Peer-reviewed version available at *Forests* **2019**, *10*, 1045; doi:10.3390/f10111045

Article

A Regularized Raking Estimator for Small Area Mapping from Forest Inventory Surveys

Nicholas N Nagle [1]*, Todd A Schroeder [2] and Brooke Rose [1,†]

[1] Department of Geography, University of Tennessee, Knoxville TN 37996
[2] USDA Forest Service, Southern Research Station, Knoxville, TN 37910
* Correspondence: nnagle@utk.edu
† Current address: Botany and Plant Sciences, University of California, Riverside, CA, 92521

Abstract: We propose a new estimator for creating expansion factors for survey plots in the USDA Forest Inventory and Analysis program. This estimator was previously used in the GIS literature where it was called Penalized Maximum Entropy Dasymetric Modeling. We show here that the method is a regularized version of the raking estimator widely used in sample surveys. The regularized raking method differs from other predictive modeling methods for integrating survey and ancillary data in that it produces a single set of expansion factors that can have general purpose use to produce small area estimates and wall-to-wall maps of any plot characteristic. This method also differs from other more widely used survey techniques, such of GREG estimation, in that it is guaranteed to produce positive expansion factors. We extend the previous method here to include cross-validation, and provide a comparison to expansion factors between the regularized raking and ridge GREG survey calibration.

Keywords: FIA; forest inventory; small area estimation; survey weight)

---

1. Introduction

Surveys for inventories of forests - such as the USDA Forest Service Forest Inventory and Analysis (FIA) program - are typically designed to provide reliable estimates of characteristics over large spatial units, such as states. Managers, however, often desire estimates over much smaller units, or even continuous ("wall-to-wall") maps of natural resources. Unfortunately, the high cost of sampling prevents the ability to create direct survey estimates at such high resolutions.

With the increasing availability of wall-to-wall, remotely-sensed data and the computational power to process these data, there is a great deal of research toward providing new estimators capable of utilizing both field-based survey data and auxiliary spatial data to produce high resolution, wall-to-wall maps of forest characteristics. Two types of methods have primarily been used to produce wall-to-wall maps from plot data, these include: 1) empirical predictive models, such as classification and regression trees and Random Forests ([1,2]), which combine satellite-derived composites, GIS layers, and FIA plot data to predict maps of FIA attributes such as forest biomass [3,4], forest type [5], and cause of disturbance [6,7] and 2) interpolation techniques such as k-Nearest Neighbors Classification [8,9], gradient nearest neighbor (GNN) [10], and kriging [11] which combine satellite-derived products and FIA plot data to impute maps of FIA attributes, such as forest biomass [12–14], stand density and volume [15] and tree species distribution [16].

Although the two methods differ statistically, the common theme is that they both are highly tailored to produce reliable estimates for a single response variable. While these approaches can certainly produce reliable results, they can suffer from two possible shortcomings for general use by public agencies. First, models that are tailored to estimating a specific variable, may not produce

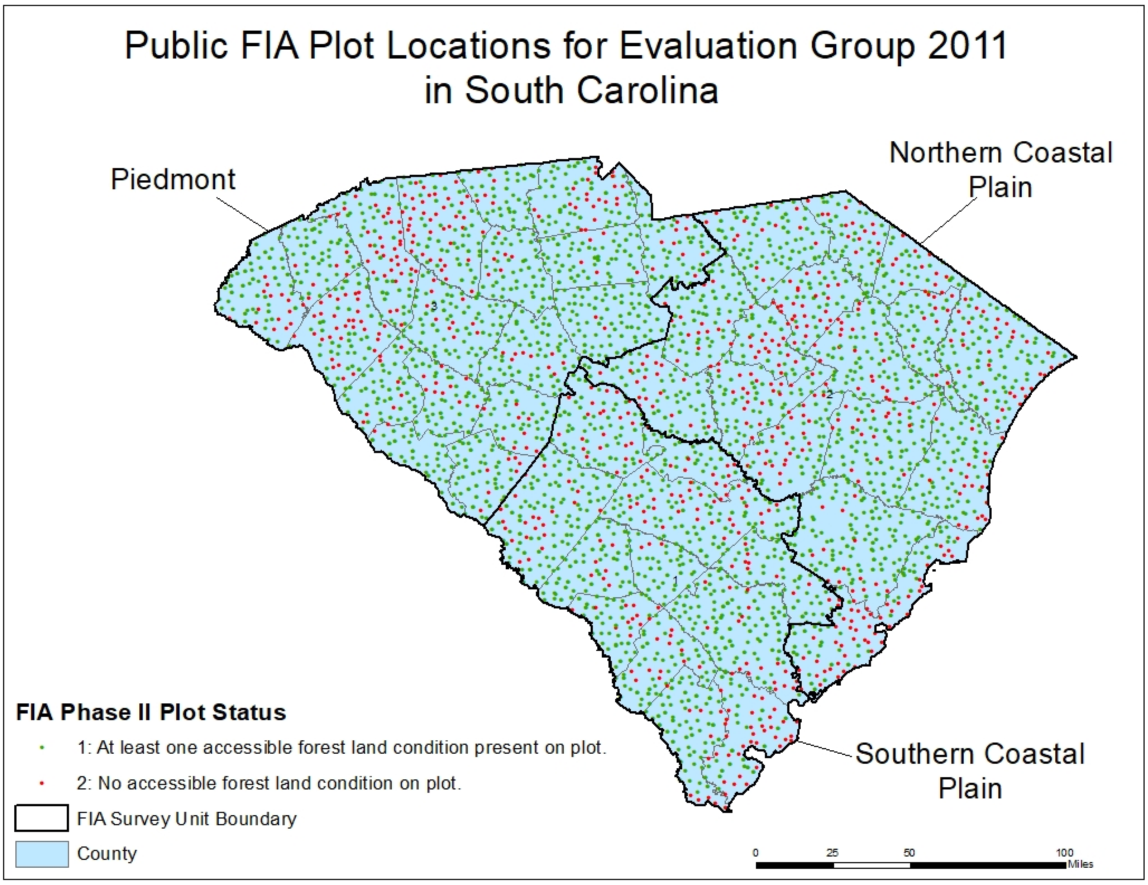


Figure 1. FIA plot locations and boundaries of counties and survey units. South Carolina has three survey units. In 2011, the smallest survey unit in South Carolina contained 672 plots with some forest use, whereas the smallest county contained only 25 such plots.

reasonable results when applied to other variables, thus greater specificity may come at a loss of generality. Second, these predictive models often produce estimates that are inconsistent across spatial scales or with "official" published estimates. For example, an empirical model that produces a wall-to-wall map of land use or basal area may not add up to state-level official estimates of land use or basal area. Such inconsistencies can create obstacles for the use of such estimates in official business.

In the Eastern United States, the USDA FIA survey uses a permanent network of plots that are visited on a rotating five-year cycle. These plots have been selected through a spatially stratified random design with a density of approximately 1 plot per 2400 ha (6000 acres). This sampling density certainly does not permit detailed resource mapping, and for most of the United States, does not even permit reliable direct estimates of county averages. For official reporting, the FIA has created custom *survey units*, which are agglomerations of counties that typically contain hundreds of forested sample plots.

To illustrate the approximate scale of these sample densities and survey units, Figure 1 shows a map of the FIA plot locations in South Carolina, as well as county and survey unit boundaries (the locations are adjusted by FIA to preserve the confidentiality of landowners). The entire state of South Carolina is divided into only three survey units. While the survey units do roughly correspond to ecological zones, they are much larger than a single county and are too large for many desired uses by forest managers.

In addition to the official publications of characteristics for survey units, the FIA also publishes a public version of the plot- and tree-level data (with suitable anonymization). Published along with these plot-level data are survey weights, also called *expansion factors*. These expansion factors allow power users inside and outside the USFS to produce custom, large-area estimates that are not otherwise published. By using these expansion factors, users are guaranteed to produce estimates that are consistent with other official estimates. Unfortunately, these expansion factors are designed for estimates at the spatial scale of the survey unit and are not appropriate for small area use. Additionally, these expansion factors are designed for forest area estimates, and may not be suitable for estimates of other FIA variables.

In this article, we show a method for producing expansion factors more suitable for producing small area estimates and wall-to-wall maps of survey characteristics. This method is designed to

- accommodate multiple sources of ancillary information,
- be applicable to all survey characteristics,
- allow mathematical consistency with published estimates at large scales,
- maintain reasonable properties regarding the increase of uncertainty and variance that can be expected with small area estimation.

In this paper, we define "small area" to be any region smaller than can be supported by a direct estimate from plot-level data. In the context of the FIA, that includes estimates for regions as large as the county, however we will produce expansion factors for homogeneous patches that are smaller than counties.

The method we use was originally developed in the geography literature as a modification of a existing *dasymetric mapping* techniques for producing wall-to-wall maps from tables published by the US Census Bureau [17]. Here, we tie that method to existing survey estimation techniques used by FIA, known as raking estimators, and call the new method "regularized raking".

The product of our regularized raking method is a set of small area expansion factors w_{it} that match each database record (plot) i to each small area t. Conceptually, these expansion factors also represent a probability distribution of records for each small area, i.e. $\frac{w_{it}}{\sum_i w_{it}}$ represents a sort of probability that that plot i represents place t. Using these expansion factors, it is possible to produce small area estimates for every characteristic in the FIA database. A wall-to-wall map can be produced by selecting small enough areas.

The research most similar in objective to ours are Ohmann and Gregory [10] and Riley et al [18], who fit GNN and random forests, respectively, to determine a single plot that most represents each pixel. Regularized Raking, however, rather than providing a *single* best plot for each place, provides a *distribution* across the sample plots for each place, to provide information not only about the "best" plot, but how likely each plot is to represent that pixel. Additionally, instead of producing a single expansion factor representing the plot's contribution to the survey unit, our method provides a map of expansion factors for each plot.

2. Background

Our proposed method is a modification of the "poststratification" and "raking estimator" widely used in survey estimation, including the FIA program. We will briefly introduce the raking estimator, but will then describe a different survey weighting approach - Generalized Regression (GREG) - that has been more fertile ground for recent innovation than has raking. Some of these innovations to GREG - particularly ridge and LASSO regression - will motivate our modified raking estimator. After presenting our survey weighting estimator, we will then discuss related work.

2.1. Survey Estimation

Consider a population and study area U of size N, and a sample survey s of size n. Let the study area be divided into small, non-overlapping pixels or patches $\{T_t\} \subset U$. Let $\{J_j\}$ be another set areas, large or small, and possibly overlapping, for which estimates are desired, such as counties, the entire study area, management areas, or the individual pixels. Let $\{y_{ik}\}$ denote the measurement of characteristic k for individual (or plot) i. We are interested in obtaining estimates of population totals $\tau_k = \sum_{i \in U} y_{ik}$ and small area totals $\tau_{jk} = \sum_{i \in J_j} y_{ik}$.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 October 2019 doi:10.20944/preprints201910.0093.v1

Peer-reviewed version available at Forests 2019, 10, 1045; doi:10.3390/f10111045

Most survey estimates of population totals can be reduced to linear estimators: $\hat{\tau}_k = \sum_{i\in s} w_i y_{ik}$ where w_i are survey weights or expansion factors. Expansion factors earn their name from the fact that they can be understood as the number of elements in the population that are represented by a sample unit. For example, expansion factors in the FIA data represent the number of acres each plot represents in the entire sample population. For small area estimation, we will require expansion factors w_{it} that allocate sample i to target patch t.

If a survey is conducted with known sampling probability, then a simple, design-unbiased estimator for a population total is the "Horvitz-Thompson" (HT) estimator [19,20]

$$\hat{\tau}_j^{HT} = \sum_{i\in s} d_i y_{ij} = \mathbf{d}'\mathbf{y}_j$$

where the weights are called *design* weights because they are related to the sample design through the inverse sampling probability, i.e. $d_i = 1/\pi_i$. In a simple random sample with equal probability, the expansion factors are simply $\frac{N}{n}$.

At the next level of complexity, survey weights are often modified to account for known ancillary data. For example, the FIA program adjusts the HT weights to account for the land area that is forest, non-forest, and forest-fringe, as derived from the National Land Cover Database [21–23].

2.1.1. The raking estimator

Two widely used methods of accounting for ancillary data are *raking* (also called "Iterative Proportional Fitting", and "poststratification" when there is only one ancillary variable) and *Generalized Regression* (GREG) [24]. Both methods modify (or calibrate) the HT expansion factors so that the survey estimates reproduce known population totals τ_ℓ. (We use 'ℓ' and '$x_{i\ell}$' to indicate plot-level characteristics that we know totals for, and 'k' and 'y_{ik}' to indicate other plot-level characteristics that we would like to estimate or map.) In particular, both methods create weights such that $\sum_{i\in s} w_i x_{i\ell} = \tau_\ell$ for each ancillary total. Thus, both raking and GREG assume that ancillary totals are precisely known. However, this is rarely the case, since most ancillary estimates are statistical predictions or imputations developed from satellite reflectance values and other imperfect spatial data sets. Since most ancillary data have uncertainty, they do not strictly satisfy the condition that ancillary data be known.

Raking is an iterative process that uses the HT estimates to initialize the expansion factors for each survey plot. Then an estimated ancillary total $\sum_i w_i x_{i\ell} = \hat{\tau}_\ell$ is calculated. Then, the expansion factors are all adjusted by the multiplicative factor $\frac{\tau_\ell}{\hat{\tau}_\ell}$. This guarantees that the known ancillary total τ_ℓ is exactly reproduced by expansion factors. This step is then repeated for each of the ancillary totals and the entire process is repeated from the beginning until the expansion factors converge or until the algorithm reaches a maximum number of iterations. The raking method is used by FIA, however it converges in one step since there is only one ancillary variable: a spatial layer of of forest, non-forest and forest fringe area. This layer is derived from NLCD land cover, NLCD Tree Canopy Cover, and land ownership.

A naive method to generate small estimates would be to simply incorporate many ancillary totals for small ares $\{T_t\}$. In practice, however, this produces erratic expansion factors and may cause the raking algorithm to fail to converge. Our regularized raking method is a natural extension of the raking estimate that does not succumb to these problems as easily. While our estimator is an extension of the raking estimator, it is motivated by recent innovations to the GREG estimator, and we thus discuss GREG next.

2.1.2. The Generalized Regression Estimator

The GREG estimator is motivated as a linear regression problem using the unknown total τ_k as the response variable, and the variables $\tau_\ell - \tau_\ell^{HT}$ as the explanatory variables. Intuitively, if the

HT estimator over- or under-predicts some of the ancillary totals, then we might want to adjust our weights accordingly. The GREG estimator is given by the equation:

$$\hat{\tau}_j^{GREG} = \hat{\tau}_j^{HT} + \sum_{\ell} (\tau_\ell - \hat{\tau}_\ell^{HT})\beta_\ell$$

where the β_ℓ are the weighted least squares regression coefficients obtained by regressing the response variable y_j on the ancillary variables $(\tau_\ell - \hat{\tau}_\ell^{HT})$, using the design weights d_i on each observation.

It may appear that the GREG estimator depends on the response variable y_j, and thus that a different estimator would be needed for each response variable, but it is possible to rewrite the GREG estimator in terms of an expansion factor as:

$$\hat{\tau}_j^{GREG} = (\mathbf{w}^{GREG})'\mathbf{y}_j = (\mathbf{d} + \boldsymbol{\lambda}'(\boldsymbol{\tau}_x - \hat{\boldsymbol{\tau}}_x^{HT}))'\mathbf{y}_j \tag{1}$$

where $\boldsymbol{\lambda}$ are coefficients that do not depend on the response y. Thus, the expansion factors $\mathbf{w}^{GREG}$ can be calculated once and stored for general purpose use.

The GREG estimator is not widely used in survey practice. In particular, the GREG estimator has a tendency to produce some negative expansion factors. This is especially common when many ancillary data are used. In these situations, the raking estimator can produce erratic expansion factors, whereas the GREG estimator tends to produce expansion factors that are slightly less erratic, but sometimes negative. Negative survey weights are clearly problematic, as they defy the interpretation of the weights as expansion factors. For this reason, GREG is rarely used in situations in which agencies must publish the expansion factors.

2.1.3. Regularized GREG

A recent expansion of GREG that motivates our estimator is that of regularized regression, including ridge and LASSO regression. As mentioned earlier, GREG and raking can often produce erratic results when there are many ancillary totals. Ridge regression modifies the GREG estimator by adding a term $\sum_\ell \beta_\ell^2$ to the least squares regression problem [25]. This has an effect of not exactly reproducing ancillary totals if doing so would cause the coefficient to get large. LASSO regression modifies the GREG estimator by the addition of the term $\sum_\ell |\beta_\ell|$ to the least squares regression [26]. Ridge and Lasso are both types of “regularization” estimators. Regularization has proven to be a powerful heuristic tool for prediction problems involving many explanatory variables because if effectively addresses common practical problems caused by the multicollinearity and overfitting problems that emerge in such situations. Regularization estimators typically trade off a little bias in a prediction in exchange for greatly reduced mean square error.

Not requiring the survey to exactly reproduce the ancillary data is an attractive property, both conceptually and statistically. Conceptually, we ought to recognize that ancillary data do not usually represent a “gold standard” of truth. For example, the FIA weighting process uses estimates of forest, not-forest and forest edge that are derived from satellite imagery. Satellites do not observe forests; they observe spectral reflectances which we then feed through models that create predictions of tree canopy and land cover. Thus, while we may have confidence that the model is - on the whole - reliable, the output by no means represents the absolute truth. By forcing our survey estimates to match these modeled outputs, we are forcing our survey estimates to reproduce the errors found in these imperfect spatial data layers.

We believe that many of the practical limitations of calibration estimators are amplified by calibrating on noisy and imprecise auxiliary variables. Intuitively, if the auxiliary data contain an element of truth (or signal) and an element of error (or noise), then the calibration will incorrectly reproduce the noise along with the signal. This places limits on the number of auxiliary variables that can be used - if too many noisy auxiliary variables are used then calibration will overfit these data

Peer-reviewed version available at *Forests* **2019**, *10*, 1045; doi:10.3390/f10111045

and the final weights will be erratic. In regression modeling, the effects of overfitting are similar to the effects of multicollinearity.

To extend these ideas to raking, we first point to the fact that GREG solves the optimization problem [27]:

$$\min_{w_i} \sum_{i \in s} \frac{1}{2} \frac{(w_i - d_i)^2}{d_i} \text{subject to} \sum_{i \in s} w_i x_{i\ell} = \tau_\ell$$

Thus, GREG can be interpreted as "find new expansion factors w_i, that are close to the HT factors but that also satisfy the ancillary constraints."

Guggemos et al. show [25] show that the ridge GREG estimator can be similarly written as:

$$\min_{w_i} \sum_{i \in s} \frac{1}{2} \frac{(w_i - d_i)^2}{d_i} + \frac{1}{2\gamma} \sum_{\ell} c_\ell (\tau_\ell - \sum_i w_i x_{i\ell})^2$$

In contrast to GREG, which exactly reproduces the ancillary totals, the ridge GREG estimator approximately reproduces the ancillary totals. The approximation can be made exact by increasing c_ℓ or decreasing γ. The factors c_ℓ provide a means to control the relative importance of reproducing certain ancillary totals.

2.1.4. Regularized Raking

To develop our regularized raking estimator, we first point to results showing that the raking and GREG estimators are similar; Deville and Sarndal [28] proposed a class of estimators that calibrate the Horvitz-Thompson design weights to ancillary data. These estimators are:

$$\min_{w_i} \sum_{i \in s} d_i G(w_i, d_i) \text{ subject to } \sum_{i \in s} w_i x_{i\ell} = \tau_\ell$$

where $G(w_i, d_i)$ is a generalized distance measure. They show that when the distance function is the Chi-Square function $(\frac{w_i - d_i}{d_i})^2$ then the GREG estimator results. Similarly, when the distance function is the entropy distance function $w_i \log(\frac{w_i}{d_i})$, then the raking estimator results. In both cases, minimizing the distance between the HT weight and the calibrated weights preserves an approximate design-unbiased property [25].

By analogy to the regularized GREG estimator, we propose a regularized raking (or regularized entropy) estimator. Let $\tau_{j\ell}$ be an ancillary total for spatial region J_j and characteristic k. The estimator we use is

$$\min_{w_{i\ell}} - \sum_{i \in s} \sum_t w_{it} \log(\frac{w_{it}}{d_{it}}) - \frac{1}{2\gamma} \sum_{\ell} \sum_j \frac{(\tau_{j\ell} - \sum_{i \in s} \sum_{t \in J_j} w_{it} x_{i\ell})^2}{\sigma_{j\ell}}$$

The output of this estimator is the set of strictly positive expansion factors w_{it}. In our case, the expansion factors have units of acres of patch t represented by plot i. If the weights are normalized by the small area size: $\frac{w_{it}}{\sum_t w_{it}}$ then they create a sort of empirical histogram or probability density across the samples for each small patch.

Nagle et al. [17] suggest setting $\sigma_{j\ell}$ to the standard deviation of the ancillary total $\tau_{t\ell}$ as this serves to distribute the accuracy in proportion to the quality of the ancillary data. The estimator will try to calibrate the weights to the ancillary exactly, but if there are conflicts between the ancillary data, then the weights will lean toward more closely reproducing the more precise ancillary data.

Nagle et al. suggest setting the prior weights d_{it} to $d_i \left(\frac{\text{area of } T_t}{\text{area of } U}\right)$. This has the effect of assuming that — in the absence of any ancillary information — the best prior estimate is that each patch is identical. Other prior weights may be possible, for example by weighting nearby samples more, but this is not explored in this article.

The regularization factor γ was not considered by Nagle et al. We include it here to allow a data-driven selection to the regularization. If $\gamma = 0$, then the final weights will be to use the HT

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 October 2019 doi:10.20944/preprints201910.0093.v1

Peer-reviewed version available at Forests 2019, 10, 1045; doi:10.3390/f10111045

weights at each patch. If γ is large, then the final expansion factors will try to reproduce the ancillary totals even if it requires large deviations from the HT weights. Later, we describe a cross-validation procedure to choose this regularization factor.

2.2. Related Work

Regularization is an approach that is experiencing growing exploration in survey research. In addition to ridge regularization, which uses a square error penalty as here, recent research has explored LASSO regularization, which uses an absolute deviation penalty [26,29,29]. McConville et al. report that LASSO estimators, while producing fewer negative expansion factors than GREG, still produces negative factors sometime. This problem renders them inadequate for creating general-purpose expansion factors.

More generally, it may be possible to use raking with a LASSO regularization rather than a ridge regularization. LASSO is often justified over ridge regression on the grounds that it is "sparse" and automatically selects some ancillary variables to use and others not to use. Sparsity is a useful feature when the regression coefficients β must be stored for later use. We find the justification less compelling in the current use case because the model only needs to be fit once, and it is not a set of regression coefficients that must be stored, but expansion factors w. LASSO has no effect on the sparsity of expansion factors w, and here is no guarantee that a sparse set of regression coefficients leads to more efficient or robust expansion factors. A sparse set of expansion factors may be desirable, but neither ridge nor LASSO methods currently provide that.

Another related approach is the previously mentioned one by Riley et al. [18]. Using ancillary data on vegetation and biophysical attributes, they build a random forest model to identify the single FIA plot that best predicts the ancillary data at each 30m pixel across the Western US. This is in contrast to our expansion factor approach that identifies the probability that each site is represented by each plot. Additionally, Riley et al. fit their model at a large scale, and their model often identifies a best plot that is thousands of km away from the pixel. In our implementation, we keep the plots limited to the most immediate FIA survey unit, however it is theoretically possible to expand our approach to larger region as do Riley et al.

A similar product is the National Tree-List Layer [30]. That product creates a geographic layer of trees for each site, by stratifying places using vegetation and biophysical characteristics, and then a nearest neighbor search from the site to a database of sample plots that includes the FIA and other samples.

3. Data

3.1. Forest Inventory and Analysis

The sampling network of the FIA is non-overlapping mesh of 2400 ha ($\sim$ 5930 acres) hexagons [31]. A permanent sample plot is randomly located within each cell, which is visited on a 5- to 7-year cycle in the Eastern US. Our study region is the state of South Carolina, which is divided into three survey units (refer back to Figure 1). Following FIA practice, we will estimate expansion factors separately for each survey unit.

When a plot is sampled, the site is first viewed in the office using aerial imagery. If the office staff determine that the site is unlikely to be in use as a forest, then its land use is recorded and the sampling process ends for that site. If the plot is regarded to possibly contain use as a forest, then the plot is visited by a field crew. Field crews divide each plot into separate "condition classes" so that each condition represents a homogeneous land use type and ownership class. In this study, we are interested in three attributes: land use, forest basal area, and forest volume. The Land Use classification is not nationally standardized and we use the classification used by the Southern Research Station (coded as LAND_USE_SRS in Burrill et al. [23]). We have collapsed the the land use classification to the categories: Forest, Other Agriculture, Urban, and Other Uses.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 October 2019 doi:10.20944/preprints201910.0093.v1

Peer-reviewed version available at Forests 2019, 10, 1045; doi:10.3390/f10111045

FIA field crews also record the dominant species type of forest in each condition class (FORTYP). The Forest Type is a three digit classification which we collapsed into a two digit classification corresponding to "Pine and other Softwood", Oak/Hickory, Oak/Pine, Oak/Gum/Cypress, Elm/Cottonwood, and Other. Owing to small sample sizes, we have further grouped together the Oak/Gum/Cypress and Elm/Cottonwood into one group, and we have included the "Other" category with the Oak/Hickory. This produces a classification of 4 different Species groups.

We have then combined the Land Use and Forest Type variables into one nested variable, so that Forest Use is divided by Forest Type, and and Non Forest Use is divided by Use class. The definitions of these 9 classes classes are shown in Table 1.

We also estimate the Forest Volume for each plot, which is defined as the (estimated) net volume of wood in timber species greater than 5 inches in diameter at breast height.

In addition to land use, forest type, and forest volume, we also used the estimate of Basal Area of Live trees (BALIVE) for each plot for prediction. We do not use basal area as an ancillary variable, but as a response variable for creating small area estimates. BALIVE represents the density of trees greater than 5 in each condition (measured in units of square feet per acre). For creating survey estimates, we set BALIVE to 0 in sites that are non forested or that do not contain any trees larger than 5 in.

Group	Class	FIA Database Definition
Forest	Eastern Softwood	LAND_USE_SRS in (1,2) AND FORTYPCD in (100-199)
	Oak/Pine	LAND_USE_SRS in (1,2) AND FORTYPCD in (400-499)
	Oak/Hickory	LAND_USE_SRS in (1,2) AND FORTYPCD in (500-599; 800-998)
	Bottomland Hardwood	LAND_USE_SRS in (1,2) AND FORTYPCD in (600-699)
Not Forest	Not Stocked	LAND_USE_SRS in (1,2) AND FORTYPCD in 999
	Agriculture	LAND_USE_SRS in (10-19)
	Urban/Developed	LAND_USE_SRS in (30-39)
	Barren	LAND_USE_SRS in (40-49)
	Water/Wetland	LAND_USE_SRS in (90-99)

Table 1. Custom Land Use/Forest Type classification. Variable names refer to those in the FIA Database [23]

In addition to the FIA variables, we use land cover and tree canopy cover maps from the 2011 National Land Cover Database (NLCD2011) to generate our ancillary totals [32].

4. Methods

To implement the weighting strategy, we require auxiliary estimates of quantities that can be estimated for the FIA plots. For example, it is possible to obtain the NLCD land cover class for each plot in the office. However, to demonstrate the flexibility of the model, we instead use the NLCD land cover and tree canopy cover layers in Bayesian generalized additive models to develop wall-to-wall maps of: (1) the 9-class land use and forest type classification in Table 1, (2) a 2-class forest/not forest use and (3) volume. For the purpose of this article, the relevant factor is that these models generate predictions and standard errors for forest characteristics measured at FIA plots. While we will create small area estimates for basal area in addition to those above, we do not include a direct measurement of basal area as an ancillary layer.

Theoretically, we could estimate expansion factors for individual pixels, however, this is computationally infeasible and unnecessary. Expansion factors for pixels that observationally equivalent (i.e. have the same predictions and standard errors), will also be identical. Thus, it not necessary to fit the model at the pixel scale, but at a scale that has the pixels grouped into patches that are homogeneous with respect to the ancillary predictions and standard errors. Since our ancillary maps are derived from NLCD land cover and tree canopy cover layers, and we also desire county-level estimates, we group pixels into patches bases on county, NLCD class, and tree canopy cover (grouped into 20 bins as 0-4 percent, 5-9 percent,... 95-100 percent). For each of the three survey units, there are 180 such patches (9 land cover classes times 20 tree canopy cover classes). When these homogeneous

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 October 2019 doi:10.20944/preprints201910.0093.v1

Peer-reviewed version available at Forests 2019, 10, 1045; doi:10.3390/f10111045

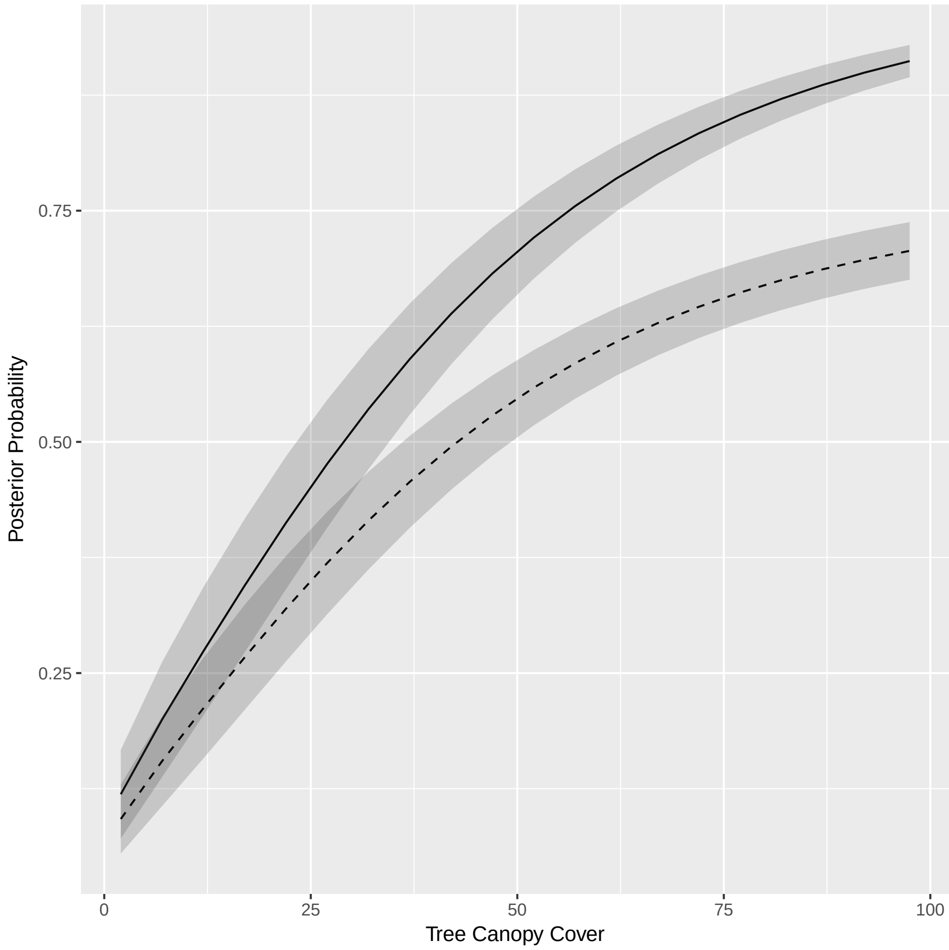


Figure 2. Posterior probabilities and standard error for Forest (solid) and Eastern Softwood (dashed).Only NLCD class 42 in the Piedmont survey unit is shown. While auxiliary data are predicted or imputed from a model such as used here, it is obvious that there is some uncertainty in the prediction.

patches are overlayed on the 45 counties of South Carolina, there are 8260 such patches across the state of South Carolina. These are the target patches T_t defined in the background section.

The mapped outputs from these models are then used as auxiliary variables in the regularized raking estimator. In all cases, the ancillary data model produces both estimate as well as standard errors of prediction, which are used to weight the ancillary variables.

Figure 2 shows an example model fit for the auxiliary variables: the probability of forest use and the probability of the Eastern Softwood Forest Type, as well as their standard errors, for one of the survey units. It is obvious that standard error of the "Forest" land use class is smaller than the standard error for the more specific "Eastern Softwood" class. This is especially so at the higher canopy cover levels that are most likely to be forested.

The complete set of auxiliary variables used across the state include:

1. Patch level - estimates of the 9-class land use/forest type classification (180 unique patches * 9 classes * 3 survey units = 4,860 ancillary totals)
2. Patch level estimates of the 2-class forest/not forest land use (180 * 2 * 3) = 1,080 ancillary totals)
3. Patch level estimates of forest volume (180 * 3 = 540 ancillary totals)
4. county size (45 ancillary totals). The variance of county size is set to 0 so that the expansion factors reproduce them exactly.

It may appear that there are some redundancies in the set of auxiliary variables. For example, the land areas of the 9-class map perfectly recreate the land areas of the 2-class map, the number of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 October 2019 doi:10.20944/preprints201910.0093.v1

Peer-reviewed version available at Forests 2019, 10, 1045; doi:10.3390/f10111045

acres in the two classes must add up to the number of acres in the county. When the uncertainties of the auxiliary data are considered during regularization, however, they are no longer redundant. For example, while there is uncertainty in the number of acres of forest and number of acres of non-forest land, there is zero uncertainty in their sum across the county. And where there is relatively large uncertainty in each of the individual 9-class land-use/forest-type proportions at each patch, there is less uncertainty in their aggregation to the 2-class forest/non forest proportions (e.g. see Figure 2). The regularization incorporates ancillary data variance, but not covariance. It is possible that the 9-class map would be sufficient if all cross-covariances are accounted for, but this would dramatically complicate the computational routines. Additionally, while it is somewhat reasonable to obtain estimates of prediction error variance from ancillary geospatial data products, it is much less reasonable to expect covariances between different ancillary data products.

Using the ancillary maps created by our predictive model, we fit both regularized raking model and a ridge GREG model to each of the three FIA survey units. The regularized raking model was fit using a custom R package available from the first author. The ridge GREG model was fit using the 'Matrix' package in R using the linear algebra equations for solving ridge regression.

For both the raking and GREG estimators, the regularization parameter γ was determined using cross validation. To conduct the cross-validation experiment, first, we simulated 30 instances of the auxiliary spatial data. Given the statistical uncertainty of the auxiliary data, each of the simulations is in some sense an equally valid sample of auxiliary data. Cross validation was conducted by withholding 1/16 of plots from the estimation. These withheld plots are a subset of the FIA plots located on a coarser hexagonal grid. Expansion factors were created using each of the simulated auxiliary datasets. The expansion factors were then used to estimate forest volume for each of the patches containing a withheld plot and we calculated the square error between the withheld plot measurement and the estimate, and then averaged across simulations and withheld plots.

Before continuing, we remark that the ancillary data were fit using the survey estimates. The ancillary data are thus endogenous to the sample. Breidt and Opsomer [33] consider endogenous post-stratification and suggest that this has little effect on the the final survey estimates. How that result translates to the small area situation we consider here, however, is unknown.

5. Results

5.1. Comparison of raking and GREG expansion factors

A benefit of regularized raking is that it produces positive expansion factors, while ameliorating erratic weights that are caused by overfitting to inconsistent and noisy ancillary data. Table 2 shows the minimum and maximum expansion factors for the ridge GREG and regularized raking estimators as a function of the regularization parameter γ. Recall that, as γ approaches zero, the normal GREG and raking estimators are obtained. In contrast, as γ gets large, both estimators converge on the HT expansion factors. The minimum expansion factor for the ridge GREG model is consistently negative for all but the largest values of the regularization factor γ. In contrast, the raking estimator always produces strictly positive expansion factors.

It is harder to interpret the absolute magnitude of the maximum expansion factor because they have units of acres and depend on the size of the patches, but it is easy to compare the maximum weights between the raking and GREG models. The maximum expansion factor is always smaller for the GREG estimator than for the raking estimator, and the largest expansion factor can get quite large for very small values of the regularization parameter, (i.e. when we try to exactly match the auxiliary variables). When the regularization parameter is very small, the regularized raking estimator converges to the raking estimator, and we note that the normal raking estimator is not guaranteed to exist when there are incompatible auxiliary variables. Our simulations revealed some explosive behavior in the regularized raking estimator for very small values of gamma ($\gamma < .01$). Regularization appears to temper these undesirable effects.

gamma	GREG weights min	GREG weights max	raking weights min	raking weights max
0.1	-151.58	2381.00	<1.00e-08	2930.64
0.2	-113.37	2029.45	<1.00e-08	2833.97
0.3	-87.17	1776.99	<1.00e-08	2744.86
0.4	-75.52	1586.54	<1.00e-08	2659.87
0.6	-76.66	1317.79	<1.00e-08	2498.54
0.8	-77.93	1136.75	<1.00e-08	2346.58
1.0	-78.73	1006.18	<1.00e-08	2203.08
2.0	-78.66	672.69	4.00e-07	1604.76
8.0	-72.25	323.54	1.49e-05	465.47
30.0	-49.23	203.38	4.51e-05	219.71
100.0	5.75e-05	161.34	5.80e-05	161.60

Table 2. Minimum and maximum expansion factors for the ridge GREG models and regularized raking. South Carolina, Survey Unit 1

Figure 3 shows the cross-validation error as function of the regularization parameter γ. The cross-validation measure is the average square error of estimating forest volume across the withheld plots. Regularized raking consistently performs better than ridge GREG when predicting forest volume out of sample. This is especially interesting since it is GREG that minimizes (within sample) square error loss, not raking. The improvement seems to be in the fact that GREG can produce errors beyond the range of the data, whereas raking can not. We also note that regularized raking prefers slightly more regularization than does ridge GREG. This is slightly unfortunate: since GREG is less computationally demanding, and cross validation is a computationally intensive process, we had hoped that the optimal regularization parameter for GREG would also be a good regularization parameter for raking. One possible explanation for why raking may prefer larger regularization parameters stems from the more extreme maximum weight of the raking procedure relative to GREG.

It should not be inferred from Figure 3 that regularized raking is always better than ridge GREG. At some point to the left of the range in these plots ($\gamma < .01$), the regularized raking error explodes as the coefficients fly apart to infinities trying to reach a raking solution that may not exist, whereas the ridge GREG line converges to the normal GREG solution. Furthermore, Figure 3 is slightly misleading as the y-axis does not extend to zero. In terms of root mean square error, the difference between regularized raking and regularized GREG is not as extreme as might appear here. The most noticeable difference to users is likely to be the presence or absence of negative expansion factors and not any differences in estimates.

5.2. Prediction Maps

Once expansion factors are calculated, they may be used to estimate wall-to-wall maps of any plot characteristic. Figures 4a and 4b show maps of (a) Basal Area for all trees and (b) Basal Area for Eastern Softwood species. At this scale, it can be seen that the maps reproduce large scale patterns in forest areas, while also capturing the dominance of pines in the Piedmont region, and relative absence of pines from coastal bottomlands.

Figure 5 shows a higher resolution comparison of the ancillary layers and the raking estimates for a 30km x 10 km swath in the South Carolina Piedmont. While the regularized raking estimate does not exactly reproduce the ancillary layers, the difference are visually indistinguishable here. It is worth pointing out that the any errors in the ancillary data will be carried through to the raking estimates. For example, the model that creates Forest Volume from NLCD tree canopy cover and NLCD land use saturates with respect to tree canopy cover; while the model is able to distinguish between low and moderate volume, the auxiliary model is unable to effectively model high volume. The range of predictions in the auxiliary model (0 to 100 sq ft per acre) is restricted relative to values of Volume on the landscape (0 to 1000 sq ft per acre), and this restricted range in the auxiliary data is duplicated by restricted range in the regularized raking estimates.

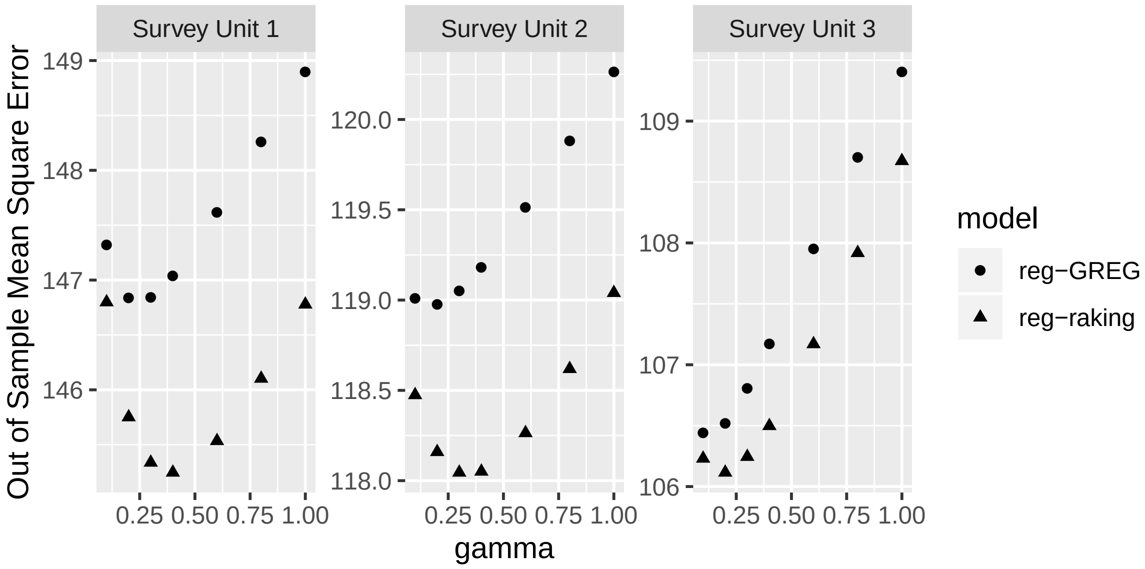
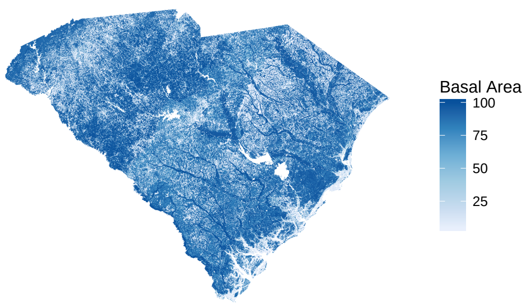
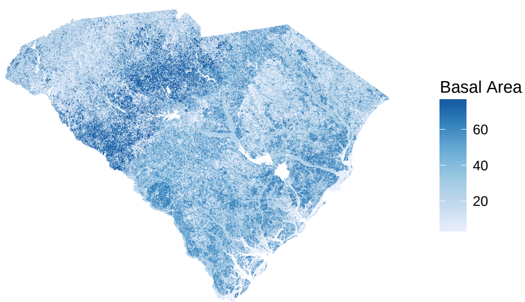


Figure 3. Cross-validation error as a function of the regularization parameter γ for regularized raking (triangle) and ridge GREG (circle), for survey units 1 (left), 2, and 3(right). The response variable is the mean square error for predicting forest volume at out-of-sample locations.



(a) All forest types.



(b) Eastern Softwood Forest Type.

Figure 4. Estimated Basal Area (sq ft per acre).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 October 2019 doi:10.20944/preprints201910.0093.v1

Peer-reviewed version available at *Forests* **2019**, *10*, 1045; doi:10.3390/f10111045

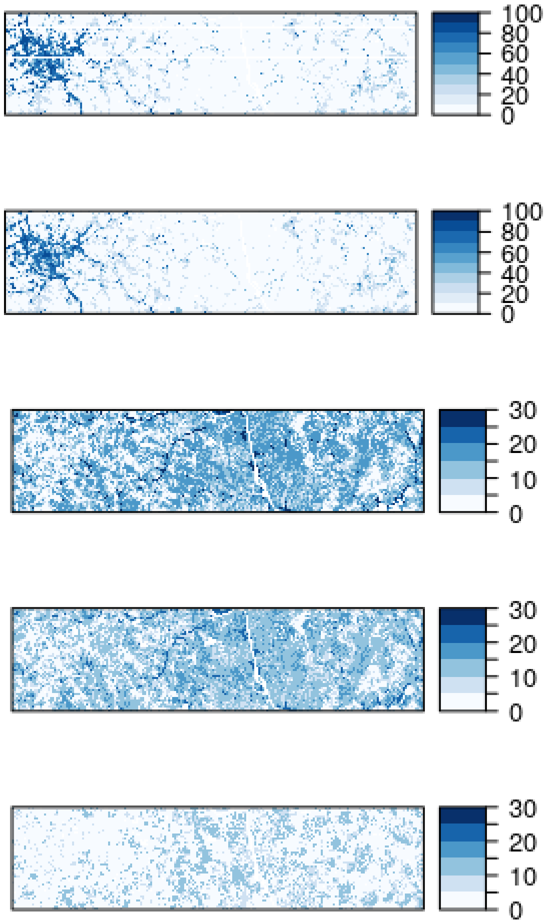


Figure 5. Ancillary and raking estimated maps. Map size is 30km x 10km. From top to bottom: Probability of developed land use - ancillary; probability of developed land use - raking estimate; Forest Volume - ancillary; Forest Volume - raking estimate; Volume of Eastern Softwood - raking estimate.

variable	plot	county
Forest Land Use	0.65	0.83
Oak Pine Forest Type	0.49	0.86
Eastern Softwood Forest Type	0.59	0.91
Volume	0.27	0.79
Basal Area	0.68	0.88

Table 3. Pearson correlation coefficient between the direct survey estimate and the regularized raking estimate. At the plot level, the direct survey estimate is plot measurement (when multiple conditions are present, the direct estimate is the weighted average across conditions). At the county level, the direct survey estimate is the Horvitz-Thompson estimate from the plots in that county.

While we are able to produce high-resolution wall-to-wall maps, perhaps the best use of the expansion factors may be to produce estimates for slightly larger areas. Table 3 shows the correlation between the direct survey estimates and the regularized raking estimates. At the plot-level, the direct estimate is just the observed value. At the county level, the direct estimate is the weighted value of all plots within with the county.

As expected, the correlation between the direct estimate and the regularized raking estimate increases as the spatial unit gets larger. Whereas the correlation at the plot level is relatively weak, the correlation increases when estimating larger units such as counties. We would not expect the county level correlation to necessarily equal 1, because the direct estimates have high sampling variance at the county level, but the results nonetheless suggest that estimates for larger units are more reliable than for the smallest areas.

The distinguishing characteristic of this method is that all estimates use the same expansion factors w_{it}. Typically, small area estimators will fit separate models for each variable of interest, for example the model for Basal Area might be different from the model for Basal Area of Pine or from the model of Volume. An advantage of using the same weights is that all maps are consistent, it is impossible to obtain more Basal Area of softwood species than Basal Area of all species, and it is impossible to obtain estimates have positive Basal Area but no volume, and vice versa. Additionally, all estimates are constrained to range of the data. Because none of the weights are negative, all estimates like strictly within the range of the survey data. In particular, negative estimates are impossible, unlike other methods, such as kriging.

6. Conclusion

In this paper we have introduced a new method for producing survey expansion factors that are suitable for small area use. This method extends poststratification and raking methods already in use by the FIA and other survey agencies. Raking estimators calibrate the survey probability weights to reproduce ancillary data. The regularization method modifies raking to allow for the approximate reproduction of ancillary data, with a relative weighting between different ancillary data based on their relative precision. Whereas the usual raking estimator often gets overwhelmed by relatively few ancillary variables, producing erratic and very large expansion factors, the regularized raking estimator in this paper utilized tens of thousands of small area ancillary totals.

Like many other methods in the literature, the regularized raking method allows the integration of FIA plot data with wall-to-wall ancillary data. Unlike those other methods, however, the regularized raking methods is relatively agnostic about the response variable. It is mathematically possible to use the expansion factors to produce wall-to-wall maps for any characteristics available for FIA plots. These wall-to-wall maps and small area estimates can also be made to be consistent with published totals (it is even possible to create the published totals from the same expansion factors).

However, the regularized raking method is by no means optimal for any characteristic. A regularized raking estimator for any characteristic is certain to be outperformed by a highly tailored predictive model. For example, even if an "optimal" data layer is used as an ancillary variable, the regularized raking estimator would not exactly reproduce this estimate unless the prediction standard error were artificially forced to zero, which could lead to erratic expansion factors from the raking estimator.

The regularized raking estimator can have an intriguing scale-dependence, in that it can transition from a direct survey estimate at the scale of the survey unit, similar to the current practice of producing expansion factors, to an "indirect" or model-based estimate at small areas, borrowing samples and statistical relations from across the entire study area to make small area estimates. This also suggests that the estimates are relatively design-unbiased at the scale of the survey unit, but are less so for small area estimation. If the path-level weights are exactly constrained to add to the survey unit-level Horvitz-Thompson weights, then the same set of expansion factors can produce exactly design-unbiased estimates at the survey unit level, and model-based estimates at the small area level.

Finally, we'd like to comment on the feasibility of disseminating these expansion factors. Publishing the expansion factors for one survey unit in South Carolina would require a table containing one row for each of the approximately 3000 plots, and one column for each of the approximately 2700 patches in the survey unit. If the expansion factors are stored at integer precision, such a table would require about 50 MB of storage, which is easily within the range of feasibility for a geospatial dataset. Storage sizes would be much smaller in a suitable binary format (such as in a SQL database table).

Further research is needed to determine a suitable set of ancillary data. These ancillary data should be correlated with as many plot-level characteristics as possible. Efforts are currently underway to explore the use of ancillary data developed from Landsat time series fitting algorithms. Potentially, combining these smoothed outputs with regularized raking will allow the construction of temporally consistent time series expansion factors, which can lead to improved estimates of forest characteristics across space and time.

Author Contributions: conceptualization, NNN and TAS; methodology, NNN; software, NNN; validation, NNN; formal analysis, NNN and TAS; data curation, BR; writing–original draft preparation, NNN; writing–review and editing, TAS and BR; visualization, NNN and BR; supervision, NNN and TAS; project administration, TAS; funding acquisition, TAS

Funding: Funding for this research was provided by the U.S. Forest Service, Southern Research Station, Forest Inventory and Analysis (FIA) program (via agreement 17-CR-11330145-057).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Breiman, L. Random forests. *Machine Learning* **2001**, *45*, 5–32.
2. Cutler, D.R.; Edwards, T.C.; Beard, K.H.; Cutler, A.; Hess, K.T.; Gibson, J.; Lawler, J.J. Random Forests for Classification in Ecology. *Ecology* **2007**, *88*, 2783–2792.
3. 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. Mapping U.S. forest biomass using nationwide forest inventory data and moderate resolution information. *Remote Sensing of Environment* **2008**, *112*, 1658–1677.
4. Powell, S.L.; Cohen, W.B.; Healey, S.P.; Kennedy, R.E.; Moisen, G.G.; Pierce, K.B.; Ohmann, J.L. Quantification of live aboveground forest biomass dynamics with Landsat time-series and field inventory data: A comparison of empirical modeling approaches. *Remote Sensing of Environment* **2010**, *114*, 1053–1068.
5. 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. Conterminous U.S. and Alaska Forest Type Mapping Using Forest Inventory and Analysis Data. *Photogrammetric Engineering and Remote Sensing* **2008**, *74*, 1379–1388.
6. Moisen, G.G.; Meyer, M.C.; Schroeder, T.A.; Liao, X.; Schleeweis, K.G.; Freeman, E.A.; Toney, C. Shape selection in Landsat time series: a tool for monitoring forest dynamics. *Global Change Biology* **2016**, *22*, 3518–3528.

7. Schroeder, T.A.; Schleeweis, K.G.; Moisen, G.G.; Toney, C.; Cohen, W.B.; Freeman, E.A.; Yang, Z.; Huang, C. Testing a Landsat-based approach for mapping disturbance causality in U.S. forests. *Remote Sensing of Environment* **2017**, *195*, 230–243.
8. Finley, A.O.; McRoberts, R.E. Efficient k-nearest neighbor searches for multi-source forest attribute mapping. *Remote Sensing of Environment* **2008**, *112*, 2203–2211.
9. McRoberts, R.E. Estimating forest attribute parameters for small areas using nearest neighbors techniques. *Forest Ecology and Management* **2012**, *272*, 3–12.
10. Ohmann, J.L.; Gregory, M.J. Predictive mapping of forest composition and structure with direct gradient analysis and nearest-neighbor imputation in coastal Oregon, U.S.A. *Canadian Journal of Forest Research* **2002**, *32*, 725–741.
11. Oliver, M.A.; Webster, R. Kriging: A method of interpolation for geographical information systems. *International journal of geographical information systems* **1990**, *4*, 313–332.
12. Freeman, E.A.; Moisen, G.G. Evaluating kriging as a tool to improve moderate resolution maps of forest biomass. *Environmental Monitoring and Assessment* **2007**, *128*, 395–410.
13. Temesgen, H.; Hoef, J.M.V. Evaluation of the spatial linear model, random forest and gradient nearest-neighbour methods for imputing potential productivity and biomass of the Pacific Northwest forests. *Forestry* **2015**, *88*, 131–142.
14. Kennedy, R.E.; Ohmann, J.; Gregory, M.; Roberts, H.; Yang, Z.; Bell, D.M.; Kane, V.; Hughes, M.J.; Cohen, W.B.; Powell, S. An empirical, integrated forest biomass monitoring system. *Environmental Research Letters* **2018**, *13*, 025004.
15. Franco-Lopez, H.; Ek, A.R.; Bauer, M.E. Estimation and mapping of forest stand density, volume, and cover type using the k-nearest neighbors method. *Remote Sensing of Environment* **2001**, *77*, 251–274.
16. Wilson, B.T.; Lister, A.J.; Riemann, R.I. A nearest-neighbor imputation approach to mapping tree species over large areas using forest inventory plots and moderate resolution raster data. *Forest Ecology and Management* **2012**, *271*, 182–198.
17. Nagle, N.N.; Buttenfield, B.P.; Leyk, S.; Speilman, S. Dasymetric Modeling and Uncertainty. *dx.doi.org.proxy.lib.utk.edu* **2013**, *104*, 80–95.
18. Riley, K.L.; Grenfell, I.C.; Finney, M.A. Mapping forest vegetation for the western United States using modified random forests imputation of FIA forest plots. *Ecosphere* **2016**, *7*, e01472.
19. Cochran, W.G. *Sampling Techniques, 3rd Ed.*; John Wiley & Sons, 2007.
20. Thompson, S.K. *Sampling*; John Wiley & Sons, 2012.
21. McRoberts, R.E.; Wendt, D.G.; Nelson, M.D.; Hansen, M.H. Using a land cover classification based on satellite imagery to improve the precision of forest inventory area estimates. *Remote Sensing of Environment* **2002**, *81*, 36–44.
22. McRoberts, R.E.; Holden, G.R.; Nelson, M.D.; Liknes, G.C.; Gormanson, D.D. Using satellite imagery as ancillary data for increasing the precision of estimates for the Forest Inventory and Analysis program of the USDA Forest Service. *Canadian Journal of Forest Research* **2005**, *35*, 2968–2980.
23. Burrill, E.A.; Wilson, A.M.; Turner, J.A.; Pugh, S.A.; Menlove, J.; Christiansen, G.; Conkling, B.L.; David, W. *The Forest Inventory and Analysis Database*, 2017.
24. Lumley, T. *Complex Surveys*; A Guide to Analysis Using R, John Wiley & Sons, 2011.
25. Guggemos, F.; Tille, Y. Penalized calibration in survey sampling: Design-based estimation assisted by mixed models. *Journal of Statistical Planning and Inference* **2010**, *140*, 3199–3212.
26. McConville, K.S.; Breidt, F.J.; Lee, T.C.M.; Moisen, G.G. Model-Assisted Survey Regression Estimation with the Lasso. *Journal of Survey Statistics and Methodology* **2017**, *5*, 131–158.
27. Deville, J.C.; Sarndal, C.E.; Sautory, O. Generalized Raking Procedures in Survey Sampling. *Journal of the American Statistical Association* **1993**, *88*, 1–9.
28. Deville, J.C.; Särndal, C.E. Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association* **1992**, *87*, 376–382.
29. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Data Mining, Inference, and Prediction, Springer Science & Business Media, 2013.
30. Drury, S.A.; Herynk, J.M. The national tree-list layer. *USDA Forest Service - General Technical Report RMRS-GTR* **2011**, *254*, 1–26.

31. Bechtold, W.A.; Patterson, P.L. The enhanced forest inventory and analysis program – national sampling design and estimation procedures. General Technical Report SRS-80. Technical report, 2005.
32. Homer, C.; Dewitz, J.; Yang, L.; Jin, S.; Danielson, P.; Xian, G.; Coulston, J.; Herold, N.; Wickham, J.; Megown, K. Completion of the 2011 National Land Cover Database for the Conterminous United States Representing a Decade of Land Cover Change Information. *Photogrammetric Engineering & Remote Sensing* **2015**, *81*, 345–354.
33. Breidt, F.J.; Opsomer, J.D. Endogenous post-stratification in surveys: Classifying with a sample-fitted model. *Annals of statistics* **2008**, *36*, 403–427.