Plant Ecol (2007) 191:85–94
DOI 10.1007/s11258-006-9216-2

ORIGINAL PAPER

Mapping pre-European settlement vegetation at fine resolutions using a hierarchical Bayesian model and GIS

Hong S. He · Daniel C. Dey · Xiuli Fan · Mevin B. Hooten · John M. Kabrick · Christopher K. Wikle · Zhaofei Fan

Received: 6 February 2006 / Accepted: 7 September 2006 / Published online: 27 October 2006

Abstract In the Midwestern United States, the General Land Office (GLO) survey records provide the only reasonably accurate data source of forest composition and tree species distribution at the time of pre-European settlement (circa late 1800 to early 1850). However, GLO data have two fundamental limitations: coarse spatial resolutions (the square mile section and half mile distance between quarter corner and section corner) and point data format, which are insufficient to describe vegetation that is continuously distributed over the landscape. Thus, geographic information system and statistical inference methods to map GLO data and reconstruct historical vegetation are needed. In this study, we applied a hierarchical Bayesian approach that combines species and environment relationships and explicit spatial dependence to map GLO data. We showed that the hierarchical Bayesian approach (1) is effective in predicting historical vegetation distribution, (2) is robust at multiple classification levels (species, genus, and functional groups), (3) can be used to derive vegetation patterns at fine resolutions (e.g., in this study, 120 m) when the corresponding environmental data exist, and (4) is applicable to relatively moderate map sizes (e.g., 792×763 pixels) due to the limitation of computational capacity. Our predictions of historical vegetation from this study provide a quantitative and spatial basis for restoration of natural floodplain vegetation. An important assumption in this approach is that the current environmental covariates can be used as surrogates for the historical environmental covariates, which are often not available. Our study showed that terrain and soil covariates least affected by past natural and anthropogenic alternations can be used as covariates for GLO vegetation mapping.

Keywords GLO · GIS · Hierarchical Bayesian models · Presettlement vegetation · Missouri

H. S. He (✉) · X. Fan · Z. Fan
School of Natural Resources, University of Missouri-Columbia, 203 ABNR Building, Columbia, MO 65211, USA
e-mail: heh@missouri.edu

D. C. Dey · J. M. Kabrick
US Forest Service, North Central Research Station, Columbia, MO, USA

M. B. Hooten
Department of Mathematics and Statistics, Utah State University, 3900 Old Main Hill Logan, Logan, UT, USA

C. K. Wikle
Department of Statistics, University of Missouri-Columbia, Columbia, MO, USA

Introduction

Historical data provide baseline information for assessing vegetation change and guiding

ecological restoration (Dey et al. 2000; Dyer 2001; Bolliger et al. 2004; Schulte et al. 2005). One of such data is the General Land Office (GLO) survey conducted from early 18th to mid-19th century in the Midwestern United States (Bourdo 1956). Under the GLO, the public land survey system (PLSS) was developed, in which land was divided into a grid of square townships, each containing 36 1-square-mile (1.6 km^2) sections. At each section corner and the mid-point between section corners (quarter corner) 2–4 witness trees were identified and measured to mark the corner location by the surveyors. Since GLO records are spatially referenced, they can be easily imported into a geographic information system (GIS) for further spatial analysis and statistical inference (He et al. 2000). Despite the survey biases including surveyor's preference and exclusion of certain species and age groups (Manies et al. 2001; Mladenoff et al. 2002), GLO data provides the only reasonably accurate data source of forest composition and tree species distribution prior to the pre-European settlement (Manies and Mladenoff 2000; Black et al. 2006).

GLO data have two fundamental limitations, associated with the PLSS structure. First, the square mile section and half mile distance between quarter corner and section corner is very coarse for a typical ecological restoration task that is often conducted at individual sites of a few hectares. Although some methods have been used to map GLO data to finer resolutions (e.g., Brown 1998), they have not been validated. Second, GLO data are point data, which alone are insufficient to describe vegetation that is continuously distributed over the landscape. GIS and statistical inference are needed to convert point data into more relevant data forms such as grids of finer resolutions and determine vegetation for places where data were not recorded. The most common statistical inference methods used to reconstruct historic vegetation by interpolating GLO data are ISOLINE, kriging or co-kriging embedded in GIS (e.g., Porter 1998; Brown 1998; Batek et al. 1999). These methods use the vegetation values of the known data points to interpolate the vegetation values for points that do not have recordings. They assume that the interpolated data are numerical and are spatially continuous, such as elevation. These methods can be problematic when applied to GLO data that are primarily in the form of tree count and tree diameter. Also, GLO data are not necessarily spatially continuous because many factors such as soil, elevation, and competition among other species often cause the spatial discontinuity of a species distribution (Bolliger and Mladenoff 2005). Therefore, the interpolation methods may ignore the ecological principles underlined by these environmental factors.

Methods have been proposed for mapping individual species pattern using known environmental covariates that account for ecological/biotic processes (e.g., Zimmermann and Kienast 1999; Lichstein et al. 2002). These methods, however, either lack the consideration of residual spatial dependence inherent in the vegetation distribution or are limited to small spatial domains (e.g., 10 ha). Hooten et al. (2003) developed a statistically rigorous method for combining species/environment relationships and explicit spatial dependence for binary response data. Their method combines information found within abiotic covariates and spatial dependence that may be used as a surrogate for various biotic covariates. They were able to provide spatial predictions for vegetation with known certainty, and map ground flora distribution using recent vegetation plot data (point) for areas ranging from 265 ha to 530 ha. However, whether this method can be applied to GLO data and much larger areas with diverse vegetation and environmental combinations is of interest.

The objective of this study was to apply the hierarchical Bayesian approach to GLO data and test its applicability in improving mapping historical vegetation mapping. More specifically, we evaluated if this approach (a) can be used to interpolate GLO data to fine spatial resolutions (e.g., smaller than 1,600 m), (b) can be applied to large landscapes (e.g., 10^3–10^6 ha), and (c) is robust at the species, genus or functional plant group levels. The Bayesian hierarchical approach differs from traditional statistical approaches (e.g., logistic regression) because it can account for uncertainty in various levels of the model including spatially correlated errors. We applied this approach by incorporating digital elevation,

slope, aspect, soil water capacity, soil depth, and soil organic matter as covariates in the model.

Material and methods

Study area and data preparation

Our study area encompasses a portion of Missouri River in the region of Columbia and Booneville, Missouri, USA (Fig. 1). The size of the study area is about 8,702 km^2, involves 14 counties, and includes 19,000 GLO tree data points. The study area has diverse terrain, soil, and hydrological features and consequently a high diversity of tree species. Thus, it provides an ideal place for evaluating our statistical approach. The bottomland of the study area includes the lower Missouri River alluvial plain land type association to the east and Missouri Grand River alluvial plain and loess woodland/forest breaks land type associations to the west (Nigh and Schroeder 2002). The bottomlands have flood tolerant species such as American elm (*Ulmus americana* L.), hackberry (*Celtis occidentalis* L.), and green ash (*Fraxinus pennsylvanica* Marsh.), cottonwood (*Populus deltoides* Bartr. ex Marsh.), sycamore (*Platanus occidentalis* L.), boxelder (*Acer negundo* L.) and pin oak (*Quercus palustris* Muenchh.). The uplands of the study area are dominated by white oak (*Quercus alba* L.), black oak (*Quercus velutina* Lam.), and hickory (*Carya* spp). There are a total of 19,000 individual GLO trees recorded in the GLO data for the study area, among which white oak is most abundant (30%), followed by black oak (21%), hickory (11%), elm (8%), hackberry (5%), and ash (2%). Cottonwood, sycamore, boxelder and pin oak are at 1–2%. Native Americans and European settlers have modified floodplain vegetation for hundreds of years. The greatest alternation occurred in the past 100 years including river channelization for flood control, forest clearing for framing, and recent urban sprawl. These activities have eliminated up to 95% of bottomland forests in the Missouri River basin and greatly altered the hydrologic regimes and species composition (Dey et al. 2000).

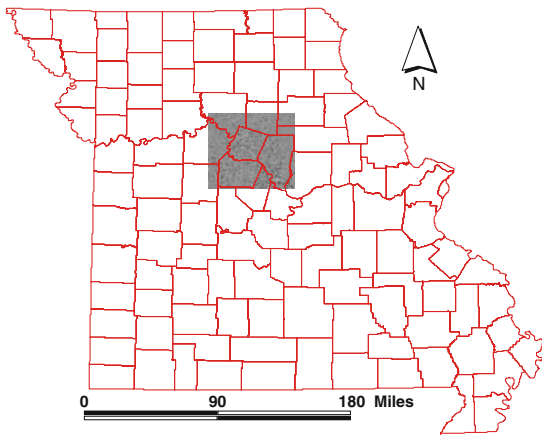

Fig. 1 The study area encompasses a portion of Missouri River in the Columbia and Booneville region. The size of the study area is 8,702 km^2 and involves 14 counties. There are 605,059 prediction locations (pixels) for 120 m resolution. There are a total of 19,000 individual GLO trees recorded in the GLO data for the study area

To evaluate the robustness of the statistical approach, we processed the GLO tree data at three classification levels: individual tree species, genus, and functional groups. For individual tree species we chose black oak since it was one of the most abundant upland species. For the genus level classification, we chose bottomland oaks (*Quercus* Spp.) that included primarily pin oak, white oak, and red oak. Functional groups were defined based upon the successional stages for the bottomland tree species and nut-producing capability for tree species in the whole area. Bottomland tree species were grouped into (a) early successional, including primarily sycamore, cottonwood, and willow (*Salix* Spp.), and (b) mid and late successional, including elm, boxelder, silver maple (*Acer saccharinum* L.) and ash. For the nut-producing functional group tree species we grouped all oak species, black walnut (*Juglans nigra* L.), and hickory.

We identified the seven most significant terrain and soil covariates for the statistical model (Table 1). They were elevation, slope, aspect, soil water capacity, soil organic matter, soil depth, and depth to bedrock. These covariates were chosen because (1) they are the determinants of the availability of basic energy, water, and nutrients influencing tree species establishment and growth, (2) they were available at the scale of

Table 1 Estimated parameters (standard errors) of the logistic models to predict species presence using covariates

	Bottom land oak	Early succession	Middle late succession	Nut producing	Black oak
Intercept	14.117 (0.939)	–5.899 (1.097)	–0.447 (0.914)	7.219 (0.208)	5.434 (0.184)
Soil water capacity	–3.453 (0.704)	3.993 (0.732)	1.806 (0.661)	–2.841 (0.458)	–1.468 (0.574)
Soil depth		0.009 (0.004)	–0.013 (0.003)	0.006 (0.002)	
Slope	–0.0267 (0.003)	0.066 (0.008)	0.019 (0.004)	–0.021 (0.002)	
Soil organic matter	0.0414 (0.018)			0.085 (0.010)	0.0380 (0.010)
Depth to bedrock				–0.005 (0.002)	–0.003 (0.001)
Elevation	–0.0206 (0.002)	0.011 (0.002)	0.003 (0.002)	–0.011 (0.000)	–0.005 (0.000)
Aspect		0.164 (0.070)			

our study, and (3) they remain relatively consistent from between the time GLO data were surveyed to the time these data were measured. The terrain data were from USGS 7.5 min DEM at 30-m horizontal resolution with about 7 m vertical resolution.[1] The DEM is the finest possible to cover the whole study area. The data were resampled into 120 m resolution using the Bilinear option in ArcGIS 9.1, which performs a bilinear interpolation and determines the new value of a cell based on a weighted distance average of the four nearest input cell. At 120 m horizontal resolutions, the vertical resolution of DEM is also improved. Thus distortion of slope calculation for certain pixels can be reduced from as much as 23% at 30 m resolution to less than 5% at 120 m resolution. The soil data was derived from the county soil survey (SSURGO) for each of the 14 counties and then merged for the study area. SSURGO database provides the most detailed level of soil information in the US. In SSURGO, maps of soil polygons are made at scales mostly ranging from 1:12,000 to 1:24,360 (USDA Natural Resource Conservation Service 1995). Each polygon may contain multiple soil components. The minimum map unit is smaller than 1.14 ha (120 m × 120 m), which is the finest resolution used in this study. In this study, we followed the standard procedures of SSURGO (USDA Natural Resource Conservation Service 1995) to derive the aggregated soil organic matter, soil depth, and soil to bedrock depth values for each polygon from the multiple soil components within the polygon. Each derived soil attribute was raterized to 120 m resolution using ArcGIS 9.1.

[1] Interdisciplinary Center for Research in Earth Science Technologies at the University of Missouri (http://icrest.missouri.edu/Projects/Infomart/Hi-resDEMs/index.htm)

Statistical modeling

We adopted the hierarchical Bayesian model developed by Hooten (2001) and Hooten et al. (2003) to predict the probability of occurrence for the three vegetative classification groups mentioned in the previous section. This generalized linear mixed model provides a method of probabilistically predicting a binary response variable based on several environmental covariates and residual spatial structure. The hierarchical modeling framework teamed with an explicitly defined spatial random effect allows us to incorporate various sources of uncertainty at many levels of the ecological process (Wikle 2002; Hooten et al. 2003; Wikle 2003).

Specifically, using the notation in Hooten et al. (2003), we let Y_i represent the presence/absence of the vegetation group of interest and be defined as:

$$Y_i = \begin{cases} 1 & \text{if} \quad Z_i > 0, \\ 0 & \text{if} \quad Z_i \leq 0, \end{cases} \tag{1}$$

where Z_i represents an underlying (latent) continuous process analogous to that derived by Albert and Chib (1993). In our case, it is composed of a covariate component, $\mathbf{X}\boldsymbol{\beta}$, and a spatially correlated component, $\boldsymbol{\eta}$. Where $\mathbf{X}$ is comprised of the covariate data (e.g., slope, aspect, elevation, soil depth, and organic matter). The model can then be summarized in matrix notation:

$$\boldsymbol{Z}|\boldsymbol{\beta},\boldsymbol{\eta} \sim N(\boldsymbol{\beta}+\boldsymbol{\eta},\boldsymbol{I}), \quad (2)$$

$$\boldsymbol{\eta} \sim N(\boldsymbol{0},\sigma_{\eta}^{2}\boldsymbol{R}_{\eta}), \quad (3)$$

$$\boldsymbol{\beta} \sim N(\boldsymbol{\beta}_{0},\Sigma_{\beta}) \quad (4)$$

In this setting, Eqs. 3 and 4 represent our prior distributions for the spatial random effect and the covariate parameters, respectively. Additionally, an inverse gamma prior distribution was specified for σ_{η}^{2}. The matrix $\boldsymbol{R}_{\eta}$ is a spatial correlation matrix although not directly parameterized. That is, we adopted the same spectral decomposition of $\boldsymbol{\eta}$ using Fourier basis functions as in Hooten et al. (2003). Although not detailed here, this Fourier (and inverse Fourier) transform allow for implementation of this model on very large spatial domains. For additional details see Wikle (2003) and Royle and Wikle (2005).

Ultimately, our interest lies with the predictions of Y_j, where j can exist on some other set of spatial location than our original data. Perhaps more informative are the predictions of probability of occurrence (p) at location j. That is, the predictions of $E(Y_j) = p_j = \Phi(x_j'\beta + \eta_j)$. In this case Φ represents the probit transform (standard normal CDF); note that the probit transform behaves similarly to the more common logit transform used in logistic regression.

Preliminary spatial analysis of the dataset was used to inform the prior for the spatial random effect. An efficient sampling algorithm similar to that used in Hooten et al. (2003) was used to sample from the posterior distribution via the full conditionals.

Result validation

Due to the dimension of the dataset and prediction grid, a repeated hold-out version of cross-validation (as in Hooten et al. 2003) was not feasible. In this study, however, we employed three approaches for result validation. The first approach was model-based validation. One benefit of using a rigorous statistical model is that it provides for model-based validation methods. This method was used to create maps of prediction error that allowed us to visualize the uncertainty across the spatial domain as well as assess specific areas of high and low variability (Hooten et al. 2003). Second, we employed the use of Receiver Operating Characteristic curves (ROC) to evaluate model accuracy and compare against similar but simpler model specifications. Assessing the sensitivity versus specificity of a model, the ROC curve is a plot of the true positive rate against the false positive rate based on the predictions (Hosmer and Lemeshow 2000; Pontius and Schneider 2001). The closer the curve follows the left-hand and top border of the ROC space, the better the predictions. Third, we constructed logistic regression models using the same response and covariate data, and directly compared their results to those of hierarchical Bayesian models. Logistic regression models are frequently used by plant ecologists to map species distribution in response to environmental variables (Franklin 1995; McDonald and Dean 2006).

Results

Spatial resolution and extent

Analysis of GLO data using the hierarchical Bayesian model at the 120-m resolution yielded about 600,000 prediction locations. At this resolution, the floodplain and islands in the river could be delineated finer than that of GLO data (~1,600 m), thereby allowing GLO data inference at more than 10 times finer than the 1,600 m resolution inhabited in the GLO data set. In addition, the study showed that the model can be applied to the study area of 8,702 km^2, which is much larger than previously reported (2.65–5.30 km^2) using the hierarchical Bayesian model (Hooten et al. 2003).

The number of prediction locations increase exponentially with increasing resolutions. At 60- and 30 m resolution, prediction locations are about 2.4 million and 10 million, respectively, which require computational capacities beyond the current state-of-art personal computers (3.7 GHz and 2 GB memory). Our current computational capability limited our analysis from being applied to finer resolutions. In addition, at the finer resolution (e.g., 60 m and 30 m), the proportions of witness-tree data points decreases,

and consequently, the potential mapping uncertainty increases.

Oak genus and functional group at bottomland

Maps that display information about the posterior distribution based on the prediction process can be viewed as probabilities of species occurrence. Such values range from 0 to 1 (100% probability of species occurrence). Results showed that as a genus group, bottomland oaks have medium probability (0.4–0.6) of occurring on the northwestern areas of Missouri River floodplain and other tributaries where floodplains are wide and islands are high in elevation. These areas are flooded less frequently compared to narrower and lower floodplains from northwest to southeastern portion of the study area. The prediction showed that the occurrence probability of bottomland oaks decreased to low (0.2–0.4) in the mid section of the Missouri River valley and very low (<0.2) in the southwestern section of the river basin within the study area (Fig. 2).

The predicted distributions for the early and late successional functional groups showed a strong association with the physical variations. The early successional group had medium probability of occurring in the mid to southeastern section of the Missouri River where the valley is narrow and frequently flooded, while the mid and late successional groups had a medium probability of occurring in the northwestern section of the Missouri River and small floodplains along other tributaries (Fig. 2). Overall, the predictions for the bottomland areas reflect the ecological dynamics coinciding with floodplain hydrological processes of erosion and deposition. Species of the early successional group are adapted to readily establish on recently deposited and exposed alluvium. Mid and late successional group species typically succeed the early successional group because they do not need mineral soil for germination and open (full sun) environments to grow to maturity. Their seed is large enough to permit seedling establishment in leaf litter and they are more tolerant of shade than the cottonwoods and willows.

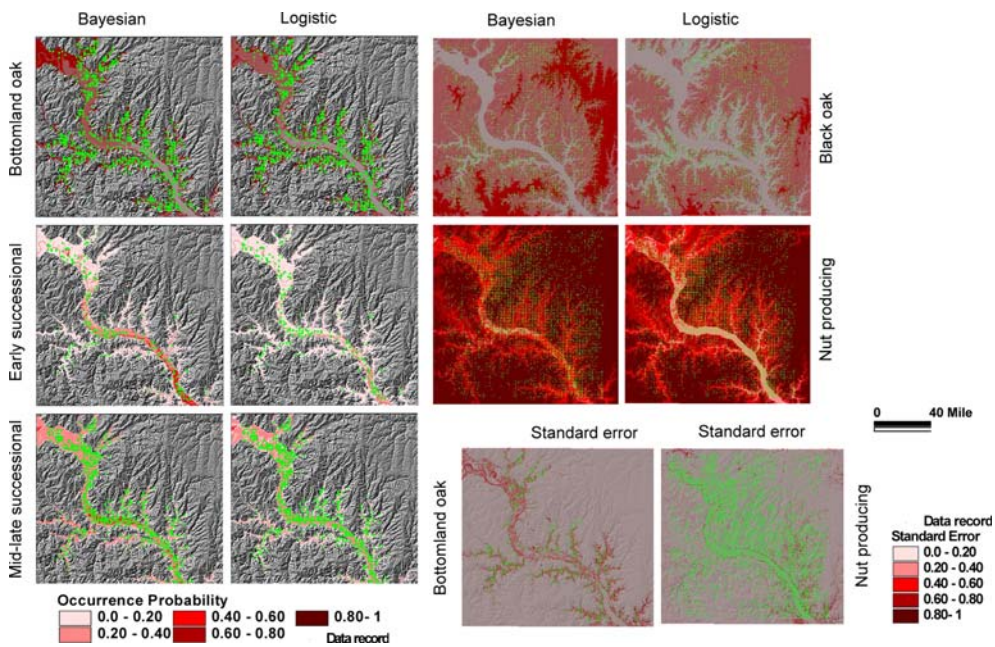

Fig. 2 Predicted probabilities for various vegetation types. The prediction maps are draped on the digital elevation model to show a 3D perspective. The strength of the Bayesian hierarchical approach over logistic regression is its ability to provide an error estimate at each pixel location (last page)

Individual species and functional group at whole study area

Predictions for black oak and the nut-producing group were made for the entire study area including both bottomlands and uplands. Predictions showed a high probability of black oak occurring on most upland areas and very low probabilities on most bottomland areas (Fig. 2). This pattern also agrees with ecological and biological characteristics of black oak as a common upland species that is not tolerant of flooding. The nut-producing functional group contained the most abundant upland tree species in this area including white oak, black oak, and hickory. The predictions showed that this group had a very high probability of occurrence (0.8–1.0) on most upland areas (Fig. 2). This pattern agrees with the ecological traits of these nut-producing species (Burns and Honkala 1990). Today they are still the most common tree species in the uplands.

Result validation

A partial validation of the predicted results was performed by presenting maps of prediction error in the form of marginal posterior standard deviations for grid cells and by ROC analysis. Posterior standard deviations for grid cells were derived for all predicted classes. Overall, they suggested that the predictions have general agreement with the GLO data records. Low prediction errors were small in places where data were abundant and relatively high in places where data were sparse (Fig. 2).

The ROC analysis suggested that the statistical method was effective in modeling historical vegetation at this fine resolution (Fig. 3). Recall however, that the hierarchical model has taken into account many other sources of uncertainty including an explicit spatial random effect. Although it may not improve on the prediction in this setting, it is providing us with a more accurate portrayal of the variability in the predictions (Fig. 3).

Comparing results from logistic regression models with those from the hierarchical Bayesian models suggest that the latter have higher goodness-of-fit at all levels (Table 2). For the hierarchical Bayesian models, when prediction accuracy is between 0.20 and 0.40, the lowest prediction probability can reach 0.50 (early successional group) and the highest prediction probability can reach 0.99 (nut-producing group). Prediction probabilities of both logistic regression and the hierarchical Bayesian models decreased with increasing prediction accuracies. However, the hierarchical Bayesian models always showed a better prediction probability than the logistic regression models (Table 2). The results also showed that when the sample data were abundant, prediction results of both types of models was similar. This was seen for the nut-producing group, most abundant class in this study area (Table 2). However, when the sample data were rare (e.g., 413 for early successional group, 672 for mid-late successional group, and 871 for bottomland oak), the hierarchical Bayesian models showed superiority.

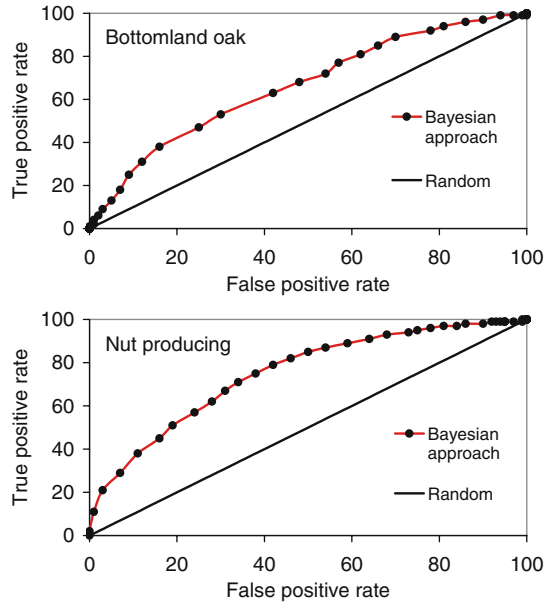

Fig. 3 ROC analysis for bottomland oak and nut-producing species group. The ROC analysis suggests that the Bayesian hierarchical model yields posterior predictions similar to the predictions from the logistic regression. However, ROC cannot illustrate the uncertainties in the predictions

Table 2 Prediction probabilities from the hierarchical Bayesian models and logistic regression models

			Prediction accuracies by thresholds			
		Number of points	0.20–0.40	0.40–0.60	0.60–0.80	0.80–1.0
Bottomland oak	Bayesian	871	0.90	0.46	0.01	0.00
	Logistic		0.25	0.00	0.00	0.00
Early succession	Bayesian	413	0.50	0.18	0.06	0.00
	Logistic		0.23	0.02	0.00	0.00
Mid-late succession	Bayesian	672	0.77	0.10	0.00	0.00
	Logistic		0.61	0.00	0.00	0.00
Black oak	Bayesian	3202	0.90	0.16	0.00	0.00
	Logistic		0.67	0.01	0.00	0.00
Nut producing	Bayesian	8938	0.99	0.99	0.93	0.63
	Logistic		0.99	0.96	0.81	0.47

Discussion

Implications

We demonstrated how GLO data could be analyzed to reveal the probable distribution of historical vegetation using a hierarchical Bayesian approach and GIS. We showed that this approach was effective in predicting historical vegetation distribution and robust at multiple classification levels (species, genus, and functional groups). The prediction probabilities for each vegetation class graphically depict the relationships between the vegetation and environment. Such predictions are based on a combination of environmental factors and surrogates for biotic factors, making them a very complete and robust reflection of species response on a continuous spatial domain. Such relationships can be easily validated against the existing ecological and biological knowledge about the predicted vegetation.

An important result of the study is that pre-settlement vegetation distributions are derived at fine scales, which, in our study, was 120 m, more than 10 times finer than the 1,600 m resolution inhabited in the GLO data set. We showed that it is feasible to incorporate empirical knowledge of the historical vegetation and environment into a modeling framework to improve vegetation mapping from sparsely recorded witness-tree points. This is especially true when the study area is large enough that it provides an adequate number of witness-tree points and when the corresponding fine scale environmental data exist. The map size in this study (e.g., 792×763 pixels) covers 8,702 km^2 (involving 14 counties and 19,000 GLO tree data points), representing a significant improvement to previous map sizes (e.g., Hooten et al. 2003; Brown 1998). At this size it is possible for forest managers and planners to go beyond the traditional site scales (a few acres) and include necessary broader and landscape-scale perspectives in ecological restoration planning efforts (Gutzwiller 2002).

Our results show that the hierarchical Bayesian models have higher prediction probabilities than those of the traditional logistic regression models when sample data are sparse. This is because the hierarchical Bayesian models are better equipped to capture the randomness and spatial dependence inherited in the sparse data. This finding has important implications to the disciplines of forestry and ecology, which are often confronted with the issue of data scarcity. The hierarchical Bayesian models present a step forward in dealing with the data scarcity problem. In addition, the strength of the Bayesian hierarchical approach over other conventional approaches (e.g., logistic regression) is its ability to provide a prediction error estimate at each pixel location while accounting for various sources of uncertainty. Maps of prediction error can be very useful for forest managers and planners so that they can consider spatially varying uncertainty in their planning processes.

Finally, the results from this study are useful for guiding ecological restoration efforts. Floodplain forests reflecting the natural processes of floodplain erosion and deposition, normally include a wide range of seral communities, however many of the natural seral communities such as

young and mature floodplain forests are now scarce due to harvesting and farming (Bragg and Tatschl 1977; Nelson 1997). Losses of extensive and continuous floodplain forest communities to cultivation, or losses of particular seral stages to the catastrophic, stand replacing disturbances may lead either to the disruption of the natural occurrence of organisms along the Missouri River or to the destruction of particular habitats required by certain species. While the debate of how to restore floodplains continues, quantitative and spatial predictions from this study provide a scientific basis for identifying appropriate seral stages for ecological restoration of floodplains.

Limitations

The study quantified the contribution of each of the seven environmental covariates to the occurrence of historical vegetation. One limitation in our approach was that we had to use the current environmental covariates as surrogates for the historical environmental covariates, which were not available for the time when GLO data were surveyed. Therefore, some historical situations could not be reconstructed. Change of the Missouri River channel has occurred in the past 100–150 years due to both natural and human causes (Dey et al. 2000). At certain locations, especially in the lower Missouri River, the river channel has shifted up to a couple kilometers (unpublished data). River channel change causes a change in hydrological regimes in the floodplains and affects soil erosion and deposition, and hence forest vegetation and succession. This can render the use of current terrain and soil variables (e.g., elevation, slope, and soil depth) irrelevant at places where significant changes have occurred. The change of river channels also limited our modeling approaches from being applied to fine resolutions (e.g., 30 m), at which greater prediction uncertainty exists due to the mismatch of the current and historic environmental data. However, in our study area most of the channel changes were less dramatic and are within the resolution of the current study (120 m). In addition, this study is conducted at a much broader spatial extent where the seven historical and current environmental variables generally agreed. Therefore, the results provide a broad spatial context for which ecological restoration can be referenced to.

Land use history is another issue that challenged the use of current environmental covariates as surrogates for the historical environmental covariates. Forest clearing for farming has eliminated up to 95% of bottomland forests in Missouri River basin. Farming can have a direct effect on changing soil physical and chemical properties and reducing soil organic matter. However, the terrain and soil covariates selected in this study were least affected by the past clearing and farming activities. Among these covariates, soil organic matter is probably more subject to change than the others. Soil organic matter generally decreases with farming since crop rotations, applying organic fertilizers, and other soil conservation measures were uncommon floodplain farming practices in the past. However, compared to most soil physical and chemical variables (e.g., available soil water content and soil nitrogen), organic matter has very slow change rates. The seven terrain and soil variables represent a conscious approximation to the historic conditions.

This study showed that readily available terrain and soil data were effective for mapping GLO vegetation at a fine resolution (120 m), as shown by the result and model validations discussed above. However, we did not study the degree of effectiveness that these terrain and soil data can be used for the fine resolution GLO mapping. Direct answers to this question require comparing mapping results derived using the terrain and soil data at different scales. We were limited from doing so because of the data availability problems. Soil survey data from SSURGO has the most detailed level of soil information and is for applications such as site-level soil erosion control in farmland planning. SSURGO is created using the 7.5 min USGS base map and thus, has a comparable scale to the DEM used in this study. The next level soil data is the state soil geographic data base (STATSGO), which is designed primarily for regional, multi-state, and multi-county resource planning and management, and is obviously too coarse for this study. It appears that DEMs of 120 m resolution is adequate in this GLO vegetation mapping. Although we did not map GLO data at other resolutions, future studies may include

evaluating terrain and soil data for mapping GLO vegetation at resolutions larger than 120 m.

Acknowledgments Funding support is from US Forest Service North Central Research Station, RWU 4154 Ecology and Management of Central Hardwood Ecosystem, and University of Missouri GIS Mission Enhancement Program.

References

Albert J, Chib S (1993) Bayesian analysis of binary and polychotomous response data. J Am Stat Assoc 88:669–679

Batek MJ, Rebertus AJ, Schroeder, WA, Haithcoat TL, Compas E, Guyette RP (1999) Reconstruction of early nineteenth century vegetation and fire regimes in the Missouri Ozarks. J Biogeogr 26:397–412

Black BA, Ruffner CM, Abrams MD (2006) Native American influences on the forest composition of the Allegheny Plateau, northwest Pennsylvania. Can J Forest Res 36:1266–1275

Bolliger J, Schulte LA, Burrows SN, Sickley TA, Mladenoff DJ (2004) Assessing ecological restoration potentials of Wisconsin (USA) using historical landscape reconstructions. Restor Ecol 12:124–142

Bolliger J, Mladenoff DJ (2005) Quantifying spatial classification uncertainties of the historical Wisconsin landscape (USA). Ecography 28:141–156

Bourdo EA (1956) A review of the General Land Office Survey and of its use in quantitative studies of former forests. Ecology 37:754–768

Bragg TB, Tatschl AK (1977) Changes in flood-plain vegetation and lands use along the Missouri River from 1826 to 1972. Environ Manage 4:348–353

Brown DG (1998) Mapping historical forest types in Baraga County Michigan, USA as fuzzy sets. Plant Ecol 134:97–111

Burns RM, Honkala BH (1990) Silvics of North America, Vol. 2. Hardwoods. Agriculture handbook 654. US Department of Agriculture Forest Service, Washington, DC, 877 pp

Dey D, Burhans D, Kabrick J, Root B, Grabner J, Gold M (2000) The Missouri River floodplain: history of oak forests and current restoration efforts. Glade 3:2–4

Dyer JM (2001) Using witness trees to assess forest change in southeastern Ohio. Can J Forest Res 31:1708–1718

Franklin J (1995) Predictive vegetation mapping: geographic modeling of biospatial patterns in relation to environmental gradients. Progress Phys Geogr 19:474–499

Gutzwiller KJ (ed) (2002) Applying landscape ecology in biological conservation. Springer-Verlag, New York

He HS, Mladenoff DJ, Sickley T, Gutensburgen GG (2000) GIS interpolation of witness tree records (1839–1866) of northern Wisconsin. J Biogeogr 27:1031–1042

Hooten MB (2001) Modeling the distribution of ground flora on large spatial domains in the Missouri Ozarks. Master's thesis, University of Missouri, Columbia, MO, USA

Hooten MB, Larsen DR, Wikle CK (2003) Predicting the spatial distribution of ground flora on large domains using a hierarchical Bayesian model. Landscape Ecol 18:487–502

Hosmer DW, Lemeshow S (2000) Applied logistic regression. Wiley, New York, USA

Lichstein J, Simmons T, Shriner S, Fanzreb K (2002) Spatial autocorrelation and autoregressive models in ecology. Ecol Monogr 72:446–463

Manies KL, Mladenoff DJ (2000) Testing methods to produce landscape-scale presentment vegetation maps from the U. S. public land survey records. Landscape Ecol 15:741–754

Manies KL, Mladenoff DJ, Nordheim EV (2001) Surveyor bias in forest data of the U. S. General Land Office records for northern Wisconsin. Can J Forest Res 31:1719–1730

McDonald RI, Dean LU (2006) Spatially varying rules of landscape change: lessons from a case study. Landscape Urban Plan 74:7–20

Mladenoff DJ, Dahir SE, Nordheim EV, Schulte LA, Guntenspergen GG (2002) Narrowing historical uncertainty: probabilistic classification of ambiguously identified tree species in historical forest survey data. Ecosystems 5:539–553

Nelson JC, (1997) Presettlement vegetation patterns along the 5th Principal Meridian, Missouri territory, 1815. Am Midland Nat 137:79–94

Nigh TA, Schroeder WA (2002) Atlas of Missouri ecoregions. Missouri Department of Conservation Publication, 212 pp

Porter S (1998) Modeling historic woody vegetation in the lower Ozarks of Missouri. Master's Thesis. University of Missouri, Columbia, MO, USA

Pontius RG, Schneider LC (2001) Land-cover change model validation by an ROC method for the Ipswich Watershed, Massachusetts, USA. Agricult Ecosyst Environ 85:239–248

Royle JA, Wikle CK (2005) Efficient statistical mapping of avian count data. Ecol Environ Stat 12:225–243

Schulte LA, Mladenoff DJ, Burrows SN, Sickley TA, Nordheim EV (2005) Spatial controls of Pre-Euro-american wind and fire in Wisconsin (USA) forests: a multiscale assessment. Ecosystems 8:73–94

USDA Natural Resources Conservation Service (1995) Soil Survey Geographic (SSURGO) Data Base, Data Use Information. USDA Natural Resources Conservation Service, Miscellaneous Publication 1527, 31 pp

Wikle CK (2002) spatial modeling of count data: a case study in modeling breeding bird survey data on large spatial domains. In: Spatial cluster modelling. Chapman and Hall/CRC, London/Boca Raton, FL, pp 199–209

Wikle CK (2003) Hierachical Baysian models for predicting the spread of ecological processes. Ecology 84:1382–1394

Zimmermann N, Kienast F (1999) Predictive mapping of alpine grasslands in Switzerland: species versus community approach. J Veget Sci 10:469–482