. Zhang. 2008. A single-chain-based multidimensional Markov chain model for subsurface characterization. Environ. Ecol. Statist. 15: 157-174.
8. Park, E., A.M. Elfeki, Y. Song, and K. Kim. 2007. Generalized coupled Markov chain model for characterizing categorical variables in soil mapping. Soil Sci. Soc. Am. J. 71(3): 909-917.
9. Zhang, C., and W. Li. 2008. A comparative study of nonlinear Markov chain models in conditional simulation of categorical variables from regular samples. Stoch. Environ. Res. Risk Assess. 22: 217-230.

Weidong Li & Chuanrong Zhang
Department of Geography,
University of Connecticut,
Storrs, CT 06269
weidong.li@uconn.edu & chuanrong.zhang@uconn.edu

Sept. 2011, Storrs, CT