DATA REPORT
published: 11 July 2022
doi: 10.3389/ffgc.2022.769268

RegRake: A Web-Based Application for Custom Small Area Estimation and Mapping of Forest Survey Data With Regularized Raking

Todd A. Schroeder[1*†], *Nicholas N. Nagle*[2†] *and Joseph M. McCollum*[1]

[1] United States Department of Agriculture (USDA) Forest Service, Southern Research Station, Knoxville, TN, United States,
[2] Department of Geography & Sustainability, University of Tennessee, Knoxville, Knoxville, TN, United States

Keywords: small area estimation (SAE), regularized raking estimator, Forest Inventory and Analysis (FIA), R Shiny application, expansion factor map

OPEN ACCESS

Edited by:
Steve Prisley,
National Council for Air and Stream Improvement, Inc. (NCASI), United States

Reviewed by:
Edson Vidal,
University of São Paulo, Brazil
Philip Radtke,
Virginia Tech, United States

***Correspondence:**
Todd A. Schroeder
todd.schroeder@usda.gov

†*These authors share first authorship*

Specialty section:
This article was submitted to Forest Management, a section of the journal Frontiers in Forests and Global Change

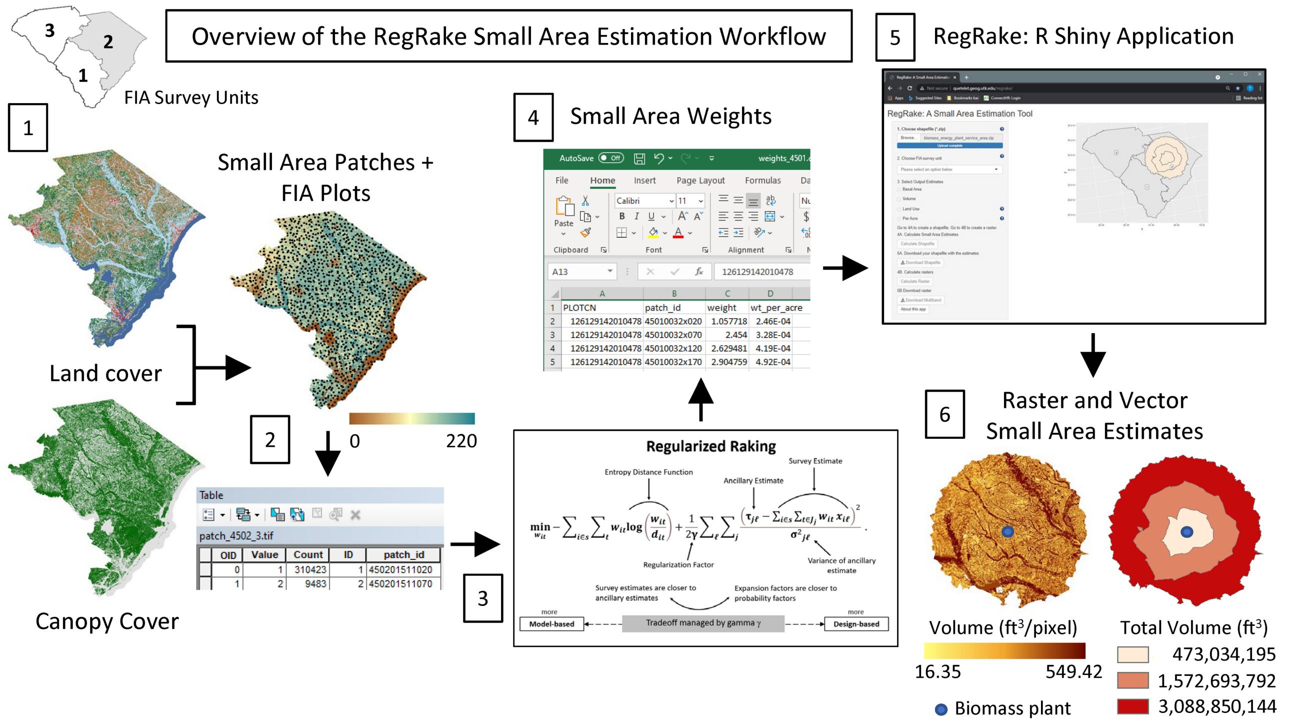
Received: *01 September 2021*
Accepted: *13 June 2022*
Published: *11 July 2022*

Citation:
Schroeder TA, Nagle NN and McCollum JM (2022) RegRake: A Web-Based Application for Custom Small Area Estimation and Mapping of Forest Survey Data With Regularized Raking. Front. For. Glob. Change 5:769268. doi: 10.3389/ffgc.2022.769268

INTRODUCTION

The U.S. Forest Service, Forest Inventory and Analysis (FIA) program manages and implements a design-based network of permanent sample plots that are used to derive a suite of estimates describing the current extent, status and condition of the nation's forest resources (Bechtold and Patterson, 2005). Like many national forest inventory (NFI) systems, the FIA sample is designed for strategic-level estimation of forest characteristics to meet broad scale monitoring and reporting requirements. FIA produces official estimates for entire states or multiple county areas (referred to as survey units), which are large enough to guarantee sufficient sample sizes for direct estimation (USDA, 2008). Users of FIA data often desire estimates for smaller areas (e.g., single counties, watersheds, burn perimeters, etc.) however, in most cases these smaller areas have insufficient sampling densities to support direct estimation alone. In addition, the expansion factors reported by FIA (Bechtold and Patterson, 2005) have been designed to estimate the area of forest within a survey unit, and thus are not appropriate for use on small areas and may not be suitable for use with all forest attributes (e.g., basal area, volume, biomass, etc.). To overcome these limitations, we use a previously published regularized raking algorithm (Nagle et al., 2019) to develop spatial expansion factors which can be combined with FIA plot data to derive statistically valid estimates in areas that are too small to support direct estimation with FIA data alone.

One advantage of regularized raking is that instead of producing a single set of expansion factors representing a plot's contribution to the survey unit (such as those published by FIA), a wall-to-wall map of expansion factors is produced, which describes the probability that any location can be represented by a particular FIA plot. This map of design weights facilitates mapping of survey estimates for small areas and allows post-stratified estimates to be derived for any forest attribute in the FIA database. Furthermore, because the expansion factors are map-based they can be delivered and used on the backend of web-based applications, allowing end users the ability to interactively derive small area estimates for specific areas of interest without having to directly interact with the FIA database. To demonstrate the potential utility of this approach we present RegRake, a new, web-based application that uses pre-developed expansion factor maps to support on-the-fly estimation of FIA forest attributes for user defined small areas in the state of South Carolina, USA. In this article we briefly describe the regularized raking algorithm and ancillary input data used to produce the small area expansion factor maps. We also use RegRake to develop a series of raster and vector-based forest attribute estimates to help evaluate the amount of wood supply surrounding a proposed biomass energy plant.



OID	Value	Count	ID	patch_id
0	1	310423	1	450201511020
1	2	9483	2	450201511070

	A	B	C	D
1	PLOTCN	patch_id	weight	wt_per_acre
2	126129142010478	45010032x020	1.057718	2.46E-04
3	126129142010478	45010032x070	2.454	3.28E-04
4	126129142010478	45010032x120	2.629481	4.19E-04
5	126129142010478	45010032x170	2.904759	4.92E-04

$$\min_{w_{it}} -\sum_{i\in s}\sum_{t} w_{it}\log\left(\frac{w_{it}}{d_{it}}\right) + \frac{1}{2\gamma}\sum_{\ell}\sum_{j}\frac{\left(\tau_{j\ell} - \sum_{i\in s}\sum_{t\in J_j} w_{it}\,x_{i\ell}\right)^2}{\sigma^2{}_{j\ell}}.$$

Total Volume (ft³)	
🟧	473,034,195
🟧	1,572,693,792
🟥	3,088,850,144

FIGURE 1 | Overview of the six main steps in the RegRake small area estimation workflow, which include: 1. Combining maps of land cover and canopy cover to produce small area patches; 2. Assigning FIA plot identifiers to the small area patches; 3. Running the regularized raking algorithm; 4. Developing a table of small area expansion factors; 5. Using the RegRake application to combine small area patches and small area expansion factors for user supplied GIS shapefile; and 6. Outputting small area estimates in raster and vector format.

SURVEY ESTIMATION WITH REGULARIZED RAKING

First, we provide a brief overview of the six main steps (shown in **Figure 1**) involved with developing custom small area estimates with the RegRake R (R Core Team, 2020) Shiny application (Chang et al., 2020). The reader is assumed to have a basic understanding of forest survey statistics (Köhl, 2004), including the use of sample plots and estimators using survey weights or expansion factors (e.g., Horvitz-Thompson, or HT; Cochran, 2007; Thompson, 2012) to derive population totals (e.g., means and variances). As a strategic-level inventory FIA uses spatially balanced, randomly selected sample plots to derive estimates for a variety of forest attributes, however at a density of one plot per 2,400 ha the sampling intensity requires the use of multi-county areas (referred to as survey units) to derive valid statistical estimates for reporting (see **Figure 1** for the three FIA survey units in SC). Expansion factors (or survey weights) published by FIA are designed for large-area estimation of forest area at the scale of the survey unit, therefore, to expand the use of FIA data to smaller areas (or regions that can't support direct estimation with plot data alone) and other attributes, we use a modification of the dasymetric mapping technique (Nagle et al., 2014) known as "regularized raking" (Nagle et al., 2019). Regularized raking develops a new set of expansion factors, w_{it} for each survey unit by matching each FIA plot (i) to a map of homogenous small area patches (t). Unlike traditional expansion factors that produce a single weight representing each plot's contribution to the survey unit (or in the case of FIA the number of acres each plot represents in the entire sample population), our approach results in a map describing the probability that each pixel or patch can be represented by a particular FIA plot. This expansion factor map, which is the main output of the regularized raking approach, can be used to produce both small area estimates and wall-to-wall maps for every attribute in the FIA database.

To develop new expansion factor maps for SC, Nagle et al. (2019) combined land cover and tree canopy cover maps (binned into 11 and 20 classes, respectively) from the 2011 National Land Cover Database (Homer et al., 2015) to form a series of up to 220 homogenous patches for each county in SC (**Figure 1**, step 1). Next, they developed predictive models for basal area and volume within each patch. Finally, they used these predictive values as ancillary information and developed expansion factors for FIA plots sampled between 2007 and 2011 (Cycle 7) using the regularized raking estimator. Both Deville et al. (1993) and Nagle et al. (2019) discuss the use of unregularized raking and calibration to adjust survey weights to match population totals derived from ancillary data. This process, also known as post-stratification, is currently used by FIA to calibrate survey design weights to land- and canopy cover maps, however this is done at the survey unit level. This approach cannot be directly applied to the large volume of ancillary data considered here for several reasons. First, ancillary data derived from predictive models violate the assumption of raking and calibration that ancillary data are perfectly known. Second, when there are large numbers of ancillary data sets, raking and calibration algorithms often produce erratic and unreliable expansion factors, or the raking algorithm can even fail to converge. The regularized raking algorithm of Nagle et al. (2019) provides a feasible solution as it estimates the expansion factors w_{it} by solving the minimization problem (shown in **Figure 1**, step 3):

$$\min_{w_{it}} - \sum_{i \in s} \sum_{t} w_{it} \log\left(\frac{w_{it}}{d_{it}}\right) + \frac{1}{2\gamma} \sum_{\ell} \sum_{j} \frac{\left(\tau_{j\ell} - \sum_{i \in s} \sum_{t \in J_j} w_{it} x_{i\ell}\right)^2}{\sigma^2{}_{j\ell}} \tag{1}$$

where, $x_{i\ell}$ are the survey data for plot i and attribute ℓ, and the ancillary data for region j and attribute ℓ has the estimated value τ_{jl} and variance σ^2_{jl}, and d_{it} are prior weights determined by the sample design probabilities.

Deville and Särndal (1992) showed that the only difference between raking and the generalized regression estimator (or GREG) is the use of the entropy distance function (i.e., the logarithmic term in Equation 1) instead of the chi-square function. Although GREG is a commonly used model-assisted approach that can produce lower squared errors than the more generalized raking approach, it can produce negative design weights which can be problematic. In addition to being asymptotically design unbiased (Guggemos and Tille, 2010), the output of our generalized raking estimator (shown in **Figure 1**, step 4) is a set of strictly positive expansion factors w_{it} in units of acres of patch t represented by plot i. Since the weights are normalized by the small-area's size $\frac{w_{it}}{\sum_t w_{it}}$, they form a vector representing the probability density across all the samples found in each small patch, allowing for broader use with all attributes in the FIA database.

Calibrating on too many ancillary variables can result in erratic expansion weights or non-convergence (in the case of raking) or negative weights (in the case of GREG). Negative weights are problematic for survey estimation, especially for agencies that publish them, thus, to ensure weights stay positive and to avoid overfitting a regularization approach is employed. Inspired by ridge GREG and LASSO regression (McConville et al., 2017), we use a global regularization parameter gamma (γ in Equation 1), which allows the raking estimator to converge when the design weights "approximately" fit the ancillary data. Essentially the regularization parameter strikes a balance between producing expansion factors that closely match the unbiased design weights vs. finding factors that closely match the ancillary totals. While this tradeoff tends to sacrifice some small amount of finite-sample bias in the predictions it significantly reduces the overall mean square error. One challenge is finding a suitable value for γ. In the limit as γ gets large, the resulting weights will match the survey-unit HT weights without regard for the patch-level data. At the other extreme, as γ approaches zero, the resulting weights are a purely model-based estimate without regard to the sample design. Here, a cross-validation procedure was used to find γ. This process (described in full in Nagle et al., 2019) involved running the regularized raking algorithm on a series of simulated auxiliary totals, then using the resulting expansion factors to predict forest volume. Errors from these simulations were then compared across a range of regularization values and the γ with the lowest mean square error was selected. We recognize our use of the survey data to determine γ is considered endogenous, and while these approaches have been shown to be unbiased (Breidt and Opsomer, 2008) more work is needed to determine their legitimacy for use in survey estimation with FIA data. Lastly, to proportionally assign the errors to the quality of the various ancillary data sets we set the denominator σ^2_{jl} in Equation (1) to the variance of the ancillary total (τ_{jl}) as proposed in Nagle et al. (2014).

DELIVERING CUSTOM SMALL AREA ESTIMATES VIA WEB-BASED APPLICATION

Although the regularized raking equation can produce a map of small area expansion factors, these weights (in acres) can also be stored in tabular form once they have been merged with the FIA plot identifiers. This table of small area weights (shown in **Figure 1**, step 4) can then be used with the map of small area patches (**Figure 1**, step 2) to produce small area estimates. Here, instead of delivering these inputs as separate products, we opt to distribute them on the backend of the RegRake R Shiny application (shown in step 5, **Figure 1**), which can combine both inputs on the fly to produce vector and raster-based estimates for user defined areas of interest (shown in step 6, **Figure 1**). Disseminating the small area weights and patch raster map *via* the R Shiny application simplifies the process of developing small area estimates, giving users the ability to derive small area estimates by simply uploading a polygon shapefile into the web-based interface.[1] Although the small area weights developed with regularized raking can be used to develop small area estimates for any attribute in the FIA database, currently RegRake only supports estimation of a limited number of key variables including total basal area of live trees >1.0 inch (based on the FIA condition variable, BALIVE in ft^2/acre), total net volume of wood in the merchantable stem of sample trees > 5.0 inches (based on the FIA tree variable, VOLCFNET in ft^3), and total acres of Forest, Agriculture, Developed and Other land use classes (derived by simplifying the FIA condition variable LAND_USE_SRS according to the cross-walk table found in the RegRake user's guide available at the above url). In addition, RegRake also offers the option to run these estimates on a per acre basis. For basal area and volume, the per acre estimates are derived by dividing the estimates of total basal area and total volume by the estimated number of forest acres in the user's polygon shapefile, while for land use, the per acre estimates are reported as a percentage of the polygon's area found in each land use class (summing to 100% across all land uses). In addition to these tabular

[1] http://quetelet.geog.utk.edu/regrake/

TABLE 1 | Population totals and per acre estimates for FIA basal area, net volume and land use variables derived for a series of concentric rings representing driving distances to a hypothetical biomass energy plant.

Distance to biomass plant (miles)	0–20	20–40	40–80
ACRES_UNADJ	480,391	1,565,802	2,694,707
BALIVE_TOT (ft^2)	26,744,795	88,923,486	175,528,455
VOL_TOT (ft^3)	473,034,195	1,572,693,792	3,088,850,144
FOREST_TOT (acres)	283,896	945,211	1,890,851
AG_TOT (acres)	126,524	371,062	428,540
URBAN_TOT (acres)	61,831	223,308	329,006
OTHER_TOT (acres)	8,140	26,221	46,309
BALIVEperAC (ft^2/acre)	94.21	94.08	92.83
VOLperAC (ft^3/acre)	1,666.23	1,663.85	1,633.58
FOREST_percent (%)	0.59	0.60	0.70
AG_percent (%)	0.26	0.24	0.16
URBAN_percent (%)	0.13	0.14	0.12
OTHER_percent (%)	0.02	0.02	0.02

and vector-based estimates, RegRake can also produce raster-based estimates, which represent the estimated amount of each variable found in each 30 m pixel (matching the resolution of the patch raster map). Therefore, when the per pixel values are multiplied by the area represented by all the pixels in each volume, basal area and land use bin and summed, the resulting values match the estimates reported in the RegRake tabular output.

As an example, we used RegRake to derive tabular, vector and raster-based small area estimates for a set of concentric rings representing various driving distances from a proposed biomass energy plant. First, the individual files making up the polygon shapefile are zipped and uploaded into the RegRake application (shown in step 5, **Figure 1**). Note, shapefiles must be in polygon format and can include one or many polygons in the same file. Here, our shapefile contains 3 polygons representing driving distances of 0–20, 20–40, and 40–80 miles from the proposed biomass energy plant. We recommend that polygons be at least 10,000 acres to ensure enough FIA plots are used to produce valid estimates. We also note that if the uploaded shapefile overlaps multiple survey units, each one must be run individually and later combined to produce estimates for the full area. After running the population totals and per acre estimates the results appear directly in the table tab of the RegRake application. These estimates (shown in **Table 1**) can then be copied and pasted into a spreadsheet or downloaded as a new shapefile with the results appended to the attribute table (e.g., the shapefile showing total volume for the three driving distances in our example is shown in **Figure 1**, step 6). As a final step the user can also develop raster-based estimates, which can be downloaded as a single or multi-band file depending on the number of variables selected (e.g., the raster-based total volume map for our example shapefile is shown in **Figure 1**, step 6). In this hypothetical example, a plant manager could overlay the raster total volume estimates with other GIS data sets (e.g., roads, land use, protected areas, ownership, etc.) to help pinpoint areas where supply is plentiful, allowing cost estimates associated with accessing and delivering the necessary raw materials to be considered when evaluating potential sites for building a new bioenergy plant.

ADVANTAGES AND FUTURE OPPORTUNITIES OF REGULARIZED RAKING

In this article we describe an approach for using a web-based, R Shiny application called RegRake to deliver a new set of expansion factors that can be used with U.S. Forest Service FIA data to develop on the fly, custom small area estimates for several forest attributes in the state of South Carolina, USA. The regularized raking estimator used to develop these new expansion factors has several appealing qualities that help facilitate the development of small area estimates. These include automated calibration of survey design weights using multiple ancillary data sets that can be applied to all FIA variables (instead of just 1 variable, as is common with the GREG estimator). The new, strictly positive expansion factors also produce consistent spatial and tabular estimates, that scale seamlessly across domains of interest, and which maintain relatively low levels of uncertainty despite the tendency for variance to increase when small area estimates are made with FIA data alone. Although RegRake is only available in South Carolina we anticipate adding new locations as the requisite expansion factor maps are developed for other states across the southeastern U.S. We also plan to add other attributes and estimates of uncertainty, giving users even greater flexibility to develop and evaluate small area estimates for a more robust suite of forest characteristics for their area of interest.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: forest inventory data - https://apps.fs.usda.gov/fia/datamart/ RegRake R Shiny Application - http://quetelet.geog.utk.edu/regrake/.

AUTHOR CONTRIBUTIONS

NN and TS: conceptualization and supervision. NN: methodology. NN and JM: software, data curation, R programming, and writing—review and editing. TS: writing—original draft, visualization, project administration, and funding acquisition. All authors contributed to the article and approved the submitted version.

FUNDING

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

REFERENCES

Bechtold, W. A., and Patterson, P. L. (2005). "The enhanced forest inventory and analysis program—National sampling design and estimation procedures", in *General Technical Report SRS-80* (Asheville, NC: US Department of Agriculture, Forest Service, Southern Research Station).

Breidt, F. J., and Opsomer, J. D. (2008). Endogenous post-stratification in surveys: classifying with a sample-fitted model. *Ann. Stat*. 36, 403–427. doi: 10.1214/009053607000000703

Chang, W., Cheng, J., Allaire, J. J., Xie, Y., and McPherson, J. (2020). *shiny: Web Application Framework for R. R package Version* 1.5.0. Available online at: https://shiny.rstudio.com/reference/shiny/1.4.0/shiny-package.html

Cochran, W. G. (2007). *Sampling Techniques, 3rd Edn*. Hoboken, NJ: John Wiley & Sons.

Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *J. Am. Stat*. 87, 376–382. doi: 10.1080/01621459.1992.10475217

Deville, J.-C., Särndal, C.-E., and Sautory, O. (1993). Generalized raking procedures in survey sampling. *J. Am. Stat*. 88, 1013–1020. doi: 10.1080/01621459.1993.10476369

Guggemos, F., and Tille, Y. (2010). Penalized calibration in survey sampling: design-based estimation assisted by mixed models. *J. Stat. Plan. Inference*. 140, 3199–3212. doi: 10.1016/j.jspi.2010.04.010

Homer, C., Dewitz, J., Yang, L., Jin, S., Danielson, P., Xian, G., et al. (2015). Completion of the 2011 National Land Cover Database for the Conterminous United States representing a decade of land cover change information. Photogramm. *Eng. Rem. S*. 81, 345–354. Available online at https://cfpub.epa.gov/si/si_public_record_report.cfm?Lab=NERL&dirEntryId=309950

Köhl, M. (2004). Inventory | forest inventory and monitoring. *Enc. For. Sci*. 403–409. doi: 10.1016/B0-12-145160-7/00154-X

McConville, K. S., Breidt, F. J., Lee, T. C. M., and Moisen, G. G. (2017). Model-assisted survey regression estimation with the LASSO. *J. Surv. Stat. Methodol*. 5, 131–158. doi: 10.1093/jssam/smw041

Nagle, N. N., Buttenfield, B. P., Leyk, S., and Spielman, S. (2014). Dasymetric modeling and uncertainty. *Ann. Assoc. Am. Geogr*. 104, 80–95. doi: 10.1080/00045608.2013.843439

Nagle, N. N., Schroeder, T. A., and Rose, B. (2019). A regularized raking estimator for small-area mapping from forest inventory surveys. *Forests*. 10, 111045. doi: 10.3390/f10111045

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Thompson, S. K. (2012). *Sampling, 3rd Edn*. Hoboken, NJ: John Wiley & Sons.

USDA (2008). *Chapter 10 Operational Procedures*. United States Department of Agriculture (USDA), Forest Service Handbook No. 4809.11, 24. Available online at: https://www.fs.fed.us/dirindexhome/fsh/4809.11/4809.11_10.doc (accessed June 24, 2022).

Author Disclaimer: This study was subject to agency review and approved for publication. Any use of product names is for descriptive purposes only and does not imply endorsement by the U.S. Government. This paper was written and prepared by U.S. Government employees on official time, and is therefore, in the public domain and not subject to copyright.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.