Forest Ecology and Management 271 (2012) 182–198

Contents lists available at SciVerse ScienceDirect

Forest Ecology and Management

journal homepage: www.elsevier.com/locate/foreco

A nearest-neighbor imputation approach to mapping tree species over large areas using forest inventory plots and moderate resolution raster data

B. Tyler Wilson *, Andrew J. Lister, Rachel I. Riemann

Forest Inventory and Analysis, Northern Research Station, USDA Forest Service, 1992 Folwell Avenue, Saint Paul, MN 55108, USA

ARTICLE INFO

Article history:
Received 7 December 2011
Received in revised form 2 February 2012
Accepted 5 February 2012

Keywords:
Nearest-neighbor imputation
Canonical correspondence analysis
MODIS
Vegetation phenology
Forest inventory
Species distribution

ABSTRACT

The paper describes an efficient approach for mapping multiple individual tree species over large spatial domains. The method integrates vegetation phenology derived from MODIS imagery and raster data describing relevant environmental parameters with extensive field plot data of tree species basal area to create maps of tree species abundance and distribution at a 250-m pixel size for the entire eastern contiguous United States. The approach uses the modeling techniques of k-nearest neighbors and canonical correspondence analysis, where model predictions are calculated using a weighting of nearest neighbors based on proximity in a feature space derived from the model. The approach also utilizes a stratification derived from the 2001 National Land-Cover Database tree canopy cover layer. Data pre-processing is also described, which includes the use of Fourier series transformation for data reduction and characterizing seasonal vegetation phenology patterns that are apparent in the MODIS imagery.

A suite of assessment procedures is applied to each of the modeled dataset presented. These indicate high accuracies, at the scales of assessments used, for total live-tree basal area per hectare and for many of the most common tree species found in the study area. The end result is an approach that enables the mapping of individual tree species distributions, while retaining much of the species covariance found on the forest inventory plots, at a level of spatial detail approaching that required for many regional management and planning applications. The proposed approach has the potential for operational application for simultaneously mapping the distribution and abundance of numerous common tree species across large spatial domains.

Published by Elsevier B.V.

1. Introduction

1.1. Background

National and regional maps of forest characteristics have long been of interest in the United States. Types and levels of sophistication of large area forest maps have varied through time, ranging from crude choropleth maps of timber volume produced by Sargent (1884), to satellite-based, national maps of land cover (Homer et al., 2007) and forest biomass (Blackard et al., 2008). Forest maps have been and are being used for many things, including for assessing the risks and impacts associated with pest outbreaks (Krist et al., 2010; Pontius et al., 2010), for assessment of areas that are impacted by large disturbances (Olthof, 2004; Wang and Xu, 2009) or by slow decline (Rogers et al., 2010) for examining the potential impact of urbanization (Riemann et al., 2008) for mapping species richness for conservation planning (McPherson and Jetz, 2007; Ohmann et al., 2007) for studying the impacts of differences in forest management and policy in different ownership types (Ohmann et al., 2007; Uuttera et al., 1998) for landscape scale assessment of suitable habitat for wildlife species (Riordan and Rundel, 2009; Weber and Wolf, 2000) and for state and regional resource assessments (Falk and Mellert, 2011).

Several common themes emerged from review of these and other mapping efforts. First, land use policy decisions are often made at the landscape (broad) scale (McCarter et al., 1998) because landscape processes, like risk of forest pests (Krist et al., 2010) or fire (Rollins et al., 2006), occur over large areas. Second, distribution and abundance information is often needed for individual species as opposed to forest types because individual species can play significant roles in natural systems, may have high economic impact, or may be indicators for ecosystem health (Iverson and Prasad, 1998; Poland and McCullough, 2006; Riemann Hershey, 2000). Third, the maintenance of a realistic species covariance structure across a set of maps of individual species is important because species assemblage information is used in coarse scale modeling of ecosystem processes like response to disturbance (McPherson and Jetz, 2007), urbanization, and climate change (Iverson and Prasad, 1998; Woodall et al., 2009). Finally, maps should include accuracy assessments that indicate utility for multiple scales of use.

* Corresponding author. Tel.: +1 651 649 5189.
E-mail address: barrywilson@fs.fed.us (B.T. Wilson).

0378-1127/$ - see front matter Published by Elsevier B.V.
doi:10.1016/j.foreco.2012.02.002

1.2. Goals, objectives, and justification of methods used in the study

The main goal of the current study is to develop a systematic and operational approach to mapping individual tree species distributions across large spatial domains, along with several example maps and associated accuracy assessments. The broad objectives of the study are to build upon previously documented modeling approaches by taking advantage of a combination of satellite imagery and other raster datasets, forest inventory data, statistical methods, and automated workflows. Specific objectives include the following:

1. *Use extensive, standardized, and well-established training data sources that adequately represent conditions in the analysis area for model building and validation*. Field plot data from the United States Department of Agriculture Forest Service Forest Inventory and Analysis (FIA) program represent a comprehensive, unbiased, and consistent sample dataset of known accuracy that has long been used for mapping (Blackard et al., 2008; Riemann Hershey, 2000; Zhu and Evans, 1994). Similar use of forest inventory plot data has occurred in Canada (Gillis and Leckie, 1993), Europe (Casalegno et al., 2011), parts of the Asia–Pacific region (Bystriakova et al., 2003) and elsewhere.

Using FIA data allows for the incorporation of detailed information on individual trees and site factors, use of well-defined variables, development of consistent mapping workflows, and utilization of the forest inventory system infrastructure such as field guides and reporting tools. FIA data are updated regularly, making possible future map updates using consistent methods.

2. *Incorporate information that reflects landscape composition and configuration in a cost-effective way*. Several mapping studies have used univariate, interpolative approaches like inverse distance weighting, geostatistical approaches like kriging or stochastic simulation (e.g. Hershey and Reese, 1999), or approaches based on highly-detailed, resource-intensive satellite imagery (Homer et al., 2007; Ohmann and Gregory, 2002; Rollins et al., 2006). Univariate approaches often do not show underlying land cover patterns in the final maps, and maps made using highly detailed, resource-intensive satellite imagery for very large areas can be expensive and time-consuming to produce and update, as indicated, for example, by the multi-year time lag between the Landsat-based NLCD project's scheduled updates.

To avoid such difficulties, data from the Moderate Resolution Imaging Spectroradiometer (MODIS) satellite-borne sensor (Justice et al., 1998), coupled with other moderate resolution raster datasets, were used in the current study, following Blackard et al. (2008) and Ruefenacht et al. (2008). The large swath width, moderate pixel size, and high temporal frequency of MODIS data allow users to efficiently produce maps, overcome issues of cloud contamination, and incorporate intra-annual vegetation phenology by using temporal composites (Bradley et al., 2007; DeFries et al., 1997; Moody and Johnson, 2001; Wolter et al., 1995) and time series analysis (Spruce et al., 2011). While some spatial detail is lost by using MODIS compared to Landsat data (250-m vs. 30-m pixel resolution), logistical efficiencies, fewer issues with cloud contamination, and potential benefits of using phenological time series data outweighed the loss of spatial resolution (Nelson et al., 2009). Finally, using coarser spatial resolution raster data mitigates concerns about uncertainty in the location of field data when integrated with these data during modeling. Based on Forest Service field tests, typical Global Positioning System instrumentation used to determine the coordinates of forested sites have an average measurement error ranging from roughly 1–5 m under open canopy to roughly 5–20 m under closed canopy. Again using the earlier comparison to a Landsat pixel, a MODIS pixel has an area that is approximately 70 times larger, but an edge that is only about eight times larger. Therefore, given the uncertainty associated with field plot coordinates, co-registration errors between plots and pixels (i.e. where a field plot is incorrectly associated with a neighboring pixel) are more likely to occur with finer resolution raster datasets.

3. *Base methods on functional relationships seen in ecological systems*. "Black box" empirical approaches like neural networks can often outperform more traditional mapping methods, but Tu (1996) reports several disadvantages of these methods, including the inability to interpret the models, proneness to overfitting, and the empirical nature of model development, which does not necessarily maintain process–level relationships. Ohmann and Gregory (2002) developed mapping methods using canonical correspondence analysis (CCA) (ter Braak, 1986), a constrained ordination technique, to exploit multivariate relationships among species and between species and environmental data to incorporate process–level relationships into predictive modeling. CCA is one of several multivariate ordination methods used by ecologists to order the data found in a contingency table, typically one consisting of multiple species and environmental observations for a list of sites. For such a contingency table, ordination methods arrange species along a set of environmental gradients. CCA is used to create linear combinations of predictors that represent the environmental gradients that explain the greatest amount of variability (inertia) in the multivariate species data (McGarigal et al., 2000). By "tuning" the set of predictor data to the species composition data, the set of maps of predictions of individual species should be more likely to retain their original covariance structure and maintain logical consistency.

Another way to raise the likelihood of maintaining natural relationships in a set of species-level maps is to use a k-nearest neighbor (kNN) approach. The kNN approach to estimation uses similarity (proximity) in a set of predictor variables, the values of which are known for all data points, in order to impute a set of response variables, the values of which are known only for a set of reference data points, to the set of target data points for which the values of the response variables are not known. The response variables from the set of k reference data points that are most similar (nearest) to a given target data point are then summarized (typically, averaged) in order to generate a predicted value. In some cases, the values of the response variables associated with each reference data point are weighted based on their similarity (proximity) to the target data point. The kNN approach is nonparametric, intuitive, and has been broadly applied in forestry and other mapping applications (McRoberts et al., 2002, 2007; Eskelson et al., 2009). For the example where the reference data points consist of FIA plots and the target data points consist of pixels, the kNN approach would utilize the same small pool of k reference plots for each estimate for each species for a given target pixel, thus raising the likelihood of maintaining logical consistency in the species covariance structure of the set of response variables. In theory at least, using kNN together with a set of predictor variables generated from the output of a CCA model (similar to Ohmann and Gregory, 2002) would be expected to do even better at maintaining these complex, natural relationships in the species data.

2. Materials and methods

2.1. Study area

The study area for this investigation is roughly the eastern half of the contiguous United States, which corresponds with the area surveyed by the Northern and Southern FIA programs (Fig. 1). This area was selected for several reasons, including data availability, as well as the challenges the region poses. It includes the states that were the first in the national FIA program in which annualized

Fig. 1. Map of the study area (shaded in gray).

inventories were implemented and for which there is an extensive set of current field plot data. The area spans several ecological regions and represents a variety of landscapes and climatic zones. The states included in the study have sampling intensities that range from one to three times the FIA standard of approximately one plot per 2400 hectares.

2.2. Data sets

2.2.1. Response variables

Field plot data used for model calibration and accuracy assessment were provided by the FIA program. FIA collects data annually from a series of field plots that are located in cells created by a uniform hexagonal tessellation of the country. The FIA inventory is conducted in three phases (Bechtold and Patterson, 2005). In Phase 1, all plots are characterized with respect to land use and ownership class and assigned a stratum that will later be used for post-stratification for variance reduction, using some combination of aerial photography, satellite imagery, and ancillary data. In Phase 2, field crews visit permanent plots that contain a forested land use and collect information on individual trees and site variables. In Phase 3, field crews collect additional information on forest health variables on a 1/16th subset of the Phase 2 plots. The FIA data used in this study included 190,888 Phase 2 plots in the study area collected during the period of 2001–2006, with the exception of Louisiana (2001–2005), and Florida and Texas (2001–2007).

Per-species basal area per hectare of live trees (referred to as live-tree basal area throughout the manuscript), with a diameter measured at breast height (DBH) greater than or equal to 1", was selected as the response variable for two reasons. First, when compared to other measurements such as height, which is difficult to collect precisely, or volume, which is modeled from various tree measurements, basal area is calculated from one of the most reliable field measurements made on a tree, namely DBH. Second, since basal area is generally correlated with crown area (e.g. Lister et al., 2000), it is assumed that it is also correlated with the spectral characteristics of forests that are measurable by satellite-based sensors.

2.2.2. Predictor variables

The study used several raster datasets assumed to be correlated with tree species distribution and vegetation composition. Seasonal vegetation phenology data were derived from the Enhanced Vegetation Index (EVI) of the 250-m MOD13Q1, Version 4, MODIS Terra data product. EVI (Liu and Huete, 1995) is an adjustment to the Normalized Difference Vegetation Index (NDVI) that compensates for soil and atmospheric effects and has been shown to be more sensitive to canopy type and structure than NDVI (Gao et al., 2000). The MODIS data included a subset of 100 typically snow-free, growing season images (i.e. late March through November) from 134 16-day maximum value composite (MVC) images that were collected between January of 2001 and October of 2006, a period that roughly coincides with the field plot data collection period. Climate data were extracted from the Daymet 18-year dataset of mean monthly growing-degree days and total precipitation (Thornton et al., 1997).

Topographic data were derived from the US Geological Survey Elevation Derivatives for National Applications (EDNA) dataset and included elevation, compound topographic index (CTI), and slope-aspect index (SAI). CTI (Moore et al., 1991) is a measure of site wetness based on landscape position and is correlated with several soil characteristics (Gessler et al., 1995). SAI (Frank, 1988) is a measure of site orientation and is correlated with solar insolation and exposure to prevailing winds. Geospatial location data in the form of coordinate eastings and northings were also used. Finally, the study utilized the US Environmental Protection Agency's Level III Ecoregions that were derived from Omernik (1987). Ecoregions were represented by a set of associated dummy variables, whereby each pixel was assigned a value of 1 or 0 based on its location relative to each ecoregion. These were included to account for the effects of any residual spatial trends in the attribute of interest that were not explained by the other predictor variables, i.e. non-stationarity in the underlying model. All rasters were projected to Albers Equal Area projection and resampled, if necessary, to 250-m (6.25 ha) pixels.

2.2.3. Differences in spatial resolution

Comparing values for a 6.25 ha pixel using reference data from a 0.067 ha FIA field plot (see Bechtold and Patterson (2005) for a

complete description of the cluster–plot design) involves relating two datasets describing very different spatial units of land. This difference will be most noticeable when local environmental heterogeneity causes fine-scale spatial variability within the larger 6.25 ha area. In such a case, the 0.067 ha plot may sample a condition that is quite different from the predominant condition associated with the larger area sampled by the pixel. A mixture of forest and non-forest land cover can also occur together, introducing another type of heterogeneity within a 6.25 ha pixel. Both situations will result in an increase in the variability observed between pixel values and field plot values in some areas, whether due to environmental heterogeneity, land use history, or current forest fragmentation.

2.3. Description of the methodology

2.3.1. Pre-processing the predictor variables

In an effort to reduce the dimensionality of the EVI dataset, as well as correct for residual atmospheric effects in the imagery, a slightly modified version of the Fourier-based adjustment step of the FASIR methodology described by Sellers et al. (1994) was conducted. A Fourier series was fit individually for each pixel to the full set of 100 EVI MVC images for the region using a reweighted least squares algorithm. A Fourier series can be expressed as a 2π-periodic polynomial of the form

$$y = a_0 + \sum_{n=1}^{h}(a_n \cos(nt) + b_n \sin(nt)), \quad (1)$$

where t is the time period expressed in radians, h is the number of harmonics in the series, and a_0, a_n, and b_n are coefficients. For the purposes of the current study, a Fourier series with two harmonics was found to be adequate to represent the general form of the time series data.

Because EVI is negatively biased by clouds and snow, the same Fourier series was fit again using weighted regression, giving spurious low values less weight in the regression than high values. A simplified weighting function

$$W_t = 1.5^{r_t/\max(r)} \quad (2)$$

where r_t is the residual from the initial model for compositing period t, was used rather than the step-wise weighting function described by Sellers et al. (1994). The weighted regression resulted in the estimation of five Fourier coefficients for each pixel, in effect reducing by twentyfold the number of variables required to describe the general shape of the corrected EVI time series data.

A similar Fourier procedure transformation was used for the monthly climate data. In this case all 12 months of data were used in a simple regression with equal weighting. Since the Daymet data were originally generated at 1-km pixel resolution, the resultant Fourier coefficient data were resampled to 250-m pixels using a bilinear interpolation algorithm. The EDNA data were originally generated at 30-m pixel resolution. Prior to resampling, a focal median function was applied using a 9 by 9-pixel moving window. These focal median values of elevation, CTI, and SAI were then resampled to 250-m pixels using a nearest-neighbor algorithm (Nelson et al., 2009). Finally, a layer stack was produced of the 21 predictor datasets: five phenology coefficients, five temperature coefficients, five precipitation coefficients, five topographic metrics (including geospatial coordinates), and ecoregion (see Table 1).

2.3.2. Gradient nearest neighbor

Generally following the gradient nearest neighbor (GNN) methodology described by Ohmann and Gregory (2002), the FIA field plot data and the predictor datasets were used together in a canonical correspondence analysis (CCA) model. The original formulation of GNN is a two-step process, and those steps were followed in the current study. First, a CCA model was fit to the data that related the multivariate response variable (i.e. live-tree basal area per hectare of each individual tree species) from the field plots to the set of predictor variables in the raster layer stack. Second, the results from the CCA model were used to impute field plot values of live-tree basal area to pixels based on measures of nearness in the weighted canonical variate space.

Table 1
Overview of predictor variables.

Layer	Category	Variable		
1	Phenology	ma0	← fundamental	Fourier coefficients
2	Phenology	ma1	1st harmonic	
3	Phenology	mb1		
4	Phenology	ma2	2nd harmonic	
5	Phenology	mb2		
6	Temperature	ta0	← fundamental	
7	Temperature	ta1	1st harmonic	
8	Temperature	tb1		
9	Temperature	ta2	2nd harmonic	
10	Temperature	tb2		
11	Precipitation	pa0	← fundamental	
12	Precipitation	pa1	1st harmonic	
13	Precipitation	pb1		
14	Precipitation	pa2	2nd harmonic	
15	Precipitation	pb2		
16	Topography	elevation		
17	Topography	compound topographic index		
18	Topography	slope-aspect index		
19	Topography	coordinate easting		
20	Topography	coordinate northing		
21	Ecoregion	Level III code		

The ade4 package for the R statistical language (Dray and Dufour, 2007) was used to conduct the CCA modeling via eigenanalysis. Due to computational constraints, a model was constructed using a random sample of 25,000 from the total set of field plots. Each of the 56 Level III ecoregion codes was included in the model as a factor. The resultant estimated loadings from the model were applied to the raster layer stack to construct a set of canonical variates. Each canonical variate was then weighted by its respective eigenvalue (i.e. a measure of its relative explanatory power) in order to create a stack of weighted canonical variates that would be used during imputation. Only the first eight canonical variates were used, since the smaller number captured most of the inertia explained by the model and also reduced processing time during the imputation step. Both the CCA and the imputation steps were conducted globally across all states rather than individually by smaller mapping units.

2.3.3. Grouping pixels to reduce processing time

A minimum-distance classification was performed to group together and label similar pixels. The plot identification number of the single nearest-neighboring plot (here called the seed plot) was assigned to each pixel based on proximity in the feature space defined by the eight weighted canonical variates. In this way, pixels are assigned a group label, based on similarity of predictor characteristics. There are as many groups as there are plots, and each pixel in the map is assigned a single group label. It is these groups that are used in the imputation step below.

2.3.4. Sub-pixel stratification during imputation

As discussed earlier in Section 2.2.3, a 6.25 ha MODIS pixel may include both forested and non-forested conditions, while a 0.067 ha field plot would more frequently fall within one condition or the other simply by virtue of its smaller size. Rather than assuming that the entire 6.25 ha pixel can be represented by a single plot

and imputing the plot's attributes to each pixel directly, sub-pixel information was derived from the 30-m NLCD 2001 tree canopy cover dataset. This dataset has been used by some FIA regions during Phase 1 post-stratification for variance reduction, as described by Gormanson et al. (2009). A multi-step procedure was implemented to incorporate this sub-pixel information.

1. *Field plots were assigned to forest or non-forest strata.* The NLCD tree canopy product, which is a 30-m model-based map of estimated tree canopy cover percentage, was classified into two strata: one with less than 25% canopy cover (called the "non-forest" stratum and assigned a value of 0) and the other with 25% or greater canopy cover (called the "forest" stratum and assigned a value of 1). It was then determined to which stratum each field plot belonged by spatial intersection of the plots with this stratification layer. This threshold value was chosen based on a survey of national and international definitions of forest based on canopy cover that generally range from 10–40% (Lund, 2002).

2. *kNN imputation was performed using plots within each stratum to attach basal area estimates to each group.* For plots in each stratum, the k plots that were nearest to each group (i.e. the seed plot for each group) in the feature space defined by the weighted canonical variates from the CCA model, excluding the seed plot itself, were identified using the yaImpute package for R (Crookston and Finley, 2008). The basal area values associated with these nearest neighboring plots were used to compute the weighted mean total live-tree and per-species live-tree basal areas. The weight for each neighboring plot within a stratum assigned to a given plot was calculated using an inverse distance weighting function, and was based on Euclidean distance in feature space, according to

$$W_p = d_p^{-x} \Big/ \sum_{p=1}^{k} d_p^{-x}, \tag{3}$$

where d_p is the distance between the given plot and the neighboring plot p and x is the weighting exponent. Euclidean distance was an appropriate distance metric because the axes of the canonical variate-based feature space are constrained to be orthogonal to one another. This step produced two estimates for each group, i.e. one derived from the k nearest-neighboring plots in each of the two strata (see Section 2.3.6 for a description of the methodology for determining the values for k and x).

3. *A forest stratum proportion layer was developed.* A focal mean function was performed on the 30-m stratification layer from step 1, using a kernel size of 9 by 9 pixels. The focal mean layer was then resampled to 250-m pixel resolution using a nearest-neighbor algorithm in order to determine the approximate proportion of each 250-m pixel that belonged to the "forest" stratum.

4. *The forest stratum proportion layer was used for sub-pixel stratification to compute the basal area estimate assigned to each pixel, within each group, to produce the final map.* The results from step 2 (i.e. the imputation results computed for each group by stratum) were joined with each pixel using the group label as the key in order to calculate a weighted mean estimate of live-tree basal area. The weight was based on the resampled forest stratum proportion layer. For example, if the forest stratum proportion layer indicated that a pixel with a given group label was 75% forest, then the "forest" stratum basal area estimate from step 2 for the group label assigned to that pixel received a weight of 0.75 and the "non-forest" stratum basal area estimate received a weight of 0.25.

2.3.5. Assessment of the results

Each output dataset produced in this study was assessed using a variety of methods designed to identify the location, magnitude and type of error, across a range of scales at which the dataset might be used. The methods used in the assessment protocol are described in full detail in Riemann et al. (2010). Rather than relying on a single metric, the protocol employs a suite of assessments, including multi-scale comparisons of data distributions and spatial agreement of estimates, as well as an examination of spatial and distribution patterns of local differences. All assessments were conducted for maps of total live-tree basal area, and a selection of three species that represent different ecological and distribution patterns. A smaller set of summary metrics was calculated for all 273 tree species found in the study area.

In this study, modeled results were compared to data collected on FIA plots. Although all FIA plots were used in model development, only the 2nd through 8th nearest-neighboring plots were used to estimate basal area for each associated pixel. This assures that each plot contributes only minimally, via the CCA model but not the imputation, to the estimate for the corresponding pixel. Thus, for comparisons between the field plot values and the modeled pixel values with which they are associated, it is assumed that the two datasets are independent.

Assessments for total live-tree and per-species live-tree basal area estimates were performed at three scales defined by tessellation of the area by a hexagonal mesh with various spacing of hexagon centers (area): 50 km (216,500 ha), 100 km (866,000 ha), and 200 km (3.5 million ha). Additionally, assessments for total live-tree basal area were also conducted using a finer hexagonal mesh with hexagon centers spaced at 30 km (78,100 ha). For each scale, model-based and field plot-based estimates were calculated for each hexagon and compared. Field plot-based estimates were calculated from the plots occurring within a hexagon, and model-based estimates were calculated from the value of those pixels that intersected with the FIA plots. Assessing agreement across a range of scales allows a map user to better understand the changes in observed accuracy, i.e. the difference between field plot and modeled estimates, at different scales of observation and use. Fig. 3c indicates the number and distribution of plots within hexagons at each scale. All hexagons used in the assessment were completely within the boundaries of the region.

Each assessment was conducted using several metrics. First, model-based and field plot-based data distributions were compared using empirical cumulative distribution curves. The Kolmogorov–Smirnov statistic (KS) was calculated to quantify the agreement between the distributions of the two datasets, in terms of the maximum distance between their empirical distribution functions. KS is a robust statistic that makes no assumptions about the distribution of data and is independent of scale changes (e.g. Feller, 1948).

Second, model-based and field plot-based local area estimates (i.e. per hexagon) of basal area were compared. Scatterplots were generated for each scale of analysis, and several indices of agreement were calculated to summarize the relationship between the field plot-based and model-based estimates at each scale: agreement coefficient (AC), systematic agreement (AC_{sys}), unsystematic agreement (AC_{uns}), and root mean square error (RMSE) (Ji and Gallo, 2006; Riemann et al., 2010). The agreement coefficient metrics devised by Ji and Gallo (2006) are particularly helpful because they are symmetric (assume error in both datasets), standardized (value range does not change with size of data values), and with the terms AC_{sys} and AC_{uns}, they describe independently both the systematic agreement (proximity to the $y = x$ line) and unsystematic agreement (level of scatter about the reduced major axis (RMA) regression line) present in the scatterplot. The RMA regression line, sometimes referred to as the geometric mean functional relationship, is calculated in a similar way to the ordinary least squares regression line, but with the assumption that there is error in both the x and y axes and is therefore symmetrical regardless of the ordering of the axes. The AC is a combination of AC_{sys} and AC_{uns}. RMSE was also included to provide information on the magnitude of difference in data units (m^2/ha).

Third, confidence intervals were calculated for the estimates based on the field plots found in each hexagon at the 216,500 ha scale (the finest scale used), and the distribution of the frequency at which the model-based estimates fall within these confidence intervals was both mapped and plotted relative to increasing plot mean. These metrics provide information that map consumers can use to gauge the appropriateness of the overall map products for various uses at different scales.

2.3.6. Determining the values of k and x

Because of the innumerable potential combinations of values for k and the inverse distance weighting exponent (x in Eq. (3)), both with and without stratification during the imputation step, the authors adopted an objective and systematic approach to optimize these parameters and determine their relative impact on the final results (e.g. Eskelson et al., 2009). First, ignoring the stratification layer, for each value of k from 1 to 10 a baseline root mean squared error (RMSE) in the predicted mean live-tree basal area for each of the fine-scale hexagonal estimation units (216,500 ha) was computed by using a weighting exponent of 0, resulting in equal weighting of each of the k neighbors. Second, this process was repeated using the stratification layer. Finally, a range of weighting exponents was tested, in increments of .5 in order to find evidence of a local minimum in the associated RMSE. For the set of three points nearest to the local minimum associated with each value of k, a 2nd degree polynomial was fit and the local minimum was estimated to determine the "optimal" weighting exponent. The estimated optimal RMSE and weighting exponent were then calculated for each value of k. These results were used to determine the appropriate weighting factor, value for k, and whether or not to utilize the stratification layer for the mapping procedure.

3. Results and discussion

3.1. Canonical correspondence analysis model

The total inertia in the multivariate response variable, live-tree basal area by tree species, was computed by correspondence analysis, with the CCA model explaining approximately 11% of that total. Given the fact that the predictor variables associated with each pixel were summarized over a 6.25 ha area while the response variables associated with each plot were measured over a .067 ha area somewhere within a pixel, one would expect substantial noise in the relationship between the two, particularly in heterogeneous landscapes. Given these facts, the inertia explained by the CCA model was considered meaningful, since ter Braak and Verdonschot (1995) suggest that for ecological data, the percentage of inertia explained by a CCA model is typically less than 10%, particularly for those data exhibiting a strong gradient or with a strong presence/absence aspect, as is the current case when modeling tree species over large areas. In order to facilitate a better understanding of the relative explanatory power of the predictor variables, an inertia partitioning methodology similar to the one described in Borcard and Legendre (1994) was followed. Because a complete analysis of all combinations of the predictor variables was impractical, they were grouped into the five categories listed in Table 1: phenology, temperature, precipitation, topography, and ecoregion. The inertia partitioning was conducted using these five categories, with the results depicted in Table 2.

Overall, the results suggest that for the current study area, the ecoregion category provides by far the greatest explanatory power of the set of predictor variable categories, with 61% of the explainable inertia, either alone or in combination with other predictor categories. Fully 42% of the explainable inertia due to the model can be attributed to this category alone. This should not be surprising, given that ecoregions are intended to depict spatial patterns in the composition of the biotic and abiotic factors used to define ecosystems. However, it should also be noted that the ecoregion category contains the largest number of predictor variables, with 56 individual ecoregions spanning the study area. Taking this into account, the average total contribution of each predictor variable in the ecoregion category is roughly 1% of the explainable inertia, with an average of 0.75% attributed to each of these variables alone. The average total contribution of predictor variables from the other categories, which each contain only five variables, is 2.04% (0.10% alone) for temperature, 1.95% (0.56%) for phenology, 1.95% (0.21%) for topography, and 1.80% (0.12%) for precipitation. Given that climate, topography, and vegetation are some of the central attributes used to delineate ecoregions, the average explanatory power attributed to the vegetation phenology variables alone is noteworthy in that it is substantially higher the others. These results suggest that the ecoregion and phenology variables, on average, contribute relatively greater explanatory power that cannot be attributed either wholly or in part to other predictor variables in the model.

Table 2
Partitioning of inertia.

	Inertia	Total (%)	Variables	Relative (%)
Overall inertia	69.108	100.00	–	–
Inertia explained by model	7.311	10.58	–	–
Explainable inertia attributed to category:				
Phenology	0.712	9.74	5	1.95
Temperature	0.745	10.19	5	2.04
Precipitation	0.657	8.99	5	1.80
Topography	0.714	9.76	5	1.95
Ecoregion	4.483	61.32	56	1.09
Explainable inertia attributed exclusively to category:				
Phenology	0.204	2.78	5	0.56
Temperature	0.036	0.49	5	0.10
Precipitation	0.044	0.60	5	0.12
Topography	0.077	1.05	5	0.21
Ecoregion	3.057	41.82	56	0.75

3.2. Sub-pixel stratification and the values of k and x

The proposed methodology was implemented to produce modeled results using several combinations of values for k and the inverse distance weighting exponent x, as well as with and without stratification, in order to optimize parameter selection for the imputation step. The results are shown in Fig. 2. As expected, the three graphs of RMSE each show evidence of reaching a minimum for some small value of k relative to the large number of plots used in the imputation. When using equal weighting, a minimum of 1.07 (square meters per hectare) occurred when k equals 4, amounting to a 13.9% reduction in RMSE from k equals 1. The graph of RMSE using equal weights with stratification shows an overall improvement with a minimum of 0.99 at k equals 6, representing an overall reduction in RMSE of 20.5%. With optimal weighting and stratification, an estimated minimum of 0.97 was achieved at k equals 8 and a weighting exponent of 1.88, giving a 22.1% reduction in RMSE. These results indicate that both the value of k and the use of the stratification layer had substantial impacts on the overall RMSE, while the weighting of neighboring plots did not. The modeled results that are assessed in the next section were produced using parameters near the estimated minimum achieved above, with k equals 7, a weighting exponent of 1.75, and stratification, resulting in an RMSE of roughly 0.97.

3.3. Comparative assessment of estimates

Assessment metrics at the 216,500, 816,500, and 3.5 million hectare scales were calculated for all 273 species and results for

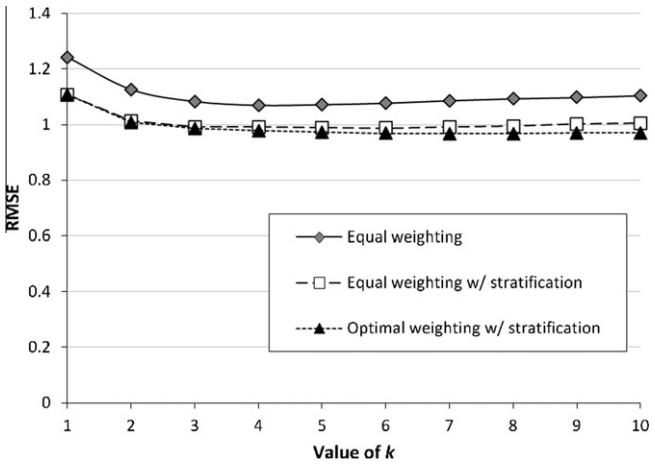

Fig. 2. Graph of root mean squared error (RMSE) in mean total live-tree basal area (m^2/ha) for the 216,500 ha hexagons as a function of the number of neighbors (i.e. value of *k*), weighting exponent, and stratification.

the 100 most common species, based on estimated total live-tree basal area in the study area, are presented in Table 3. In order to facilitate interpretation of the metrics found in the table, detailed results for total live-tree basal area and for three of the individual species are discussed here. The three species selected for detailed reporting (sugar maple, flowering dogwood, and river birch) were chosen to represent a range of distribution patterns, overall prevalence, and ecological niches, all of which can affect different aspects of map production and interpretation. For example, sugar maple represents a relatively common canopy species, although it occurred on only 7.5% of the plots, and can represent a substantial proportion of the basal area in stands where it occurs. In contrast, river birch is a rare species, found on only 0.3% of the plots, and typically constitutes a minor proportion of the stands in which it occurs. Flowering dogwood tends to have very low basal area per hectare values on the plots where it is found, and represents a species that occurs as an understory, rather than a canopy, tree. Therefore, each of the tree species illustrated in the assessment figures and graphs that follow represents a particular condition that could impact modeling results, as well as assessment of those results using FIA plot data.

A map of modeled total live-tree basal area for all 273 species is presented in Fig. 4a and maps for sugar maple, flowering dogwood, and river birch are presented in Figs. 5a, 6a, and 7a, respectively. Little's (1971) species range boundaries were overlaid on the maps of individual species (shown in red on the figures) and generally show good agreement in terms of the spatial distribution of these species.

3.3.1. Total live-tree basal area per hectare

Fig. 3a shows scatterplots, along with associated assessment metrics, of model-based vs. plot-based per-hexagon estimates of total live-tree basal area at four scales: 78,100, 216,500, 866,000, and 3,500,000 ha. Each point on the scatterplot corresponds with the total live-tree basal area estimate for one of the hexagons at the given scale. In each scatterplot, the black line is the $y = x$ line and the green line is the RMA regression line. Similarity measures between plot-based and model-based estimates increase with coarser scales of assessment, as indicated by the scatterplots, cumulative distribution functions (CDFs), and their associated summary metrics (AC, AC_{sys} and AC_{uns}, and RMSE).

The results for the KS metric (Fig. 3b) are more ambiguous, showing improvement with coarser scales (0.13 at 78,100 ha down to 0.036 at 866,000 ha), but slightly poorer results at the coarsest scale (0.064 at 3,500,000 ha). This is likely due to the fact that, because of the irregular shape of the study area boundary, the total area assessed by the mesh of hexagons changes slightly from one scale to the next. This effect is most noticeable near the Florida peninsula, but also to a lesser degree around the Great Lakes and New England (Fig. 3c). Also, at the coarsest scale there are many fewer hexagons used to compute the metric and it is therefore estimated with greater uncertainty than at the finer scales. Because of the strong similarity between the two distributions at all scales, it should be noted that KS values were found always to correspond with the difference between the two distributions at the *y*-intercept.

Fig. 4b and c present differences in model-based and plot-based estimates strictly at the 216,500 ha scale. This is the finest scale of assessment at which there is reasonable confidence in the FIA-derived estimate for each hexagon, with sampling errors averaging 34% of the plot estimate for all hexagons in the study area. Each hexagon at this scale contains an average of 89 plots, and ranges from 0% to 100% forested, based on the NLCD stratification layer. The model-based estimates fall within the plot-based 90% confidence interval over 74 percent of the time at this scale, for hexagons with both high and low basal area values. Hexagons with modeled estimates that are below the confidence interval (underestimation) are much less common than those with modeled estimates that are above (overestimation), at 3% and 23%, respectively. Seven percent of the hexagons had plot-based estimates of zero and modeled estimates that were greater than zero. This is most likely due to the fact that the stratification layer used to determine the forest stratum proportion did not account for all aspects of FIA's definition of forest (Woudenberg et al., 2010). In particular, it did not account for minimum area or width, nor did it account for non-forest land uses where trees occur such as orchards, shelterbelts, and developed areas (Riemann, 2003).

3.3.2. Individual species basal area per hectare

While forest trees occur on 41% of the FIA plots in the study area, individual tree species occurrence on FIA plots ranges from 0% to 17%. Figs. 5–7 present detailed assessment results for sugar maple, flowering dogwood, and river birch. Overall agreement (part b in the Figs. 5–7), as measured by AC values for these three species, ranges from −0.06 to 0.93 at the 216,500 ha scale, 0.68–0.97 at the 866,025 ha scale, and 0.85–0.99 at the 3.5 million ha scale. In general terms, AC values less than 0.5 represent poor agreement, while 1.0 represents perfect agreement. Values for systematic agreement (AC_{sys}) tend to be better, with sugar maple and flowering dogwood having values approaching 1.0 at all scales reported, and river birch having values ranging from 0.79 to 0.99, from finer to coarser scales of assessment. This represents substantial systematic agreement, particularly considering that even sugar maple, a relatively common species in the study area, occurs on only 7.5% of the plots. High systematic agreement indicates that the modeled estimates are not showing any strong bias, from lower to higher basal area, when considering the dataset as a whole. Unsystematic agreement, reflecting the level of scatter about the RMA regression line (Ji and Gallo, 2006), was almost always lower than the other two AC metrics. This indicates that, despite the lack of bias, there is still considerable variability in the basal area estimates not explained by the model, particularly at finer scales. Given the low overall inertia explained by the model at the scale of individual plots, this result was not surprising.

The modeled estimates for the three species examined in detail indicated low levels of species presence at the edges of their respective ranges, while the plot data reported no species presence (part c in the Figs. 5–7). The lack of agreement in these areas could be due in part to poor model performance at the boundaries of

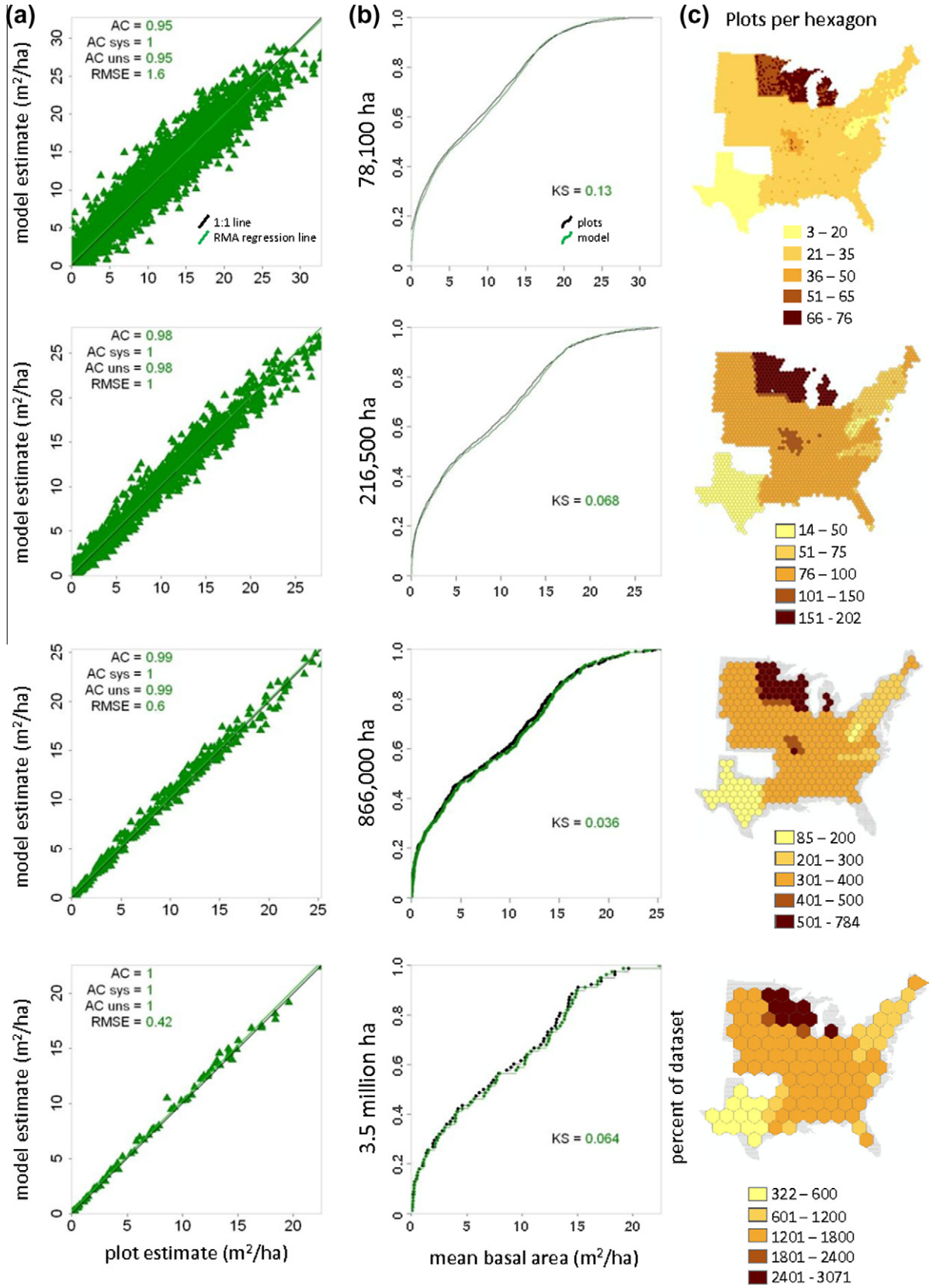

Fig. 3. Scatterplots (a) and cumulative distribution functions (b) of map vs. plot-based estimates of total live-tree basal area (m^2/ha), with maps of plot counts (c) by hexagon for each of four scales. Agreement/error metric values associated with each scale appear on the figures.

species occurrence, likely a result of the way the forest stratum layer was created, as discussed in the previous section on the total live-tree basal area results. Another contributing factor could be the relatively poor accuracy assessment dataset available in these areas, i.e. where the plot-based estimate is based on only a few plots that contain the species. With the individual species datasets,

Table 3
Assessment results.

Tree species	Scale/metric										
	Plot-level		216,500 ha			866,000 ha			3.5 million ha		
	Plots (%)[a]	ba[b]	AC[c]	AC_{sys}[d]	KS[e]	AC	AC_{sys}	KS	AC	AC_{sys}	KS
American basswood	2.6	3.35	0.76	0.97	0.25	0.92	0.99	0.18	0.98	0.99	0.06
American beech	4.1	3.10	0.89	0.99	0.15	0.95	1.00	0.14	0.99	1.00	0.19
American elm	5.6	1.48	0.68	0.97	0.19	0.88	0.99	0.09	0.97	1.00	0.06
American holly	1.3	1.12	0.76	0.95	0.10	0.93	1.00	0.09	0.98	1.00	0.12
American hornbeam	2.2	1.11	0.71	0.99	0.19	0.93	1.00	0.15	0.98	1.00	0.14
American sycamore	1.0	2.81	0.34	0.89	0.23	0.81	0.98	0.15	0.96	0.99	0.14
Ashe juniper	0.3	8.77	0.92	0.98	0.04	0.95	0.96	0.05	0.94	0.95	0.05
baldcypress	0.5	7.57	0.77	0.97	0.10	0.82	1.00	0.10	0.91	0.99	0.13
balsam fir	4.7	3.77	0.96	1.00	0.04	0.97	1.00	0.04	0.99	1.00	0.08
balsam poplar	0.8	2.65	0.88	0.97	0.06	0.97	1.00	0.07	0.88	0.99	0.09
bigtooth aspen	1.9	3.65	0.82	0.99	0.13	0.92	1.00	0.13	0.96	0.99	0.12
bitternut hickory	1.6	1.69	0.39	0.93	0.21	0.81	0.99	0.11	0.95	1.00	0.06
black ash	2.0	3.22	0.83	0.99	0.14	0.95	0.98	0.17	0.98	0.98	0.23
black cherry	7.8	1.69	0.81	0.98	0.10	0.93	0.99	0.08	0.98	1.00	0.09
black hickory	1.2	1.68	0.95	1.00	0.11	0.99	1.00	0.13	1.00	1.00	0.12
black locust	1.3	2.06	0.72	0.98	0.26	0.91	0.99	0.23	0.98	1.00	0.14
black oak	4.6	3.21	0.89	1.00	0.19	0.96	1.00	0.13	0.99	1.00	0.09
black spruce	1.6	4.48	0.92	1.00	0.04	0.97	1.00	0.04	0.98	0.99	0.05
black walnut	1.9	1.96	0.55	0.96	0.24	0.90	1.00	0.21	0.96	1.00	0.13
black willow	0.6	3.14	-0.24	0.71	0.40	0.24	0.72	0.24	0.82	0.92	0.15
blackgum	5.7	1.33	0.85	0.99	0.14	0.96	1.00	0.14	0.99	1.00	0.18
blackjack oak	0.7	1.60	0.65	0.98	0.18	0.91	1.00	0.17	0.96	1.00	0.15
boxelder	1.5	2.15	0.36	0.89	0.33	0.75	0.97	0.17	0.90	0.98	0.08
bur oak	1.3	3.85	0.65	0.96	0.25	0.89	0.99	0.23	0.97	0.99	0.14
cabbage palmetto	0.2	8.35	0.88	0.97	0.01	0.20	0.22	0.01	0.95	0.95	0.01
cedar elm	0.2	2.72	0.55	0.98	0.09	0.86	0.99	0.09	0.98	1.00	0.12
cherrybark oak	1.0	2.46	0.80	0.99	0.13	0.94	1.00	0.12	0.98	1.00	0.12
chestnut oak	2.4	5.74	0.90	0.99	0.09	0.98	1.00	0.11	0.99	1.00	0.10
Chinese tallowtree	0.3	2.41	0.90	0.98	0.08	0.91	1.00	0.09	0.62	0.72	0.09
chinkapin oak	0.8	2.09	0.74	0.97	0.23	0.94	1.00	0.22	0.98	1.00	0.19
common persimmon	1.4	0.70	0.59	0.98	0.12	0.88	1.00	0.12	0.98	1.00	0.18
eastern cottonwood	0.4	6.36	-1.05	0.53	0.39	0.18	0.91	0.31	0.80	0.98	0.15
eastern hemlock	2.5	5.60	0.86	0.99	0.09	0.95	0.99	0.10	0.98	0.99	0.12
eastern hophornbeam	3.3	1.12	0.50	0.93	0.20	0.84	0.98	0.09	0.97	1.00	0.08
eastern red cedar	3.2	2.30	0.81	0.99	0.33	0.93	1.00	0.22	0.99	1.00	0.12
eastern redbud	1.0	0.78	0.57	0.97	0.20	0.88	1.00	0.14	0.98	1.00	0.14
eastern white pine	3.4	5.13	0.87	0.98	0.16	0.94	0.99	0.19	0.97	0.99	0.21
flowering dogwood	3.8	0.84	0.87	1.00	0.13	0.96	1.00	0.11	0.99	1.00	0.13
green ash	4.0	2.45	0.59	0.94	0.23	0.81	0.99	0.11	0.90	1.00	0.09
hackberry	1.5	1.93	0.47	0.94	0.31	0.82	0.99	0.19	0.97	1.00	0.08
hawthorn spp.	0.8	1.00	0.19	0.91	0.36	0.76	0.97	0.26	0.84	0.99	0.13
honey mesquite	0.9	3.62	0.90	1.00	0.02	0.96	1.00	0.01	1.00	1.00	0.01
honeylocust	0.7	1.92	0.32	0.95	0.28	0.85	1.00	0.23	0.98	1.00	0.12
jack pine	0.8	4.87	0.78	0.97	0.11	0.91	1.00	0.18	0.84	0.97	0.24
laurel oak	1.5	3.16	0.85	0.99	0.08	0.98	1.00	0.09	0.98	0.99	0.06
live oak	0.7	4.34	0.84	1.00	0.07	0.96	1.00	0.09	0.98	0.99	0.09
loblolly bay	0.2	3.88	0.28	0.89	0.03	0.84	0.88	0.03	0.97	0.97	0.04
loblolly pine	8.0	9.77	0.97	1.00	0.09	0.99	1.00	0.08	1.00	1.00	0.05
longleaf pine	0.9	4.55	0.86	0.99	0.06	0.95	1.00	0.07	0.99	1.00	0.08
mockernut hickory	4.0	1.47	0.85	1.00	0.15	0.96	1.00	0.13	0.99	1.00	0.13
northern pin oak	0.6	3.81	0.61	0.92	0.14	0.86	0.98	0.19	0.85	1.00	0.17
northern red oak	6.5	3.75	0.87	0.99	0.13	0.95	1.00	0.12	0.99	1.00	0.06
northern white-cedar	2.0	9.09	0.86	0.99	0.08	0.95	1.00	0.09	0.98	1.00	0.10
Osage-orange	0.5	2.56	0.40	0.97	0.25	0.82	1.00	0.26	0.96	0.99	0.23
overcup oak	0.4	3.98	0.55	1.00	0.15	0.82	0.99	0.19	0.94	0.99	0.28
paper birch	4.4	2.30	0.91	0.99	0.09	0.94	0.99	0.11	0.97	0.99	0.19
pecan	0.3	2.62	0.23	0.87	0.21	0.82	0.99	0.15	0.93	1.00	0.12
pignut hickory	3.7	1.77	0.86	0.99	0.16	0.96	1.00	0.11	0.99	1.00	0.13
pin oak	0.2	3.62	0.09	0.83	0.25	0.74	0.96	0.28	0.87	0.97	0.24
Pinchot juniper	0.1	3.69	0.74	0.93	0.03	0.92	0.95	0.02	0.97	0.99	0.01
pitch pine	0.4	3.43	0.92	1.00	0.09	0.78	0.96	0.08	0.94	0.99	0.05
pond pine	0.2	4.76	0.48	0.96	0.04	0.73	0.98	0.04	0.98	0.99	0.06
pondcypress	0.4	7.30	0.87	0.94	0.04	0.97	1.00	0.03	0.99	0.99	0.03
ponderosa pine	0.2	14.05	0.99	1.00	0.08	0.96	0.96	0.13	0.99	0.99	0.19
post oak	3.6	2.80	0.89	1.00	0.14	0.97	1.00	0.13	0.99	0.99	0.17
quaking aspen	5.0	4.65	0.95	1.00	0.11	0.99	1.00	0.13	1.00	1.00	0.21
red maple	16.8	3.51	0.93	1.00	0.12	0.98	1.00	0.13	1.00	1.00	0.13
red mulberry	0.8	1.13	-0.12	0.92	0.28	0.70	0.99	0.17	0.96	1.00	0.12
red pine	1.2	8.55	0.76	0.98	0.17	0.93	0.99	0.20	0.97	0.99	0.22
red spruce	1.4	4.21	0.90	0.99	0.03	0.97	0.99	0.04	0.99	1.00	0.04
redberry juniper	0.1	4.80	0.81	1.00	0.02	0.87	0.97	0.02	0.95	0.95	0.03

Table 3 (*continued*)

Tree species	Scale/metric										
	Plot-level		216,500 ha			866,000 ha			3.5 million ha		
	Plots (%)[a]	ba[b]	AC[c]	AC_{sys}[d]	KS[e]	AC	AC_{sys}	KS	AC	AC_{sys}	KS
river birch	0.4	2.37	-0.06	0.79	0.29	0.68	0.94	0.25	0.85	0.99	0.13
sand pine	0.1	9.26	0.92	1.00	0.03	0.54	0.90	0.03	0.56	0.76	0.03
sassafras	2.4	1.22	0.69	0.99	0.18	0.90	1.00	0.18	0.96	1.00	0.19
scarlet oak	2.3	2.90	0.89	1.00	0.15	0.97	1.00	0.16	0.99	1.00	0.19
shagbark hickory	2.4	1.94	0.64	0.99	0.16	0.87	1.00	0.13	0.97	1.00	0.10
shortleaf pine	2.6	3.77	0.95	1.00	0.12	0.97	1.00	0.10	0.97	1.00	0.12
silver maple	0.6	7.10	-0.14	0.74	0.31	0.59	0.93	0.24	0.79	0.96	0.13
slash pine	1.7	8.23	0.96	1.00	0.06	0.99	1.00	0.05	1.00	1.00	0.04
slippery elm	2.1	1.23	0.60	0.97	0.25	0.88	1.00	0.16	0.97	1.00	0.13
sourwood	2.4	1.40	0.90	0.99	0.11	0.95	1.00	0.12	0.98	1.00	0.13
southern red oak	3.3	2.12	0.85	1.00	0.09	0.96	1.00	0.11	0.99	1.00	0.17
striped maple	0.9	0.99	0.86	0.99	0.09	0.96	1.00	0.11	0.99	1.00	0.12
sugar maple	7.5	5.09	0.93	1.00	0.16	0.97	1.00	0.16	0.99	1.00	0.13
sugarberry	0.9	2.47	0.79	0.97	0.19	0.95	1.00	0.12	0.97	1.00	0.08
swamp tupelo	1.3	5.15	0.78	0.99	0.10	0.95	1.00	0.11	0.99	1.00	0.08
sweet birch	1.5	2.60	0.83	0.98	0.06	0.91	0.98	0.04	0.98	0.99	0.03
sweetbay	1.1	2.51	0.76	0.97	0.07	0.95	1.00	0.11	1.00	1.00	0.13
sweetgum	8.1	3.06	0.94	1.00	0.10	0.98	1.00	0.13	0.99	1.00	0.17
tamarack	1.1	3.65	0.89	1.00	0.08	0.96	1.00	0.08	0.88	0.95	0.10
Virginia pine	1.5	3.53	0.81	1.00	0.14	0.95	1.00	0.12	0.99	1.00	0.12
water oak	4.8	2.41	0.92	1.00	0.07	0.98	1.00	0.07	1.00	1.00	0.06
water tupelo	0.3	11.52	0.71	0.97	0.11	0.81	0.99	0.09	0.85	0.96	0.12
white ash	4.9	2.22	0.82	0.99	0.18	0.94	1.00	0.14	0.98	1.00	0.10
white oak	8.4	3.75	0.92	1.00	0.13	0.97	1.00	0.09	0.98	1.00	0.09
white spruce	1.7	2.27	0.79	1.00	0.10	0.92	1.00	0.11	0.99	0.99	0.18
willow oak	1.0	2.81	0.64	0.98	0.11	0.91	1.00	0.09	0.96	1.00	0.17
winged elm	3.2	1.07	0.85	1.00	0.13	0.95	1.00	0.14	0.98	1.00	0.22
yellow birch	3.0	2.77	0.89	0.98	0.10	0.94	0.98	0.12	0.96	1.00	0.17
yellow-poplar	5.3	3.95	0.88	0.99	0.14	0.97	1.00	0.16	0.99	1.00	0.21
Median of 100 species	1.5	2.98	0.82	0.99	0.13	0.94	1.00	0.12	0.98	1.00	0.12

[a] Percentage of plots on which the species occurs.
[b] Mean basal area (m^2/ha) for the species on plots where the species occurs.
[c] Agreement coefficient (larger values indicate better agreement, max=1).
[d] Systematic agreement (larger values indicate better agreement, max=1).
[e] Kolmogorov–Smirnov statistic (smaller values indicate better agreement, min=0).

plot-based sampling errors at the 216,500 ha scale were much higher relative to those of the total live-tree basal area estimates. These ranged from a mean sampling error of 62% for sugar maple to 112% for river birch, as compared to 34% for total live-tree basal area. This suggested that much of the disagreement between model-based and plot-based estimates could be due to uncertainty with the FIA plot-derived estimates as much as due to inaccuracies in the modeled dataset.

The covariance of a purposive sample of eight of the species, representing a range of spatial distribution patterns, was assessed by a series of joint distribution scatterplots generated at the 216,500 ha scale (Fig. 8). These scatterplots indicate that the covariance structure amongst these species that is apparent in the plot-based estimates was generally retained by the model-based estimates at this scale of assessment. Species that are often observed on the same plot in the field, such as sugar maple and beech, are more likely to be predicted to co-occur within a pixel by the model. Similarly, species that are seldom observed together in the field, such as black cherry and river birch, are less likely to co-occur within a pixel.

Each of the 273 tree species present in the field plot data used in the study has different spatial and ecological characteristics that can affect the certainty with which it can be modeled at a given scale from the set of input data used in the study. Species which occur primarily as understory species, or occur as a minor proportion of the stands in which they occur, will have a less distinct spectral signature than species that occur as dominant in the canopy. Rare species occurred on fewer plots and thus probably did not provide adequate training data to build a robust model because existing training data did not necessarily represent the full range of conditions in which the species occurs. Similarly, species that do not have strong relationships with either biotic or abiotic factors are also difficult to model. As the results from this study seem to indicate, modeling many species at once as covariate response variables (i.e. species assemblages) may have the added benefit of improving the accuracy of some species by taking advantage of relationships between species in their distributions.

Given these per-species uncertainties, caution should be shown in interpreting and utilizing the results for each species. As a conservative starting point, the results in Table 3 are limited to the 100 most abundant species in the study area, and these generally show encouraging results, particularly at the coarsest scales of assessment. In general, the trend for almost all of these species is for AC metrics to increase with coarser scales of assessment, with median values of 0.82 for AC and 0.99 for AC_{sys} at the finest scale rising to 0.98 and 1.0, respectively at the coarsest scale. There are just a few species on the list that do not follow this pattern, e.g. cabbage palmetto and sand pine. However, in all such cases, the species in question have ranges located near the boundary of the study area and their results are therefore affected by differences in the area assessed by the hexagons at different scales.

The species results for the KS metric suggest that it is relatively insensitive to the scale of assessment, with median values ranging from 0.13 at the finest scale to 0.12 at the coarsest scale. There are individual species on the list that indicate either a positive or negative trend across scale; however, these again appear to be due to where the species range falls relative to the study area boundary. Species that have ranges located near the study area boundary, even abundant ones such as yellow-poplar, will generally show a trend of increasing KS values with coarser scales, while those with

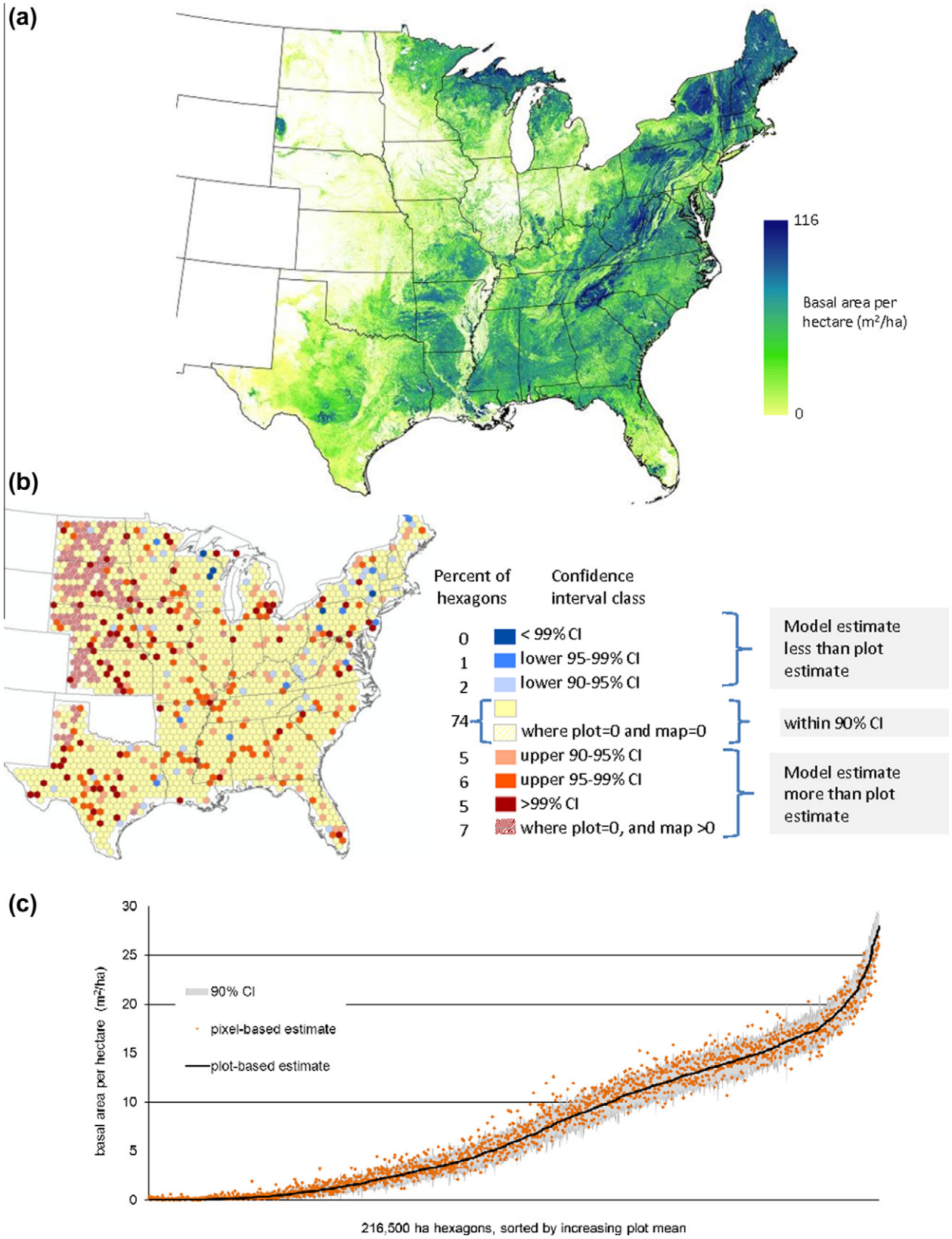

Fig. 4. Map of total live-tree basal area, ranging from low (light yellow) to high (dark green) (a), with map of differences in estimates (b) and graph of 90% confidence intervals and model estimates (c) at 216,500 ha scale.

ranges that fall more centrally in the study area will show the opposite trend, such as northern red oak. The KS metric, which measures the maximum deviation between the plot-based and model-based CDFs, was always found, regardless of species, to correspond with proportion of hexagons where the plot-based estimate was zero and the model-based estimate was non-zero.

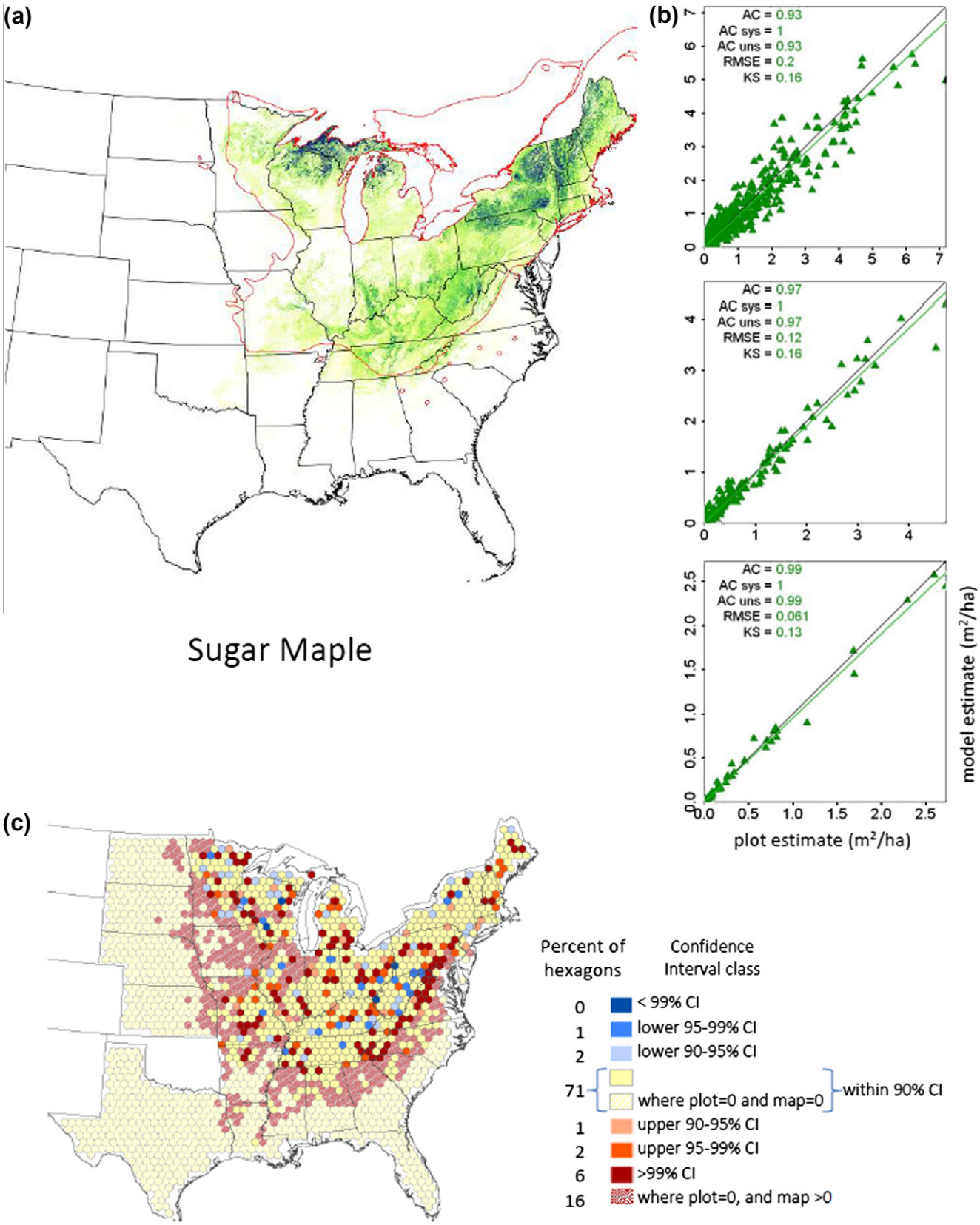

Fig. 5. Map of sugar maple live-tree basal area, ranging from low (light yellow) to high (dark green), and range boundary (red outline) (a), with scatterplots of estimates at three scales (b) and map of differences at 216,500 ha scale (c).

This last result, coupled with the finding that KS is generally insensitive to scale, suggests a systematic issue with the methodology. There are a number of potential explanations for this. One is the difference in spatial resolution between the plots and pixels,

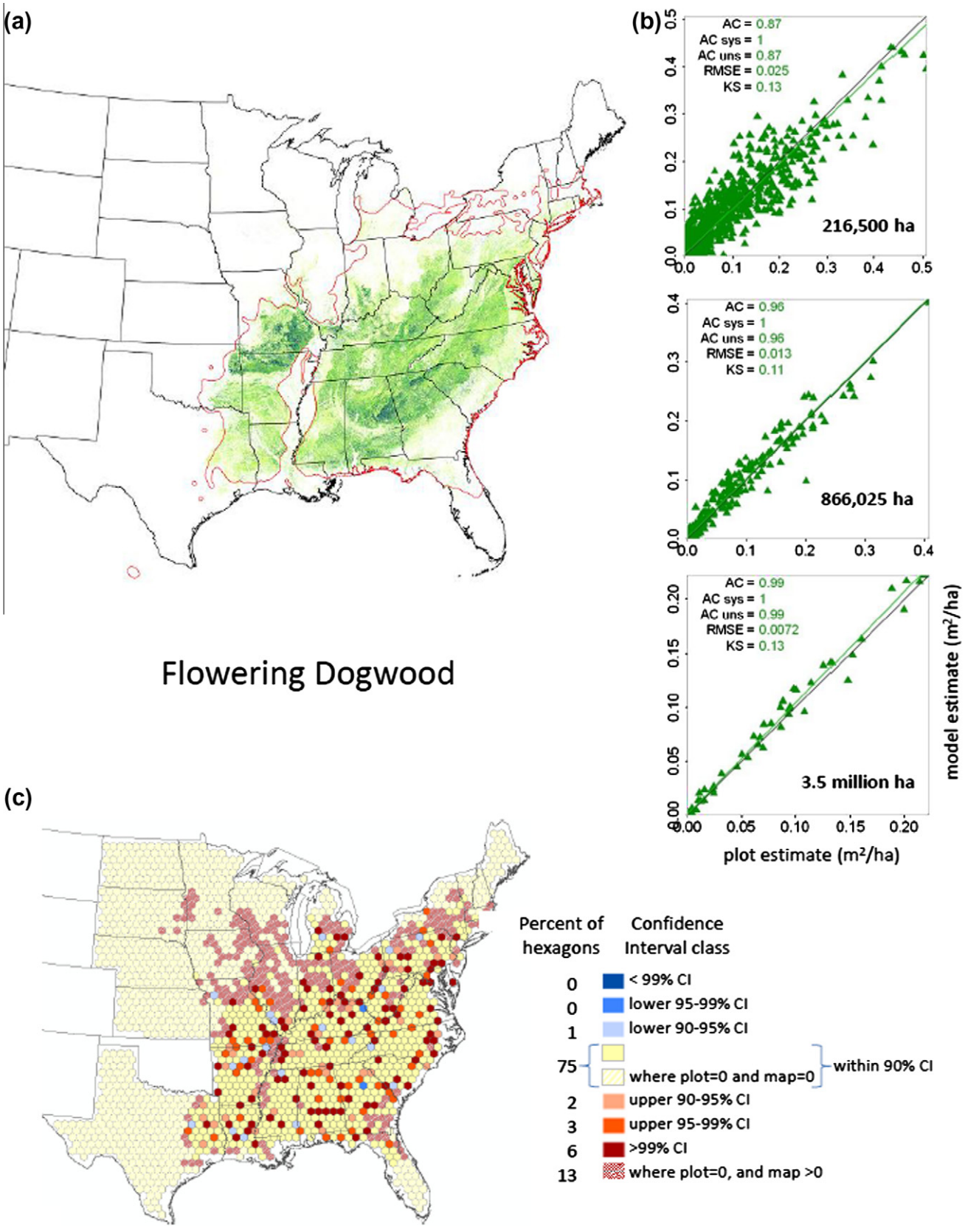

Fig. 6. Map of flowering dogwood live-tree basal area, ranging from low (light yellow) to high (dark green), and range boundary (red outline) (a), with scatterplots of estimates at three scales (b) and map of differences at 216,500 ha scale (c).

as discussed in Section 2.2.3, where fine-scale spatial variability within an individual pixel results in poorer agreement between the plot-based and model-based estimate for that pixel. Another is the choice of the value of *k*. Using kNN for estimation represents

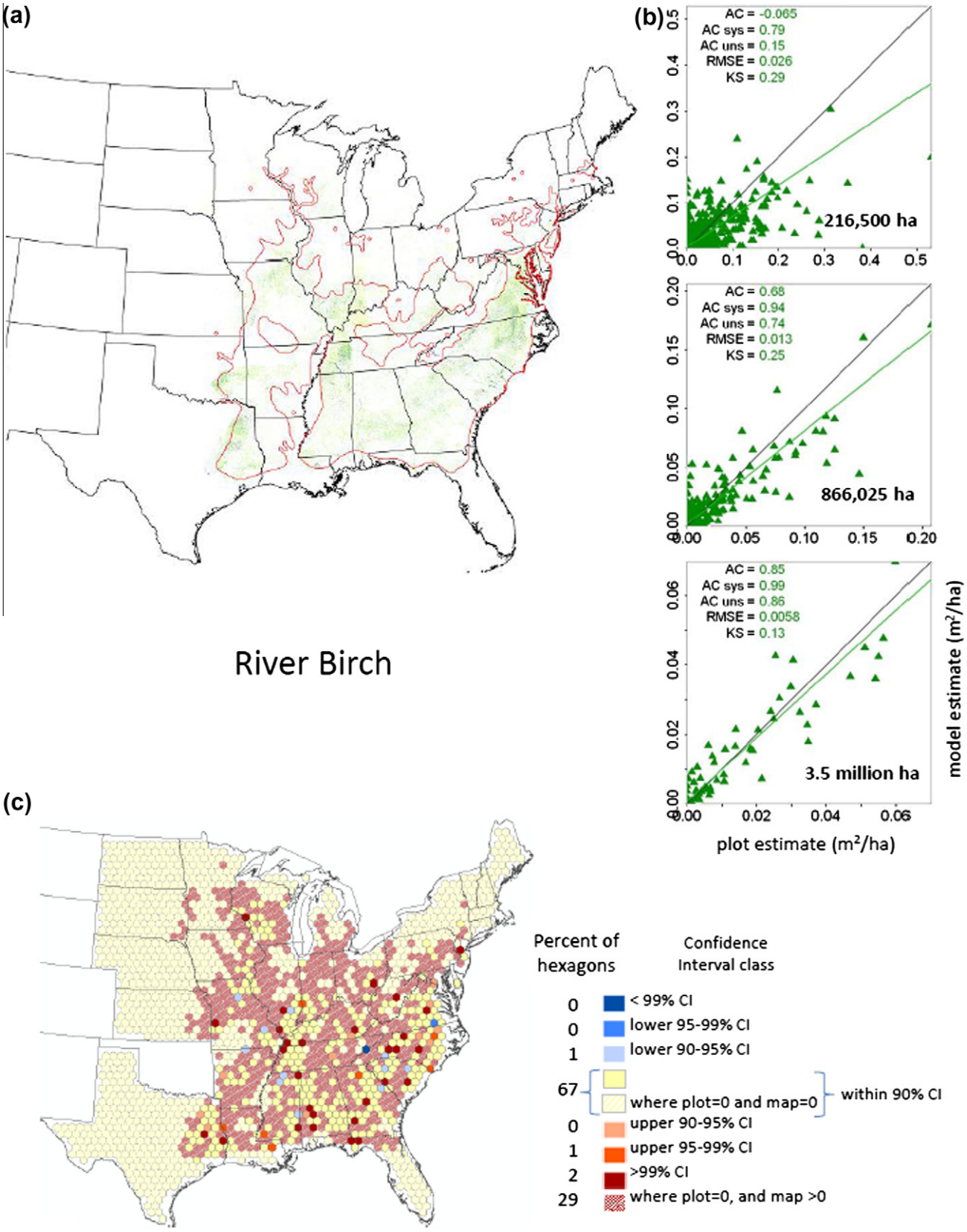

Fig. 7. Map of river birch live-tree basal area, ranging from low (light yellow) to high (dark green), and range boundary (red outline) (a), with scatterplots of estimates at three scales (b) and map of differences at 216,500 ha scale (c).

a necessary trade-off between bias and variance. While the optimization procedure suggested in this study is designed to minimize RMSE (i.e. the square root of the sum of the squared bias plus the variance), bias alone will be minimized when k equals one. This

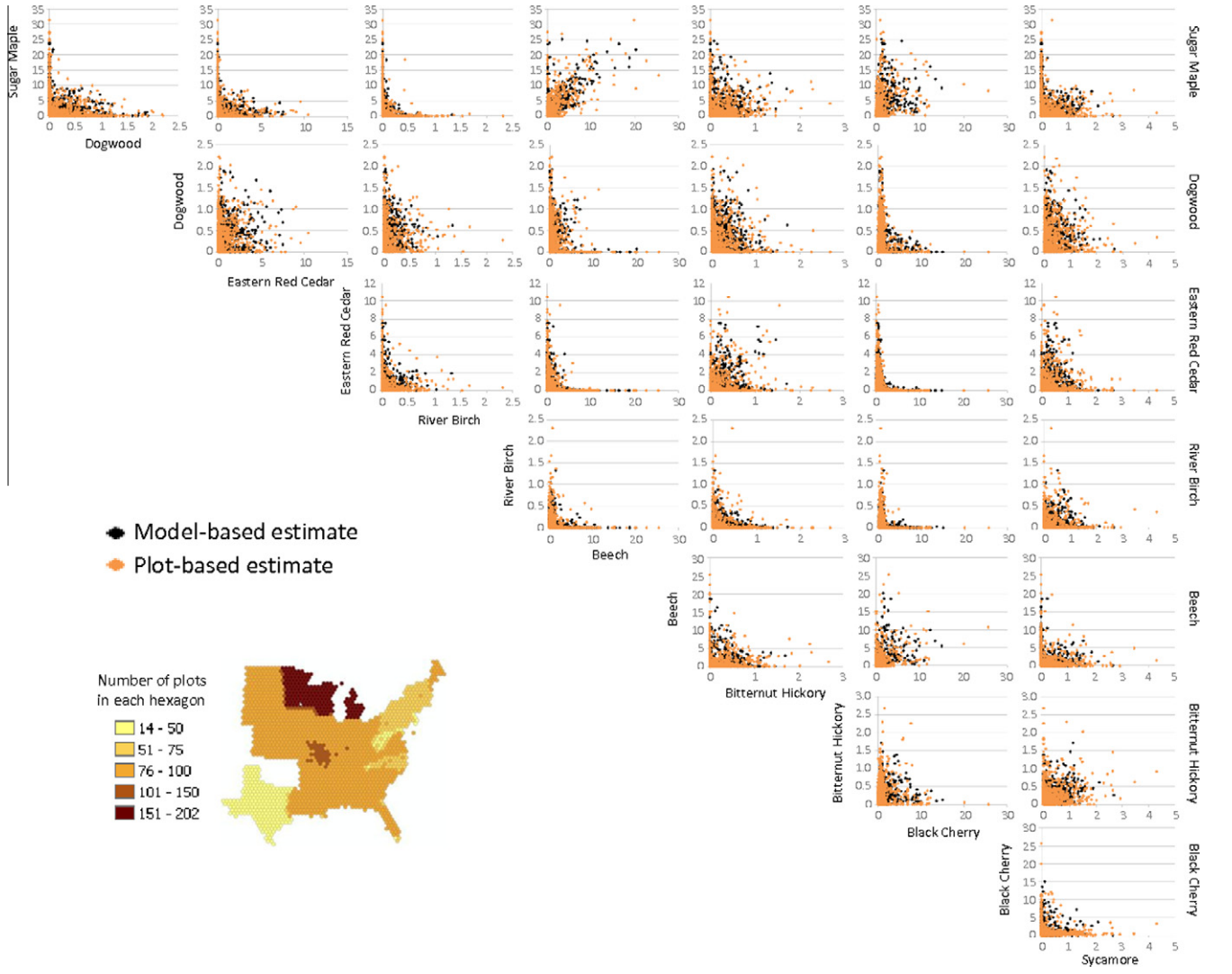

Fig. 8. Scatterplots of basal area covariance amongst eight species. Each point depicts mean live-tree basal area (m^2/ha) values for two species for a single 216,500 ha hexagon. Each scatterplot contains two sets of points, one derived from the plots and the other from the model. The inset map indicates the number of plots or pixels that contributed to the plot-based and model-based means for each hexagon.

bias will be most apparent near the boundaries of the data range, in particular near the lower limit of zero. A final explanation, mentioned earlier in the discussion of the detailed assessment of total live-tree basal area and the three highlighted species, may be the fact that there is a definitional mismatch between the forest stratum proportion layer and the FIA forest definition, which defined the areas over which tree measurements were taken. The maps produced are thus of a slightly different population than the one sampled by the FIA plots, i.e. they depict basal area of all trees, not just those found on forest land as defined by FIA. This could be tested only if the FIA sampling frame were expanded to include non-forest areas with trees such as urban parks and rural windbreaks. Accounting for definitional differences between the forest stratum proportion layer and the FIA sampling frame is therefore a possible area of improvement in the proposed mapping methodology that remains in alignment with the stated objectives established for the study.

There are many univariate and multivariate mapping approaches from which to choose, and an almost infinite combination of these methods and associated parameters. Clearly, no single method is ideal for all situations and data types, nor meets all possible efficiency criteria. Methods that are optimized to meet one goal (e.g. high spatial resolution) can be suboptimal for another (e.g. computing time and storage requirements, temporal frequency). The proposed approach was thus designed using a series of tradeoffs to take into account desirable output attributes, statistical considerations, and a desire to leverage efficiencies that would permit the efficient production of large area maps that could meet user needs. Preliminary results from this study have already been used by some state agency analysts to assist in the identification of priority forest landscape areas, such as those likely to be impacted by certain forest pests and diseases, as required by the 2008 Farm Bill.

The methodology presented in the current study could easily be adapted to produce maps of any field plot attribute that can be summarized to the plot level, because the approach in effect uses a relational database concept. The fundamental structures upon which all of the per-species maps are based are the imputed map of plot identification numbers (i.e. group labels), the associated table of stratified neighboring plots and weights, and the forest stratum proportion layer. All other post-processing steps are essentially database operations that join summarized field plot attributes to the imputed plot identification number map on a per-pixel basis. It seems reasonable to assume that the proposed methodology could be used for the production of maps of other attributes, especially those expected to be highly correlated with

total live-tree and per-species live-tree basal area. These include tree-level attributes such as volume, biomass, and standing dead-tree attributes, as well as site-level attributes like proportion forest, forest type, or possibly stand age and height. The exceptions are non-ecological field plot attributes, such as ownership, because these bear little correlation either to the plot's forest or tree characteristics or the ecological predictor layers used in the modeling. Rigorous accuracy assessment is required, including the types of multi-scale assessments presented in the current study. Though such an examination is beyond the scope of the current study, related studies are underway to determine the utility of such datasets for regional ecosystem services valuation and forest carbon modeling efforts.

4. Conclusions

The methodology presented here provides a cost-effective means for conducting regional to continental-scale mapping of tree species abundance and distribution. Its production efficiency arises from integrating the FIA database of field plots with readily-available moderate resolution raster datasets in a straightforward ecologically-based modeling framework. The explanatory power of the CCA model used in the study was bolstered by the inclusion of both ecological zone and vegetation phenology data. For the 100 most abundant tree species in the study area and at the scale of estimation units assessed, the nearest-neighbor imputation methodology produced results that compared favorably with estimates generated from FIA field plot data. Accuracy of the results improved slightly with the use of a weighting function based on distance in the canonical variate space, but substantially more so by using a stratification layer based on the 2001 NLCD tree canopy cover dataset and by selecting an optimal value for *k*. The issue of model overestimation at the edges of species ranges seen in some individual tree species may stem from an insufficient sample of field plots being used to produce the field plot-based estimates or from a stratification layer that includes trees on all lands while the FIA sampling frame does not.

Acknowledgements

The authors would like to thank Janet Ohmann and Matthew Gregory of the collaborative USFS/OSU LEMMA project for their work on GNN that planted the germ of this project; Mark Hansen, Mark Nelson, and Ron McRoberts of the FIA program for their thoughtful comments that helped flesh out many of the concepts presented; Mark Finco and Vicky Johnson of the USFS Remote Sensing Applications Center and Doug Griffith of FIA for their assistance in pulling together the raster datasets used in the project; as well as the anonymous reviewers whose comments greatly improved the clarity and focus of the manuscript.

References

Bechtold, W.A., Patterson, P.L. (Eds.), 2005. The Enhanced Forest Inventory and Analysis program – National Sampling Design and Estimation Procedures (GTR-SRS-80). US Department of Agriculture Forest Service, Southern Research Station, Asheville, NC.

Blackard, J.A., Finco, M.V., Helmer, E.H., Holden, G.R., Hoppus, M.L., Jacobs, D.M., Lister, A.J., Moisen, G.G., Nelson, M.D., Riemann, R., Ruefenacht, B., Salajanu, D., Weyermann, D.L., Winterberger, K.C., Brandeis, T.J., Czaplewski, R.L., McRoberts, R.E., Patterson, P.L., Tymcio, R.P., 2008. Mapping US forest biomass using nationwide forest inventory data and moderate resolution information. Remote Sens. Environ. 112, 1658–1677.

Borcard, D., Legendre, P., 1994. Environmental control and spatial structure in ecological communities: an example using oribatid mites. Environ. Ecol. Stat. 1, 37–61.

Bradley, B.A., Jacob, R.W., Hermance, J.F., Mustard, J.F., 2007. A curve fitting procedure to derive inter-annual phenologies from time series of noisy satellite NDVI data. Remote Sens. Environ. 106, 137–145.

Bystriakova, N., Kapos, V., Lysenko, I., Stapleton, C.M.A., 2003. Distribution and conservation status of forest bamboo biodiversity in the Asia-Pacific Region. Biodivers. Conserv. 12 (9), 1833–1841.

Casalegno, S., Amatulli, G., Bastrup-Birk, A., Houston-Durrant, T., Pekkarinen, A., 2011. Modelling and mapping the suitability of European forest formations at 1km resolution. Eur. J. For. Res. 130 (6), 971–981.

Crookston, N.L., Finley, A.O., 2008. YaImpute: an R package for kNN imputation. J. Stat. Soft. 23 (10), 1–16.

DeFries, R., Hansen, M., Steininger, M., Dubayah, R., Sohlberg, R., Townshend, J., 1997. Subpixel forest cover in central Africa from multisensor, multitemporal data. Remote Sens. Environ. 60, 228–246.

Dray, S., Dufour, A.B., 2007. The ade4 package: implementing the duality diagram for ecologists. J. Stat. Softw. 22 (4), 1–20.

Eskelson, B.N.I., Temesgen, H., LeMay, V., Barrett, T.M., Crookston, N.L., Hudak, A.T., 2009. The roles of nearest neighbor methods in imputing missing data in forest inventory and monitoring databases. Scand. J. For. Res. 24, 235–246.

Falk, W., Mellert, K.H., 2011. Species distribution models as a tool for forest management planning under climate change: risk evaluation of Abies alba in Bavaria. J. Veg. Sci. 22 (4), 621–634.

Feller, W., 1948. On the Kolmogorov–Smirnov limit theorems for empirical distributions. Ann. Math. Statist. 19 (2), 177–189.

Frank, T.D., 1988. Mapping dominant vegetation communities in the Colorado Rocky Mountain Front Range with Landsat Thematic Mapper and digital terrain data. Photogramm. Eng. Remote Sensing 54 (12), 1727–1734.

Gao, X., Huete, A.R., Ni, W., Miura, T., 2000. Optical–biophysical relationships of vegetation spectra without background contamination. Remote Sens. Environ. 74, 609–620.

Gessler, P.E., Moore, I.D., McKenzie, N.J., Ryan, P.J., 1995. Soil–landscape modeling and spatial prediction of soil attributes. Int. J. Geogr. Inf. Syst. 9 (4), 421–432.

Gillis, M.D., Leckie, D.G., 1993. Forest inventory mapping procedures across Canada (Information Report PI-X-114). Forestry Canada, Petawawa National Forestry Institute, Chalk River, ON.

Gormanson, D.D., Merriman, C., Hansen, M.H., 2009. Forest Service stand-size-class maps enhance FIA volume estimates. Proceedings of the IUFRO Division 4 meeting (May 19–22, 2009), Quebec City, Canada.

Hershey, R.R., Reese, G., 1999. Creating a "first-cut" species distribution map for large areas from forest inventory data (GTR-NE-256). US Department of Agriculture Forest Service, Northeastern Research Station, Radnor, PA.

Homer, C., Dewitz, J., Fry, J., Coan, M., Hossain, N., Larson, C., Herold, N., McKerrow, A., Vandriel, J.N., Wickham, J., 2007. Completion of the 2001 national land cover database for the conterminous United States. Photogramm. Eng. Remote Sensing 73 (4), 337–341.

Iverson, L.R., Prasad, A.M., 1998. Predicting abundance of 80 tree species following climate change in the eastern United States. Ecol. Monogr. 68, 465–485.

Ji, L., Gallo, K., 2006. An agreement coefficient for image comparison. Photogramm. Eng. Remote Sensing 72, 823–833.

Justice, C., Vermote, E., Townshend, J.R.G., Defries, R., Roy, D.P., Hall, D.K., Salomonson, V.V., Privette, J., Riggs, G., Strahler, A., Lucht, W., Myneni, R., Knjazihhin, Y., Running, S., Nemani, R., Wan, Z., Huete, A., van Leeuwen, W., Wolfe, R., 1998. The moderate resolution imaging spectroradiometer (MODIS): land remote sensing for global change research. IEEE Trans. Geosci. Remote Sens. 36 (4), 1228–1249.

Krist, F.J.J., Sapio, F.J., Tkacz, B., 2010. A multicriteria framework for producing local, regional, and national insect and disease risk maps. In: Pye, J.M., Rauscher, H.M., Sands, Y., Lee, D.C., Beatty, J.S. (Eds.), Advances in Threat Assessment and Their Application to Forest and Rangeland Management (GTR-PNW-802). US Department of Agriculture Forest Service, Pacific Northwest and Southern Research Stations, Portland, OR, pp. 621–636.

Lister, A.J., Mou, P., Jones, R.H., Mitchell, R.J., 2000. Spatial patterns of soil and vegetation in a southern pine forest. Can. J. For. Res. 30, 145–155.

Little Jr., E.L., 1971. Atlas of United States Trees, vol. 1, Conifers and Important Hardwoods (Misc. Publ. 1146). US Department of Agriculture, Washington, DC.

Liu, H.Q., Huete, A.R., 1995. A feedback based modification of the NDVI to minimize canopy background and atmospheric noise. IEEE Trans. Geosci. Remote Sens. 33, 457–465.

Lund, H.G., 2002. When is a forest not a forest? J. Forest. 100 (8), 21–28.

McCarter, J.B., Wilson, J.S., Baker, P.J., Moffett, J.L., Oliver, C.D., 1998. Landscape management through integration of existing tools and emerging technologies. J. Forest. 96, 17–23.

McGarigal, K., Cushman, S.A., Stafford, S., 2000. Multivariate Statistics for Wildlife and Ecology Research. Springer Verlag, New York, NY.

McPherson, J.M., Jetz, W., 2007. Effects of species' ecology on the accuracy of distribution models. Ecography 30 (1), 135–151.

McRoberts, R.E., Nelson, M.D., Wendt, D.G., 2002. Stratified estimation of forest area using satellite imagery, inventory data, and the k-Nearest neighbor technique. Remote Sens. Environ. 82, 457–468.

McRoberts, M.E., Tomppo, E.O., Finley, A.O., Heikkinen, J., 2007. Estimating areal means and variances of forest attributes using the k-nearest neighbors technique and satellite imagery. Remote Sens. Environ. 111, 466–480.

Moody, A., Johnson, D.M., 2001. Land-surface phenologies from AVHRR using the discrete fourier transform. Remote Sens. Environ. 75, 305–323.

Moore, I.D., Grayson, R.B., Ladson, A.R., 1991. Digital terrain modelling: a review of hydrological, geomorphological, and biological applications. Hydrol. Process. 5, 3–30.

Nelson, M.D., McRoberts, R.E., Holden, G.R., Bauer, M.E., 2009. Effects of satellite image spatial aggregation and resolution on estimates of forest land area. Int. J. Remote Sens. 30 (8), 1913–1940.

Ohmann, J.L., Gregory, M.J., 2002. Predictive mapping of forest composition and structure with direct gradient analysis and nearest-neighbor imputation in coastal Oregon, USA. Can. J. For. Res. 32, 725–741.

Ohmann, J.L., Gregory, M.J., Spies, T.A., 2007. Influence of environment, disturbance, and ownership on forest vegetation of Coastal Oregon. Ecol. Appl. 17 (1), 18–33.

Olthof, I., 2004. Mapping deciduous forest ice storm damage using landsat and environmental data. Remote Sens. Environ. 89 (4), 484–496.

Omernik, J.M., 1987. Ecoregions of the conterminous United States. Map (scale 1:7500,000). Ann. Assoc. Am. Geogr. 77, 118–125.

Poland, T.M., McCullough, D.G., 2006. Emerald ash borer:invasion of the urban forest and the threat to North America's ash resource. J. Forest. 104, 118–124.

Pontius, J.A., Hallett, R., Martin, M., Plourde, L., 2010. A landscape-scale remote sensing/GIS tool to assess eastern hemlock vulnerability to hemlock woolly adelgid-induced decline. In: Pye, J.M., Rauscher, H.M., Sands, Y., Lee, D.C., Beatty, J.S. (Eds.), Advances in Threat Assessment and Their Application to Forest and Rangeland Management (PNW-GTR-802). US Department of Agriculture Forest Service, Pacific Northwest and Southern Research Stations, Portland, OR, pp. 657–671.

Riemann Hershey, R., 2000. Modeling the spatial distribution of ten tree species in Pennsylvania. In: Mowrer, H.T., Congalton, R.G. (Eds.), Quantifying Spatial Uncertainty in Natural Resources. Ann Arbor Press, Chelsea, MI, pp. 119–135.

Riemann, R., 2003. Pilot Inventory of FIA Plots Traditionally Called 'Nonforest' (NERS-GTR-312). US Department of Agriculture Forest Service, Northeastern Research Station, Newtown Square, PA.

Riemann, R., Lister, T., Lister, A., Meneguzzo, D., Parks, S., 2008. Development of issue-relevant state level analyses of fragmentation and urbanization. In: McWilliams, W., Moisen, G., Czaplewski, R. (Comps.), Forest Inventory and Analysis (FIA) Symposium 2008; October 21–23, 2008. Park City, UT. Proceedings (RMRS-P-56CD). US Department of Agriculture, Forest Service, Rocky Mountain Research Station, Fort Collins, CO.

Riemann, R., Wilson, B.T., Lister, A., Parks, S., 2010. An effective assessment protocol for continuous geospatial datasets of forest characteristics using USFS Forest Inventory and Analysis (FIA) data. Remote Sens. Environ. 114, 2337–2352.

Riordan, E.C., Rundel, P.W., 2009. Modelling the distribution of a threatened habitat: the California sage scrub. J. Biogeogr. 36, 2176–2188.

Rogers, P.C., Leffler, A.J., Ryel, R.J., 2010. Landscape assessment of a stable aspen community in southern Utah, USA. For. Ecol. Manage. 259 (3), 487–495.

Rollins, M.G., Keane, R.E., Zhu, Z., Menakis, J.P., 2006. Executive summary. In: Rollins, M.G., Frame, C.K. (Eds.), The LANDFIRE Prototype Project: Nationally Consistent and Locally Relevant Geospatial Data for Wildland Fire Management (RMRS-GTR-175). US Department of Agriculture, Forest Service, Rocky Mountain Research Station, Fort Collins, CO, pp. 1–4.

Ruefenacht, B., Finco, M.V., Nelson, M.D., Czaplewski, R., Helmer, E.H., Blackard, J.A., Holden, G.R., Lister, A.J., Salajanu, D., Weyermann, D., Winterberger, K., 2008. Conterminous US and Alaska forest type mapping using forest inventory and analysis data. Photogramm. Eng. Remote Sensing 74 (11), 1379–1388.

Sargent, S., 1884. Report on the Forests of North America (Exclusive of Mexico). US Department of the Interior, Government Printing Office, Washington DC.

Sellers, P.J., Los, S.O., Tucker, C.J., Justice, C.O., Dazlich, D.A., Collatz, G.J., Randall, D.A., 1994. A global 1∗1 degree NDVI data set for climate studies. Part 2: the generation of global fields of terrestrial biophysical parameters from the NDVI. Int. J. Remote Sens. 15 (17), 3519–3545.

Spruce, J.P., Sader, S., Ryan, R.E., Smoot, J., Kuper, P., Ross, K., Prados, D., Russell, J., Gasser, G., McKellip, R., Hargrove, W., 2011. Assessment of MODIS NDVI time series data products for detecting forest defoliation by gypsy moth outbreaks. Remote Sens. Environ. 115, 427–437.

ter Braak, C.J.F., 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology 67, 1167–1179.

ter Braak, C.J.F., Verdonschot, P.F.M., 1995. Canonical correspondence analysis and related multivariate methods in aquatic ecology. Aquat. Sci. 57 (3), 255–289.

Thornton, P.E., Running, S.W., White, M.A., 1997. Generating surfaces of daily meteorological variables over large regions of complex terrain. J. Hydrol. 190, 214–251.

Tu, J.V., 1996. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. J. Clin. Epidemiol. 49 (11), 1225–1231.

Uuttera, J., Maltamo, M., Kurki, S., Mykrä, S., 1998. Differences in forest structure and landscape patterns between ownership groups in central Finland. Boreal Environ. Res. 3, 191–200.

Wang, F.G., Xu, Y.J., 2009. Hurricane Katrina-induced forest damage in relation to ecological factors at landscape scale. Environ. Monit. Assess. 156 (1–4), 491–507.

Weber, T., Wolf, J., 2000. Maryland's green infrastructure – Using landscape assessment tools to identify a regional conservation strategy. Environ. Monit. Assess. 63 (1), 265–277.

Wolter, P., Mladenoff, D., Host, G., Crow, T., 1995. Improved forest classification in the Northern Lake States using multi-temporal Landsat imagery. Photogramm. Eng. Remote Sensing 61 (9), 1129–1143.

Woodall, C.W., Oswalt, C.M., Westfall, J.A., Perry, C.H., Nelson, M.D., Finley, A.O., 2009. An indicator of tree migration in forests of the eastern United States. For. Ecol. Manage. 257, 1434–1444.

Woudenberg, S.W., Conkling, B.L., O'Connell, B.M., LaPoint, E.B., Turner, J.A., Waddell, K.L., 2010. The Forest Inventory and Analysis Database: Database Description and Users Manual Version 4.0 for Phase 2 (RMRS-GTR-245). US Department of Agriculture, Forest Service, Rocky Mountain Research Station, Fort Collins, CO.

Zhu, Z., Evans, D.L., 1994. US forest types and predicted percent forest cover from AVHRR data. Photogramm. Eng. Remote Sensing 60 (5), 525–531.