A Bayesian Markov Chain Approach for Land Use Classification Based on Expert Interpretation and Auxiliary Data

Weidong Li and Chuanrong Zhang

Department of Geography and Center for Environmental Sciences and Engineering, University of Connecticut, Storrs Mansfield, CT 06269, USA
Email: weidongwoody@gmail.com & zhangchuanrong@gmail.com

Introduction

Remotely sensed satellite images have been the main data source for deriving land use maps. However due to various reasons, land use maps derived from remotely sensed imagery are usually of insufficient quality for many application purposes. In addition, many conventional classification methods are mainly based on spectral data and cannot classify high level land use classes (e.g., Level III and Level IV classes) (Jensen, 2005). Expert knowledge may play an important role in determining image pixel classes through human-computer interactions because experts may integrate all available information (particularly visible patterns on images) together and discern the class labels of some pixels or clustered areas according to the specific land use classification purpose. Recently, we proposed a Markov chain geostatistical framework for land use/cover classification with uncertainty assessment based on expert-interpreted pixels from high resolution remotely sensed imagery (Li and Zhang, 2011). The approach was based on a Markov chain random field (MCRF) simulation algorithm. MCRF models are essentially extensions of Markov mesh random field models (Abend et al., 1965) to conditional simulation on sample data of categorical fields. Markov mesh models were also often called (two-dimensional) Markov chain models in image analysis and statistics due to their unilateral nature (Gray et al., 1994), and as a subclass of Markov random fields they were expressed in Bayesian inference formulations, thus representing a typical Bayesian inference approach.

In this study, we propose a Bayesian Markov chain approach for land use classification based on expert interpretation and auxiliary data. It extends the method of Li and Zhang (2011) which did not incorporate auxiliary data. Thus the proposed approach represents a method for land use classification based on diverse knowledge and data, such as expert knowledge, image information, multi-source and auxiliary data.

Methods

From the viewpoint of Bayesian inference, a MCRF defines such a spatial Markov chain: It moves or jumps in a space and at any uninformed location its probability distribution is decided by its state at its last stay location and the new evidence (i.e., nearest data) around the current uninformed location (Li, 2007). If no nearest data is available other than its last stay location, its probability distribution at the current location is solely decided by the state at its last stay location (i.e., the prior probability distribution); but if nearest data around the current location (other than the last stay location) is available as new evidence, its probability distribution at the current location will be updated based on the new evidence and the prior probability through likelihood functions. To incorporate data from other sources such as auxiliary variables, the MCRF approach needs further extension. We consider a hierarchical Bayesian approach (Gelman et al., 2004): For a spatial Markov chain $Z$, if $E$ is the state-informed data, $S$ represents auxiliary data, and $C$ refers to related parameters, their joint probability distribution can be generally expressed as $[Z,E,S,C] = [E \mid Z,S,C][S \mid Z,C][C \mid Z][Z]$. If we ignore $C$ at this stage and recognize the conditional independence between $E$ and $S$ given $Z$, we have $[Z,E,S] = [E \mid Z][S \mid Z][Z]$. For the conditional probability distribution of $Z$, we have

$$[Z \mid E,S] = [E \mid Z][S \mid Z][Z]/[E,S].$$

(1)
Reducing Equation (1) to a single uninformed location \( u_0 \), its four nearest neighbors in four quadrants and the co-located datum of an auxiliary variable, and assume the conditional independence of the four nearest neighbors, we have

\[
p[i_0(u_0) | i_1(u_1), \ldots, i_m(u_m); r(u_0)] = \frac{\prod_{i=1}^{m} p_{i|0}(h_{1,0}) \prod_{r=2}^{4} p_{i|r}(h_{rg})}{\sum_{i'|0}^{m} \prod_{i'=1}^{m} p_{i'|0}(h_{1,0}) \prod_{r=2}^{4} p_{i'|r}(h_{rg})},
\]

where, \( i \) refers to the state of the pixels in the primary variable, \( r \) the state of the co-located datum in the auxiliary variable, and \( h \) the lag distance between two locations. In addition, \( p_{i|r}(h_{rg}) \) represents a transiogram within a multi-class spatial variable, and \( p_{i|0}(h_{1,0}) \) a transition probability from the primary variable to the auxiliary variable.

We used the expert-interpreted pixel labels from a high resolution remote sensing image (or multiple or a time series of related images) as sample data, and used the classified pixel data of a good remote sensing image using the supervised maximum likelihood (SML) classifier as auxiliary data. We considered three land use classes (farmland, built area, and waterbody) in the case study. We interpreted a large pixel label data set from the image using ArcGIS. The extra-dense data set was then split into one dense data set for conditional co-simulation and the other data set for validation. A medium, a sparse, an extra-sparse, and an extreme-sparse data set were further extracted from the dense one. Unconditional co-simulation was regarded as conditional on a zero data set. These sixe data sets accounted for 0% to 1.81% of the total image pixels. The same image was classified into similar three land use classes using the SML classifier, and the classified data served as an exhaustive auxiliary data set. Auxiliary data may be different in class number and the classes may not need to have physical meaning, but should be correlated with the primary variable.

**Figure 1.** The original remotely sensed image (a), classified image using the supervised maximum likelihood classifier (b), and optimal classification maps based on maximum occurrence probabilities, by the Bayesian Markov chain approach, conditioned on an auxiliary data set and different expert-interpreted sample data sets (c–f).

**Results**

The accuracy of the land use map classified using the SML classifier is 77.63%, which is normal. Optimal classification maps from the proposed approach show that all three land use classes are exactly captured with dense sample data, but the built areas gradually deviate from their real shapes at boundaries with decreasing density of sample data; however, the general pattern has little changes with the change of sample density, and the water bodies are always well classified except for two shallow ponds missed by the SML classifier. Even when no sample data is used for conditioning, the general class pattern and the shapes of water bodies are still captured in classification. This should be attributed to the contribution of the auxiliary data, that is, the classified image by SML, and the correlation information conveyed by transiogram models. These optimal classification maps have very high accuracies, ranging from 86.12% to 98.59%, which demonstrate much improvement over the SML.
classification. Even when very sparse (e.g., 0.2% of the total pixels) or zero sample data are available, the suggested method still can improve the classification accuracy a lot (10% or so). Simulated realizations indicate more details than corresponding optimal classification maps. The averaged classification accuracy of realizations decreases from 97.76% to 84.35% with increasing density of interpreted sample data from 0% to 1.8% of total pixels. Interpreting a sparse sample data set from remotely sensed imagery is not difficult and the gain in classification accuracy is huge.

**Conclusions**

The incorporation of auxiliary data is a large boost to its practicality of MCRFs. This study proves the value of the suggested approach in land use classification from remotely sensed imagery. Through incorporating auxiliary data such as a classified image by a conventional classification method, the required density of interpreted sample data is largely reduced given a required classification accuracy level. A few of expert-interpreted pixel label data may obviously improve the classification accuracy, and a sparse informed pixel data set may support the classification of some high level land use classes that could not be recognized by conventional non-expert-based classification methods. Of course, as a human-computer interactive approach, data collection always needs some time. Parameter estimation, simulation, and data processing also spend time. This labor and time cost is the trade-off for high accuracy classification with uncertainty assessment. The development of a specific software tool may largely reduce the cost.

**References**


