Biomass and Bioenergy 121 (2019) 64–77

Contents lists available at ScienceDirect

Biomass and Bioenergy

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

Research paper

Comparison of small area estimation methods applied to biopower feedstock supply in the Northern U.S. region

Michael E. Goerndt[a,*], Barry T. Wilson[b], Francisco X. Aguilar[c]

[a] Missouri State University, William H. Darr College of Agriculture, 901 S. National Ave., Springfield, MO, 65897, USA
[b] USDA Forest Service, Northern Research Station, 1992 Folwell Ave., St. Paul, MN, 55108, USA
[c] Swedish University of Agricultural Sciences, Department of Forest Economics, Umea, 90183, Sweden

ARTICLE INFO

Keywords:
Small area estimation
Biomass
Feedstock
Biopower
Imputation
MODIS

ABSTRACT

Increasing interest in utilization of forest biomass for bioenergy has prompted extensive contemporary research regarding costs, supply and technology for efficiently producing electricity and other forms of renewable energy. One challenge facing both researchers and users is obtaining precise estimates of available forest biomass within plausible supply areas for individual power plants. Due to the wide distribution of power plants poised to co-fire with forest biomass, assessing its availability requires methods that can yield precise and low-bias estimates of aboveground forest biomass and other key attributes at varying spatial scales. Small area estimation (SAE) methods have high potential to accomplish this due to the availability of national forest inventory data, combined with satellite imagery and other forms of remotely-sensed auxiliary information. The study assessed several indirect, direct and composite estimators of four forest attributes: aboveground tree biomass, biomass of small-diameter trees, biomass of tops and limbs, and volume at the county-level and within the estimated supply areas around power plants across 20 states in the contiguous Northern U.S. Composite estimators using both k-nearest neighbors imputation and multiple linear regression provided superior estimates of indicators of forest biomass availability based on both precision and bias at the county-level at sampling intensities as low as 10–20%, compared to the other SAE methods examined. The composite estimator using k-nearest neighbors imputation was subsequently shown to produce precise estimates of forest biomass availability for selected power plant supply areas.

1. Introduction

Estimates of forest biomass and related forest attributes provide information that can support management decisions regarding forest health, forest structure, and potential resource utilization in conjunction with forest management. Such estimates are also important to the renewable energy industry as they provide information on the physical availability of forest biomass resources. Forest biomass comprises a large portion of annual renewable energy generation in Europe and the U.S. For example, in Sweden biomass accounted for 65% of all renewable energy consumed in 2016 [1], 77% of which was from forest biomass. In Finland, forest biomass comprised approximately 50% of all bioenergy consumption in 2016 [2]. Similarly, forest biomass currently accounts for the greatest share of bioenergy generation in the U.S. at about 53% [3]. Potential methods of procuring forest biomass for bioenergy include removal of logging by-products (slash) following traditional roundwood harvests, removal of small-diameter trees (tree < 25 cm dbh) during pre-commercial thinning and roundwood harvest, and through an integrated harvest (roundwood + small-diameter trees + slash). Regarding its conversion to useable energy, arguably, the technological platform with the greatest potential for utilizing largest amounts of forest biomass in the U.S. and other countries highly dependent upon coal electricity is co-firing (i.e. combusted biomass with coal) or utilization in dedicated boilers for electricity generation [4–7]. One of the greatest challenges to assessing supply potential for multiple power plants is deriving spatial estimates of forest biomass availability within supply areas where transport and utilization are technologically and economically feasible [5,7,9,10]. Spatial estimation in this sense refers to the inference of forest attributes across the landscape using spatially explicit information.

In recent studies by Goerndt et al. [6] and Goerndt et al. [10], the spatial limitation of design-based estimates for supply circles surrounding power plants locations was addressed by using a weighted mean of forest biomass estimates for all counties intersecting supply

* Corresponding author.
E-mail addresses: mgoerndt@missouristate.edu (M.E. Goerndt), barrywilson@fs.fed.us (B.T. Wilson), francisco.aguilar@slu.se (F.X. Aguilar).

https://doi.org/10.1016/j.biombioe.2018.12.008
Received 11 July 2018; Received in revised form 19 November 2018; Accepted 6 December 2018
Available online 24 December 2018
0961-9534/ Published by Elsevier Ltd.

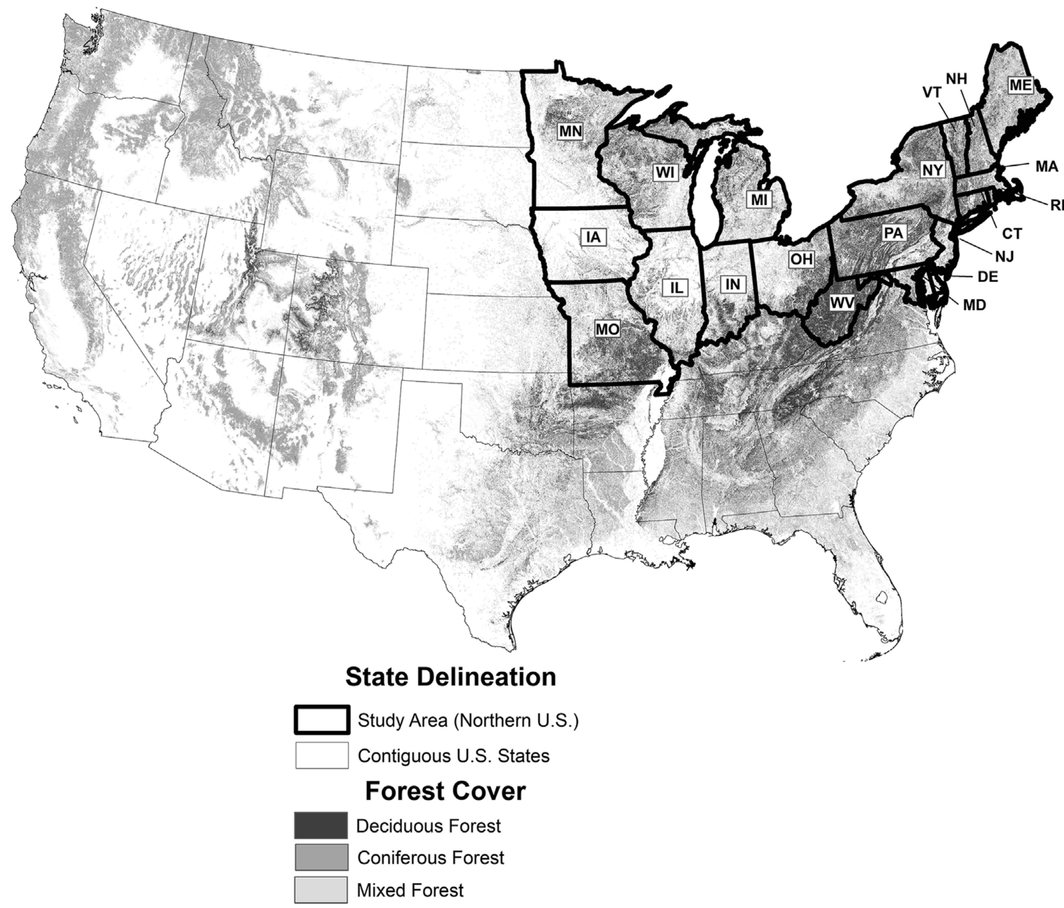

Fig. 1. Location of study area (Northern U.S.) within the contiguous U.S delineated by individual states and showing forest coverage by forest type.

circles. One limitation to this approach was the assumption that the forest biomass resource in each applicable county is uniformly distributed across the county. The potential for bias in this estimation method, combined with the inability of most researchers and managers to determine which specific national forest inventory plots are located within such areas due to confidentiality policies, stresses the need to assess spatial estimation methods that do not rely on an assumption of uniformity. Such estimation methods also must yield low-bias and precise estimates with small sample sizes due to the small size of many supply areas combined with limitations in data availability. For instance, the national forest inventory of the United States is conducted by the USDA Forest Service Forest Inventory and Analysis (FIA) program using a probability sample of field plots, similar to several national forest inventory systems used in other countries such as Sweden and Finland and Canada, just to name a few [11–14]. FIA plots for most eastern states have recently been measured at a base sampling intensity of one plot per 2400 ha and on a 5-year cycle, meaning that 20% of the plots are measured every year for five years, after which time the measurement cycle starts over. Due to budgetary restrictions, FIA has shifted to a 7–10 year measurement cycle. Therefore, future estimates and monitoring of forest biomass availability for bioenergy may have to be conducted using partial data panels, depending on the year or group of years being analyzed and the temporal resolution required. A set of estimation techniques well suited to the spatial challenges of using FIA data and the issue of reduced sample size is small area estimation (SAE). SAE techniques can increase the precision of forest attribute estimates in cases where the sample size for an area of interest is insufficient to achieve the desired level of precision through design-based direct estimation [15–18].

The main goal of this study was to compare SAE methods for estimating select forest attributes for supply areas around coal-fired power plants in the North-central and Northeastern contiguous U.S. Three specific objectives supported our overall goal. The first objective was to assess the performance of several different estimators: indirect estimators based on multiple linear regression (MLR) and *k*-nearest neighbors imputation (KNN); a stratified indirect estimator based on MLR; and composite predictors using each of the indirect estimators as a component for all counties in the study area. Precision and bias for each estimator were evaluated by simulating (via subsampling) smaller samples as a proxy for limited data availability as a pre-text for SAE. The second objective was to use the optimal county-level estimators to estimate forest attributes of interest for supply areas surrounding coal-fired power plant locations.

2. Methods

The two types of SAE techniques implemented in this study are indirect estimation and composite estimation. Indirect estimators typically utilize a model-based approach to statistical inference (e.g. through regression or imputation) to obtain estimates for areas of interest by "borrowing strength" from auxiliary information within the entire population [15,16]. As such, indirect estimation does not require that an area of interest contains sample units. Composite estimation encompasses a weighted mean between a model-based indirect estimate and a design-based direct estimate (DE). Composite estimators are designed to balance the higher precision of an indirect estimator with the low bias of a direct estimator [15,17].

2.1. Study area

The population of interest for this study consisted of every county in the U.S. Northeastern and North-central states, hereafter referred to as the Northern U.S. Northern U.S. states included: Connecticut (CT), Delaware (DE), Illinois (IL), Indiana (IN), Iowa (IA), Maine (ME), Maryland (MD), Massachusetts (MA), Michigan (MI), Minnesota (MN), Missouri (MO), New Hampshire (NH), New Jersey (NJ), New York (NY), Ohio (OH), Pennsylvania (PA), Rhode Island (RI), Vermont (VT), West Virginia (WV), and Wisconsin (WI) (Fig. 1). Forest biomass combustion and co-firing with coal were identified as major potential sources of renewable energy in the region [4,19] as it hosts more than 350 coal-fired power plants [20], many of which could utilize forest biomass as a feedstock for renewable energy generation [10].

2.2. Field data

For regional-level analyses of forest biomass and other attributes, the most authoritative and readily available forest inventory data come from the U.S. Department of Agriculture Forest Service's Forest Inventory and Analysis (FIA) program [14]. FIA's sample of plots provide design-based estimates of forest attributes at varying spatial scales (e.g. state-level, county-level). However, it can often be challenging to derive FIA estimates for areas that do not fit standard state-level or county-level geopolitical boundaries because explicit FIA plot locations are not made available to the general public due to security and confidentiality considerations. Consequently, the smallest spatial observational unit for which estimates of forest attributes can be reliably obtained is at the county level. This limitation creates a challenge when trying to obtain biomass supply estimates for procurement areas defined by economic over political boundaries. Additionally, county-level direct (design-based) estimates of forest attributes based on FIA data often have low precision due to the relatively small sample size within individual counties.

The ground data used in this study included 95,351 FIA plots collected throughout 1037 counties in the study area during the period of 2004–2008. County-level design-based means for the four variables of interest were calculated from the FIA plot data, including total aboveground biomass of live trees (Mg/ha), aboveground biomass of trees with dbh $<$ 25 cm (Mg/ha), aboveground biomass of tops and limbs (Mg/ha), and total cubic volume of live trees (m^3/ha). These forest attributes were chosen because of their high importance in assessing forest biomass availability for bioenergy generation, and comprise the dependent variables for the SAE methods analyzed at the county-level [4,6]. Table 1 shows summary statistics for each forest attribute of interest across the Northern U.S. as defined by the study area (Fig. 1).

There are two characteristics that stand out in the statistical summary across the counties of the Northern U.S. First, the coefficient of variation is the same between aboveground biomass and volume, reinforcing an assumption that there is a close relationship between aboveground biomass and cubic volume, even when summarized by county. Second, all coefficients of variation are very high. This emphasizes the need to assess SAE methods that utilize auxiliary information and do not rely solely upon direct estimates within counties or among counties.

Table 1
Summary statistics for forest attributes of interest across the Northern U.S. region (n = 1037 counties).

Attribute	Mean	Standard Deviation	[a]CV%
Aboveground Biomass (Mg ha^{-1})	16.7	14.8	89
Small-diameter Biomass (Mg ha^{-1})	5.1	4.7	92
Top and Limb Biomass (Mg ha^{-1})	2.9	2.6	90
Volume (m^3 ha^{-1})	606.5	538.6	89

[a] Coefficient of variation = Standard Deviation/Mean.

2.3. Auxiliary information

The study used several auxiliary variables that have been shown in numerous studies to be correlated with vegetation composition [15,21–24]. The final set of auxiliary variables consists of remotely-sensed metrics related to the presence of green vegetation, seasonal vegetation phenology, climate, topography, land cover, and ecological region. Values for all auxiliary variables described in this section were organized into a stack of raster files for use in the analysis (Table 2).

2.3.1. Landsat imagery

Vegetation indices were derived from annual Landsat 7 ETM + composite images for 2007. All Landsat data were obtained through the Web-Enabled Landsat Database (WELD) maintained by the U.S. Geological Survey (USGS), National Aeronautics and Space Administration, and South Dakota State University. Geometric rectification, radiometric correction, and mosaicking of individual Landsat scenes were conducted by the WELD team. Due to a Scan-line Corrector (SLC) malfunction in late 2003, Landsat 7 ETM + images contain strips of missing pixels for each data collection. Consequently, the annual Landsat 7 ETM + composite images from WELD display the same SLC artifacts. To adjust for this, we adapted methods described by Scaramuzza et al. [23] by which SLC gaps are filled by assuming a linear relationship between two Landsat scenes (Appendix).

From the Landsat data, six of the original bands including bands one through five (B1-B5) and seven (B7) were extracted. The normalized difference vegetation index (NDVI) is a very simple function that can be important in describing forest health and vigor [21]. More specifically, NDVI has been shown to be effective in describing factors such as crown closure, forest density, and tree species diversity [22]. NDVI was calculated using the following equation:

$$NDVI = \frac{B4 - B3}{B4 + B3} \tag{1}$$

In addition to the aforementioned Landsat bands and transformations, there are several simple band ratios that can be useful for describing forest characteristics. The band ratios used in this study were B4/B3 (R1), B5/B4 (R2), and B7/B5 (R3). The combined purpose of these ratios is to enhance the features of forest cover, bare soil, and water in order to better contrast them on the landscape [24,25]. The original spatial resolution of all Landsat bands used in this study was 30 × 30 m. Subsequently, a focal mean filter was used to resample each Landsat coverage to 250 m resolution to match the resolution of the MODIS EVI data described in the next section. These resampled values of all Landsat variables were obtained for every pixel located within the population (Table 2).

2.3.2. Land coverage

Forest cover, imperviousness, and percent canopy cover were derived from the 2001 and 2006 National Land Cover Database [26,27]. Percent canopy and imperviousness pixels were resampled from 30-m to 250-m using the focal window mean. Forest cover pixels were resampled to 250-m resolution using the focal window majority and assigned a dichotomous indicator (0,1) based on whether the majority of 30-m pixels were classified as forested or non-forested (Table 2).

2.3.3. Seasonal phenology and climate

Seasonal vegetation phenology data were derived from the Enhanced Vegetation Index (EVI) of the 250-m pixel resolution MOD13Q1, Version 4, MODIS Terra data product. EVI is similar in form to NDVI, but also uses the blue band to correct for aerosols in the atmosphere. The original 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.

Table 2
Description of category and source for all auxiliary variables used to derive indirect SAE estimates in this study.

Source	Category	Variable Name	Source	Category	Variable Name
Landsat	Vegetation	Band 1	Daymet	Temperature	tb1
Landsat	Vegetation	Band 2	Daymet	Temperature	ta2
Landsat	Vegetation	Band 3	Daymet	Temperature	tb2
Landsat	Vegetation	Band 4	Daymet	Precipitation	pa0
Landsat	Vegetation	Band 5	Daymet	Precipitation	pa1
Landsat	Vegetation	Band 7	Daymet	Precipitation	pb1
Landsat	Vegetation	NDVI	Daymet	Precipitation	pa2
Landsat	Vegetation	R1	Daymet	Precipitation	pb2
Landsat	Vegetation	R2	USGS	Topography	elevation
Landsat	Vegetation	R3	EDNA	Topography	CTI
MODIS	Phenology	ma0	EDNA	Topography	SAI
MODIS	Phenology	ma1	US Census	Topography	Coordinate Easting
MODIS	Phenology	mb1	US Census	Topography	Coordinate Northing
MODIS	Phenology	ma2	NLCD	Land Cover	Imperviousness %
MODIS	Phenology	mb2	NLCD	Land Cover	Canopy %
Daymet	Temperature	ta0	NLCD	Land Cover	Forest Cover
Daymet	Temperature	ta1	EPA	Ecoregion	Level III

Because the MODIS EVI dataset was intended to represent seasonal vegetation phenology, it was important to pre-process the data to reduce dimensionality and adjust for atmospheric effects. Pre-processing of EVI was performed following Wilson et al. [28] and consisted of pixel-level harmonic regression of all MODIS time series images for the region using a re-weighted least squares algorithm [29]. The original Fourier series is expressed as follows:

$$y = a_0 + \sum_{n=1}^{h} (a_n \cos(nt) + b_n \sin(nt)), \tag{2}$$

where t is the time period in radians, h is the number of harmonics in the series, and a_0, a_n, and b_n are coefficients. Wilson et al. [28] used a 2nd order Fourier series to represent the general form of the time series data. To adjust for clouds and snow, the same Fourier series was fit again using weighted regression with weights defined as:

$$W_t = .1.\ 5^{r_t/\max(r)}, \tag{3}$$

where r_t is the residual from the initial model for compositing period t. The weighted regression resulted in five estimated Fourier coefficients for each pixel, labeled with the prefix *m*, i.e. ma_0, ma_1, mb_1, ma_2, and mb_2.

Climate data were extracted from the Daymet 18-year dataset of mean monthly growing-degree days and total precipitation [30]. Because temperature and precipitation also follow seasonal patterns, a similar harmonic regression was used for the monthly climate data, but without using weight regression. Due to the original 1 km pixel resolution of the Daymet data, the Fourier coefficients for climate were resampled to 250-m pixel resolution using bilinear interpolation. The five estimated Fourier coefficients (a_0, a_1, b_1, a_2, and b_2) associated with each pixel's mean monthly growing-degree days and total precipitation are labeled with the prefix *t* and *p* respectively (Table 2).

2.3.4. Topographic metrics and ecoregions

Several topographic metrics were utilized as predictor variables in this study. Elevation was derived from the 2008 USGS National Elevation Dataset (NED) [31]. Compound topographic index (CTI) and slope-aspect index (SAI) were derived from the 2008 USGS Elevation Derivatives for National Applications (EDNA) dataset. CTI [32] is a measure of site wetness based on landscape position and is correlated with several soil characteristics [33]. SAI [34] is a measure of site orientation and is correlated with solar insolation and exposure to prevailing winds. CTI and SAI were resampled to 250-m pixels by applying a 9 × 9-pixel moving window and then using a nearest neighbor algorithm [28]. The geospatial location of the county centroids in the study area were obtained from the 2010 U.S. Census Bureau TIGERline data. Finally, the study utilized the U.S. Environmental Protection Agency's (EPA) Level III Ecoregions [35]. 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 (Table 2).

2.3.5. Resampling

The primary assumption for SAE in this study is that there is an insufficient sample size within each small area of interest to obtain direct estimates with high precision, an issue common to county-level direct estimates of forest attributes with FIA. For example, within the U.S. state of Minnesota (MN), county-level direct estimates of merchantable volume range in relative standard error from 1% to 97%, with a mean of 41% [14]. Again, the base sampling intensity of the spatially balanced FIA sample is one field plot per 2400 ha. Without knowing the true variability of the domain of interest within the population, and therefore assuming as a rule-of-thumb (arbitrarily) that a minimum sample size of 60 is required for use with a direct estimator, this suggests a minimum domain size of approximately 72,000 ha, which would immediately exclude 40 counties within the study region. In this study we used the full sample of FIA plots per county for validation, because census data for the entire population of forest variables for each county does not exist. We then simulated the effect of smaller sample sizes by sub-sampling from this full sample of plot data. This allowed us to assess the performance of the different SAE techniques as the sample size decreased and to facilitate validation of the estimators using direct county estimates of the forest attributes from the full sample as a surrogate for census information. Therefore, the simulation of reduced sample sizes was done by randomly selecting FIA plots from each county using 5%, 10%, 20%, 30%, and 40% sub-sampling intensities, with 500 replications at each sub-sampling intensity. Direct estimates (DE) in the form of county-level means were calculated for each attribute at all sub-sampling intensities as well as for the full samples. All of the estimation methods of interest were assessed for the five sub-sampling intensities separately. County-level direct estimates for each attribute based upon the full sample were retained as validation data for each estimator in the absence of census information.

2.4. Statistical analysis

2.4.1. Multiple linear regression (MLR)

MLR models were developed both as an indirect estimation method for the forest attributes of interest and as a necessary component for composite estimators developed later in the analysis. These models were designed to estimate county-level means of the forest attributes of interest using the DE as the response variable and the variables described in Table 2 summarized (mean) by county as the predictor variables. Because it would have been impractical to reassess the MLR models for each of the 500 replicated sub-samples, an MLR model was developed for each attribute of interest using the full sample DE as the response values for each attribute. Models were selected using a subset regression technique that identifies the independent variables which create the best fitting linear regression models according to Bayesian Information Criteria (BIC) using an exhaustive search [15]. This was performed using the '*regsubsets()*' tool available in the '*leaps()*' package for '*R ()*' [36,37]. The subsets regression stage of the analysis was set to identify the best fitting models with a maximum of seven independent variables for each model, to avoid potential issues of high multi-collinearity between variables [15].

2.4.2. Imputation through k-nearest neighbors (KNN)

The KNN approach to imputation was assessed as an alternative indirect estimator of forest attributes in this study. KNN is a multivariate nonparametric method that imputes sample units to observational units based on distance in feature space. Unlike many other imputation methods, KNN imputation uses either a simple or weighted mean of distance metrics for multiple neighbors to impute values to observational units on the landscape. The distance metric used for KNN imputation had the following form [28,38–40].

$$d_{ij}^2 = (X_i - X_j)'\Gamma\Lambda^2\Gamma'(X_i - X_j), \tag{4}$$

where X_i is the vector of auxiliary variables for the ith county, X_j is the vector of auxiliary variables for the jth reference county, Γ is a matrix of standardized canonical coefficients for the X variables, and Λ^2 is a diagonal matrix of squared canonical correlations between the X and Y datasets [38,39]. The set of X variables for this analysis consisted of a matrix of the remote-sensed auxiliary variables (Table 2), whereas the set of Y variables was a matrix of values for all four forest attributes of interest. The canonical correlations are created using both the X and Y variables.

One of the most important aspects of the KNN analysis was deciding how many (k) neighbors to use. Tuominen et al. [41] stated that "The higher the value of k, the more averaging that occurs in the estimates. Thus, the optimal value of k is a trade-off between the accuracy of estimates and the variation retained in the estimates". Additionally, it has been observed that as the value of k increases, average bias tends to increase for extreme response variable values [42]. Therefore, it is considered beneficial to choose the lowest value of k possible while still maintaining a relatively small root mean squared error between observed and predicted values. For this study, k was chosen by observing root mean squared errors for KNN models using 1 to 20 neighboring counties. For this study, we opted to use the traditional approach to KNN using Euclidean distance and including all independent variables (Table 2) in the feature space [15,43,44]. Therefore, while the final MLR models used a reduced number of independent variables based on the results of subsets regression model fit, KNN output utilized all independent variables. Final imputation using KNN was done by calculating the simple mean (i.e. equal weighting) of the response variables over the k nearest neighbors [45,46]. All calculations for KNN were performed using the *yaImpute()* tool for *R()* [36,47].

2.4.3. Stratification

Past studies such as Goerndt et al. [40], Breidenbach and Astrup [48] and Mauro et al. [49] have shown that SAE techniques have high potential for precise estimation of forest attributes. However, in most cases, these estimation methods have been applied to populations that were predominantly forested. One feature of this study is that it includes data from both forested and non-forested areas within each small area (county) of interest and within the study area as a whole. A necessary feature due to the fact that county-level estimates of forest attributes have to incorporate "zero" values of forest attributes as valid observations for areas within counties with little to no forest cover. As such, it was desirable to analyze effects of stratification between forested and non-forested areas on the MLR estimators. We used the indicator of dominant forest cover (i.e. based on NLCD forest in Table 2) to classify the individual counties into one of two strata: Stratum 1 = primarily forested and Stratum 2 = primarily non-forested. A county was grouped into one stratum or the other if > 50% of land was of that classification. For the MLR indirect estimation method, models were developed for each stratum individually, meaning that only data belonging to that stratum were used. Estimates were then obtained for each county in the study area using the models from both strata across all sub-samples. Final stratified estimates were derived for each county using the following:

$$\hat{Y}_i = \hat{Y}_{fi}P_{fi} + \hat{Y}_{ni}P_{ni}, \tag{5}$$

where $\hat{Y}_i$ is the stratified estimate for county i, $\hat{Y}_{fi}$ and $\hat{Y}_{ni}$ are the indirect estimates for county i based on the forested and non-forested strata respectively, P_{fi} and P_{ni} are the proportions of NLCD land cover pixels classified as forested and non-forested respectively for county i. This method is a variation on a stratification procedure used by Wilson et al. [28] and Lund [50]. This weighted mean method was used, as opposed to simply using indirect estimates for each county from the stratum in which it resides, due to that fact that forested and non-forested land could not be spatially matched to certain FIA plots due to unknown plot locations. This may also lead to less precise indirect estimates due to a lower number of observational units for either forested or non-forested counties relative to the dataset including all counties. Note that while we initially applied the same stratification technique to KNN, we had to drop this component from the analysis due to an inherent tendency for KNN to produce spurious estimates, which was greatly a result of drastically reduced county "pools" from which the KNN procedure could impute values for non-forested areas. Consequently, KNN had a tendency to drastically overestimate when using this method of stratification.

2.4.4. Composite predictors

Two composite estimators, or predictors, were analyzed in this study. For every sampling intensity, each composite prediction was calculated using the reduced sample DE and one of the aforementioned indirect estimates for each county. Individually, the composite predictors (CP) will be referred to hereafter as MLR_CP and KNN_CP to coincide with the respective indirect components. The composite predictors for each attribute of interest were developed using the following equation [17].

$$\hat{Y}_{iCP} = \varphi_i\hat{Y}_{i2} + (1 - \varphi_i)\hat{Y}_{i1}, \tag{6}$$

where $\hat{Y}_{iCP}$ is the composite predictor of the attribute for the ith county, $\hat{Y}_{i1}$ is the direct estimate of the attribute for the ith county, $\hat{Y}_{i2}$ is the indirect estimate of the ith county, and φ_i is the weight calculated for the ith county. The weights were calculated using a variation of the James-Stein estimator provided in the Appendix [17]. The ultimate purpose of calculating composite predictors is to balance the potential bias of an indirect estimator against the potentially large variance of a direct estimator [15,40].

2.4.5. Estimation of forest attributes by power plant supply area

Although the forest attributes of interest were estimated using each selected SAE technique by county, we were ultimately interested in estimating forest attributes for selected supply areas around coal-fired power plants. Aguilar et al. [5] identified 219 counties in the Northern U.S. with relatively high county-level probability of co-firing using econometric models which incorporated key independent variables pertaining to infrastructure (roads, railways, waterways), population, availability of standing forest biomass, and state-level bioenergy policies, among others. We decided to focus our analysis on circular supply areas around the 219 power plant locations identified by Aguilar et al. [5] based on an assumed 60 km transport radius, which coincides with the economically feasible maximum transport distance for biomass feedstocks identified in Goerndt et al. [6]. A 60 km transport radius results in supply areas that are considerably larger than most individual counties, and in most cases encompass several entire counties and county fragments. For this analysis, the number of power plant locations was reduced to 164 by removing plants with supply areas which crossed into states outside the study area. The omitted supply areas were subsequently added back in later on for actual estimation of forest biomass availability for power plants, using the best estimators identified in the SAE analysis, and for comparison to the results from the earlier study. In the absence of having access to actual FIA plot locations, Goerndt et al. [6] obtained weighted predictions (WP) of forest biomass availability for supply areas were obtained by applying the following formula:

$$WP = \sum_{i=1}^{m} a_i b_i, \tag{7}$$

where b_i is the total annually available forest biomass for county i and a_i is the percentage of the area of county i that falls within the supply area (circle). As such, equation (7) relies on the assumption that the forest biomass resource within each county is uniformly distributed across the entire county [6].

In order to make a direct assessment of the potential of SAE to improve estimation of forest attributes relative to methods of direct estimation (i.e. using design-based inference), we calculated supply area-level WP for each of the 500 replicated sub-samples of each forest attribute using equation (7). Indirect estimates using either MLR or KNN were derived by predicting values for supply areas based on all 500 replicated sub-samples of county-level data [15,40]. In doing so, we used the database of county-level values as the prediction database for supply areas. This was intended as a feature of this study, but also necessary as each of the supply areas includes whole or partial counties already included in the county-level database.

2.4.6. Validation for SAE

Validation at the county-level for each estimator was done using the county-level full sample DE. Subsequently, validation at the supply area-level was accomplished by using the supply-area full sample DE. Note that in a real-world situation, the FIA plot location data necessary to estimate full sample supply area-level DE would probably not be available, but is used in this study for validation only. In order to compare the estimators for each sampling intensity, summary statistics were calculated based upon the county-level validation including relative root mean square error (RRMSE) and relative bias (RB). Precision was assessed using RRMSE [49,51]:

$$RRMSE = \frac{1}{m} \sum_{i}^{m} (RMSE_i / \hat{Y}_{iO}), \tag{8}$$

with

$$RMSE_i = \sqrt{\frac{1}{R} \sum_{r=1}^{R} (\hat{Y}_{iP} - \hat{Y}_{iO})^2}, \tag{9}$$

where $RMSE_i$ is the root mean squared error for county i, m is the number of counties, R is the number of subsamples (iterations), $\hat{Y}_{iP}$ is the predicted value for county i, and $\hat{Y}_{iO}$ is the full sample direct estimate for county i, which is taken as an unbiased estimator of the population total. Similarly, overall bias was assessed using relative bias calculated as

$$RB = \frac{1}{m} \sum_{i=1}^{m} (RB_i), \tag{10}$$

with

$$RB_i = \frac{1}{R} \sum_{r=1}^{R} \left(\frac{\hat{Y}_{iP} - \hat{Y}_{iO}}{\hat{Y}_{iO}} \right), \tag{11}$$

where RB_i is the relative bias for county i.

2.4.7. Potential forest biomass availability estimates for power plants

The practical aspect of this research is to obtain estimates of forest biomass availability for co-firing in power plants located in the U.S. North. Therefore, the final stage of this analysis was to use the best estimator(s) identified through the SAE analysis and validation to obtain useful estimates of forest biomass availability for bioenergy feedstocks within power plant supply areas. Two major estimates of biomass availability provided applicable to forest biomass feedstocks included availability of small diameter trees and availability of logging by-products. Ultimately, we utilized the estimators WP and KNN_CP to derive estimates supply areas with a 60 km procurement radius around the same 219 power plant locations featured in Aguilar et al. [5] and Goerndt et al. [6].

In order to derive estimates representing available biomass from small diameter trees and logging by-products using SAE methods, we needed to simulate a reduced sample size that would be plausible in a practical situation. Therefore, data from only one FIA measurement year from within the original panel was used for the study, which is logical given that one premise of this study is the strong possibility that full FIA panels will often not be available to many individuals due to time frame and the recently increased number of measurement years (e.g. 7 years) to complete one full FIA panel. For this analysis, we chose FIA plots measured in 2005.

One limitation of this analysis was that we were unable to differentiate between timberland and forestland in estimation as in due to lack of necessary spatial delineation of timberland areas with which to coincide the remotely sensed auxiliary data. Forest attributes of particular interest for this assessment were top and limb biomass and small diameter tree biomass. The two metrics, estimated by SAE, provided the basis for calculating available biomass for energy feedstocks from logging by-products and small diameter trees. Deriving final estimates required several assumptions, due to inherent differences in the data available and estimation strategies between this study and others [4–6,10]. To derive final estimates for logging by-products, we used a ratio of volume of annual logging removals [52] to the KNN_CP estimated volume, which was then multiplied by KNN_CP estimated top and limb biomass by power plant supply area. Similarly we needed to adjust the KNN_CP estimates of small diameter biomass to mimic the use of Stand Density Index (SDI) selection, which assumes that area-level removal of small diameter biomass cannot reduce relative SDI to below 30% within a 30 year cutting cycle [6,10,53]. To accomplish this, we used the ratio of full available small diameter biomass to SDI selected biomass as multiplicative adjustment to small diameter biomass availability estimated using KNN_CP.

As in Goerndt et al. [6,10], the SAE estimates of small diameter tree biomass were adjusted using a 65% retention rate of biomass left on site to account for ecological stability, and a subsequent 20% reduction in available biomass to account for inaccessibility. Additionally, an adjustment of 1/30 was applied to small diameter biomass to simulate a 30 year cutting cycle.

3. Results and discussion

The final MLR models for the full sample county-level data were selected using the subsets regression process which identified the seven best variables for each model using BIC and adjusted R^2 as primary selection criteria. Table 3 shows the coefficients and performance statistics of the selected county level models for aboveground biomass, small-diameter biomass, top and limb biomass and total cubic volume.

The MLR models for each of the four forest attributes of interest showed a very high degree of fit, with a minimum adjusted R^2 of 87%

Table 3
Significant regression coefficients for selected county-level models, with adjusted R^2 for each model.

Coefficient	Biomass (Mg ha^{-1})	Small Diameter Biomass (Mg ha^{-1})	Top & Limb Biomass (Mg ha^{-1})	Volume (m^3 ha^{-1})
Intercept	16537.51 (2377.2)	6606.05 (568.47)	9.16 (1.25)	61.79 (12.63)
R1	777.18 (41.4)		0.44 (0.059)	3.13 (0.59)
ta1	−274.31 (41.42)	54.64 (8.41)	−0.16 (0.022)	−0.59 (0.22)
Canopy %	236.71 (9.3)	80.95 (3.02)	0.13 (0.005)	0.98 (0.049)
Forest Cover		736.38 (112.34)		
ma1	81.39 (8.3)	18.42 (2.04)	0.047 (0.004)	0.29 (0.044)
pb1	1452.24 (108.87)		0.71 (0.057)	3.84 (0.58)
pa2		−254.31 (62.51)		
Imperviousness %		−41.24 (4.04)		
SAI	160.94 (22.59)		0.087 (0.012)	0.32 (0.12)
Ecoregion 53	5666.87 (547.63)	825.56 (128.39)	2.88 (0.29)	23.9 (2.9)
R^2 (adj), %	89.2	90.7	90.4	87.1

Coefficient standard errors provided in parentheses.

for Volume and a maximum of 90.7% for small-diameter tree biomass. Each model included at least one of each of the following variable types: seasonal vegetation phenology, climate, land cover (canopy cover, imperviousness), topography (SAI, CTI, DEM), and ecological region (Tables 2 and 3). One notable difference between the models was high significance of pa2 and imperviousness in the model for small-diameter biomass, which was lacking from the other models. This difference most likely denotes a contrast in how certain remotely-sensed variables relate to both presence and abundance of small-diameter biomass compared to the other three forest attributes, which rely much heavier on presence of larger trees. For example, high values for imperviousness relate directly to intensive land management (particularly agricultural) and urbanization. Therefore, it is logical that high imperviousness would have a significant negative effect on predictions of small-diameter biomass. One variable that was consistently significant across all models was Ecoregion 53. This ecoregion represents a relatively small region comprised mainly of the state of Wisconsin (WI) (Fig. 1) [35]. Due to the small size and placement, this ecoregion very explicitly represents an area with a unique combination of high percentage forest cover, high frequency of timber harvest, and relatively low degree of urbanization when compared with many of heavily forested states further to the east. This uniqueness is the mostly likely reason for the significant area-level effect that Ecoregion 53 has in the models for all forest attributes. Including both forested and non-forested FIA plots in the samples meant that many sub-samples produced a DE of zero or at least a very low number, meaning that the MLR models occasionally produced negative values for forest attributes if the values of some auxiliary variables were on the outer boundaries of the ranges of values observed across all counties. This difference in scaling of predicted values compared to the imputation methods is an important factor of the analysis as will be shown in Section 3.

The KNN models were assessed for different values of *k* to determine the optimal number of neighbors to use for imputation. The analysis of root mean square error and bias between models identified k = 15 as being optimal for imputation of the forest attributes of interest. Performance statistics (RRMSE, RB) for each attribute of interest by each of the estimation methods and by sampling intensity are provided in Table 4.

3.1. County-level comparison

3.1.1. Indirect estimation

One of the most readily apparent results in Table 4 is that MLR and KNN outperformed DE in terms of precision at low sampling intensities, with RRMSE values for MLR being 22%–25% lower than DE for 10% sampling intensity, a relative difference often doubled at 5% sampling intensity. Notice that precision tends to increase at a faster rate for DE than for the indirect estimation methods as sampling intensity increases. This is expected, as DE is a design-based estimate using a sub-sample of the full FIA sample per county. MLR had RRMSE levels similar to that of DE at sampling intensity as high as 20% for both aboveground biomass and top and limb biomass.

3.1.2. Composite prediction

Composite prediction showed improvement over indirect estimation with regard to precision and bias for all forest attributes at all sampling intensities. While this is not altogether surprising, it is interesting how much of an improvement in precision over that of DE was achieved by some of the composite predictors. KNN_CP yielded higher precision than DE for most forest attributes at 5%, 10%, and 20% sampling intensity. Even though MLR_CP had higher precision than KNN_CP for most forest attributes and sampling intensities, the opposite was often true for bias. While most composite predictors outperformed their corresponding indirect estimate for most sampling intensities, this was not true at 5% sampling intensity. This is partially an artifact of lower precision of indirect estimation at low sampling intensities, but also the dependence of composite predictors on DE. Notice that this relative increase in RRMSE for composite predictors at 5% sampling intensity corresponds with a similar increase of RRMSE for DE. This comparison highlights the fact that a composite predictor's dependence on DE can be both an advantage and disadvantage depending on the weight placed on DE and the performance of DE relative to indirect estimation.

3.1.3. Optimal county-level estimators

The previous section summarized the descriptive statistics of RRMSE and RB for all county-level estimators at all sampling intensities. It was found that composite prediction generally outperformed all forms of indirect estimation for both precision and bias except at the smallest sampling intensity. Interestingly, composite prediction based on KNN, while of somewhat lower precision compared to MLR-based composite prediction, was superior overall in terms of bias [54,55].

Given these observations regarding precision and bias of all SAE methods, MLR, MLR_ CP, and KNN_CP were selected as the methods to assess for estimation of forest attributes to supply areas surrounding coal-fired power plants.

3.2. Power plant supply area-level estimation

Estimates from MLR, MLR_ CP, and KNN_CP were derived for each of the 164 selected power plant supply circles. Recall that WP was calculated in place of DE in the supply area analysis due to the assumption that FIA plot locations are unknown. Performance statistics (RRMSE and RB) for each attribute of interest by selected estimation method and sampling intensity are shown in Table 5.

Of the selected composite predictors, MLR_ CP consistently produced RRMSE estimates lower than WP and the other SAE methods for all forest attributes at 5% and 10% sampling intensity. Although MLR_

Table 4
Estimated RRMSE and RB from validation of indirect county-level estimation methods for each attribute of interest by sampling intensity (%).

	5%		10%		20%		30%		40%	
	RRMSE	RB	RRMSE	RB	RRMSE	RB	RRMSE	RB	RRMSE	RB
Biomass (Mg ha^{-1})										
MLR	72.4	−37.9	67.8	−38.4	64.8	−38.1	63.4	−37.5	57.2	−37.8
MLR_CP	84.8	−18.5	66.1	−18	51.5	−16.2	43.8	−14.7	39.3	−13.8
KNN	86.2	−34.6	75.6	−34.6	69.4	−35.2	66.9	−40.1	68.3	−32.2
KNN_CP	100.4	−16.1	73.6	−15.8	54.6	−14.7	45.3	−13.7	39.4	−13.1
DE	129.8	NA	91.1	NA	62.9	NA	49.6	NA	41.2	NA
Small Diameter Biomass (Mg ha^{-1})										
MLR	79.9	−35.3	74.4	−35.1	71.1	−34.2	67.5	−35.2	66.2	−35.2
MLR_CP	76.2	−16.5	63.5	−15.2	53.8	−14.1	45.1	−14	40.4	−13.5
KNN	89.7	−38.7	80.1	−39.2	74	−39.2	71.1	−39.4	69.8	−39.1
KNN_CP	126.1	−16.5	68.1	−16.2	53.6	−15.6	46.6	−14.9	42.5	−14.3
DE	110.6	NA	80.4	NA	58.9	NA	49.3	NA	43.4	NA
Top & Limb Biomass (Mg ha^{-1})										
MLR	68.5	−33.2	64.5	−33.8	61.8	−33.5	54.2	−32.9	53.6	−33.2
MLR_CP	83.4	−15.9	64.6	−15.8	50.1	−14.5	41.5	−13.3	37.2	−12.5
KNN	82.5	−40.1	73.2	−39.3	67.5	−39.1	65.2	−39.5	64	−39.4
KNN_CP	97.3	−15.4	71.5	−15.7	53.2	−15.1	44.4	−14.2	39.1	−13.8
DE	127.5	NA	89.6	NA	61.9	NA	48.9	NA	40.6	NA
Volume (m^3 ha^{-1})										
MLR	70.8	−35.2	66.7	−35.7	63.8	−35.4	63.2	−34.9	62.7	−35.2
MLR_CP	85.4	−17.6	66.5	−17.5	51.8	−16.1	48.2	−14.7	43.7	−13.9
KNN	89.4	−42.5	78.5	−43.2	71.9	−43.1	69.4	−42.5	68.1	−43.1
KNN_CP	102.1	−17.2	75.1	−17.5	56.1	−16.8	46.8	−15.9	41.2	−15.5
DE	131.2	NA	92.2	NA	63.7	NA	50.4	NA	41.8	NA

CP was superior to all other estimators in terms of precision, KNN_CP actually resulted in lower RB for all forest attributes and at all sampling intensities. Results of the analysis for supply areas show that 5–10% sampling intensity is the threshold at which indirect and composite estimators begin to outperform DE with regard to precision. This emphasizes the SAE aspect of such a comparison, in that indirect and composite estimators gain prediction strength relative to DE as the available sample size decreases. Notice that MLR_ CP outperforms WP and all other SAE methods at low sampling intensity. This indicates that, even in the absence of specific plot locations, composite predictors using adequate auxiliary data can produce estimates of higher precision than direct estimation methods.

3.2.1. *Potential biomass availability for power plants*

Up to this point, our analysis has demonstrated that SAE methods can provide precise estimates of forest attributes relating to forest biomass availability using relatively small sample sizes. For these methods to be applied to a particular issue, e.g. bioenergy, it is useful to demonstrate how the SAE estimators featured in this study can be used to calculate available forest biomass feedstocks for coal-fired power plants. Available biomass from logging by-products and small diameter trees was estimated using KNN_CP. KNN_CP was used because it was one of the strongest estimators with regard to precision and bias. These estimates were derived for all 219 original power plant locations, assuming a procurement radius of 60 km. Estimates of potential electrical

Table 5
Estimated RRMSE and RB from validation of selected supply area-level estimation methods for each attribute of interest by sampling intensity (%).

	5%		10%		20%		30%		40%	
	RRMSE	RB	RRMSE	RB	RRMSE	RB	RRMSE	RB	RRMSE	RB
Biomass (Mg ha^{-1})										
MLR	28	−11.4	25.3	−11.5	23.9	−11.5	23.4	−11.9	22.8	−11.8
MLR_ CP	25.7	−10.9	22.5	−10.8	20.1	−10.5	18.9	−10.4	18.1	−10.2
KNN_CP	38.3	−0.1	29.8	−0.1	23.8	−0.8	20.6	−1.5	19.1	−1.9
WP	38.2	−7.1	29.5	−7.1	24.1	−7.1	21.6	−7.1	20.1	−7.2
Small Diameter Biomass (Mg ha^{-1})										
MLR	31.5	−10.8	27.7	−10.6	24.9	−11.1	24.1	−11.2	23.2	−11
MLR_ CP	28.4	−9.4	24.1	−8.7	21.4	−7.9	20.2	−7.4	19.4	−6.6
KNN_CP	41.9	4.7	33.1	4.5	26.9	4.7	24.2	4.1	22.5	3.7
WP	41.8	−7.2	33.1	−7.4	27.4	−7.1	24.8	−7.2	23.4	−7.2
Top & Limb Biomass (Mg ha^{-1})										
MLR	27.6	−10.2	25.2	−10.4	23.8	−10.5	23.3	−10.9	22.7	−10.7
MLR_ CP	25.5	−9.9	22.3	−9.8	19.8	−9.6	18.8	−9.6	18.1	−9.3
KNN_CP	37.6	0.1	29.3	0.1	23.5	−0.6	20.6	−1.2	18.9	−1.7
WP	37.3	−6.5	37.3	−6.5	24.1	−6.5	21.6	−6.6	20.3	−6.7
Volume (m^3 ha^{-1})										
MLR	27.2	−11.5	24.6	−11.5	23.1	−11.7	22.6	−12.0	22.1	−11.8
MLR_ CP	25.1	−10.9	22.0	−10.7	19.7	−10.4	18.7	−10.3	17.9	−9.9
KNN_CP	38.2	−0.2	29.5	−0.1	23.3	−0.7	20.4	−1.2	18.6	−1.7
WP	38.1	−6.3	29.7	−6.3	24.1	−6.3	21.5	−6.3	20.1	−6.5

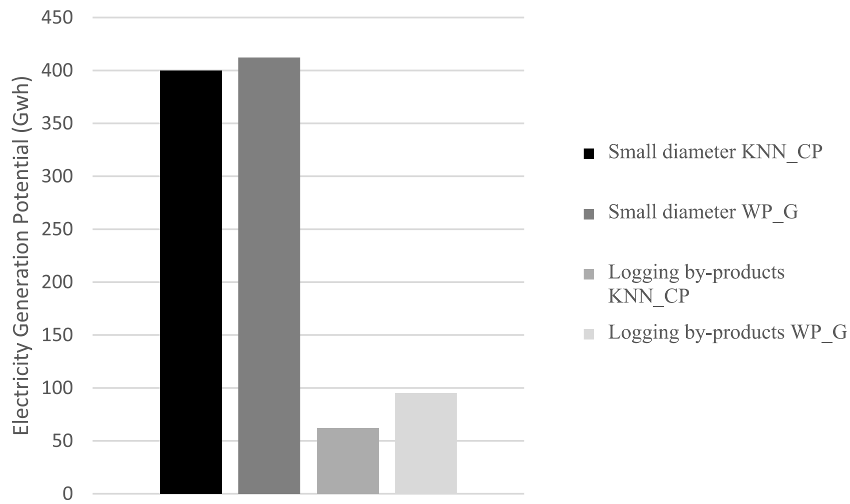

Fig. 2. Comparison of average annual electricity generation potential between KNN_CP estimates and WP_G.

power generation were obtained by applying an adjustment of 1.7 Mwh per dry metric ton [4,6].

Based on the KNN_CP estimates, the average forest biomass annually available across all power plants was approximately 35,000 dry metric tons for logging by-products and 235,000 dry metric tons for small diameter trees. This equates to an average of 400 Gwh potential annual electricity generation per power plant for small diameter biomass and 62 Gwh for logging by-products. The WP estimates from Goerndt et al. [6], hereafter referred to as WP_G yielded an average of 412 Gwh potential annual electricity generation per power plant for small diameter biomass and 95 Gwh for logging by-products. Fig. 2 compares average annual electricity generation potential between the estimates from this study and WP_G.

From the standpoint of direct comparison, Fig. 2 highlights two important features. First, KNN_CP provides estimates for small diameter biomass per power plant supply area that are very similar to WP_G Second, there is a clear tendency for KNN_CP to provide lower estimates for logging by-products. The latter is not altogether surprising, given that our method of obtaining final estimates of logging by-products is somewhat different from WP_G in that we do not rely on logging residues data from TPO [6,10]. What is interesting is the tendency to provide *lower* estimates, highlighting the variation between inventory data and industry data with regard to potential annual removals of forest biomass.

Prior to discussing observed plant-by-plant differences in the KNN_CP and WP_G estimates, it is useful to assess the spatial trends in estimation for WP_G, as it will be used as a baseline for comparing and discussing major trends in estimation by KNN_P. Fig. 3 shows annual electricity generation potential per plant based on the WP_G estimates for logging by-products and small diameter trees [4].

While it was expected that there would be higher levels of co-firing potential generation from small-diameter biomass compared to logging residues, there are apparent spatial trends in generation potential across the Northern U.S. More specifically, power plants in Northeast portion of the region such as (e.g. Pennsylvania (PA) and West Virginia (WV)) and power plants in the Great Lakes Region (e.g. Minnesota (MN), Wisconsin (WI) and Michigan (MI)). These spatial trends are a result of multiple factors pertaining to availability of biomass feedstocks, infrastructure, renewable resource policy and supply chain logistics. Recall that selection of power plant locations observed in this analysis stemmed from an econometric analysis conducted by Aguilar et al. [5], which reported high importance of variables describing location-specific factors affecting supply-chain logistics as well as renewable energy policy, which were also identified as significant factors in analyses by Ko and Lautala [56], Eksioglu et al. [57] and Berggren et al. [58]. Of equal importance to transport logistics, renewable energy policy and local economics when it comes to co-firing generation potential is the abundance and sustainable potential supply of biomass feedstocks from forestland within economically feasible transport distances, which was a primary focus of the WP_G estimates and is emphasized studies conducted by Hansson et al. [59] and Roni et al. [8]. WP_G for both small-small diameter trees and logging residues serves as a baseline for analyzing the spatial trends in KNN_CP estimates obtained from our analysis.

As will be illustrated shortly, the tendency for KNN_CP to provide lower estimates for potential logging by-products (Fig. 2) persists for most individual power plant supply areas, but varies in extremity across the region. To better illustrate power plant level variation between the KNN_CP estimates and WP_G, we calculated the percentage difference (KNN_CP − WP_G) relative to WP_G. Fig. 4 shows the resulting percentage difference in estimated available biomass for logging by-products and small-diameter trees.

While there was an overall tendency for KNN_CP to yield lower estimates for logging by-products than WP_G there is a noticeable spatial trend for high estimation in the Northcentral States (e.g. Minnesota (MN), Iowa (IA), Illinois (IL)). Supply areas with a high estimation occur primarily in areas with a high percentage of nonforested land. Interestingly, most supply areas where KNN_CP estimates lower logging by-products occur in the same states or adjacent states. The most likely reason for this was variability in the same state or subregion with regard to forested land. Conversely, KNN_CP had a tendency to produce higher estimates of small diameter biomass in states with a high percentage of forestland such as Pennsylvania (PA) and West Virginia (WV), which is most likely an effect of the indirect estimation (KNN) component of KNN_CP. Areas with a much higher percentage of forest land coincided with notably different values for many of the remote sensing metrics used during the modeling process.

For both types of biomass feedstock, the spatial trends observed across the region were greatly an effect of spatial variation of both forest characteristics and auxiliary information influencing the indirect estimation component (WP) of KNN_CP. More specifically, Fig. 4 illustrates the influence of using an estimator that does not rely on an assumption of uniformity across large areas. A major objective of this study was to obtain precise estimates of forest attributes pertaining to biomass availability without having to assume uniformity of forest attributes by county. This is crucial because as the complexity of SAE methods increases (e.g. composite prediction), one is more likely to violate the assumption of county- or area-level uniformity of forest attributes. Deviation from the assumption of uniformity is not only desired, but also necessary to utilize SAE methods that provide much

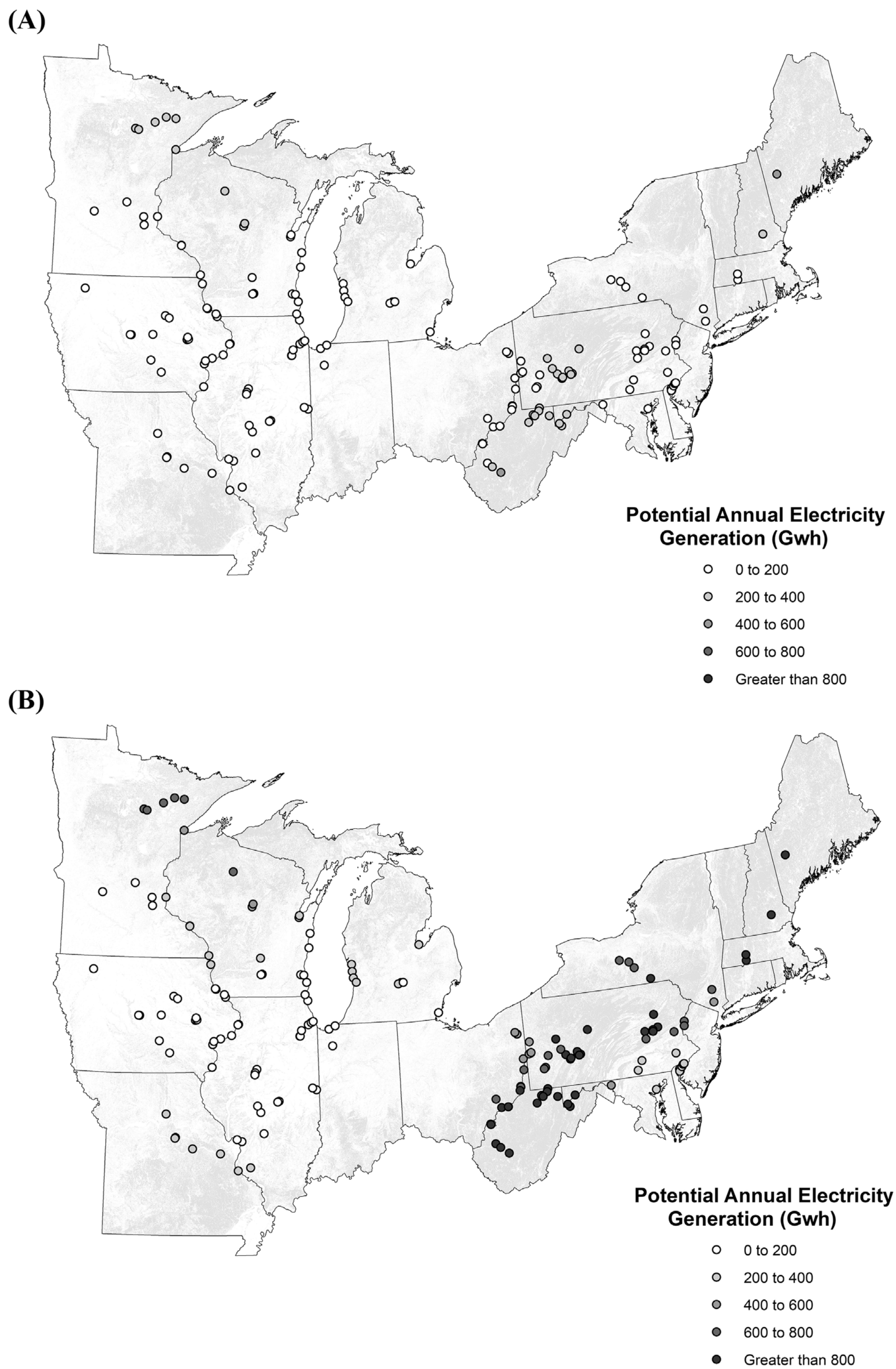

Fig. 3. Estimated potential electricity generation from forest biomass as estimated by WP_G in Gwh for each selected power plant assuming a 60 km concentric procurement radius for (A) logging by-products and (B) small-diameter trees. Background shading represents forest cover.

more power for capturing spatial variation across the landscape. The most practical outcome of this study is an ability to estimate forest biomass availability by power-plant and to illustrate how much electricity could be annually generated relative to current generation rates. Fig. 5 shows the percentage of coal generated electricity per power plant that could potentially be generated with forest biomass based on the KNN_CP estimates [20].

Aside from the noticeably higher levels of co-firing generation potential from small-diameter trees compared to logging by-product estimates, there were noticeable trends in spatial distribution of power plants with high co-firing potential for both sources of biomass. Power plants with high generation potential from logging by-products were more sporadically distributed when compared to small-diameter trees, but one can see that many plants in the Great Lake states and the Northeast have high co-firing potential [5,6]. Many of the same areas denoted high co-firing potential with small-diameter trees, but there

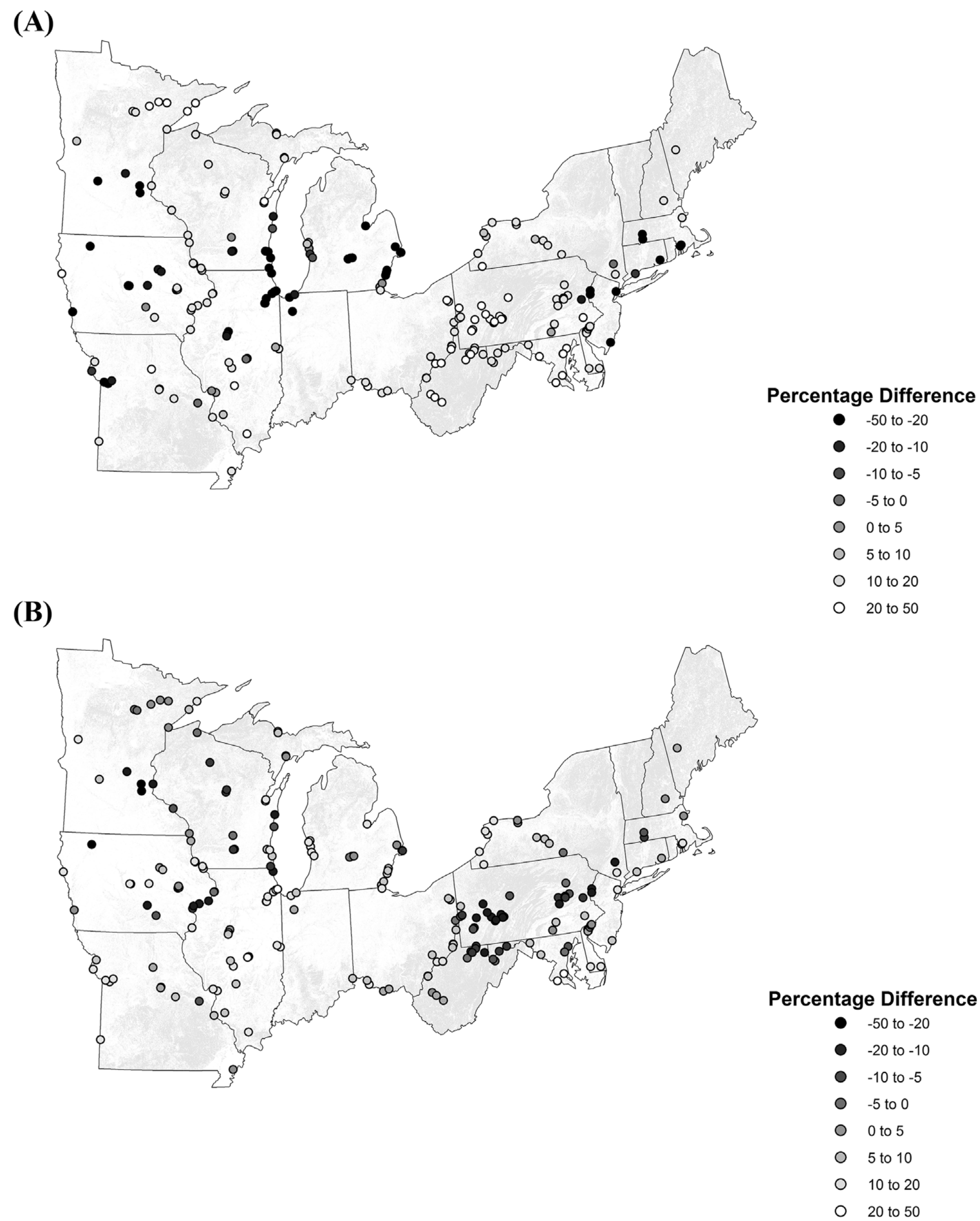

Fig. 4. Percentage difference between estimates from KNN_CP and WP_G for (A) logging by-products, and (B) small diameter trees. Background shading represents forest cover.

were evident differences in co-firing potential between the two biomass sources in states such as Ohio (OH), Pennsylvania (PA) and West Virginia (WV), where most power plants show a high co-firing potential with small diameter trees and a low co-firing potential with logging by-products. This trend actually contrasts with WP_G, which is most likely due to lower dependence of the SAE methods on the uniformity assumption of forest attributes across counties. This is logical due to the tight clustering of power plants within these particular states, leading to little variation in forest attribute estimates between power plant supply areas when using direct estimation. The results of this study show a relatively high degree of variation in biomass estimates within such areas compared to simple direct estimates such as WP_G.

4. Conclusions

Precise estimates of forest attributes at varying spatial scales are essential for forest managers and forest industry to make important decisions. In particular to the use of forest biomass in co-fired systems, past estimates of forest biomass availability have relied on assumptions of forest resource uniformity across entire counties and other geo-political boundaries due to limitations in spatial estimation using FIA and other forest inventory systems. Additionally, limitations in sample size may make it more difficult in the future to obtain precise direct estimates of forest attributes in potential supply areas that can include many counties and county fragments. SAE can utilize estimators with potential for precise and low-bias estimates of forest attributes at both the county- and supply area-levels using very small sample sizes.

At the supply-area level, composite predictors were the most precise estimation methods at low (5–10%) sampling intensity. Interestingly, while the composite predictor using an MLR indirect component (MLR_CP) was generally the best estimator at the county-level, the composite predictors using a KNN indirect component (KNN_CP) tended to be superior in terms of reduced bias. This was particularly apparent for estimation of forest attributes for supply-areas, demonstrating an interesting trade-off between precision and bias when assessing the use of SAE. Therefore, KNN_CP is deemed the ideal estimator for obtaining

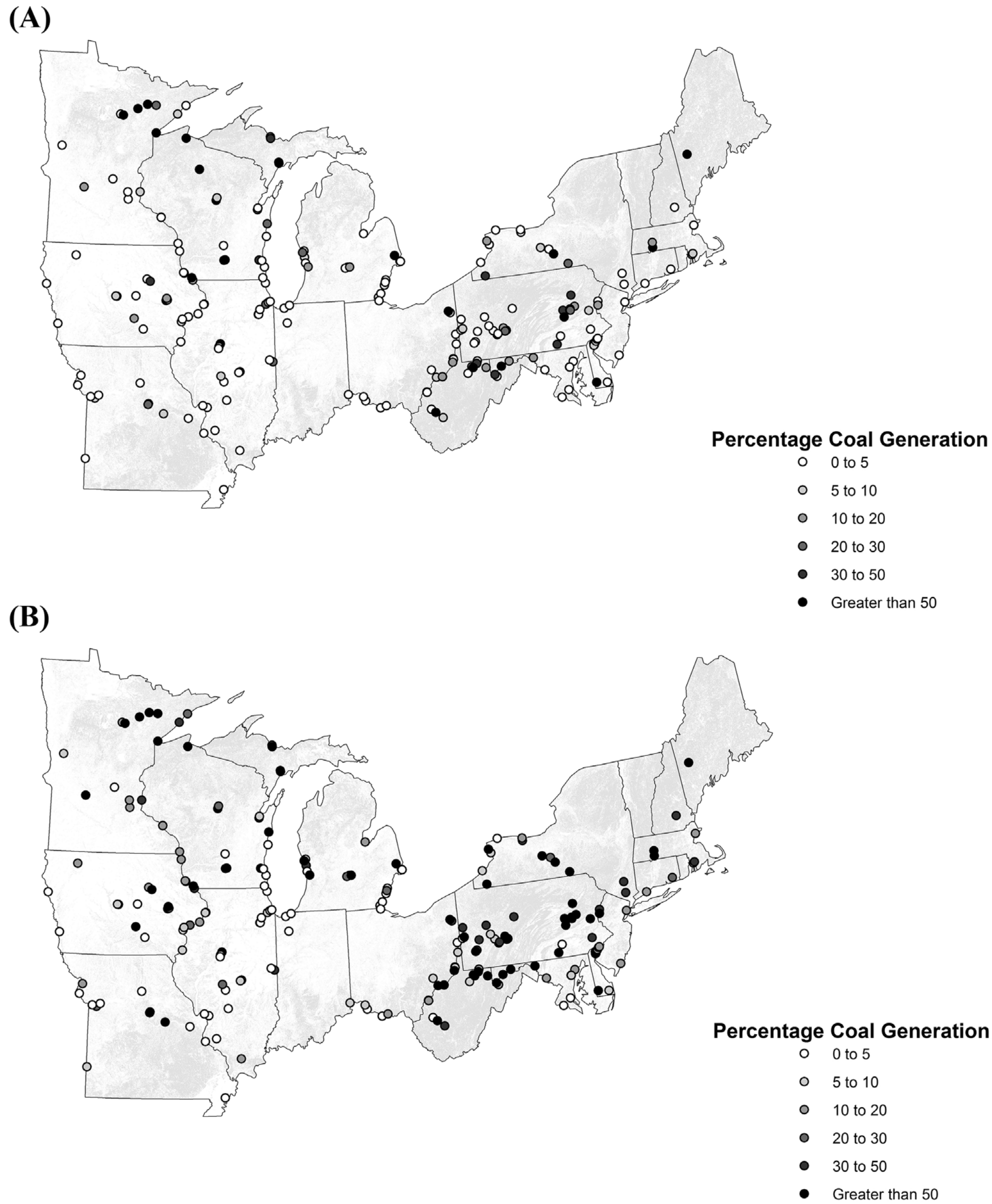

Fig. 5. Estimated percentage of annual coal electricity generation that could potentially be replaced by forest biomass for each selected power plant at a 60 km procurement radius based on the KNN_CP estimates for (A) logging by-products, and (B) small diameter trees. Background shading represents forest cover.

estimates of biomass availability for power plants derived from logging by-products, or harvest of small diameter trees.

The estimates of forest biomass availability for co-firing highlighted the ability of the SAE methods to capture variability in forest attributes pertaining to biomass availability at varying spatial scales. In particular, we found that the lack of dependency on uniformity assumption of forest attributes by county enables SAE methods to capture spatial variability even for power plants that are clustered closely together, which tend to have relatively homogeneous biomass estimates when only using direct estimation. By comparing several SAE methods at varying sampling intensities, we were ultimately able to identify several potential estimators that can be useful for estimating forest attributes even when inventory data for the specific supply area of interest is either limited or non-existent. These estimators will be extremely useful for assessing forest biomass availability for facilities that are utilizing biomass for energy generation or are considering such activity in the future.

Acknowledgements

We thank the faculty and staff of the William H. Darr College of Agriculture at Missouri State University for their support in completing this project. We are grateful to Bill Dijak of the USFS Northern Research Station for his assistance and excellent advice on processing and utilizing digital elevation models. We are also grateful to Elizabeth Burrill of the USFS for helping to facilitate acquisition of FIA plot locations, as well as Pat Miles of the USFS Northern Research Station for his assistance with plot-level direct estimation of forest attributes. This project was partially funded under USFS and University of Missouri Joint Venture Agreement 12-JV-11242305-120. This publication is not intended to reflect the opinions of these organizations. Any errors remain the responsibility of the authors.6. Literature Cited.

Appendix

2007 annual Landsat composite coverage was adjusted by assuming the following:

$$Y \approx GX + B,$$

where Y is the primarily SLC-off primary scene (2007), X is the primarily SLC-on fill scene (2002), G is the gain used to histogram match the fill scene to the primary scene, and B is the bias used to histogram match the fill scene to the primary scene.

These are defined as:

$$G = \frac{\sigma_Y}{\sigma_X}, \tag{A1}$$

and

$$B \approx \bar{Y} + G\bar{X}, \tag{A2}$$

where σ_X is the standard deviation of data in the primary scene, σ_Y is the standard deviation of data on the fill scene. Predicted values of Y were applied to the 2007 annual WELD composite for pixels omitted by the SLC malfunction. σ_Y and σ_X were estimated by using a focal moving window in ArcGIS for the 17 × 17 grid of pixels surrounding each pixel of interest.

Variation of composite predictor weight derivation by Costa et al. [9].

$$\hat{\varphi}_i = \frac{\hat{\psi}_i}{\sum_i^m \hat{\psi}_i + (\hat{Y}_{i2} - \hat{Y}_{i1})^2}, \tag{A3}$$

with

$$\hat{\psi}_i = \frac{V_e}{n_i}, \tag{A4}$$

and

$$V_e = \frac{\sum_i^m a_i s_i^2}{\sum_i^m a_i}, \tag{A5}$$

where a_i is the total area (ha) in county *i*, s_i^2 is the sample variance for county *i*, V_e is the weighted sample variance for county *i*, and n_i is the ground sample size for county *i*. Note that $\hat{\psi}_i$ is a form of "smoothed" variance used in place of raw sample variance in the composite estimator to provide stability to the weights [15].

References

[1] Swedish Energy Agency, Swedish Energy Statistics, (2018) Available at: http://www.energimyndigheten.se/en/facts-and-figures/statistics/ , Accessed date: 12 November 2018.

[2] OSF (Official Statistics of Finland), Energy Supply and Consumption, Statistics Finland, Helsinki, 2018 Available at: http://www.stat.fi/til/ehk/tau_en.html , Accessed date: 12 November 2018.

[3] Energy Information Administration, Annual energy outlook, Available at: https://www.eia.gov/outlooks/aeo/pdf/0383(2014).pdf, (2014) , Accessed date: 23 April 2018Washington, D.C. (2014).

[4] M.E. Goerndt, F.X. Aguilar, P. Miles, S. Shifley, N. Song, H. Stelzer, Regional assessment of forest biomass as an energy feedstock for combined combustion in the U.S. Northern region, J. For. 110 (3) (2012) 138–148.

[5] F.X. Aguilar, M. Goerndt, N. Song, S. Shifley, Internal, external and location factors influencing cofiring of biomass with coal in the U.S. Northern region, Energy Econ. 34 (2012) 1790–1798.

[6] M.E. Goerndt, F.X. Aguilar, K. Skog, Resource potential for renewable energy generation from co-firing of forest biomass with coal in the Northern U.S, Biomass Bioenergy 59 (2013) 348–361.

[7] L. Baxter, Biomass-coal co-combustion: opportunity for affordable renewable energy, Fuel 84 (2005) 1295–1302.

[8] M. Roni, S. Eksioglu, E. Searcy, K. Jha, A supply chain network design model for biomass co-firing in coal-fired power plants, Transport. Res. E Logist. Transport. Rev. 61 (2014) 115–134.

[9] A. Zaliwski, A. Faber, P. Rafal, M. Biberacher, S. Gadocha, M. Borzecka, Biomass supply for co-firing in main-network power stations in Poland, J. Food Agric. Environ. 11 (2013) 2031–2035.

[10] M.E. Goerndt, F.X. Aguilar, K. Skog, Potential for Coal Power Plants to Co-fire with Woody Biomass in the U.S. North, 2010-2030: a Technical Document Supporting the Northern Forest Futures Project. USDA Forest Service Gen. Tech. Rep. FPL-GTR-237, USDA Forest Service Forest Products Laboratory, 2015 19p.

[11] J. Fridman, S. Holm, M. Nilsson, P. Nilsson, A. Hedstrom Ringvall, G. Stahl, Adapting national forest inventories to changing requirements - the case of the Swedish National Forest Inventory at the turn of the 20th century, Silva Fenn. 48 (3) (2014) 29p.

[12] E. Tomppo, The Finnish multisource national forest inventory: small-area estimation and map production, in: R. McRoberts, G. Reams, P. Van Deusen, W. McWilliams (Eds.), Proceedings of the Eighth Annual Forest Inventory and Analysis Symposium, USDA Forest Service, Monterey CA, 2006October 16-19 pg. 341-349. Gen. Tech. Report WO-79.

[13] M. Gillis, Canada's national forest inventory (responding to current information needs), Environ. Monit. Assess. 67 (2001) 121–129.

[14] P.D. Miles, Forest Inventory EVALIDator Web-application Version 1.5.00, U.S. Department of Agriculture, Forest Service, Northern Research Station, St. Paul, MN, 2012 Available at: https://apps.fs.usda.gov/Evalidator/evalidator.jsp , Accessed date: 23 April 2018.

[15] M.E. Goerndt, V. Monleon, H. Temesgen, A comparison of small-area estimation techniques to estimate selected stand attributes using LiDAR-derived auxiliary variables, Can. J. Res. 41 (2011) 1189–1201.

[16] P. Heady, P. Clarke, G. Brown, K. Ellis, D. Heasman, S. Hennell, J. Longhurst, B. Mitchell, Small Area Estimation Project Report. Model-based Small Area Estimation Series, Office of National Statistics, London, UK, 2003, p. 15.

[17] A. Costa, A. Sattora, E. Ventura, An empirical evaluation of small area estimators, SORT 27 (2003) 113–135.

[18] A. Costa, A. Sattora, E. Ventura, Improving small area estimation by combining surveys: new perspectives in regional statistics, SORT 30 (2006) 101–122.

[19] F.X. Aguilar, H. Garrett, Perspectives of forest biomass for energy: Survey of state foresters, state energy biomass contacts, and National Council of Forestry Association executives, J. For. 107 (2009) 297–306.

[20] Environmental Protection Agency, eGRID2012 Version 1.0, (2012) Available at: https://www.epa.gov/energy/egrid2012-version-10-summary-tables-year-2009-data , Accessed date: 23 April 2018.

[21] T.W. Gillespie, Predicting forest plant species richness in tropical dry forests: a case study from south Florida, USA, Ecol. Appl. 15 (1) (2005) 27–37.

[22] K. Feeley, T. Gillespie, J. Terborgh, The utility of spectral indices from Landsat ETM + for measuring the structure and composition of tropical dry forests, Biotropica 37 (4) (2005) 508–519.

[23] P. Scaramuzza, E. Micijevic, G. Chander, SLC Gap-filled Products: Phase One Methodology, (2004) Available at: https://landsat.usgs.gov/sites/default/files/documents/SLC_Gap_Fill_Methodology.pdf , Accessed date: 23 April 2018.

[24] E.H. Wilson, S.A. Sader, Detection of forest type using multiple dates of Landsat TM imagery, Rem. Sens. Environ. 80 (2002) 385–396.

[25] I. Olthof, D. King, R. Lautenschlager, Mapping deciduous forest ice storm damage using Landsat and environmental data, Remote Sens. Environ. 89 (2004) 484–496.

[26] J. Dewitz, J. Fry, M. Coan, N. Hossain, C. Larson, N. Herold, A. McKerrow, J.N. VanDriel, J. Wickham, Completion of the 2001 national land cover database for the conterminous United States, Photogramm. Eng. Rem. Sens. 73 (4) (2001) 337–341 https://catalog.data.gov/dataset/usgs-national-land-cover-dataset-nlcd-downloadable-data-collection.

[27] J. Fry, G. Xian, S. Jin, J. Dewitz, C. Homer, L. Yang, C. Barnes, N. Herold,

J. Wickham, Completion of the 2006 national land cover database for the conterminous United States, Photogramm. Eng. Rem. Sens. 77 (9) (2006) 858–864 https://catalog.data.gov/dataset/usgs-national-land-cover-dataset-nlcd-downloadable-data-collection.
[28] B. Wilson, A. Lister, R. Riemann, A nearest-neighbor imputation approach to mapping tree species over large areas using forest inventory plots and moderate resolution raster data, For. Ecol. Manag. 271 (2012) 182–198.
[29] P.J. Sellers, S.O. Los, C.J. Tucker, C.O. Justice, D.A. Dazlich, G.J. Collatz, D.A. Randall, 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. Rem. Sens. 15 (17) (1994) 3519–3545.
[30] P.E. Thornton, S.W. Running, M.A. White, Generating surfaces of daily meteorological variables over large regions of complex terrain, J. Hydrol. 190 (1997) 214–251.
[31] USGS (United States Geographical Survey), National elevation dataset (NED), Available at: https://lta.cr.usgs.gov/NED, (2008) , Accessed date: 23 April 2018.
[32] I.D. Moore, R.B. Grayson, A.R. Ladson, Digital terrain modelling: a review of hydrological, geomorphological, and biological applications, Hydrol. Process. 5 (1991) 3–30.
[33] P.E. Gessler, I.D. Moore, N.J. McKenzie, P.J. Ryan, Soil–landscape modeling and spatial prediction of soil attributes, Int. J. Geogr. Inf. Syst. 9 (4) (1995) 421–432.
[34] T.D. Frank, Mapping dominant vegetation communities in the Colorado Rocky mountain front range with Landsat thematic mapper and digital terrain data, Photogramm. Eng. Rem. Sens. 54 (12) (1988) 1727–1734.
[35] Level III and IV Ecoregions of the Continental United States, US Environmental Protection Agency, 2017 Available at: https://www.epa.gov/eco-research/level-iii-and-iv-ecoregions-continental-united-states , Accessed date: 12 November 2018.
[36] R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2008.
[37] T. Lumley, Leaps: Regression Subset Selection. R Package Version 2.9, (2009) Available at: http://cran.rproject.org/web/packages/leaps/index.html , Accessed date: 23 April 2018.
[38] V. Lemay, H. Temesgen, Comparison of nearest neighbor methods for estimating basal area and stems per acre using aerial auxiliary variables, For. Sci. 51 (2) (2005) 109–119.
[39] B. Eskelson, H. Temesgen, V. Lemay, T. Barrett, N. Crookston, A. Hudak, The roles of nearest neighbor methods in imputing missing data in forest inventory and monitoring databases, Scand. J. For. Res. 24 (2009) 235–246.
[40] M.E. Goerndt, V. Monleon, H. Temesgen, Small-area estimation of county-level forest attributes using ground data and remote sensed auxiliary information, For. Sci. 59 (5) (2013) 536–548.
[41] S. Tuominen, S. Fish, S. Poso, Combining remote sensing, data from earlier inventories, and geostatistical interpolation in multisource forest inventory, Can. J. Res. 33 (2003) 624–634.
[42] R. McRoberts, M. Nelson, D. Wendt, Stratified estimation of forest area using satellite imagery, inventory data, and the k-nearest neighbors technique, Rem. Sens. Environ. 82 (2002) 457–468.
[43] T. Cover, Estimation by the nearest neighbor rule, IEEE Trans. Inf. Theor. 14 (1) (1968) 50–55.
[44] G. Chirici, M. Mura, D. McInerney, N. Py, E. Tomppo, L. Waser, D. Travaglini, R. McRoberts, A meta-analysis and review of the literature on the K-nearest neighbors technique for forestry applications that use remotely sensed data, Rem. Sens. Environ. 176 (2016) 282–294.
[45] M. Maltamo, A. Kangas, Methods based on k-nearest neighbor regression in the prediction of basal area diameter distribution, Can. J. Res. 28 (8) (1998) 1107–1115.
[46] H. Temesgen, T. Barrett, G. Latta, Estimating cavity tree abundance using nearest neighbor imputation methods for western Oregon and Washington forests, Silva Fenn. 42 (3) (2008) 337–354.
[47] N.L. Crookston, A. Finley, yaImpute: an R package for k-NN imputation, J. Stat. Software 23 (10) (2007) 1–16.
[48] J. Breidenbach, R. Astrup, Small area estimation of forest attributes in the Norwegian National Forest Inventory, Eur. J. For. Res. 131 (2012) 1255–1267.
[49] F. Mauro, V. Monleon, H. Temesgen, K. Ford, Analysis of area level and unit level models for small areas estimation in forest inventories assisted with LiDAR auxiliary information, PLoS One 12 (12) (2017) e0189401.
[50] H. Lund, When is a forest not a forest? J. For. 100 (8) (2002) 21–28.
[51] J.N.K. Rao, Small Area Estimation, Wiley Series in Survey Methodology. John Wiley and Sons, Hoboken, NJ, 2003 283 p.
[52] TPO (Timber Products Output Database), USDA Forest Service Southern Research Station, 2012 Available at: https://www.fia.fs.fed.us/program-features/tpo/ , Accessed date: 23 April 2018.
[53] S. Hall, P. Marchand, Effects of stand density on ecosystem properties in subalpine forests of the southern Rocky Mountains, Ann. For. Sci. 67 (2010) 102.
[54] A. Gjertsen, Accuracy of forest mapping based on Landsat TM data and KNN based method, Rem. Sens. Environ. 110 (2007) 420–430.
[55] A. Finley, R. McRoberts, A. Ek, Applying an efficient k-nearest neighbor search to forest attribute imputation, For. Sci. 52 (2) (2006) 130–135.
[56] S. Ko, P. Lautala, Advanced woody biomass logistics for co-firing in existing coal power plant: case study of the Great Lakes States, Transport. Res. Rec. (2018), https://doi.org/10.1177/0361198118797806.
[57] S. Eksioglu, A. Acharya, L. Leightley, S. Arora, Analyzing the design and management of biomass-to-biorefinery supply chain, Comput. Ind. Eng. 57 (4) (2009) 1342–1352.
[58] M. Berggren, E. Ljunggren, F. Johnsson, Biomass Co-firing Potentials for Electricity Generation in Poland - Matching Supply and Co-firing Opportunities vol. 32, (2008), pp. 865–879 (9).
[59] J. Hansson, G. Berndes, F. Johnsson, J. Kjrstad, Co-firing biomass with coal for electricity generation - an assessment of the potential in EU27, Energy Pol. 37 (4) (2009) 1444–1455.