Statistically-estimated tree composition for the northeastern United States at the time of European settlement

April 14, 2015

Christopher J. Paciorek 1*, Simon J. Goring 2&, Andrew Thurman 3&, Charles V. Cogbill 4,
John W. Williams 3,5, David J. Mladenoff 6, Jody A. Peters 7, Jun Zhu 8,
and Jason S. McLachlan 7
[author order to be discussed]
[please tell me if you do/don't want your middle initial]

[let me know if there are any issues with how I've listed your affiliation]

1 Department of Statistics, University of California, Berkeley, California, USA

2 Department of Geography, University of Wisconsin, Madison, Wisconsin, USA

3 Department of Statistics, University of Iowa, Iowa City, Iowa, USA

4 Harvard Forest, Harvard University, Petersham, Massachusetts, USA

5 Center for Climatic Research, University of Wisconsin, Madison, Wisconsin, USA

6 Department of Forestry and Wildlife Ecology, University of Wisconsin, Madison, Wisconsin, USA

7 Department of Biological Sciences, University of Notre Dame, Notre Dame, Indiana, USA

8 Department of Statistics, University of Wisconsin, Madison, Wisconsin, USA

* Corresponding author

E-mail paciorek@stat.berkeley.edu (CJP)

& These authors contributed equally to this work.

TODO:

focus on PLOS One, but Nature's Scientific Data is a possibility

check journal requirements for archiving and presentation; talk to Jody/Ann about archiving/DOI

Abstract

We present a data product of the estimated composition of tree taxa at the time of European settlement of the northeastern United States and the statistical methodology used to produce the product. Composition is defined as the proportion of stems larger than approximately 20 cm diameter at breast height in 22 taxonomic groupings, generally at the genus level. The data come from settlement survey records that provide raw data that are transcribed and then aggregated spatially, giving count data. The domain is divided into two regions, eastern (Maine to Ohio) and western (Indian to Minnesota). Public Land Survey point data in the western region are aggregated to a regular 8 km grid, while data in the eastern region, from Town Proprietor Surveys, is aggregated at the township level in irregularly-shaped local administrative units. The product is based on a Bayesian statistical model fit to the count data that estimates composition on a regular 8 km grid. The statistical model allows us to estimate composition at locations with no data and to smooth over noise caused by limited counts in locations with data. Critically, it also allows us to quantify uncertainty in our composition estimates. We expect this data product to be useful for understanding the state of vegetation in the northeastern United States prior to large-scale European settlement. In addition to specific regional questions, the data product can also serve as a baseline against which to investigate how forests and ecosystems change after intensive settlement.

1 Introduction

Historical datasets provide critical context to understand forest ecology. They allow researchers to define 'baseline' conditions for conservation management, to understand ecosystem processes at decadal and centennial scales, and, particularly in regions with widespread land use change, to understand the extent to which forests after conversion and regeneration differ from the original forest cover.

Euro-American settlement and subsequent land use change occured in a time transient fashion across North America, and land surveys that provide vegetation information followed a similar pattern. Early surveys (from 1620 until 1825) in the northeastern United States provide areally-aggregated data at the township level (Cogbill et al., 2002b; Thompson et al., 2013); we refer to these as the Town Proprietor Survey (TPS). These surveys provide composition of major tree species using sometimes inconsistent common names. Later surveys, after the establishment of the U.S. Public Land Survey System (PLS) - from 1832 to 1907 - provide point-level data along a regular grid, with one mile spacing Bourdo (1956); Schulte and Mladenoff (2001); Goring et al. (prep), again with sometimes inconsistent common names. Survey instructions during the PLS varied through time and by point type. This requires the application of spatially-varying correction factors (Cogbill et al., prep; Goring et al., prep) to accurately assess stem density, basal area and biomass from the early settlement records but has limit impact when estimating composition.

Logging, agriculture, and abandonment have left an indelible mark on forests in the northeastern United States Foster et al. (1998); Rhemtulla et al. (2009b); Thompson et al. (2013); Goring et al. (prep). However most studies have assessed these effects in individual states or smaller domains (Rhemtulla et al., 2009a; Friedman and Reich, 2005) and with varying spatial resolution, from townships (36 square miles) to forest zones of hundreds or thousands of square miles. Goring et al. (prep) provide a dataset for the upper Midwest that is resolved to an 8 x 8 km grid size, providing broad spatial coverage at a spatial scale that can be compared to the Forest Inventory and Analysis products Gray et al. (2012). Combined with additional, coarsely-sampled PLS data

from Illinois and Indiana and with the TPS data, this gives us raw data for much of the northeastern United States. However, there are several limitations of using the raw data that can be alleviated by the use of a statistical model to develop a statistically-estimated data product. First, the PLS and TPS data provide only estimates of within-cell variance that do not account for information from nearby locations. Second, the available digitized data from Illinois and Indiana represent a small fraction of those states and missing townships are common in the TPS data.

Properly assessing uncertainty in ecological data is imperative to understanding and modelling ecological processes (Cressie et al., 2009). In this way, a model that can account for the spatial structure of the underlying PLS and TPS data, and provide reliable estimates of uncertainty across the northeastern United States, provides a valuable tool for researchers interested in the ecological structure and function of forests at longer time scales.

2 Data

The raw data were obtained from survey records collated from across the northeastern U.S. by a number of researchers. For the states of Minnesota, Wisconsin, Illinois, Indiana, and Michigan (the western subdomain), data are available at PLS survey point locations and have been aggregated to a regular 8 km grid in the Albers projection. (Note that for Indiana and Illinois, at the moment trees are associated with township centroids and then assigned to 8 km grid cells based on the centroid but in the near future we will have point locations available for each tree.) For the states of Ohio, Pennsylvania, New Jersey, New York and the six New England states (the eastern subdomain), data are aggregated at the township level. There are also data from a single township in Quebec and a single township in northern Delaware. Data are essentially complete in Minnesota, Wisconsin and Michigan, but data in Illinois and Indiana represent a sample of the full set of grid cells, with survey record transcription ongoing. Data for the remaining states are available for a subset of the full set of townships covering the domain. Fig. 2 [Fig. 1 - Lyx is having a numbering problem] shows the domain, indicating the grid cells and townships with data.

Note that surveys occurred over a period of more than 200 years as European colonists (before U.S. independence) and the United States settled what is now the northeastern United States. Our estimates are for the period of settlement represented by the survey data and therefore are time-transgressive; they do not represent any single point in time across the domain, but rather the state of the landscape at the time just before settlement occurred (Whitney, 1996; Cogbill et al., 2002a).

Extensive details on the data are available in Goring et al. (2015) and the raw data are available at XXX [ask Simon/Charlie about public access to Paleon composition data]. The aggregation into taxonomic groups is primarily at the genus level, but is at the species level in some cases of monospecific genera. We model the following 22 taxa plus an "other hardwood" category: Atlantic white cedar (*Chamaecyparis thyoides*), Ash (*Fraxinus spp.*), Basswood (*Tilia americana*), Beech (*Fagus grandifolia)*, Birch (*Betula spp.*), Black gum/sweet gum (*Nyssa sylvatica* and *Liquidambar styraciflua*), Cedar/juniper (*Juniperus virginiana* and *Thuja occidentalis*), Cherry (*Prunus spp.*), Chestnut (*Castanea dentata*), Dogwood (*Cornus spp.*), Elm (*Ulmus spp.*), Fir (*Abies spp.*), Hemlock (*Tsuga canadensis*), Hickory (*Carya spp.*), Ironwood (*Carpinus caroliniana* and *Ostrya virginiana*), Maple (*Acer spp.*), Oak (*Quercus spp.*), Pine (*Pinus spp.*), Poplar/tulip poplar (*Populus spp.* and *Liriodendron tulipifera*), Spruce (*Picea spp.*), Tamarack (*Larix laricina*), Walnut (*Juglans nigra*). Note that in several cases (e.g., black gum/sweet gum, ironwood, poplar/tulip

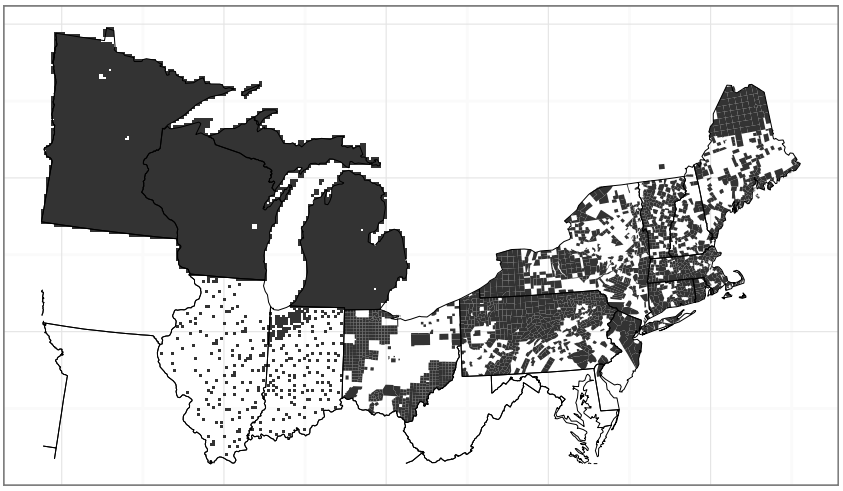


Figure 1: Spatial domain, with locations with data shown in gray. Locations are grid cells in western portion and townships in eastern portion. In addition to locations without data being indicated in white, grid cells completely covered in water are white (e.g., a few locations in Minnesota and Wisconsin).

poplar, cedar/juniper), because of ambiguity in the common tree names used by surveyors, a group represents trees from different families and even orders. For the western subdomain we do not fit statistical models for Atlantic white cedar and chestnut as these have 0 and 6 trees present, respectively. The taxa grouped into the other hardwood category are those for which fewer than roughly 2000 trees were present in the dataset; however we include Atlantic white cedar explicitly despite it only having 336 trees in the dataset. Given the nature of the data, there is not a firm cutoff on the sizes of trees included, but the data roughly represent trees greater than 8 inches (~20 cm) diameter at breast height (dbh), but with some trees as small as 4 inches. [Charlie,Simon, please check this wording] In Indiana and Illinois there are also 402 trees and 29 trees, respectively, that are between 1 and 4 inches dbh, but these will be excluded in the next version of the product.

There are approximately 860,000 trees from the western subdomain and 420,000 trees from the eastern subdomain. In the western subdomain, oak is the most common taxon and pine the second most common, while in the eastern subdomain oak is the most common and beech the second most common.

Our domain is a rectangle covering all of the states, in the Albers projection (NAD 1983 Albers Great Lakes and St. Lawrence), with the rectangle split into 8 km cells, arranged in a 296 by 180 grid of cells, with the centroid of the cell in the southeast corner located at (-71000, 58000). For the modeling of the western subdomain we use the western-most 146 by 180 grid of cells. For the modeling of the eastern subdomain we use the eastern-most 180 by 180 grid of cells and then omit 23 cells in the north and 17 cells in the south outside the states containing data.

3 Statistical model

We fit a Bayesian statistical model to the data, with two primary goals:

1. To estimate composition on a regular grid across the entire domain, filling gaps where no data are available, and

2. To quantify uncertainty in composition at all locations. Even in grid cells and townships with data, we wish to quantify uncertainty because the empirical proportions represent estimates of the true proportions that could be calculated using full population of all the trees in an areal region.

The result of fitting the Bayesian model via Markov chain Monte Carlo (MCMC) is a set of representative samples from the posterior distribution for the composition in the 23 taxonomic groupings at each of the grid cells. These samples are the data product (described further in the Data Product section), and can then be used in subsequent analyses. The mean and standard deviation of the samples for each cell by taxon pair represent our best estimate (i.e., prediction) of composition and a Bayesian "standard error" quantifying the uncertainty in the estimate.

3.1 Data model

We start by describing the basic model for those states for which we have raw data on the 8 km grid and in Section 3.5 we describe the extension of the model to accommodate data aggregated at the township level.

The statistical model treats the observations as coming from a multinomial distribution with a (latent) vector of proportions for each grid cell.

$$y_i \sim \text{Multi}(n_i, \theta(s_i))$$

where y_i is the vector of counts for the P taxa at the ith cell, n_i is the number of trees counted in the cell, and $\theta(s_i)$ is the vector of unknown proportions for those taxa at that cell. Note that we use a standard multinomial distribution without overdispersion as the set of trees in the dataset is roughly uniformly sampled across the cells or townships (Goring et al., 2015).

The proportions, $\theta_p(s_i)$, $p = 1, \ldots, P$, are modeled spatially by a set of P Gaussian spatial processes, one per taxon, $\alpha_p(s_i)$, $p = 1, \ldots, P$. This collection of processes defines a multivariate spatial process for composition. The $\alpha_p(s)$ processes are defined on the 8 km grid, $\alpha_p = \{\alpha_p(s_1), \ldots, \alpha_p(s_m)\}$ for the m grid cells. In Section 3.4 we introduce a multinomial probit model that relates the $\alpha_p(s)$ processes to the proportion processes, $\theta_p(s)$, via the introduction of latent variables, with an implicit sum-to-one constraint, $\sum_{p=1}^{P} \theta_p(s) = 1$.

The critical component of the statistical model is the representation of $\alpha_p(s)$ as a spatial process. This process is a prior structure that serves to smooth across noise in the observations and allows for interpolation to locations with no data. Apart from the sum to one constraints, the taxa are considered to be independent in the prior. We did not want to impose any structure that ties the different taxa together, as any correlation will likely vary across space.

In the next section, we consider two spatial models to define the structure of the $\alpha_p(s)$ processes, a standard conditional autoregressive model (Banerjee et al., 2003) and a Gaussian Markov random field (MRF) approximation to a Gaussian process with Matern covariance (Lindgren et al., 2011).

3.2 Spatial process models

MRF models work directly with the precision matrix of the values of the spatial process, so calculation of the prior density of $\alpha_p(s)$ is computationally simple (Rue and Held, 2005), but in situations where the likelihood is not normal, it can be difficult to set up effective MCMC algorithms that are able to move in the high-dimensional space of α_p. The latent variable representation of Section 3.5 helps to alleviate this problem.

3.2.1 Standard conditional autoregressive models

Our first model is a standard conditional autoregressive (CAR) model (Banerjee et al., 2003). We use a standard form of this model, which treats the four cardinal neighbors of each grid cell as the neighbors of the grid cell. The corresponding precision matrix has diagonal elements, Q_{ii}, equal to the number of neighbors for the ith area (i.e., four except for cells on the boundary of the domain), while $Q_{ik} = -1$ (the negative of a weight of one) when areas i and k are neighbors and $Q_{ik} = 0$ when they are not. This gives the following model for the values of $\alpha_p(s_i)$ collected as a vector across all of the grid cells, $i = 1, \ldots, m$:

$$\alpha_p \sim \mathrm{N}(0, \sigma_p^2 Q^-)$$

The use of the generalized inverse notation indicates that Q is not full-rank, but is of rank $m - 1$; this gives an improper prior on an implicit overall mean for the process values. This specification is called an *intrinsic conditional autoregression (ICAR)* and we can write $Q = D - C$ where C is the $m \times m$ adjacency matrix defining the neighborhood relation of the locations; that is $(C)_{ik} = 1$ if locations i and k are neighbors. The matrix D is an $m \times m$ diagonal matrix containing the row sums of matrix C as the diagonal entries $(D)_{ii} = \sum_{k=1}^{m} (C)_{ik}$.

We refer to this as the *CAR model*.

3.2.2 Gaussian process approximation

Gaussian processes (GP) are also standard models for spatial processes (Banerjee et al., 2003). GP models are computationally challenging for large datasets because of manipulations involving large covariance matrices. Given this, Lindgren et al. (2011) proposed a new framework for using Gaussian MRFs (GMRFs) as approximations to GPs, based on the use of stochastic partial differential equations (SPDE).

We consider Gaussian processes in the Matérn class, using the following parameterization of the Matérn correlation function as

$$R(d) = \frac{1}{\Gamma(\nu) 2^{\nu-1}} \left(\frac{2\sqrt{\nu} d}{\rho} \right)^{\nu} \mathcal{K}_\nu \left(\frac{2\sqrt{\nu} d}{\rho} \right), \tag{1}$$

where d is Euclidean distance, ρ is the spatial range parameter, and $\mathcal{K}_\nu(\cdot)$ is the modified Bessel function of the second kind, whose order is the smoothness (differentiability) parameter, $\nu > 0$. $\nu = 0.5$ gives the exponential covariance. For any pair of locations, $R(d)$ defines the correlation of the process, (i.e., $\alpha_p(s)$ in our context), as a function of the distance between the locations. Considering all pairs of locations, this defines a correlation matrix for all locations of interest.

The Lindgren et al. (2011) approach allows us to consider MRF approximations to the Matern-based GP for $\nu = 1$ and $\nu = 2$. Our second model is the Lindgren approximation for Matern-based GPs with $\nu = 1$. To implement the Lindgren model, one modifies the Q matrix defined previously as follows. Let $a = 4 + \frac{1}{\rho^2}$. The diagonal elements of Q are $4 + a^2$. The entries corresponding to cardinal neighbors are $-2a$. Those for diagonal neighbors are 2, and those for 2nd-order cardinal neighbors are 1. This extends the neighborhood structure relative to the CAR model and parameterizes it as a function of ρ.

The primary difference between the CAR and Lindgren models is that the Lindgren model provides an additional degree of freedom by estimating ρ. In particular ρ allows us to estimate the locality of the smoothing. As ρ decreases, the model uses more and more local data to estimate the compositional proportions at a given location, effectively averaging the empirical proportions over smaller neighborhoods. In general, the Lindgren et al. (2011) model will generally provide for a smoother estimate than the CAR model (Paciorek, 2013).

To ensure that the σ^2 parameter is mathematically equivalent between the two models, we reparameterize, producing our second model:

$$\alpha_p \sim \mathrm{N}\left(\mu_p, \sigma_p^2 \cdot \frac{4\pi}{\rho_p^2} Q(\rho_p)^{-1}\right)$$

We refer to this model as the *SPDE model*.

3.3 Prior Distributions

The ICAR specification contains a set of hyperparameters σ_p^2 for $p = 1, ..., P$. Following Gelman (2006) we use a uniform distribution on each σ_p parameter, with upper bound of [check max values]. For the SPDE model we also have parameters μ_p, which we give flat, non-informative priors (truncated at ± 10), and ρ_p which we give uniform priors on the interval $(0.1, \exp(5))$.

3.4 Latent Variable Model

It is well-known that devising an effective MCMC algorithm for models with latent Gaussian process(es) and a non-Gaussian likelihood is difficult (cite recent cent/noncent papers Rue:Held:2005, Ormerod and Wand 2012 JCGS 21:2Lele et al JASA Dec10Yu:Meng:2011, Tan:Nott:2013). To develop an algorithm, we make use of a latent variable representation for the multinomial probit model (McCulloch and Rossi, 1994). The representation introduces latent variables that allow one to develop a MCMC sampling strategy that takes advantage of closed form full conditional distributions (so-called Gibbs sampling steps) for α_p.

Suppose that compositional counts are available at a number of locations. At location i, a sample size of n_i observations are collected, and each observation (i.e., each tree) can be classified into P distinct categories. For a given tree j at location i, let Y_{ij} denote the response variable indicating the category. Let Y_{ij} be associated with P latent variables $W_{ij1}, ..., W_{ijP}$ such that $Y_{ij} = p$ if and only if $W_{ijp} = \max_{p'} \left\{W_{ijp'}\right\}$; in other words, the maximum of the set of latent variables $\{W_{ijp}\}_{p=1}^{P}$ determines the category of observation j at location i.

The final piece of the latent variable representation is the relationship between the W variables and the $\alpha_p(s)$ processes. We have that

$$W_{ijp} \sim \mathrm{N}(\alpha_p(s_i), 1)$$

independently for all of the W_{ijp} values.
Consider the following example with two locations that are neighbors and $P = 2$ categories. Each tree j at location i is associated with two variables W_{ij1} and W_{ij2}, governed by the latent variables α_{1i} and α_{2i}, respectively. Suppose that $\alpha_{1i} > \alpha_{2i}$ for a given location i. Then this model implies that any tree j is more likely to be labeled 1 than 2 at location i. The difference between α_{1i} and α_{2i} explains the *difference* in probability of *categories* 1 and 2 at location i, and the similarity between α_{p1} and α_{p2} explains the *correlation* between the probabilities at *locations* 1 and 2 for category p.

3.5 Model for township data

We developed an extension of the model described in previous sections to account for data at a different aggregation than our core 8 km grid. This extension introduces a new set of latent variables, one per tree, that indicate the grid cell in which the tree is located. These latent 'membership' variables, c_{tj}, for tree j in township t, $t = 1, \ldots, T$ can be sampled within the MCMC as additional unknown parameters. The prior for c_{tj} is a discrete distribution that puts mass proportional to the areal overlap between the township in which the tree is located and the m grid cells, with the priors across different trees in a township being independent and with most of the .

$$c_{tj} \sim \text{Multinom}(1, \{\psi_1, \ldots, \psi_m\}).$$

Because the townships overlap a limited number of grid cells, most of the $\psi_1, \ldots, \psi_m$ values are zero.

In updating the other parameters in the model during the MCMC, we condition on the current values, c_{tj}, which provides a "soft" assignment of trees to grid cells that respects both the township in which the tree was sampled and the uncertainty in which grid cell the tree was sampled.

Note that this prior has the unrealistic feature that it does not represent our knowledge that the trees in a township would be distributed more regularly across the area of the township than expected by such an independence prior.

3.6 Computation

The McCulloch and Rossi (1994) representation is convenient for MCMC sampling, particularly in this high-dimensional spatial context, as it allows us to draw from the posterior conditional distributions of the W_{ijp} variables (these distributions are truncated normal) in closed form and to draw the entire vector of latent process values for each taxon, α_p, as a single sample that respects the spatial dependence structure for each taxon.

While the latent variable representation provides great advantages in the MCMC sampling for each α_p compared to joint Metropolis updates or updating each location individually, there is still strong dependence between the hyperparameters, $\{\sigma_p^2, \mu_p, \rho_p\}$ and the latent process values (as well as between the latent process values and the latent W_{ijp} variables). To address the first, we developed a 'cross-level' joint updating strategy in which we propose $\phi \in \{\sigma_p\}, p = 1, \ldots, P$ (and for the SPDE model, $\phi \in \{\{\sigma_p\}, \{\mu_p\}, \{\rho_p\}\}$), via a Metropolis-style random walk and then conditional on the proposed value of ϕ^* propose α_p from its full conditional distribution given $\phi^*$and the latent W variables. This is equivalent to sampling from the marginalized (with respect

to α_p) distribution of ϕ conditional on W_p. Note that in sampling $\sigma_p$and ρ_p in the SPDE model, for each p, we jointly propose $\{\sigma_p, \rho_p\}$ using an adaptive Metropolis proposal and then use the cross-level strategy to also propose α_p as part of a single accept-reject step. For these various joint samples of hyperparameters and α_p, we use adaptive Metropolis sampling (Shaby and Wells, 2011).

The full description of the MCMC sampling steps is provided in the Appendix. In addition, in the latent variable representation, $\theta_p(s)$ never appears explicitly and cannot be calculated in close form. Instead we use Monte Carlo integration over W_{ijp}, $p = 1, \ldots, P$ to estimate $\theta_p(s_i)$, also described in the Appendix.

The model is implemented in R R Core Team (2014) with core computational calculations coded in C++ using the *Rcpp* package (Eddelbuettel and Francois, 2011). We also make extensive use of sparse matrix representations and algorithms, using the *spam* package in R (Furrer and Sain, 2010). All code is available on Github, including pre- and post-processing code, at https://github.com/Paleon-Project/composition.

4 Model comparison

4.1 Design

We compared the CAR and SPDE models using cross-validation by holding out data from the fitting process and assessing the fit of the model on the held-out data. This cross-validation used a subregion containing most of Minnesota and a small amount of western Wisconsin, defined to be the cells whose x-coordinate was less than 300,000 (this defines a north-south line that approximately goes through Duluth, Minnesota) and hereafter referred to as the "Minnesota subregion" . We used two types of experiments:

1. We held out all the data from 95% of the cells in the Minnesota subregion, with cells selected at random. This was meant to assess the ability of the model to interpolate from a sparse set of cells/townships and mimics the limited data in Illinois and Indiana.

2. We held out 5% of the trees from all of the trees in the dataset (leaving aside the held-out Minnesota subregion cells). This was meant to assess the ability of the model to estimate the composition in cells in which data were available.

Finally, in a separate sensitivity analysis we instead left out 80% of the cells in Minnesota subregion at random. This variation on the first test above was meant to indicate whether our model comparison conclusions would be robust as the digitization process for Illinois and Indiana progresses and provides us with increasingly dense data.

There has been extensive work in the statistical literature on good metrics to use to compare the predictive ability of models; these metrics are referred to as scoring rules. In particular it is well-recognized that predictive distributions should maximize sharpness subject to calibration. That is the predictive distribution should be as narrow as possible while being calibrated such that the observations are consistent with the distribution (Gneiting et al., 2007). When thinking in terms of prediction intervals as summaries of the predictive distribution, we seek intervals that are as narrow as possible while still covering the truth the expected proportion of the time.

Following the suggestions in Gneiting et al. (2007), we considered the following metrics. For experiment 1, $Y_i = \{Y_{i1}, \ldots, Y_{iP}\}$ is the count of all trees in the held-out cell and for experiment 2, Y_i is the count of held-out individual trees in the cell. For each of these metrics, we calculated the metric in two ways. First, we used the posterior mean composition estimates (as a measure of our core predictions), with $\tilde{\theta}_p(s)$ being the posterior mean. Second, we averaged the metric over the posterior samples (as a measure of our full data product, including uncertainty), taking $\tilde{\theta}_p(s)$ to be an individual MCMC sample and then averaging the metric over all the posterior samples.

1. Brier score: Gneiting et al. (2007) suggest this measure, which has been in use for decades. For multi-category as opposed to binary outcomes, this takes the form

$$\frac{1}{n_i}\sum_{j=1}^{n_i}\sum_{p=1}^{P}(y_{ijp} - \tilde{\theta}_p(s_i))^2$$

   for each cell. The full Brier score for the complete set of held-out locations sums over all locations.

2. Log predictive density: This measure takes the log of the density of held-out observations under the fitted model, $Y_i \sim \text{Multinom}(n_i, \{\tilde{\theta}_1(s_i), \ldots, \tilde{\theta}_P(s_i)\})$, summing on the log scale across all of the held-out data.

   While in principle, this metric should be optimal (Krnjajić and Draper, 2014), it suffers from a lack of robustness in being very sensitive to small predictions near zero (Gneiting et al., 2007). Even worse, our Monte Carlo estimation of θ used 10000 samples, so in some cases $\tilde{\theta}_p(s) = 0$. When a tree is present in a cell but its corresponding proportion is 0, this gives a log density of $-\infty$, preventing use of the metric. As an informal solution to this we set $\tilde{\theta}_p(s) = \frac{1}{100000}$ in such cases, but given these issues we treat the log predictive density as a secondary measure.

3. (Experiment 1 only) Weighted root mean square prediction error (RMSPE) and mean absolute error (MAE), averaging over cells and taxa. We weight by the number of held-out trees in each cell to account for the greater variability in the empirical proportions in locations with few held-out trees.

4. (Experiment 1 only) Coverage and length of 95% prediction intervals for Y_{ip}. We considered only cells with at least 50 trees to focus our assessment on cases where the raw estimates, $\hat{\theta}_p(s)$, were reasonably certain and avoid being strongly influenced by predictive inference for cells where observational variability dominates.

Note that for the calculation of the metric applied to the individual posterior samples, we can estimate the posterior probability that one model has a lower (better) value of the metric than the other model by simply calculating the proportion of samples for which each model has a lower value of the metric.

In our initial exploratory fitting, we noticed that the SPDE model produced boundary effects in the predicted composition near the edges of the convex hull of the observations. To attempt to alleviate this, we added a buffer zone of six grid cells around our entire original domain, although the boundary effects were still evident even after inclusion of the buffer. For the model comparison, for comparability, we included this buffer for both the SPDE and CAR models.

Table 1: Predictive ability based on several predictive metric criteria for the CAR and SPDE spatial models when holding out 95% of entire cells of data in Minnesota. Smaller values are better.

	Posterior mean of metric			Metric of posterior mean predictions	
	CAR model	SPDE model	Posterior Prob. CAR < SPDE	CAR model	SPDE model
Brier	180688	186045	0.98	162813	161559
Negative Log Density	468397	513144	1.00	395685	395998
Mean Absolute Error	0.0364	0.0383	0.98	0.0275	0.0269
Root Mean Square Error	0.0897	0.0959	0.96	0.0647	0.0626

Table 2: Coverage and length of prediction intervals for the CAR and SPDE spatial models when holding out 95% of entire cells of data in Minnesota.

	CAR model	SPDE model
Coverage	0.976	0.978
Mean Interval Length	0.129	0.142
Median Interval Length	0.037	0.033

For the model comparison, we used an earlier version of the dataset [include text if we stick with 0.5 for CV]. Given that part of the goal of the exercise is to understand the ability of the models to interpolate to areas without data and the ongoing nature of digitization, we believe this difference in datasets is not an issue.

We ran each model for 150,000 iterations. After discarding 25,000 iterations for burn-in, we retained a posterior sample of 250 subsampled iterations to reduce post-processing computations and storage needs.

4.2 Results

Here we summarize the results of our cross-validation analyses that inform the choice between the CAR and SPDE models.

4.2.1 Full cell hold-out experiment

Table 1 shows predictive ability for Experiment 1: those cells in the Minnesota subregion held out of the fitting process, for both the mean of the metric applied to the individual posterior samples and for the metric applied to the posterior mean predictions. Table 2 shows performance of the uncertainty estimates. For the posterior distribution over the predictive metric values, the CAR model outperforms the SPDE model, while for the posterior mean predictions, the SPDE model appears to outperform the CAR model, but we do not have any uncertainty estimates for this comparison. Coverage and interval lengths are similar between the two models. From a practical perspective, based on the difference in mean absolute error, which is on a readily interpretable scale, the differences between the models are small.

Table 3 shows the results when the proportion of cells that are held out decreases from 95% to 80%, indicating performance on less sparse data. Table 4 shows performance of the uncertainty

Table 3: Predictive ability based on several predictive score criteria for the CAR and SPDE spatial models when holding out 80% of entire cells of data in Minnesota. Smaller values are better, except for coverage, where values near 0.95 are optimal.

	Posterior mean of score			Score of posterior mean predictions	
	CAR model	SPDE model	Posterior Prob. CAR < SPDE	CAR model	SPDE model
Brier	141546	140211	0.10	130006	129910
Negative Log Density	356985	355640	0.31	312509	312950
Mean Absolute Error	0.0309	0.0297	0.10	0.0225	0.0222
Root Mean Square Error	0.0763	0.0740	0.04	0.0531	0.0528

Table 4: Coverage and length of prediction intervals for the CAR and SPDE spatial models when holding out 80% of entire cells of data in Minnesota.

	CAR model	SPDE model
Coverage	0.981	0.971
Mean Interval Length	0.112	0.103
Median Interval Length	0.028	0.021

estimates. For denser data, unlike the results when data are more sparse, the SPDE model generally outperforms the CAR model, but again differences from a practical perspective, based on mean absolute error, are limited.

4.2.2 Individual tree hold-out experiment

Table 5 shows the predictive ability for Experiment 2. Here we have evidence (posterior probability of 0.94) that the SPDE model is better based on the Brier score.

4.2.3 Choice of spatial model

The differences between models are not consistent across the various comparisons, so there is not a clear choice. In our final data product we use the CAR model, for three reasons. First, the CAR model has modestly better performance when data are sparse, as is still the case for Illinois and Indiana. Second, the model is simpler and easier to explain and computations can be done more quickly. Third, predictions from the SPDE model showed boundary effects, with some taxa

Table 5: Predictive ability based on several predictive score criteria for the CAR and SPDE spatial models when holding out 5% of trees. Smaller values are better.

	Posterior mean of metric			Metric of posterior mean predictions	
	CAR model	SPDE model	Posterior Prob. CAR < SPDE	CAR model	SPDE model
Brier	21296	21277	0.06	21146	21143
Negative Log Density	52199	52052	0.01	51141	51156

showing non-negligible posterior mean values at the edges of the domain, well away from where the taxa were present in the empirical data. This included non-negligible values within (but near the edge of) the convex hull of locations with data.

5 Data product

For the final data product, we ran the model using the CAR specification for 150,000 iterations, discarding the first 25,000 iterations for burn-in and retaining 250 subsampled iterations to limit storage needs and limit the size of the data product. Mixing was generally reasonable, but for some of the hyperparameters was relatively slow, particularly for less common taxa. Despite this, mixing for the variables of substantive interest, the proportions was good. [Report ESS values]

In Figure 5, we show maps of estimated composition for the full domain for several taxa of substantive interest to illustrate the results: beech, cherry, chestnut, elm, hemlock, oak, and pine. These maps contrast the raw data proportions, the posterior means and posterior standard deviations as pointwise estimates of uncertainty. Fig. 5 shows posterior means for all 23 taxa.

The data product is publicly available at XX (insert accession number) under YY license as version 0.3 as of April 2015. The product is in the form of a netCDF-4 file, with dimensions x, y, and MCMC iteration. There is one variable per taxon. The PalEON project will continue to maintain this product, releasing new versions as additional data in Illinois, Indiana and Ohio are digitized.

[comment on stability of hosting once get better feel for this]

6 Discussion

Note that digitization of data from Illinois and Indiana is ongoing, and digitization of additional data from Ohio is planned as well. As a result, at some point we expect to have complete data for the western half of the domain. There will remain some missing townships in the eastern half of the domain.

Given the density of data and the limited differences seen between the CAR and SPDE models, we expect the data product to be reasonably robust to the choice of spatial model, particularly in those areas with complete data. However, additional investigation of other statistical representations is of interest, in particular nonstationary spatial models and use of covariates. The biggest shortcoming of the current model is its inability to account for local features such as rivers (note the Minnesota River floodplain in evidence in the raw data). The current model, by using a simple stationary spatial model that smooths as a function of Euclidean distance, does not account for topographic, soil, or other features.

Other related data products that are under development include:

- The gridded raw count data for the western subdomain and data aggregated to township for the eastern subdomain. The raw data product is available at XXX and that product provides the input data for the product described in this work. Note that we cannot provide the raw data at point locations because of limitations specified in the data use agreements under which we have access to the data.

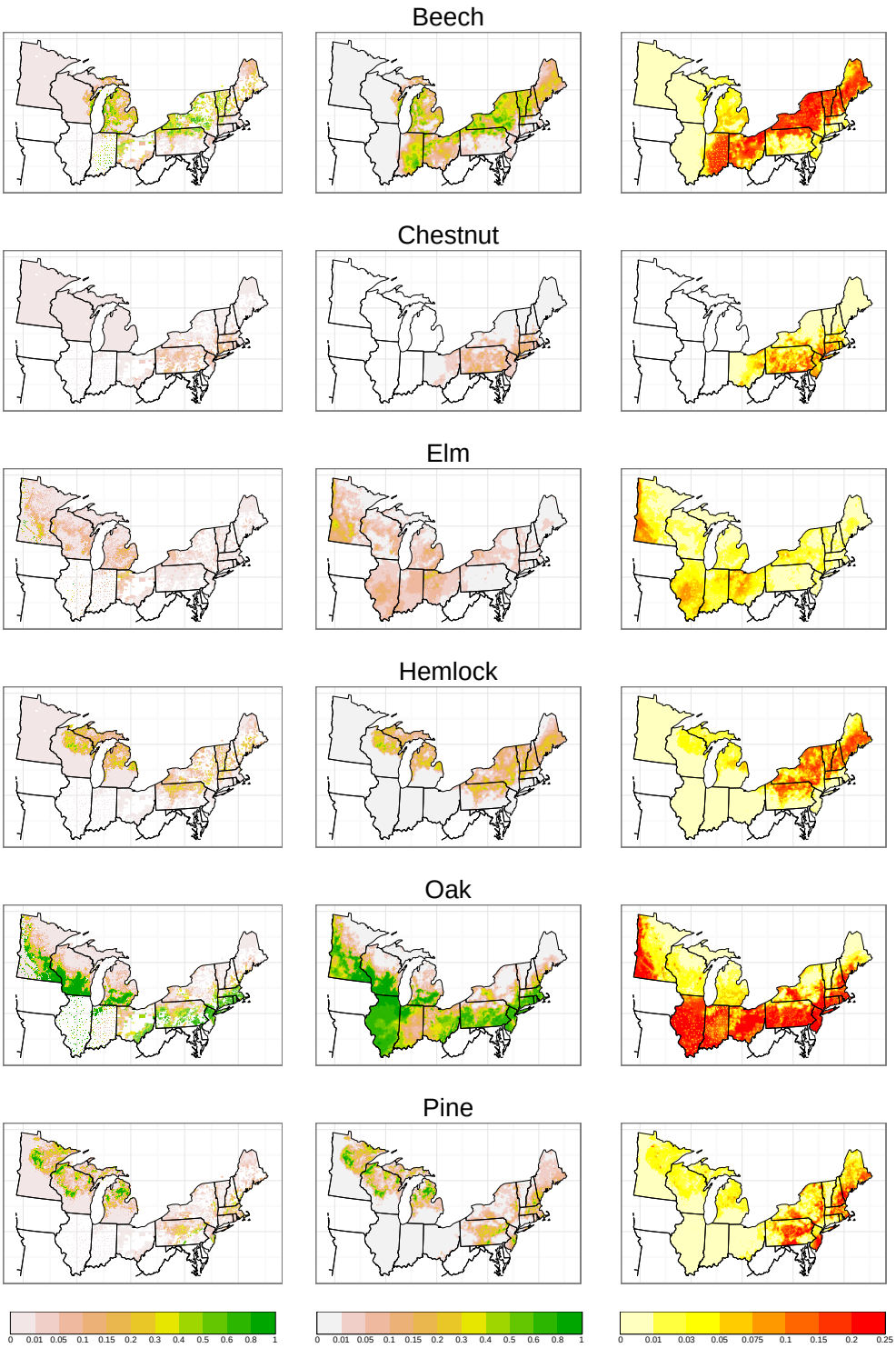
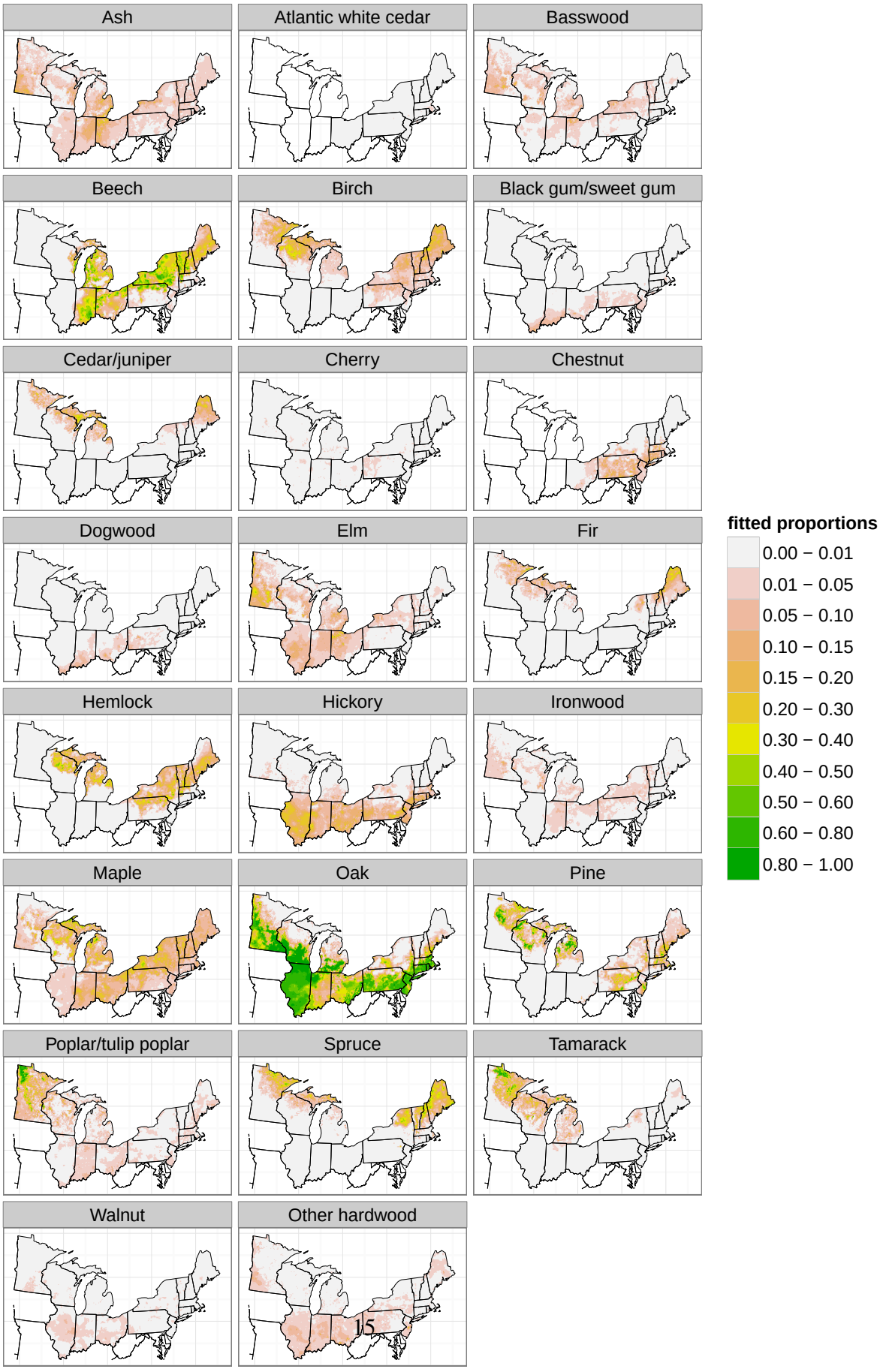


Figure 2: Empirical proportions from raw data (column 1), predictions in the form of posterior means (column 2) and uncertainty estimates in the form of posterior standard deviations – representing standard errors of prediction (column 3) for select taxa.


Ash
Atlantic white cedar
Basswood
Beech
Birch
Black gum/sweet gum
Cedar/juniper
Cherry
Chestnut
Dogwood
Elm
Fir
Hemlock
Hickory
Ironwood
Maple
Oak
Pine
Poplar/tulip poplar
Spruce
Tamarack
Walnut
Other hardwood
fitted proportions
0.00 − 0.01
0.01 − 0.05
0.05 − 0.10
0.10 − 0.15
0.15 − 0.20
0.20 − 0.30
0.30 − 0.40
0.40 − 0.50
0.50 − 0.60
0.60 − 0.80
0.80 − 1.00

- Gridded raw biomass estimates for Minnesota, Wisconsin, and Michigan based on the PLS data, with extension to Illinois and Indiana planned.
- Statistically estimated biomass for Minnesota, Wisconsin, and Michigan using a statistical model applied to the raw biomass estimates, with extension to Illinois and Indiana planned.

An additional drawback of the product is its focus on composition, which does not directly tell us about vegetation structure, in particular does not distinguish between closed forest, savanna, and prairie, of particular note in Minnesota, Wisconsin, Illinois, and into Indiana. The statistically-estimated biomass product mentioned just above will directly inform questions about vegetation structure. Extensions of that product will also estimate basal area and stem density.

Acknowledgments

The authors are deeply indebted to all of the researchers over the years who have preserved, collected, and digitized survey records, in particular Jim Dyer, Peter Marks, Robert McIntosh, and Ed Schools {David/Simon please add names here}{Jason/Jody, should we mention Bowles and Stuckey; if so I need full names}. We thank Madeline Ruid, Ben Seliger, Morgan Ross and Daniel Handel for processing of the southern Michigan data. Indiana and Illinois data were made possible through the hard work of many Notre Dame undergraduates in the McLachlan lab. This work was carried out by the PalEON Project with support from the National Science Foundation MacroSystems Program through grants EF-1065656, EF-1241868, DEB-1241874 and DEB-1241868 and from the Notre Dame Environmental Change Initiative.

References

Banerjee, S., A. Gelfand, and C. Sirmans (2003). Directional rates of change under spatial process models. *Journal of the American Statistical Association 98*(464), 946–954.

Bourdo, E. A. (1956). A review of the general land office survey and of its use in quantitative studies of former forests. *Ecology*, 754–768.

Cogbill, C., J. Burk, and G. Motzkin (2002a). The forests of presettlement New England, USA: spatial and compositional patterns based on town proprietor surveys. *Journal of Biogeography 29*, 1279–1304.

Cogbill, C. V., J. Burk, and G. Motzkin (2002b). The forests of presettlement new England, USA: spatial and compositional patterns based on town proprietor surveys. *Journal of Biogeography 29*(10-11), 1279–1304.

Cogbill, C. V., S. J. Goring, and A. Thurman (in prep). Estimation of robust correction factors for public land survey data.

Cressie, N., C. A. Calder, J. S. Clark, J. M. V. Hoef, and C. K. Wikle (2009). Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. *Ecological Applications 19*(3), 553–570.

Eddelbuettel, D. and R. Francois (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software 40*, 1–18.

Foster, D. R., G. Motzkin, and B. Slater (1998). Land-use history as long-term broad-scale disturbance: regional forest dynamics in central new England. *Ecosystems 1*(1), 96–119.

Friedman, S. K. and P. B. Reich (2005). Regional legacies of logging: departure from presettlement forest conditions in northern minnesota. *Ecological applications 15*(2), 726–744.

Furrer, R. and S. R. Sain (2010). spam: A sparse matrix R package with emphasis on MCMC methods for Gaussian Markov random fields. *Journal of Statistical Software 36*(10), 1–25.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis 1*(3), 515–534.

Gneiting, T., F. Balabdaoui, and A. E. Raftery (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69*(2), 243–268.

Goring, S., J. Wlliams, D. Mladenoff, C. Cogbill, S. Record, C. Paciorek, S. Jackson, M. Dietze, J. Matthes, and J. McLachlan (2015). Changes in forest composition, stem density, and biomass from the settlement era to present in for the upper midwestern united states. *in preparation 0*, 0.

Goring, S. J., J. W. Williams, D. J. Mladenoff, C. V. Cogbill, S. Record, C. J. Paciorek, S. T. Jackson, M. C. Dietze, J. H. Matthes, and J. S. McLachlan (in prep.). Changes in forest composition, stem density, and biomass from the settlement era to present in for the upper midwestern united states. *in prep.*.

Gray, A. N., T. J. Brandeis, J. D. Shaw, W. H. McWilliams, P. D. Miles, et al. (2012). Forest inventory and analysis database of the United States of America (fia). *Vegetation databases for the 21st century.–Biodiversity & Ecology 4*, 255–264.

Krnjajić, M. and D. Draper (2014). Bayesian model comparison: Log scores and DIC. *Statistics and Probability Letters 88*, 9–14.

Lindgren, F., H. Rue, and J. Lindström (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society B 73*, 423–498.

McCulloch, R. and P. E. Rossi (1994). An exact likelihood analysis of the multinomial probit model. *Journal of Econometrics 64*(1), 207–240.

Paciorek, C. (2013). Spatial models for point and areal data using Markov random fields on a fine grid. *Electronic Journal of Statistics 7*, 946–972.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rhemtulla, J. M., D. J. Mladenoff, and M. K. Clayton (2009a). Historical forest baselines reveal potential for continued carbon sequestration. *Proceedings of the National Academy of Sciences 106*(15), 6082–6087.

Rhemtulla, J. M., D. J. Mladenoff, and M. K. Clayton (2009b). Legacies of historical land use on regional forest composition and structure in wisconsin, USA (mid-1800s-1930s-2000s). *Ecological Applications 19*(4), 1061–1078.

Rue, H. and L. Held (2005). *Gaussian Markov Random Fields: Theory and Applications*. Boca Raton: Chapman & Hall.

Schulte, L. A. and D. J. Mladenoff (2001). The original us public land survey records: their use and limitations in reconstructing presettlement vegetation. *Journal of Forestry 99*(10), 5–10.

Shaby, B. and M. Wells (2011). Exploring an adaptive Metropolis algorithm. Technical Report 2011-14, Department of Statistics, Duke University.

Thompson, J. R., D. N. Carpenter, C. V. Cogbill, and D. R. Foster (2013). Four centuries of change in northeastern United States forests. *PloS one 8*(9), e72540.

Whitney, G. G. (1996). *From coastal wilderness to fruited plain: a history of environmental change in temperate North America from 1500 to the present*. Cambridge University Press.

7 Appendix

7.1 MCMC details

Define $\bar{w}_{i\cdot p} = n_i^{-1} \sum_{j=1}^{n_i} W_{ijp}$ be the average of the W values for the pth taxon in the ith grid cell. Let A be a diagonal matrix where A_{ii} is the number of trees in the ith grid cell. When there are no trees in a grid cell, $\bar{w}_{i\cdot p} = 0$ and $A_{ii} = 0$. For the township data, at each iteration, based on the current values of the grid cell membership variables, c_{tj}, trees are aggregated into grid cells and the calculations above can then be carried out.

The conditional distribution for W_{ijp} given the other unknowns in the model and the data is as follows. Let $\text{TN}(a, b, \mu, \tau^2)$ denote the truncated normal distribution with mean parameter μ and variance parameter τ^2, truncated below by a and above by b.

$$W_{ijp} \sim \begin{cases} \text{TN}\big(\max\limits_{p^* \neq y_{ij}} w_{ijp^*}, \infty, \alpha_{y_{ij}}(s_i), 1\big), & \text{if } p = y_{ij} \\ \text{TN}\big(-\infty, w_{ijy_{ij}}, \alpha_p(s_i), 1\big), & \text{if } p \neq y_{ij} \end{cases} \tag{2}$$

In essence, the truncation value is determined by the taxon of the jth tree. For a given p, the W values for all trees in all cells can be sampled in parallel.

The conditional distribution of α_p is

$$\alpha_p \sim \mathrm{N}\left(\left(A + Q_p\right)^{-1} A\bar{w}_p, \left(A + Q_p\right)^{-1}\right). \tag{3}$$

where $Q_p = (\sigma_p^2)^{-1}Q$ for the CAR model and $\left(\sigma_p^2 \cdot \frac{4\pi}{\rho_p^2}\right)^{-1} Q(\rho_p)$ for the SPDE model. For each hyperparameter, $\phi \in \{\{\mu_p\}, \{\log\sigma_p, \rho_p\}\}$, we sampled $\{\phi, \alpha_p\}$ jointly, proposing ϕ as a random walk and, conditional on the proposed value of ϕ, sampling α_p from the distribution just above. The joint proposal is accepted or rejected as a standard Metropolis-Hastings proposal, with adaptation of the proposal (co)variance (Shaby and Wells, 2011). As mentioned previously, for the SPDE model, we jointly proposed $\phi = \{\log\sigma_p, \rho_p\}$ from a bivariate normal distribution to account for the posterior dependence of these parameters, while for the CAR model, $\phi \in \{\{\log\sigma_p\}\}$.

For the township-level data, for a given tree j in township t, we draw the latent tree membership variable, $c_{tj} \in \{1, \ldots, m\}$, from a discrete distribution by normalizing posterior weights, $\{\psi_1 L_1, \ldots, \psi_m L_m\}$, produced by multiplying the prior weights by a likelihood contribution. For cell k, L_k is the density of the latent $W_{tj1}, \ldots, W_{tjP}$ values for the given tree under the condition that $c_{tj} = k$, namely the product of independent normal densities, $W_{tjp} \sim \mathrm{N}(\alpha_p(s_k), 1)$, over $p = 1, \ldots, P$. Thus the posterior reweights the prior based on how consistent the current W values for a tree are with the α values for the candidate grid cells.

7.2 Estimating $\theta_p(s)$ via Monte Carlo integration

In the latent variable representation, $\theta_p(s)$ never appears explicitly and cannot be calculated in close form. Instead we use Monte Carlo integration over W_{ijp}, $p = 1, \ldots, P$ to estimate $\theta_p(s_i)$. The quantity $\theta_p(s_i) = P(W_{ijp} = \max_{p^*} W_{ijp^*})$ defines the probability of taxon p at grid cell i. This requires one to choose the number of Monte Carlo samples, which we set at 10000. For each of the saved MCMC samples, $k = 1, \ldots K$, we estimate $\theta_p^{(k)}(s_i)$ numerically. Specifically, for $t = 1, \ldots, 10000$, we independently draw

$$W_{itp}^{(k)} \sim \mathrm{N}(\alpha_p^{(k)}(s_i), 1),\ p = 1, \ldots, P$$

and estimate

$$\theta_p^{(k)}(s_i) \approx \frac{1}{10000} \sum_{t=1}^{10000} 1(W_{itp}^{(k)} = \max_{p^*} W_{itp^*}^{(k)})$$

where $1(\cdot)$ is the indicator function that evaluates to 1 if the expression is true and 0 if false. In other words, we calculate the proportion of times that the maximum of W_{itp}, $p = 1, \ldots, P$ corresponds to taxon p. Considering $\theta_p^{(k)}(s_i)$, $k = 1, \ldots, K$, we have a sample from the posterior of $\theta_p(s_i)$.