Statistically-estimated biomass for the northeastern United States: modern and Euro-American settlement-era

Christopher J. Pac
[in no particular order] Simon J. Goring, Charles V. Cogbill, Mike Dietze, John W. Williams,

February 9, 2018

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

[2]fill in other authors: check in with David Mladenoff, Kelly Heilman, Dave Moore

[*]Corresponding author; E-mail: paciorek@stat.berkeley.edu

NOTE: much of this text is directly from the composition paper and needs substantial rewording and addition of analogous content for FIA.

Abstract

[Edit to transform to biomass not composition] We present a gridded 8 km-resolution data product of the estimated biomass of tree taxa at the time of Euro-American settlement of the northeastern United States and the statistical methodology used to produce the product from trees recorded by land surveyors. . The data come from settlement-era public survey records that are transcribed and then aggregated spatially, giving count data. The domain is divided into two regions, eastern (Maine to Ohio) and midwestern (Indiana to Minnesota). Public Land Survey point data in the midwestern region (ca. 0.8-km resolution) are aggregated to a regular 8 km grid, while data in the eastern region, from Town Proprietor Surveys, are 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 across the entire domain. The statistical model is designed to handle data from both the regular grid and the irregularly-shaped townships and allows us to estimate composition at locations with no data and to smooth over noise caused by limited counts in locations with data. Critically, the model also allows us to quantify uncertainty in our composition estimates, making the product suitable for applications employing data assimilation. We expect this data product to be useful for understanding the state of vegetation in the northeastern United States prior to large-scale Euro-American 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. The data product is being made available at the NIS data portal as version 1.0.

Keywords: biogeography, biomass, old-growth forests, spatial modeling, Bayesian statistical model, vegetation mapping

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, to track forest responses to shifting climates, 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 occurred in a time-transient fashion across North America and were accompanied by land surveys needed to demarcate land for land tenure and use. Various systems were used by surveyors to locate legal boundary markers, usually by recording and marking trees adjacent to survey markers. These data provide vegetation information that can be mapped and used quantitatively to represent the period of settlement. Early surveys (from 1620 until 1825) in the northeastern United States provide spatially-aggregated data at the township level (Cogbill et al., 2002; Thompson et al., 2013), with typical township size on the order of 200 km^2 and no information about the locations of individual trees; we refer to these as the Town Proprietor Survey (TPS). Later surveys after the establishment of the U.S. Public Land Survey System (PLS) by the General Land Office (GLO) provide point-level data along a regular grid, with one-half mile (800 m) spacing, for Ohio and westward during the period 1785 to 1907

(Bourdo, 1956; Pattison, 1957; Schulte and Mladenoff, 2001; Goring et al., 2016). At each point 2-4 trees were identified, and the common name, diameter at breast height, and distance and bearing from the point were recorded. Survey instructions during the PLS varied through time and by point type. Accounting for this variation requires data screening to maximize consistency among points and the application of spatially-varying correction factors Goring et al. (2016) to accurately assess tree stem density, basal area and biomass from the early settlement records, but the impact on composition estimates is limited (Liu et al., 2011). Surveyors sometimes used ambiguous common names, which requires matching to scientific names and standardization (Mladenoff et al., 2002; Goring et al., 2016).

Logging, agriculture, and land 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., 2016). However most studies have assessed these effects in individual states or smaller domains (Friedman and Reich, 2005; Rhemtulla et al., 2009a) and with various spatial resolutions, from townships (36 square miles) to forest zones of hundreds or thousands of square miles. Goring et al. (2016) provide a new dataset of forest composition, biomass, and stem density based on PLS data for the upper Midwest that is resolved to an 8 km by 8 km grid cell scale, providing broad spatial coverage at a spatial scale that can be compared to modern forests using Forest Inventory and Analysis products (Gray et al., 2012). Combined with additional, coarsely-sampled PLS data from Illinois and Indiana, newly-digitized data from southern Michigan, 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 only provide estimates of within-cell variance that do not account for information from nearby locations. Second, there are data gaps: the available digitized data from Illinois and Indiana represent a small fraction of those states, and missing townships are common in the TPS data. Third, the TPS and PLS data have fundamentally different sampling design and spatial resolution. Our statistical model allows us to provide a spatially-complete data product of settlement-era tree composition for a common 8 km grid with uncertainty across the northeastern U.S.

In Section 2 we describe the data sources, while Section 3 describes our statistical models. In Section **??** we quantitatively compare competing statistical specifications, and in Section 5 we describe the final data product. In Section 6 we discuss the uncertainties estimated by and the limitations of the statistical model, and we list related data products under development.

2 Data

2.1 Public Land Survey

The raw data were obtained from land division survey records collated and digitized from across the northeastern U.S. by a number of researchers (Fig. 1). For the states of Minnesota, Wisconsin, Illinois, Indiana, and Michigan (the midwestern subdomain), digitized 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

Figure 1: Spatial domain of the northeastern United States, with locations with data shown in gray. Locations are grid cells in midwestern 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 the northwestern portion of the domain in the states of Minnesota and Wisconsin).

and the six New England states (the eastern subdomain), data are aggregated at the township level. We make predictions for all of the states listed above; these constitute our core domain. There are also data from a single township in Quebec and a single township in northern Delaware; these data help inform predictions in nearby locations within our core domain, but predictions are not made for Quebec or Delaware. Digitization of PLS data in Minnesota, Wisconsin and Michigan is essentially complete, with PLS data for nearly all 8 km grid cells, but data in Illinois and Indiana represent a sample of the full set of grid cells, with survey record transcription ongoing. Data for the eastern states are available for a subset of the full set of townships covering the domain; the TPS data for some townships were lost, incomplete, or have not been located (Cogbill et al., 2002).

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 and midwestern 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 prior to widespread Euro-American settlement and land use (Whitney, 1996; Cogbill et al., 2002). These forest composition datasets do include the effects of Native American land use and early Euro-American settlement activities (e.g., Black et al., 2006), but it is likely that the imprint of this earlier land use is highly concentrated rather than spatially extensive (Munoz et al., 2014).

Extensive details on the upper Midwest (Minnesota, Wisconsin, Michigan) data and processing steps are available Goring et al. (2016); key elements include the use of only corner points, the use of only the two closest trees at each corner point, spatially-varying correction factors for sampling effort, and a standardized taxonomy table. The lower Midwest (Illinois, Indiana) data were purchased from the Indiana State Archives (Indiana) and Hubtack Document Resources (hubtack.com; Illinois) and processed using similar steps as for the upper Midwest data. Digitization of the Illinois and Indiana data is still underway, so many grid cells contained no data at the time the statistical model was fit. Note that the number of trees per grid cell varies depending on the number of survey points in a cell, with an average of 124 trees per cell. The gridded data at the 8 km resolution for the midwest subdomain are available through the NIS data portal (Goring and University of Wisconsin, 2016). The TPS data were compiled by C.V. Cogbill from a myriad of archival sources representing land division surveys conducted in connection with local settlement and are available through the NIS data portal (Cogbill, 2016a,b).

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 balsamea*), 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*), and Walnut (*Juglans spp.*). Note that in several cases (black gum/sweet gum, ironwood, poplar/tulip poplar, cedar/juniper), because of ambiguity in the common tree names used by surveyors, a group represents trees from different genera or even families. For the midwestern subdomain we do not fit statistical models for Atlantic white cedar and chestnut as these species have 0 and 7 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 because of specific ecological interest in Atlantic white cedar wetlands.

Diameters are only recorded in the PLS data. Although surveyors avoided using small trees, there was no consistent lower diameter limit. The PLS data generally represent trees greater than 8 inches (~20 cm) diameter at breast height (dbh), but with some trees as small as 1 inch dbh (smaller trees were much more common in far northern Minnesota). TPS data have no information about dbh, but the trees were large enough to blaze and are presumed to be relatively large trees useful for marking property boundaries.

There are approximately 860,000 trees in the midwestern subdomain and 420,000 trees in the eastern subdomain. In the midwestern 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 using a metric Albers (Great Lakes and St. Lawrence) projection (PROJ4: EPSG:3175), 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 southwest corner located at (-71000 m, 58000 m). For the midwestern subdomain we use the western-most 146 by 180 grid of cells when fitting the statistical models. For the eastern subdomain we use the eastern-most 180 by 180 grid of cells and then omit 23 rows of cells in the north and 17 rows of cells in the south, as these grid cells are outside of the states containing data.

Data processing steps:

- See Goring et al. for processing of upper Midwest data; similar steps for IL/IN/So MI; result is tree dbh for trees at each PLS point
- See Goring et al. for methods to estimate density at each point based on distance to survey point and various correction factors (cite Cogbill in press)
- scale dbh to biomass per tree
- scale by point density to get biomass per unit area for each taxon present at a point
- aggregate biomass within grid cells:
  - proportion of PLS points within each cell containing a given taxon
  - average biomass per taxon across occupied points

2.2 Forest Inventory and Analysis

We use data from the most recent data collection at each FIA plot in the focal states, with no data prior to 1999 used because of changes in FIA methodology at that point in time, which make it difficult to be sure we are not using multiple surveys from a single plot. Also this limits the degree of time transgressiveness.

Data processing steps:

- download from DataMart
- based on MOU, assign plots to PalEON grid cells
- scale dbh to biomass per tree
- compute total biomass per taxon within each FIA plot
- aggregate biomass within grid cells:
  - proportion of FIA plots within each cell containing a given taxon
  - average biomass per taxon across occupied cells

2.3 Allometric scaling

[tradeoff of taxon resolution vs. ability to use random site effects difficulty of capturing spatial patterns in allometry, so we capture variability in individual tree allometry but have spatial biases and overall bias from uncertainty in allometry given limited spatial info and limited data in allometry papers use of component 6]

3 Statistical model

The major challenge of modeling biomass data is that biomass is positive-valued but continuous, for which limited statistical distributions are available.

In early efforts we considered a Tweedie model but encountered computational difficulties in model convergence and concerns about how well the model fit. Given this we developed a two-stage model to address the challenge of zero inflation in non-negatively valued distributions.

There are many zero-inflated models in the statistical literature, most focusing on count or proportional data [need some lit review]. [look for zero-inflated continuous lit]. Our model was motivated by the biological insight that local conditions may prevent a taxon from occurring in an area even though the taxon may be present at high density nearby. Thus we combine a model for "potential biomass", which reflects the large-spatial-scale patterns in biomass with a model for "occupancy", which reflects the propensity for a given forest stand to contain the taxon. This model allows for zero inflation with a small value for the occupancy model in a given location.

Let $N(s)$ be the number of PLS sample points or FIA plots in grid cell s. Let $n_p(s)$ be the number of points in the cell that have one or more trees of taxon p. Let $\bar{Y}_p(s)$ be the average biomass for taxon p calculated ONLY from the $n_p(s)$ points at which the taxon is present. In other words, $\bar{Y}_p(s) = \frac{1}{n_p(s)} \sum_{i=1}^{n_p(s)} Y_{ip}(s)$ where i indexes sites within cell s, $i = 1, \ldots, n_p(s)$. Let

$m_p(s)$ be the potential (log) biomass process, evaluated at grid cell s, and $\theta_p(s)$ be the occupancy patial process. The biomass in a cell can then be calculated as $b_p(s) = \theta_p(s)\exp(m_p(s))$, namely weighting the average biomass in "occupied patches" by the proportion of patches that have the taxon.

Let's first consider the occupancy model. The likelihood is binomial, $n_p(s) \sim \text{Bin}(N(s), \theta_p(s)$. Note that the occupancy model represents the occupancy of patches within a grid cell, and that $\sum_p \theta_p(s) > 1$ because two taxa will often "occupy" the same patch since most PLS points have two trees. Next consider the (log) biomass process. The likelihood is taken to be normal, $\log\bar{Y}_p(s) \sim \text{N}(m_p(s), \frac{\sigma_p^2}{n_p(s)})$. Note that this likelihood accounts for heteroscedasticity related to the number of points at which the taxon is observed (not the number of PLS points in the cell).

This two-stage model is able to account for structural zeros (the taxon is not present because local conditions prevent it) and the resulting zero inflation through the occupancy model while also able to capture the smooth larger-scale variation in biomass and the differential amounts of information in the face of the large number of zeros and different numbers of sampling points in each grid cell.

Note that $m_p(s)$ is likely to be quite smooth spatially, at least for the PLS data, because when a patch is occupied by a given taxon, the tree is likely to be of adult size, regardless of whether the tree is common in the grid cell. So most of the spatial variation in biomass may be determined by variability in occupancy. The potential biomass is meant to correct for the fact that density and tree size may vary somewhat, but probably not drastically, across the domain.

This model is fit to both PLS and FIA data. For PLS, we have a large number of points in each grid cell, while for FIA we have a small (1-XX FIA plots per cell) number.

We fit the two component models using the penalized splines to model the spatial variation, with the fitting done by the numerically robust generalized additive modeling (GAM) methodology implemented in the R package mgcv.

[discuss possible oversmoothing]

As discussed in Wood (XXXX), one can derive a quasi-Bayesian approach and simulate draws from the quasi-posterior as follows: xxxxx.

We combined draws from the occupancy and potential biomass processes to produce biomass draws for each taxon and for total biomass.

Note that one major drawback of this methodology is that the individual taxon estimates are not constrained to add to a reasonable total biomass because the taxa are fit individually. Further, as was the case in our related modeling of composition, we do not capture correlations between taxa in part to reduce computational bottlenecks and in part

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 the full population of all the trees in a grid cell or township.

At a high level, the statistical model estimates composition across the domain, even in locations with sparse or no data, by combining the raw composition data with the assumption that composi-

tion varies in a smooth spatial fashion across the domain. The information in the data is quantified by the data model, also known as the likelihood. The assumption of smoothness is built into the model by representing the true unknown spatially-varying composition using a statistical spatial process representation that induces smoothing of estimates across nearby locations. This spatial process representation is a form of prior distribution and is a function of model parameters called hyperparameters that determine the correlation structure of the process and are also estimated based on the data.

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 Section 5) and can be used in subsequent analyses. The mean and standard deviation of the samples for each pair of cell and taxon represent our best estimate (i.e., prediction) of composition and a Bayesian "standard error" quantifying the uncertainty in the estimate.

In the remainder of this section we provide the technical specification of the model and of the computations involved in fitting the model.

3.1 Computation

3.2 Assessment of smoothing

[metric for CV should be between log(biomass) and biomass given that we don't care about very small changes in low biomass but also don't want to overemphasize errors in large biomass - show both in terms of MSE for biomass and log(biomass)? also consider prediction coverage]

[do in terms of biomass not (a) presence and (b) biomass when non-zero given the two-step model is just a device to handle data distribution. May need to have bivariate grid of gamma values for both models or it may be that focusing on potential biomass part is enough]

4 Scientific Results

5 Data product

The final data product is a dataset that contains 250 posterior samples of the proportions of each of the 23 tree taxa at each grid cell in the states in our domain of the northeastern United States.

For this final data product, we ran the model using the CAR specification with all of the data (including the data held out in the model comparison analyses) for 150,000 iterations with the same burn-in and subsampling details as described in Section **??**. Based on graphical checks and calculation of effective sample size values, 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, with effective sample sizes for the final product generally near 250.

Maps of estimated composition for the full domain for several taxa of substantive interest illustrate the results, contrasting the raw data proportions, the posterior means, and posterior standard deviations as pointwise estimates of uncertainty (Fig. 2). We also present the posterior means for all 23 taxa (Fig. 3).

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.

Figure 3: Predictions (posterior means) for all taxa over the entire domain.

The data product is publicly available at the NIS Data Portal under the CC BY 4.0 license as version 1.0 as of January 2016 (Paciorek et al., 2016). The product is in the form of a netCDF-4 file, with dimensions x-coordinate, y-coordinate, and MCMC iteration. There is one variable per taxon. In addition, dynamic visualizations of the product using the Shiny tool are available at https://www3.nd.edu/∼paleolab/paleonproject/maps. The PalEON Project (in particular the first author) will continue to maintain this product, releasing new versions as additional data in Illinois, Indiana and Ohio are digitized. 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 midwestern half of the domain.

6 Discussion

A key advance of this work over prior reconstructions of settlement-era vegetation lies in the estimates of uncertainty across the spatial domain. These estimates of uncertainty include the sampling uncertainty within grid cells (as do the within-grid cell estimates of uncertainty available from the raw proportions), but, because this is a spatial model, predictions and their associated uncertainty estimates are also informed by the information content of nearby cells. The maps of standard errors across species (Fig. 2, third column) highlight the advantages of this approach in areas of high data coverage (Minnesota, Wisconsin, Michigan) and in areas of sparse coverage (e.g., Illinois, Indiana, parts of Ohio). Where there are not large gaps in the data, the model provides low and fairly smooth estimates of uncertainty. Uncertainty is generally higher in the eastern subdomain than in the areas of the midwestern subdomain with high data coverage because of missing townships and lower sampling density even in townships with data. In areas of sparse coverage and in areas with low tree density (e.g., southwestern Minnesota), the standard error of our estimates increases appropriately. Nevetheless, these uncertainties surround reasonable estimates of trends in composition. For example, the model does a good job of capturing the oak ecotone in Indiana and Illinois, representing a shift from oak savannas and woodlands to closed mesic forests (Fig. 2). Experiment 1 showed that both models predicted composition at cells with no data reasonably well, mimicking the case of sparsely sampled data and giving confidence in the broad spatial patterns predicted in more poorly sampled regions, particularly those with regular, but sparse sampling that mimic the experiment (Illinois and Indiana, but not Ohio). The apparent blockiness of uncertainty estimates in a few places such as Ohio is caused by spatial gaps and variations in sampling resolution. Absolute uncertainty generally increases with abundance for all taxa (Fig. 2, column 3).

Author contributions

CJP, and [JZ, XF] developed the statistical model and code. CJP carried out the model comparison and created the data product. CJP and CVC wrote the paper with feedback and editing from SJG, etc.. SJG, JAP, CVC, DJM, JSM, and JWW led the processing and analysis of the PLS and TPS data and assisted with interpretation of results.

Acknowledgments

The authors are deeply indebted to all of the researchers over the years who have preserved, collected, and digitized survey records, in particular John Burk, Jim Dyer, Peter Marks, Robert McIntosh, Ed Schools, Ted Sickley, Ronald Stuckey, and the Ohio Biological Survey. We thank Madeline Ruid, Benjamin Seliger, Morgan Ripp 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-1065702, EF-1065656, DEB-1241874 and DEB-1241868 and from the Notre Dame Environmental Change Initiative.

Thanks USFS / FIA MOU.

References

Black, B. A., C. M. Ruffner, and M. D. Abrams (2006). Native American influences on the forest composition of the Allegheny Plateau, northwest Pennsylvania. *Canadian Journal of Forest Research 36*(5), 1266–1275.

Bourdo, E. A. (1956). A review of the General Land Office survey and of its use in quantitative studies of former forests. *Ecology 37*, 754–768.

Cogbill, C. (2016a). Settlement-era tree composition, eastern US: Level 1. Technical Report http://dx.doi.org/10.6073/pasta/e40c1ad172882d5113844296611451f6, Long Term Ecological Research Network.

Cogbill, C. (2016b). Settlement-era tree composition, Ohio: level 1. Technical Report http://dx.doi.org/10.6073/pasta/d09bb56c6af8ef0783cd7e76f8113b34, Long Term Ecological Research Network.

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

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.

Goring, S. and University of Wisconsin (2016). Settlement-era gridded tree composition, midwestern US: Level 1. Technical Report http://dx.doi.org/10.6073/pasta/aa0ef9828d41569a96651b056ad89fb3, Long Term Ecological Research Network.

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 (2016). Changes in forest composition, stem density, and biomass from the settlement era to present in the upper Midwestern United States. *PLOS ONE in press*.

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

Liu, F., D. J. Mladenoff, N. S. Keuler, and L. S. Moore (2011). Broadscale variability in tree data of the historical public land survey and its consequences for ecological studies. *Ecological Monographs 81*(2), 259–275.

Mladenoff, D. J., S. E. Dahir, E. V. Nordheim, L. A. Schulte, and G. G. Guntenspergen (2002). Narrowing historical uncertainty: Probabilistic classification of ambiguously identified tree species in historical forest survey data. *Ecosystems 5*(6), 539–553.

Munoz, S. E., D. J. Mladenoff, S. Schroeder, and J. W. Williams (2014). Defining the spatial patterns of historical land use associated with the indigenous societies of eastern North America. *Journal of Biogeography 41*(12), 2195–2210.

Paciorek, C., S. Goring, A. Thurman, C. Cogbill, J. Williams, D. Mladenoff, J. Peters, J. Zhu, and J. McLachlan (2016). Settlement-era gridded tree composition, northeastern US: Level 2. Technical Report http://dx.doi.org/10.6073/pasta/8544e091b64db26fdbbbafd0699fa4f9, Long Term Ecological Research Network.

Pattison, W. D. (1957). *Beginnings of the American Rectangular Land Survey System, 1784-1800*. Chicago: University of Chicago.

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.

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.

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.