

# Response to the Letter to the Editor by W. Li and C. Zhang

D. Allard<sup>a</sup>, D. D'Or<sup>b</sup>, R. Froidevaux<sup>b</sup>

<sup>a</sup> Biostatistique et Processus Spatiaux (BioSP), INRA, Site Agroparc, 84914 Avignon, France.

<sup>b</sup> Ephesia Consult, 9, rue Boissonnas, CH - 1227 Geneva, Switzerland.

Corresponding author: D. Allard, [allard@avignon.inra.fr](mailto:allard@avignon.inra.fr)

## Summary

In their long Letter to the Editor (referred to as the Letter below), Li and Zhang raised a series of issues which, we believe, calls for a detailed response. They raised three main issues: incorrect references and statements to some of their work; incorrect equation citations; incorrect judgement. We shall address each of these points as clearly as possible. We fully acknowledge the incorrect statements and we apologize for it. However, we will make the case that the equation citation is not that incorrect and that most of our judgements are valid. We will then open a discussion about the notion of Markov Chain Random Field vs. Maximum Entropy.

## Incorrect statements and references

We must first acknowledge that in page 382 of our paper (Allard *et al.*, 2011) we mis-referenced the work by Li and co-authors. In Li *et al.* (2004) an approach called TMC (Triplex Markov Chain) was proposed, which was an extension of the coupled Markov

chain (CMC) model developed in Elfeki (1996). To overcome the shortcomings of CMC and TMC such as the underestimation of small classes, W. Li and C. Zhang presented in an impressive series of papers [Li (2007a), Li (2007b), Li and Zhang (2007), Zhang and Li (2007), Li and Zhang (2008) and Zhang and Li (2008)] an approach based on a single Markov Chain for predicting and simulating categorical variables, called Markov Chain Random Field (MCRF). Li and Zhang therefore rightly noted that we mixed the earlier papers with the later ones, leading to improper reference to their work. Such a confusion is truly embarrassing. We present our apologies.

### **Incorrect equation citation**

The point raised by Li and Zhang in this paragraph is that the equation referenced as EQ2 in the Letter and which was presented in their 2007-2008 series of papers is not equivalent to our presentation (i.e. Equation (14) in our paper), referenced as EQ1. There are two differences: (i) the number of conditioning points is fixed to 4 in our Equation (14) while it is general, say equal to  $m$ , in theirs; (ii) we singled out the directions to the four main directions (in 2D), while their general equation does not. It is actually not an error, but a deliberate choice. The question is thus: why did we make these changes?

We re-read all above-referenced papers carefully. In their papers they indeed include a rather general sentence about  $m$  being the number of considered neighbours, and  $p_{kl_i}^i(\mathbf{h}_i)$  being the transiogram in direction  $i$  computed at distance  $\mathbf{h}_i$  for categories  $k$  and  $l_i$ . But this general consideration is then systematically followed by a greatly detailed presentation (e.g. 2 full pages in Li, 2007) of the 2D implementation with four neighbours in the (left,right,above,below) directions plus the specific cases with 3,2 and 1 neighbours. It usually comes with a Figure, such as Figure 1 in Li (2007) which shows the prediction point surrounded by four neighbours on a regular cross. In (Zhang and

Li, 2008) they write “Usually considering only the nearest known neighbors in the four cardinal directions is sufficient for simulating multinomial classes in the horizontal two dimensions. Thus, Eq. 1 can be simplified to (using the notations of EQ2)

$$Pr(X(x) = k | l_1, l_2, l_3, l_4) = \frac{p_{kl_4}^4(h_4)p_{kl_3}^3(h_3)p_{kl_2}^2(h_2)p_{l_1k}^1(h_1)}{\sum_{f=1}^n p_{fl_4}^4(h_4)p_{fl_3}^3(h_3)p_{fl_2}^2(h_2)p_{l_1f}^1(h_1)} \quad (1)$$

where, 1, 2, 3, and 4 represent the four cardinal directions.” Similar quotations can be found in any of the above mentioned papers. So, although EQ2 being very general, they in fact use the less general equation EQ1. In their Letter, they acknowledge this fact: “Indeed, so far all simulation algorithms published by Li and his colleagues consider only neighbors in four cardinal directions or quadrants.”

More importantly, when they justify their Markov Chain model by a theorem by Pickard (1980), they must use the fact that the four neighbours are organized along the cardinal directions (Li, 2007; Zhang and Li, 2008). Pickard theorem cannot be invoked for  $m$  points organized freely in 2D. We will come back to this issue in the last section.

All in all, we consider that although mis-referencing the letter of the equation we did not mis-referenced its spirit. On purpose, we specified the equation to the implementation actually used in all available papers on the subject, e.g. Equation (19) in Li (2007) and Equation (2) in Zhang and Li (2008).

### **Incorrect judgement**

Our statement “Our method is thus a generalization of MCRF, allowing to consider as many neighbours in as many directions as desired” (p. 386 of our paper) was challenged by Li and Zhang on the basis that MCP would need stationarity while MCRF would not. We shall show that this statement is incorrect: MCP equation is indeed equivalent to the MCRF equation, even when local proportions are non stationary.

Let us first recall that there are two ways of deriving the MCP equations. The usual one consists in applying a conditional assumption. The conditional probability

$p_{i_1, \dots, i_n | i_0}$  is approximated using products of bivariate conditional probabilities:

$$p_{i_1, \dots, i_n | i_0}^* = \prod_{k=1}^n p_{i_k | i_0}(\mathbf{h}_{0k}).$$

Then by applying the Bayes theorem, MCP equation (Equation (7) in our paper) follows immediately.

The second one is by seeking the joint probability  $p_{i_0, i_1, \dots, i_n}^*$  that maximizes the entropy, subject to specified univariate and bivariate probabilities. It leads to Equation (12). The important point is that the maximum entropy approach does not need stationarity assumptions at any stage, see the proof in appendix of our paper. The consequence is immediate. In Equation (12), the univariate probability  $p_{i_0}(\mathbf{x}_0)$  can vary in space and the bivariate probabilities  $p_{i_0, i_k}(\mathbf{x}_0, \mathbf{x}_k)$  can depend on both points  $\mathbf{x}_0$  and  $\mathbf{x}_k$ . Equations (7) and (12) are equivalent as long as the bivariate probability can factorize into

$$p_{i_0, i_k}(\mathbf{x}_0, \mathbf{x}_k) = p_{i_k | i_0}(\mathbf{x}_k - \mathbf{x}_0)p_{i_0}(\mathbf{x}_0). \quad (2)$$

For clarity and conciseness reasons, we chose to not present the non stationary case in our paper. We thus made the unnecessary stationarity assumption, but extension to the non-stationary case was possible by examining closely our equations and how they were obtained. In the specific non stationary case of Equation (2), we get

$$p_{i_0 | i_1}(\mathbf{h}_{10}) = p_{i_0, i_1}(\mathbf{h}_{01}) / p_{i_1}(\mathbf{x}_1) = p_{i_1 | i_0}(\mathbf{h}_{10})p_{i_0}(\mathbf{x}_0) / p_{i_1}(\mathbf{x}_1).$$

The quantity  $p_{i_1}(\mathbf{x}_1)$ , being present in the numerator and in the denominator of (EQ5), cancels out. Hence (EQ4) and (EQ5) are equivalent. This non stationary case has been implemented and tested with success in D'Or *et al.* (2010).

Our statement was thus only related to the fact that in our implementation, and in our examples, we **used** a general number of neighbours (from at most 2 to at most 10), while their implementation always considers 4 fours points in the 4 cardinal directions as already discussed in the previous section.

## Discussion

Considering that, as we showed above, even for non-stationary univariate probabilities, the equations of MCRF and MCP are equivalent (EQ 4/5), what are the differences between MCP and MCRF? Or, put it differently, what is new and original in our work? A first difference lies in the theoretical foundation leading to EQ4/5. According to Li and Zhang (e.g. Li, 2007) MCRF is based on a theorem by Pickard (1980). Our approach is based on a maximum entropy principle. The other differences lie in the implementation of EQ 4/5. For detailing these differences, we split the discussion into a theoretical part and an algorithmical part.

### *Theoretical considerations*

In their letter to the Editor, Li and Zhang wrote “*The general solution of MCRFs is completely explained by the MCRF theory*”. They later wrote “*a single spatial Markov chain moves or jumps in a space and decides its state at any location by interactions with its nearest neighbors in different directions*” and “*Li (2007) found that nearest neighbours in cardinal directions could be regarded as independent in a sparse data space based on the theorem of Pickard (1980)*”. These statements raise the following theoretical points not addressed in our paper but which ought to be addressed here: is MCRF a Markov Chain at all? What is the MCRF theory? What exactly is the role of the theorem by Pickard?

- In a Markov chain, a variable, say  $X$  (a scalar or a vector), changes randomly at each time step from one state to another based only on the knowledge of the current state and a transition matrix. With MCRF the prediction points follow a random path. If the random path is not controlled, successive prediction points are thus likely to be at a distance larger than the correlation range. The influence of the previously simulated point would be negligible. We can seldom consider

this as a Markov Chain. Actually, since the chain moves from one location to the next, it is not the same variable that changes from one state to another. The probability distribution also changes from one location to the next. Strictly speaking MCRF is thus not a Markov Chain. It is in fact a sequential algorithm, as SIS or SGS are.

A correct Markov Chain framework would have been the following:  $X$  would be the entire simulation grid. Then, when moving from one location to the next, one coordinate of  $X$  would change at a time given the knowledge of all other coordinates (with a spatial Markovian assumption, knowledge of the nearest neighbours is sufficient). Changing all coordinates of the vector sequentially constitutes one iteration of the Markov Chain. The Markov Chain would then be run until convergence to the stationary distribution of  $X$ . This is nothing but the Markov Random field framework, with an iterative algorithm referred to as Markov Chain Monte-Carlo for simulation (Gilks, Richardson and Spiegelhalter, 1996). Note that  $X$  represents the same vector of variables at each iteration. Note also that the transition matrix is the same along iterations.

Neither MCRF nor MCP are Markov Chains. They are not iterative algorithms since each location is only simulated once. They are sequential algorithms based on the trivial decomposition

$$p(i_1, \dots, i_n) = p(i_n | i_1, \dots, i_{n-1})p(i_{n-1} | i_1, \dots, i_{n-2}) \cdots p(i_2 | i_1)p(i_1).$$

- Li and Zhang made reference to “the theory of MCRF”. An MCRF is a Random Field defined by the following parameters: distribution of the random path, rules for neighbourhoods, transiograms and probability distribution (i.e., EQ4/5). It is thus an algorithmic based definition rather than a model-based definition. There is no explicit mathematical formula for this random field. To Li and Zhang

question “*Is the MCRF approach equal to a simulation algorithm?*” we answer a clear and definitive “Yes”.

- In Li (2007), Equation (12) (i.e. EQ4 in their Letter to the Editor) is then obtained by applying a conditional independence assumption. In many places (Li, 2007; Li and Zhang, 2008) conditional independence is justified by invoking a theorem in Pickard (1980). This theorem is unfortunately very particular and of limited range of applications. Let us consider a square lattice equipped with a lexicographical order: from left to right, from top to bottom. In other words, at each point of the lattice, all points located above and/or to the left of that point belong to the past. On this square lattice, let

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

be a  $2 \times 2$  square of random variables. Pickard (1980) considered an unilateral Markov random field such that  $D \mid \text{past} \sim D \mid (A, B, C)$ , where  $\sim$  means “identically distributed as”. The theorem is now the following: **If**  $B$  and  $C$  are conditionally independent given  $A$ , written  $B \perp C \mid A$ , **then** any two diagonal variables are conditionally independent given their adjacent variable, e.g.  $A \perp D \mid B$ . The theorem is limited to the unilateral random field described above. It cannot prove in all generality that 4 neighbours in the 4 directions (left,right,up,down) are conditionally independent for any kind of transition probabilities without any reference to lexicographical order and unilateral fields. This is however what Li and Zhang pretend in their Letter and in Li (2007): “*Li (2007) found that nearest neighbours in cardinal directions could be regarded as independent in a sparse data space based on the theorem of Pickard (1980)*” (Letter to the Editor, already quoted), and “*Because the Conditional Independence assumption is correct regarding cardinal directions, specific SMC models that consider nearest*

*known neighbors only in cardinal directions are theoretically sound Markov chain models”* (Li, 2007, p. 328).

Clearly these statements are wrong. To put it in other words, Pickard’s construction is a Simultaneous Auto Regressive model (SAR). MCP and MCRF are examples of Conditionally Auto Regressive models (CAR). It is well known (see e.g. Gaetan and Guyon, 2010) that every SAR is also a CAR, but that the inverse is not true. The set of admissible CARs is much larger than the set of SARs. A theorem proven for all SARs cannot be true for any CAR.

Since Li and Zhang made an inappropriate use of Pickard’s theorem in all their papers, we maintain our statement saying that the theoretical properties of MCRF were not known.

In contrast, we do not claim that we were able to build explicitly a random field. But we showed that the MCP conditional probability distribution (our Equation 12) is the probability distribution minimizing the KL divergence to the real unknown conditional probability, subject to specified univariate and bivariate probabilities. As a corollary, it shows the strong ties between conditional independence and maximum entropy subject to bivariate constraints. In our opinion, this result shades an interesting light on Equation 12.

Our result is of course related to Bayesian Maximum Entropy BME (D’Or *et al.*, 2001; Bogaert, 2002; D’Or and Bogaert, 2004). Recall that BME seeks the maximum entropy distribution subject to all univariate and bivariate distributions, while MCP seeks the maximum entropy distribution subject to univariate and those bivariate distributions with one endpoint at the prediction point. It is thus an approximation of the unknown distribution of lower quality, but much faster to compute.

We can make an interesting parallel with kriging: while kriging is the predictor minimizing the Mean Square Prediction Error with the unknown value subject to specified

second order properties (variogram or covariance function), MCP is the prediction distribution minimizing the KL divergence to the unknown conditional distribution.

### *Algorithmical considerations*

On the algorithmical side, there are two differences between MCP and MCRF:

1. On simulations and on a real case we showed the advantage of using directional transiograms instead of an omnidirectional one. We proposed to use a kernel smoothing technique for estimating the bivariate probabilities in every directions at once, which is new and original in this context. Note that in our paper we presented the estimation of bivariate probabilities  $p_{ij}(\mathbf{h})$  and not transiograms  $p_{j|i}(\mathbf{h})$ . Transiograms are deduced from the knowledge of the bivariate and univariate probabilities.
2. Since MCP is a maximum entropy approximation of the unknown joint probability  $p_{i_0, i_1, \dots, i_n}$  with less constraints than BME, we chose to compare MCP to BME. For doing so, we ran MCP with at most  $n_{max} = 1, 2, \dots, 10$  neighbours (obviously, at the very beginning of unconditional simulations there might be less neighbours than  $n_{max}$ ). On our simulation exercise, we found that MCP was already a very good approximation with at most 2 neighbours, and that it was optimal with at most 5. Figure 2 illustrates that indeed a situation with clustered neighbours is not favorable for MCP and that BME copes much better with this situation. Since MCP does not honor between-neighbours bivariate probabilities, the most favorable situation is when the redundancy is the lowest, i.e. when neighbours are organized at equal distance in the 4 cardinal directions. In this case difference between BME and MCP is tiny. Since MCP and MCRF have equivalent equations, the same conclusion would hold for MCRF. In their Letter, Li and Zhang explain the numerous reasons leading to the choice of quadrant

search: historical reasons, practical reasons and reasons based on a theorem in Pickard (1980). We explained above why this theorem cannot be a theoretical reason, even though it is experimentally observed that conditional independence is a good approximation when four neighbours are selected along the four cardinal directions.

MCP can be used for prediction (map of the most likely category given the data) and for simulations (conditional or not). For prediction maps (see Figure 6 in our paper), the data location is fixed and can be unevenly distributed. We decided to consider the  $n$  nearest neighbours, whatever their arrangements in quadrant instead of imposing neighbours in the 4 quadrants regardless their distance to the prediction point. For simulations, instead of forcing a quadrant neighbourhood, we chose to organize the path  $\mathbf{x}_1, \mathbf{x}_2, \dots$  of the simulated points in order to have a regular pattern of the data points that have been already simulated. Comparing on prediction/simulation maps the implementation choices would definitely be interesting, but beyond the scope of this Response.

## Conclusion

The main emphasis of our paper is the maximum entropy derivation of Equation (12). We tried to make exhaustive relationships to other works, but we had to be concise. We certainly did not spend enough time explaining the similarities and differences between MCP and MCRF: their theoretical background, how they were derived, how they are used in the algorithms. We mis-referenced Li and Zhang work, and made citations that could give the impression that we down-played their contribution. Again, we present our apologies for this. One difficulty we faced was that Li and Zhang wrote an impressive series of papers, very similar in nature but with some differences between them. A second one is that, although their algorithm being sound and seemingly ef-

ficient, the theoretical foundation was shaky as shown above. Adding this to the fact that we certainly did not spend enough time writing our paper explain in part the mis-reference.

We do believe however that we made interesting advances in understanding the theoretical properties of MCP/MCRF equations and that estimating directional tranziograms non parametrically with kernel smoothing is an interesting proposition.

We also believe that our contribution together with those by Li and Zhang, will help scientists to adopt a new paradigm for making prediction of categorical data; instead of combining bivariate probabilities additively as in indicator kriging or SIS, combining them multiplicatively has been shown to verify an optimality criterion while at the same time providing very sensible maps.

## References

- Bogaert, P. 2002. Spatial prediction of categorical variables: the Bayesian Maximum Entropy approach. *Stochastic Environmental Research & Risk Assessment*, **16**, 425–448.
- D’Or, D. & Bogaert, P. 2004. Spatial prediction of categorical variables with the Bayesian Maximum Entropy approach: the Ooypolder case study. *European Journal of Soil Science*, **55**, 763–775.
- D’Or D., Bogaert P. & Christakos G. 2001. Application of the BME approach to soil texture mapping. *Stochastic Environmental Research & Risk Assessment*, **15**, 87–100.
- D’Or D., Allard, D., Froidevaux, R. Incorporating non stationarity in categorical variables simulations with MCS. Book of Abstracts of geoENV 2010, 8th International Conference on Geostatistics for Environmental Applications. Available at <http://www.geoenv.ugent.be/geoENV10BookofAbstractsonweb.pdf>

- Elfeki, A.M. 1996. Stochastic Characterization of Geological Heterogeneity and Its Impact on Groundwater Contaminant Transport. Ph.D. dissertation. Delft University of Technology, Balkema Publisher, The Netherlands.
- Gaetan, C. and Guyon, X. 2010. *Spatial Statistics and Modeling*, Springer.
- Gilks, Richardson and Spiegelhalter, 1996. *Markov Chain Monte Carlo in Practice*, Chapman and Hall.
- Kullback, S. & Leibler, R.A. 1951. On information and sufficiency. *Annals of Mathematical Statistics*, **22**, 76–86.
- Li, W. 2007a. A Fixed-Path Markov Chain Algorithm for Conditional Simulation of Discrete Spatial Variables *Mathematical Geology*, **39**, 159–176.
- Li, W. 2007b. Markov Chain Random Fields for Estimation of Categorical Variables. *Mathematical Geology*, **39**, 321–225.
- 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.
- Li, W. and C. Zhang. 2008. A single-chain-based multidimensional Markov chain model for subsurface characterization. *Environmental and Ecological Statistics*, **15**(2), 157-174.
- Li, W., Zhang, C. Burt, J.E., Zhu, A.-X. & Feyen, J. 2004. Two-dimensional Markov Chain Simulation of Soil Type Spatial Distribution. *Soil Science Society of America Journal*, **68**, 1479–1490.
- Pickard, D.K. 1980. Unilateral Markov fields. *Advances in Applied Probability*, **12**, 655-671.

Zhang, C. and W. Li. 2007. Comparing a fixed-path Markov chain geostatistical algorithm with sequential indicator simulation in categorical variable simulation from regular samples. *GIScience & Remote Sensing*, **44**(3), 251-266.

Zhang, C. and W. Li. 2008. A comparative study of nonlinear Markov chain models for conditional simulation of multinomial classes from regular samples. *Stochastic Environmental Research and Risk Assessment*, **22**(2), 217-230.