Updating Categorical Soil Map with Limited Survey Data by Bayesian Markov Chain Co-Simulation

Weidong Li and Chuanrong Zhang

Department of Geography, University of Connecticut, Storrs, CT 06269, USA
Email: weidongwoody@gmail.com, zhangchuanrong@gmail.com

Introduction

Categorical soil maps are widely used for resources and environmental management. Because the spatial distribution of soils may change due to various reasons, existing soil maps may be too outdated to reflect current field soil distributions; thus, periodical map update is necessary to meet the requirements of applications. However, large-scale detailed soil survey is too costly to carry out frequently for generating new high-quality maps. If a soil map is of sufficient quality and appropriately scaled, updating may not require a new full-coverage survey for revising the map because the types of soils at most places may not have changes. Consequently, we may update a legacy soil map with only limited new survey data. Recently, a Bayesian Markov chain random field (MCRF) approach was proposed for simulating categorical fields (Li, 2007). The MCRF sequential simulation (MCSS) algorithm (Li and Zhang, 2007) was further extended into a MCRF sequential co-simulation (Co-MCSS) algorithm. In this study, Co-MCSS was used to incorporate legacy map data through co-simulations into categorical soil map creation with limited survey data. A case study using synthetic data demonstrated its feasibility. The objective is to suggest a cost-efficient method for updating categorical soil maps.

Method

A MCRF refers to a spatial Markov chain that moves or jumps in a space and decides its state at any uninformed location by interactions with its nearest neighbors in different directions through sequential Bayesian updating (Li, 2007; Li and Zhang, 2013). If we assume i_1 to i_m are the states of the nearest neighbors in different directions around an uninformed location $\mathbf{u}_0$, the local conditional probability distribution of a MCRF $Z(\mathbf{u})$ can be factorized as

$$p[i_0(\mathbf{u}_0)\,|\,i_1(\mathbf{u}_1),...,i_m(\mathbf{u}_m)] = A^{-1}p[i_m(\mathbf{u}_m)\,|\,i_0(\mathbf{u}_0),\cdots,i_{m-1}(\mathbf{u}_{m-1})]\cdots p[i_2(\mathbf{u}_2)\,|\,i_0(\mathbf{u}_0),i_1(\mathbf{u}_1)]p[i_0(\mathbf{u}_0)\,|\,i_1(\mathbf{u}_1)] \quad (1)$$

where $A = p[i_1(\mathbf{u}_1),...,i_m(\mathbf{u}_m)]/\,p[i_1(\mathbf{u}_1)]$ is a normalizing constant, and $\mathbf{u}_1$ indicates the last stay location or the location that the spatial Markov chain goes through to current location $\mathbf{u}_0$ (Li, 2007). This full general solution is a multiple-point spatial statistical model, composed of two- to m+1-point statistics involving the uninformed location $\mathbf{u}_0$ and directional lag distances.

If the spatial Markov chain is stationary and its last stay location is far away from the current uninformed location, we may exclude the last stay location. Thus, the local conditional probability distribution can be factorized differently as

$$p[i_0(\mathbf{u}_0)\,|\,i_1(\mathbf{u}_1),...,i_m(\mathbf{u}_m)] = A^{-1}p[i_m(\mathbf{u}_m)\,|\,i_0(\mathbf{u}_0),\cdots,i_{m-1}(\mathbf{u}_{m-1})]\cdots p[i_1(\mathbf{u}_1)\,|\,i_0(\mathbf{u}_0)]p[i_0(\mathbf{u}_0)] \quad (2)$$

where $A = p[i_1(\mathbf{u}_1),...,i_m(\mathbf{u}_m)]$ is a normalizing constant, and $\mathbf{u}_1$ is not the last stay location but just a nearest neighbor. Equation (2) is a special case of Equation (1).

Examining Equation (1) with the Bayesian inference principle, one can find that $p[i_0(\mathbf{u}_0)\,|\,i_1(\mathbf{u}_1),...,i_m(\mathbf{u}_m)]$ is the posterior, $p[i_0(\mathbf{u}_0)\,|\,i_1(\mathbf{u}_1)]$ is the prior, and the other part of the right-hand side excluding the constant is the likelihood part, composed of multiple terms. These likelihood terms update the prior recursively:

$$posterior_1 = prior$$

$$posterior_2 \propto L_2 \times posterior_1$$

$$\cdots$$

$$posterior_m \propto L_m \times posterior_{m-1} \propto L_m \times \cdots \times L_2 \times prior$$

where L_k refers to the likelihood term for the kth nearest neighbor. When no nearest neighbor other than the last stay location is available, we get a posterior probability equal to the prior. But when there are other nearest neighbors, update begins on each datum in turn, and in each time of update the posterior of last update serves as the new prior. Therefore, the MCRF general full solution, that is, Equation (1), represents a *simultaneous sequential Bayesian updating on different nearest data in a Markov-type neighborhood*, which can be simply expressed as

$$posterior \propto likelihood[i_m(\mathbf{u}_m)] \times \cdots \times likelihood[i_2(\mathbf{u}_2)] \times prior. \quad (3)$$

Equation (2) is similarly in accordance with the Bayesian inference principle, except that its prior becomes $p[i_0(\mathbf{u}_0)]$. It also can be similarly expressed like Equation (3) but with one more likelihood term $likelihood[i_1(\mathbf{u}_1)]$. However, with Equation (1), Equation (2) is not much useful because one always can assume one of the nearest neighbors to be the last stay location of the spatial Markov chain.

Equation (1) is difficult to estimate from sample data due to the multiple-point statistics it involves. If we invoke the conditional independence assumption, a simplified general solution can be obtained as

$$p[i_0(\mathbf{u}_0) \mid i_1(\mathbf{u}_1),...,i_m(\mathbf{u}_m)] = \frac{p_{i_1 i_0}(\mathbf{h}_{10}) \prod_{g=2}^{m} p_{i_0 i_g}(\mathbf{h}_{0g})}{\sum_{f_0=1}^{n} [p_{i_1 f_0}(\mathbf{h}_{10}) \prod_{g=2}^{m} p_{f_0 i_g}(\mathbf{h}_{0g})]}, \quad (4)$$

where $p_{i_0 i_g}(\mathbf{h}_{0g})$ represents a transiogram from class i_0 at location $\mathbf{u}_0$ to class i_g at location $\mathbf{u}_g$ with the lag distance $\mathbf{h}_{0g}$; and i and f represent states in the state space S = (1, …, n). Because this simplified solution involves only two-point statistics, it is directly computable from sample data.

The contributions of auxiliary variables may be incorporated by different ways. Here we regard the auxiliary data as nearest neighbors of the uninformed location $\mathbf{u}_0$ in other variable spaces. For the co-located co-simulation case, the Co-MCRF model with one auxiliary variable can be written as

$$p[i_0(\mathbf{u}_0) \mid i_1(\mathbf{u}_1),...,i_m(\mathbf{u}_m); r_0(\mathbf{u}_0)] = \frac{b_{i_0 r_0} p_{i_1 i_0}(\mathbf{h}_{10}) \prod_{g=2}^{m} p_{i_0 i_g}(\mathbf{h}_{0g})}{\sum_{f_0=1}^{n} [b_{f_0 r_0} p_{i_1 f_0}(\mathbf{h}_{10}) \prod_{g=2}^{m} p_{f_0 i_g}(\mathbf{h}_{0g})]}. \quad (5)$$

The four nearest neighbors in four cardinal directions may be regarded as conditionally independent given the state of the surrounded central location in a sparse data space (Li 2007). Therefore, it is proper for this model to consider only the nearest neighbors in four cardinal directions to approximately meet the conditional independence assumption and increase the computation efficiency. Nearest neighbors for sparse data may not be located exactly along cardinal directions. To cover the whole search area, quadrants is used to replace cardinal directions, and we can seek one nearest neighbor from each quadrant if available.

The transiogram model estimation used linear interpolation. The cross-field transition probabilities were estimated by counting point-to-point frequencies of different class pairs from the sample data to the co-located data of the legacy map.

Data

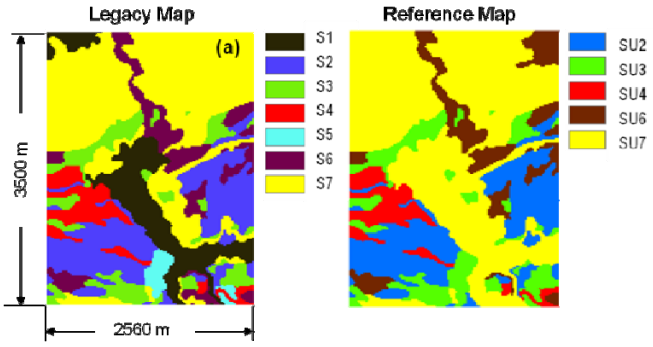

Fig. 1. The legacy soil map and the reference soil map

A piece of soil series map was used as the legacy map for case study. It has seven soil types (i.e., S1, S2, S3, S4, S5, S6 and S7). For method testing, we designed the following soil series changes: S5 is joined to S3; S1 is joined to S7; part of S6 became S7 at the bottom middle east; and part of S7 became S6 at the top-right corner. As a result, we have five new soil series: SU2 (i.e., S2), SU3 (i.e., S3 + S5), SU4 (i.e., S4), SU6 (i.e., S6 + part of S7), and SU7 (i.e., S7 + S1 + part of S6). The resulting new soil series map was used as the reference map. Because we assumed only a few of small areas were subject to soil type changes, our limited field survey was also confined to these small areas. Thus, the survey data are insufficient and also biased for estimating the parameters (e.g., transiogram models) used in the co-simulation. Our suggestion is to use pseudo sample data, that is, sample data directly extracted from unchanged

areas in the legacy soil map. Therefore, we sampled a sparse data set of 646 points from the reference soil map, which cover both the changed and unchanged areas.

Results

The optimal prediction map of the soil series and the corresponding maximum probability map were estimated from simulated realizations, conditioned on both sample data and the legacy soil map. Comparing them with the legacy soil map and the reference map shows that the unchanged S2 and S4 were exactly reproduced (as SU2 and SU4, respectively), and that the S3, which was merged with S5 without other changes, was also exactly reproduced as SU3, in the optimal prediction map. However, S6 and S7, which changed into each other in some areas, were only approximately captured (as SU6 and SU7, respectively) with apparent uncertainty. The uncertainty mainly occurred at boundary zones between these two soil series. Those areas of these two soil series that are located far away from each other were also well reproduced. Although soil type changes were confirmed by sample data only in two small areas (i.e., the top-left corner and the bottom middle east) for S6 and S7, such changes caused the uncertainty of these two soil types in other areas in the updated map. The changed areas of S6 and S7 were well captured in the optimal prediction map. The merging of S1 into S7 only increased the total area of SU7 and did not affect its uncertainty. Simulated realization maps and occurrence probability maps of single soil series further verify above judgments.

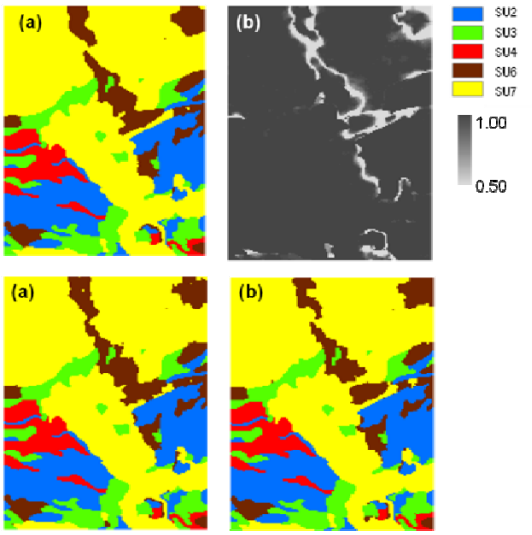

Fig. 2. Simulated results: (a) optimal prediction map, (b) maximum probability map, (c) and (d) two realization maps

Conclusions

Co-MCSS demonstrated following merits: (1) if a soil type has no changes confirmed in an update survey or if it is decided to be reclassified into another type that is deemed to have no change, it will be simply reproduced; (2) if a soil type has changes confirmed in some areas, it will be simulated with uncertainty. In general, Co-MCSS may provide a practical method for revising categorical soil maps.

References

Li, W. 2007. Markov chain random fields for estimation of categorical variables. *Mathematical Geology* 39: 321-335.

Li, W., and C. Zhang. 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. 2013. Some further clarification on Markov chain random fields and transiograms. *International Journal of Geographical Information Science*, 27: 423-430.