Notice: This publication is a Postprint

Postprint publications are not considered final versions but are citable. This publication has undergone technical and policy review, but still requires technical editing and final layout. The final version will reflect changes to wording and grammar and limited corrections to content, if needed. The most current version of the publication will always be available from the Digital Object Identifier (DOI): https://doi.org/10.2737/NRS-GTR-198.

Weighting for the USDA Forest Service, National Woodland Owner Survey

Brett J. Butler and Jesse Caputo

Abstract

The USDA Forest Service's National Woodland Owner Survey collects information on the attitudes, behaviors, and general characteristics of private forest ownerships across the United States. An area-based sample design that results in inclusion probabilities proportional to size of forest holdings is used to select ownerships to participate in the survey. In order to make accurate population-level estimates, the sample design must be incorporated into the estimators. A weighting approach for generating estimates of totals, means, proportions, and quartiles in terms of ownerships and acreages is presented, along with a bootstrapping approach for estimation of the associated variances. In addition to presenting a theoretical justification for the approach, the estimators are validated using data from a fully enumerated population. An R package for implementing the estimators is available on GitHub (https://github.com/familyforestresearchcenter/nwos).

Keywords: Forest Inventory and Analysis, National Woodland Owner Survey, estimation, bootstrapping, private forest owners, family forest owners

The Authors

BRETT J. BUTLER is a research forester with the USDA Forest Service, Northern Research Station.

Postprint

JESSE CAPUTO is a research forester with the USDA Forest Service, Northern Research Station.

Table of Contents

<Insert Table of Contents>

INTRODUCTION

Nearly six out of every ten acres of forest land in the United States are privately owned (Butler et al. In review). The USDA Forest Service's Forest Inventory and Analysis (FIA) program implements the National Woodland Owner Survey (NWOS) to generate information on the numbers of private forest ownerships, their general characteristics, attitudes, and behaviors. Statistics are generated at the national, regional, and state levels for use by forestry agencies, nongovernmental organizations, educators, the private sector, and researchers for analyzing, designing, and implementing programs, policies, and services affecting forest owners. Although the NWOS covers all private forest ownerships, most publications have concentrated on the 9.6 million (SE = 0.2) families, individuals, trusts, estates, and family partnerships, collectively referred to as family forest ownerships, who own 38.6 percent of the U.S.[1] forest land and in particular the 3.7 million (SE = 0.04) family forest ownerships with 10+ acres (4+ ha) of forest land who own 93.2 percent of the family forest land (Butler et al. In review).

The NWOS estimation methods have been the subject of previous publications (Metcalf et al. 2012, Dickinson and Butler 2013). These publications debated the then current estimation approach, but the final conclusions were that the approach and basic estimators were theoretically sound. Although sound, the approach was complex and not very transparent. In addition, the variance estimators had some shortcomings (e.g., inability to incorporate covariances) and the mathematical derivations for the variance estimators were never published.

[1] Excluding interior Alaska

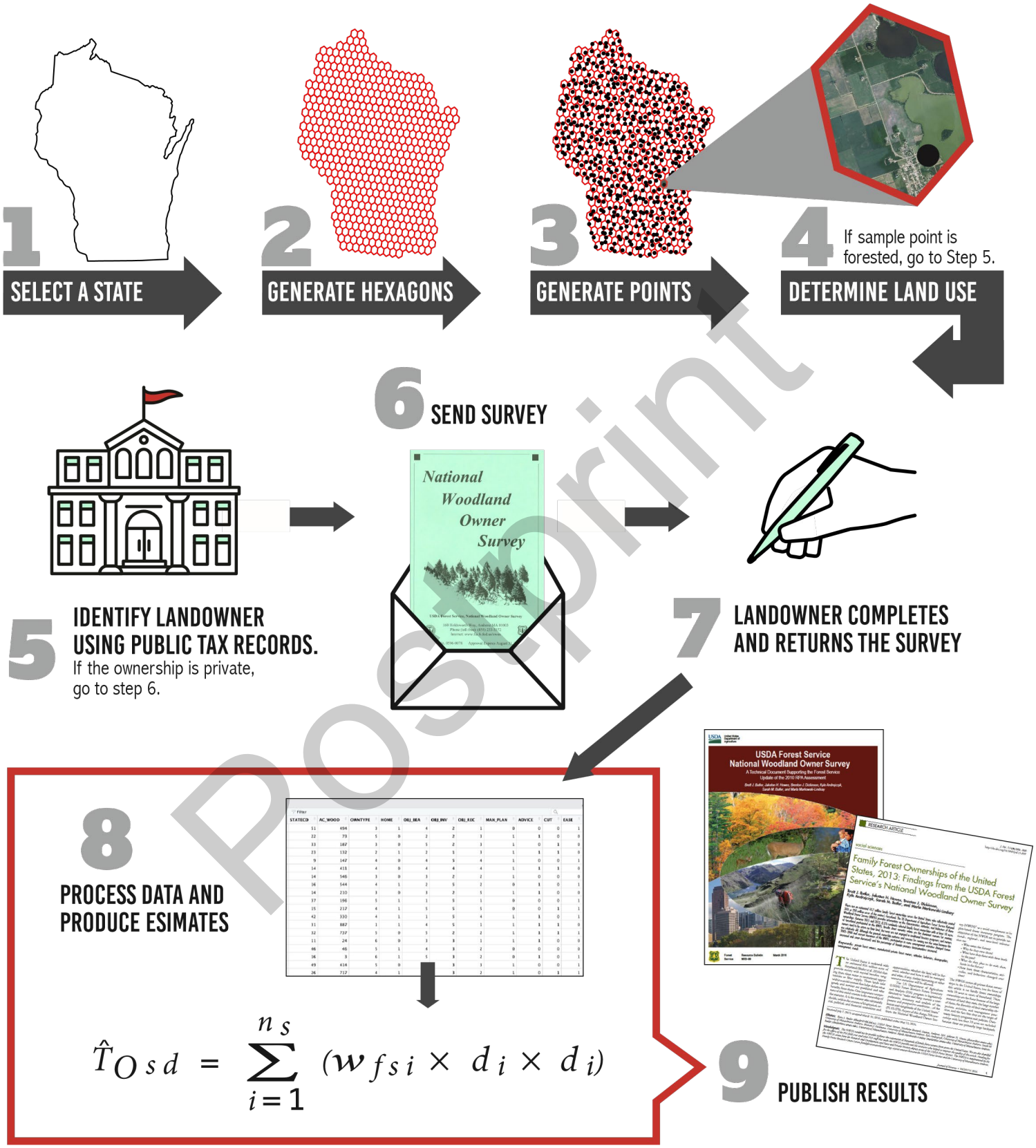
Postprint

Here a weighting approach for NWOS estimation that is simpler and more transparent is presented. This new approach also allows for easier incorporation of nonresponse adjustments. Valliant et al. (2013) provide general guidance on implementing weighting approaches. The basic steps are to draw the sample, receive responses from the respondents, generate base weights, adjust weights for response rates, calibrate weights for nonresponse bias, and generate final weights.

The NWOS sampling procedure (Fig. 1) is built on top of the sample design used by FIA for the plot-based, biophysical inventory (Bechtold and Patterson 2005). The NWOS, as with the FIA plots, is stratified by state[2]. States are the basic reporting units and the levels at which the sampling intensities are defined. Within each state, a hexagonal grid is established to spatially distribute the sample. Within each hexagon, a sample point is randomly located. For each sample point, the land use (e.g., forest or nonforest) is determined. Ownership for sample points that are determined to be forested, according to the FIA definition (USDA Forest Service 2016), is identified from publicly available property tax records. The resulting ownership list is deduplicated so that a given ownership is not included in the sample more than once for a given state and cycle. Ownerships are classified based on the names recorded in the tax records (Table 1). The private forest ownerships identified constitute the sample for the NWOS for a given survey cycle.

Postprint

[2] FIA divides Alaska, Oklahoma, and Texas into subregions and samples each of these subregions separately (e.g., different timing and/or intensities). The subregions are treated as separate "states" for estimation purposes.



Postprint

Figure 1.—Overview of the process used to generate estimates for the USDA Forest Service,

National Woodland Owner Survey.

Table 1.—Ownership categories used by the USDA Forest Service, National Woodland Owner Survey. Descriptions and FIA ownership class codes are from O'Connell et al. (2016). The use of the search terms is described below.

Group	Category	Description	OWNCD	Search term examples
Private				
	Family	Individual and family, including trusts, estates, and family partnerships	45	*
	Corporate	Corporate, including Native Corporations in Alaska and private universities	41	Corporation, Company
	Other private	Nongovernmental conservation/natural resources organization. Examples: Nature Conservancy, National Trust for Private Lands, Pacific Forest Trust, Boy Scouts of America; and unincorporated partnerships/associations/clubs. Examples: Hunting clubs that own, not lease property, recreation associations, 4H clubs, and churches	42; 43	Club, Association
Tribal				
	Tribal	Native American (Indian) - within reservation boundaries	44	Tribe
Public				
	Federal	National Forest; National Grassland and/or Prairie; Other Forest Service land; National Park Service; Bureau of Land Management; Fish and Wildlife Service; Departments of Defense/Energy; and other Federal	11; 12; 13; 21; 22; 23; 24; 25	United States of America, US Fish and Wildlife
	State	State, including state public universities (31)		State of Wisconsin, Wisconsin DNR
	Local	Local (county, municipality, etc.), including water authorities; and other non-Federal public	32; 33	Adams County, City of
Unknown				
	Unknown		--	Unknown, Not available

* If none of the full list of search terms were found, the ownership was classified as Family.

The hexagons associated with the base FIA fields plots are approximately 6,000 acres in size and each contains one, randomly located sample point. Where this sample is insufficient to

Postprint

obtain the NWOS target sample sizes, additional sample points are generated by randomly locating points within smaller hexagonal grid cells nested within the original grid cells. The land uses of the sample points collocated with FIA field plots are determined by initially assessing if a plot is likely forested and if so, verifying the land use in the field (USDA Forest Service 2019). The land uses of the augmented sample points are determined based on high resolution aerial photography with rules adapted from the field measurement designations.

Beginning in 2019, the NWOS is implemented on an annualized basis by spreading the sample across multiple years. Twenty percent of the sample is contacted each year over each 5-year cycle. The process is then repeated for subsequent cycles with the same points being reused, ownerships being resurveyed (where the ownerships have not changed), and new ownerships being surveyed where they have changed.

The area-based NWOS sample design results in inclusion probabilities proportional to size – ownerships with larger forest holdings are more likely to be included in the sample than ownerships with smaller holdings. The inclusion of appropriate weights for sample designs with inclusion probabilities proportional to size will produce unbiased estimates (Lohr 1999).

This report presents a weighting approach for generating estimates for the NWOS. Estimators presented include totals, means, proportions, and quantiles. An approach for estimating the associated variances is also included. Following the theoretical justification for the estimators, the estimators are empirically validated using data from a known, fully-enumerated population. Adjustments for response rates and known totals are discussed.

Postprint

THEORETICAL APPROACH

The objective of a statistical estimator is to use sample data to estimate the value of a population parameter (e.g., total or mean) along with a measure of its reliability. For the discussion here, the target population is family forest ownerships in a given state, but the estimators can be easily applied to other populations of interest, such as corporate forest ownerships. A family forest ownership is defined as "families, individuals, trusts, estates, and family partnerships that own forest or woodland" (Butler et al. 2016: 11). Forest/ woodland is defined as "Land that has at least 10 percent crown cover by live tally trees of any size or has had at least 10 percent canopy cover of live tally species in the past, based on the presence of stumps, snags, or other evidence. To qualify, the area must be at least 1.0 acre in size and 120.0 feet wide" (USDA Forest Service 2016). NWOS estimates are stratified by land use and ownership category within a state (e.g., family forest ownership in Wisconsin). Estimates across strata are calculated by combining the statistics for the underlying strata. As discussed below, estimates for domains of interest (e.g., ownerships with written forest management plans) are calculated by including dummy variables.

Postprint

Weights

The first step in the weighting process is to calculate the base or design weights. These are calculated for each ownership in a given stratum in the sample; weights are only calculated for respondents. The base weights are equal to the inverse of the selection probabilities. For the NWOS, the selection probabilities are a function of the acreage of forest holdings in a state for each ownership, the total land area in the state, and the sample size:

$$\omega_{bsi} = \frac{1}{\pi_{si}} p_{si} = \frac{1}{\frac{a_{si} n_s}{A_s}} p_{si} = \frac{A_s}{a_{si} n_s} p_{si}$$

(1)

Where

- ω_{bsi} = base weight for ownership i in stratum s,
- $\pi_{si} = a_{si} / (A_s / n_s) = (a_{si}\, n_s)/A_s$ = selection probability for ownership i in stratum s,
- a_{si} = area of forest land of ownership i in stratum s,
- p_{si} = number of sample points on the forest land of ownership i in stratum s,
- $A_s = (n_S/n)A$ = area of land in stratum s,
- n_s = sample size (i.e., number of sample points) in stratum s,
- n = total sample size in the state, and
- A = total area of land in the state.

It is important to note that A is for all land area in a state and n is the total sample size (i.e., number of sample points) in a state, regardless of land use or ownership category.

To force forest areas in a stratum to equal those from another data source, (e.g., area of forest land estimated from FIA plots $[A'_S]$), A_S in Equation 1 can be adjusted by A'_S / A_s. This is algebraically equivalent to replacing A_S with A'_S in Equation 1.

It is extremely rare that responses are obtained from all individuals selected to participate in a survey, accordingly the weights need to be adjusted for this unit nonresponse. For the NWOS, response rates are calculated by stratum (and state) to account for the fact that response rates vary across strata. This adjustment is equal to the inverse of the response rate:

Postprint

$$adj_{rs} = \frac{1}{r_s} = \frac{1}{n_{sr}/n_s} = \frac{n_s}{n_{sr}} \tag{2}$$

Where

- adj_{rs} = response rate adjustment for stratum s,
- r_s = response rate for stratum s, and
- n_{sr} = number of sample points owned by respondents in stratum s.

If nonresponse biases are detected, an additional case-based adjustment can be incorporated. For example, propensity score matching (Brick 2013) can be used to estimate response probabilities using ancillary data associated with all sample points. Normalized, inverse response probabilities can then be incorporated into the weights. The normalization is necessary to ensure the total acreages in strata do not change. A propensity score matching approach was used for the 2018 NWOS (Butler et al. In review).

The final weights are the product of the base weights and response and nonresponse adjustments:

$$\omega_{fsi} = \omega_{bsi} \times adj_{rs} \times adj_{nsi} \tag{3}$$

Where

- adj_{nsi} = nonresponse adjustment for ownership i in stratum s and
- ω_{fsi} = final weight for ownership i in stratum s.

Estimators

Estimators for totals, means, proportions, and quartiles, including medians, are presented. Statistics are defined in terms of ownerships (O) and area (A). The term "domain of interest" is

Postprint

used to describe a specific attribute of the ownerships that is being described, such as ownerships with a written forest management plan.

Totals

The estimated total number of ownerships in a domain of interest in a stratum is equal to the sum of the weights multiplied by a dummy variable indicating inclusion in the domain of interest, multiplied by the variable of interest (Eq. 4a). The variable of interest is set to one if the desired units are number of ownerships, which is typically the case. Alternatively, the variable of interest can be set to some other numeric variable, such as age. This may not seem immediately useful, but it can be used to calculate means or proportions, as is done below. To estimate area totals (Eq. 4b), area of land owned in the stratum is incorporated into the product calculated using the elements in Equation 4a. Equation 4b is algebraically equivalent to Equation 4a if $a_{si} = 1$. The estimated total number of family forest ownerships, or family forest acres, is calculated by setting the dummy variable to 1 for all family forest ownerships, and 0 otherwise. To estimate other totals (e.g., number of family forest ownerships with 10+ acres), the dummy variable is coded accordingly.

$$\hat{T}_{Osdv} = \sum_{i=1}^{n_{sr}} (\omega_{fsi} \times d_i \times v_i) \tag{4a}$$

$$\hat{T}_{Asdv} = \sum_{i=1}^{n_{sr}} (\omega_{fsi} \times d_i \times v_i \times a_{si}) \tag{4b}$$

Where

- $\hat{T}_{Osdv}$ = estimated total, in terms of ownerships, for variable v in domain d in stratum s,
- $\hat{T}_{Asdv}$ = estimated total, in terms of area, for variable v in domain d in stratum s,

Postprint

- d_i = dummy variable indicating inclusion in domain d, and
- v_i = variable of interest.

Means

The estimated mean in terms of ownerships of a given variable of interest is equal to the sum of the weighted values of the variable in the domain of interest divided by the total number of ownerships in the domain of interest (Eq. 5a). To calculate the mean on an area basis, the weights and variables are multiplied by the area for each ownership before being summed and divided by the total area for the domain (Eq. 5b).

$$\bar{v}_{Osd} = \frac{\sum_{i=1}^{n_{sr}} (\omega_{fsi} \times d_i \times v_i)}{\sum_{i=1}^{n_{sr}} (\omega_{fsi} \times d_i)} \tag{5a}$$

$$\bar{v}_{Asd} = \frac{\sum_{i=1}^{n_{sr}} (\omega_{fsi} \times d_i \times a_{si} \times v_i)}{\sum_{i=1}^{n_{sr}} (\omega_{fsi} \times d_i \times a_{si})} \tag{5b}$$

Where

- $\bar{v}_{Osd}$ = estimated mean value of v in domain d in stratum s in terms of ownerships and
- $\bar{v}_{Asd}$ = estimated mean value of v in domain d in stratum s in terms of area.

Proportions

Proportions are a special case of means where v is an indicator or dummy (i.e., binary) variable. To estimate a proportion, Equations 5a or 5b are used with $v_i = 1$ where the condition of interest is present and $v_i = 0$ otherwise.

Quantiles

Estimating quantiles using weights is difficult to do with a closed form equation. For the NWOS, an iterative approach is used. The midpoint value of the full range of the variable of

Postprint

interest is used as an initial starting value, and the estimated weighted proportion of the population above (and below) that value is calculated. The value is then changed incrementally until the proportion is equal to the target quantile probability. The median value is equal to a probability of 0.50. The exceptions to this approach are for probabilities of 0.00 and 1.00 where the values are equivalent to the minimum and maximum values, respectively, and are set accordingly.

Estimation of Variances

Estimates of variances associated with estimators are typically constructed using either a closed form equation or a resampling approach (Valliant et al. 2013). The complex sample design associated with the NWOS makes the variances very difficult to specify formulaically. Often variances for complex sample designs cannot be calculated exactly for sample sizes greater than two, and a linearization technique is needed to estimate them, as was done for the 2013 NWOS (Dickinson and Butler 2013). Fortunately, resampling techniques can be used to empirically estimate the variances of virtually any sample-based statistic (Efron and Tibshirani 1986).

A bootstrapping resampling technique has been adopted for estimating variance for the NWOS. Resampling has multiple advantages, including being very flexible and robust, even for complex sample designs (Efron and Tibshirani 1986). This approach also has disadvantages, including computational intensity and estimates that will vary with each run. Bootstrapping (Efron and Tibshirani 1986) is one of the most common resampling approaches. This technique works by resampling the responses, with replacement, to create a new sample with the same number of responses as the original sample. Estimates are calculated using the new sample. This process is typically repeated thousands of times, and the variance is calculated from the

Postprint

iterative estimates. To capture the full variation in the estimates for the NWOS, it is important to not force the stratum areas (A_S) to match the values from other data sources.

EMPIRICAL TESTING

To understand the behavior of the estimators, the previously outlined methods were applied to a complete dataset derived from actual parcel data. Having a complete dataset allows for the true population values to be known. However, the number of family forest ownerships in a state is not known, nor are other NWOS relevant attributes. Therefore, a surrogate population with a comparable distribution was used. Currently, only a handful of states have statewide, publicly available parcel data, but data will likely be available for more states soon. For the purposes of this report, however, only a single population was needed. The State of Wisconsin was selected because it has one of the most complete and clean parcel databases in the United States, it has extensive forest land, and there are a large number of family forest ownerships in the State, all attributes that are advantageous for the purposes of this report.

A Bayesian approach (Kruschke 2011) was used to assess if the estimated values were significantly different than the actual values. Prior values were set using the actual population values. High density intervals (HDI), which reflect the range of values representing the estimated values, and effect sizes were used to assess differences. These models were run in the R computing environment using the Bayesian First Aid package (Bååth 2014). The graphics associated with the results include predicted values, effect sizes, standard deviations, and a plot of the data versus the predicted posterior distribution. For the first three of these charts in each set, the 95 percent high density intervals and median values are included.

Postprint

Defining the Population of Interest

Family forest ownerships are the ultimate population of interest for many NWOS analyses and for the analyses presented in this report. Common attributes of interest are numbers of family forest ownerships, area owned, and characteristics such as presence of written forest management plans and age of owners. These attributes are often presented in terms of both ownerships and acres, as the two can show different patterns. For example, a small percentage of family forest ownerships may have written forest management plans, but they may own a relatively larger percentage of the family forest land.

The input data used for this analysis were parcel and forest coverage spatial layers for Wisconsin. The parcel data were from the Wisconsin State Cartographer's Office and Land Information Program (2016). The forest coverage data came from the National Land Cover Database (NLCD) 2011 Forest Service Percent Tree Canopy Analytical Product (USGS 2014). The forest canopy layer was clipped to the state boundary. All of the 98 feet by 98 feet (30 m by 30 m) pixels in the clipped forest canopy layer that had a percent forest cover of at least 10 percent, the FIA threshold for forest land (USDA Forest Service 2016), were coded 1 and all other pixels were coded 0. The number of forest and nonforest pixels was counted for every parcel. This geospatial processing was completed using the R computing environment (R Core Team 2019).

Subsequent data processing assigned ownership categories, identified unique ownerships, and summed forest area by ownership. Ownerships were classified into eight categories based on keywords in the ownership name field associated with each parcel (Table 1). If none of the

Postprint

search terms were found, the ownership was classified as family. Within each ownership group, unique ownerships were identified using fuzzy string matching. For family ownerships, the string was a concatenation of ownership name and address; for all other ownerships, it was based solely on ownership name. Names and names/addresses were standardized by removing all punctuation, removing extra spaces, and converting all text to uppercase. The Levenshtein distance (Levenshtein 1966) was calculated between pairs of names or names/addresses using a threshold value of five to identify the same ownerships. Due to computing capacity limitations caused by the large number of record pairs, calculations were run in batches. Family forest ownership matching was done within zip codes, and zip codes with more than 1,000 records were further subdivided by the first two letters of the owners names. Other ownership categories were batched based on the first two letters of the names. The final step was to sum forest acreages by ownership. All of these steps were completed in the R computing environment (R Core Team 2019).

To further test the estimators, two factor variables and one numeric variable were generated. One of the factor variables (y_1) was created by randomly assigning a 1 or 0 to each observation, i.e., $y_1 \sim Bern(0.5)$. This variable can, for example, be thought of as representing ownerships that report having harvested firewood. A second factor variable (y_2) was generated so that it correlated with size of forest holdings. This variable can, for example, be thought of as representing ownerships reporting they have a written forest management plan. Based on the observed relationship between size of forest holdings and written forest management plans for family forest ownerships who responded to the 2013 NWOS (Butler et al. 2016), this variable was generated as:

Postprint

$$p(y_2) = \frac{e^{\beta_0 + (\beta_1 ln(a_{si}))}}{1 + e^{\beta_0 + (\beta_1 ln(a_{si}))}} \tag{6a}$$

$$y_2 = Bern(p(y_2)) \tag{6b}$$

Where

- $p(y_2)$ = probability of y_2 based on a logistic regression model with β_0 = -4.0 and β_1 = 0.8 and
- $Bern(p)$ = Bernoulli distribution with a probability of p.

The numeric variable (y_3) was randomly generated using a normal distribution with a mean of 60.9 and a standard deviation of 7.5, i.e., $y_3 \sim N(60.9, 7.5)$. These values represent the mean landowner age and associated sampling error for family forest owners in Wisconsin as reported in the 2013 NWOS (Butler et al. 2016).

A response propensity variable was also created. This variable represents the likelihood of an ownership responding. The distribution was assumed to be uniformly distributed with a range of 0 to 1, i.e., $rp_i \sim U(0, 1)$. This variable can be interpreted to mean that as it nears 0, the probability of responding decreases, and as it nears 1, the probability increases. Ownerships were classified as respondents if $rp_i \geq 0.5$, and nonrespondents otherwise. This value is approximately equal to the cooperation rate reported for the 2013 NWOS (Butler et al. 2016).

Sample Selection

To test the estimators, a probability proportional to size design was used when drawing the samples to mimic the NWOS sample design. The complete list of parcels was the sampling frame. For each ownership, forest and nonforest acreage was listed as separate records to allow for the sample to be stratified by these two land cover types. Samples were selected with

Postprint

probabilities equal to the area of forest and nonforest associated with each ownership, and a record was allowed to be selected more than once (i.e., a with replacement design). This sampling procedure was repeated for 5,000 independent iterations. All procedures were conducted in the R computing environment (R Core Team 2019).

Sample Size

Selecting a target sample size involves some degree of subjectivity. Using a power equation approach (Cohen 1988), the willingness to accept Type I and II errors must be selected. From a resource dependent perspective, the value of additional samples can also be assessed. The target sample sizes will also vary depending on the specific estimate being evaluated, with the ultimate target sample size taken as the maximum across the assessments. For the NWOS, the two primary statistics used to determine sample sizes are the number and area of family forest ownerships. These two statistics are considered for all family forest ownerships with 1+ acres and 10+ acres.

Using the approaches outlined in this report, estimates of the total numbers of ownerships and acreage for family forest ownerships with 1+ acres in the empirical dataset were tested in sample sizes with increments of 100 for 10 sets of samples, with sample sizes that ranged from 100 to 10,000. These sample sizes yielded number of family forest ownership "respondents" that ranged from 12 to 1,713. In some cases, the numbers of respondents were low due to only a low percentage of the sample points being forested and family owned, low response rates, and the randomness inherent in the sample selection process. The coefficients of variation for both ownerships and acreage estimates reduced drastically after about 100 respondents, and the marginal gains in coefficient of variation reduction became much more limited after about 250

Postprint

respondents (Fig. 2). The coefficient of variation for family forest acreage reached 0.05, on average, after 146 respondents and 0.025 after 544 respondents. In contrast, even once the maximum sample size tested (average of 1,654 respondents) was reached, the average coefficient of variance for estimated number of family forest ownerships still did not reach 0.05, although it was close, reaching 0.056. Looking at family forest ownerships with 10+ acres (Fig. 3), the number of respondents needed to reach the same thresholds for numbers of ownerships is greatly reduced but is slightly increased for acreage estimates.

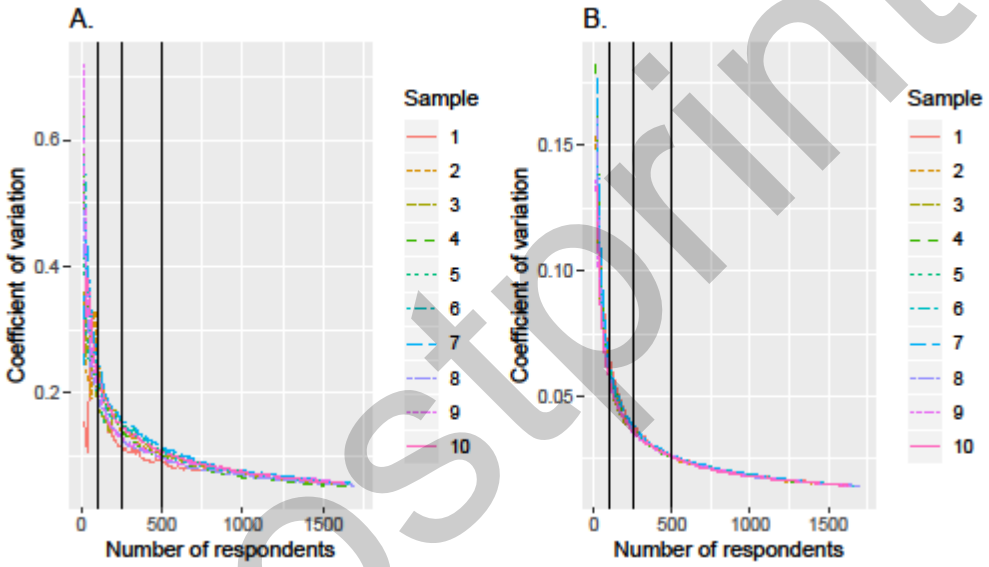


Figure 2.—Coefficients of variation for estimates of total numbers of (A) ownerships and (B) acres of family forests (1+ acres) in the empirical dataset by number of respondents. The vertical black lines indicate the critical values discussed in the text.

Postprint

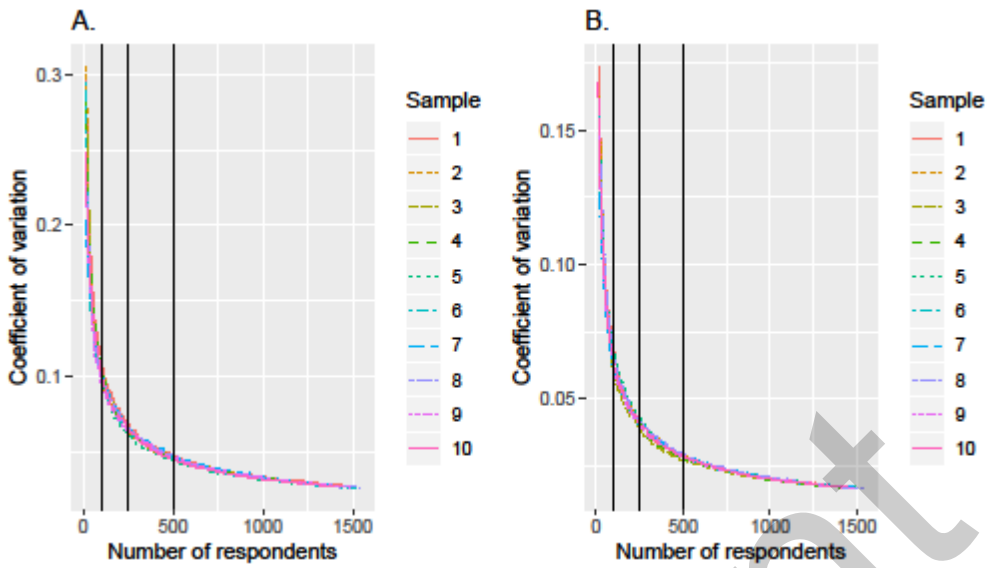


Figure 3.—Coefficients of variation for estimates of total numbers of (A) ownerships and (B) acres of family forests (10+ acres) in the empirical dataset by number of respondents. The vertical black lines indicate the critical values discussed in the text.

Bootstrap Replicates for Variance Estimation

The stability of variance estimates increase with the number of bootstrap replicates (R), but there is a point at which little additional information is gained, and computational time requirements must be taken into consideration. When the number of bootstrap replicates is small (i.e., $R < 500$), there is greater noise in the coefficients of variation (Fig. 4). The fluctuations attenuate substantially after about 1,000 replicates. It should be noted that the ranges for all of these estimates (i.e., the ranges of the vertical axes) are small.

Postprint

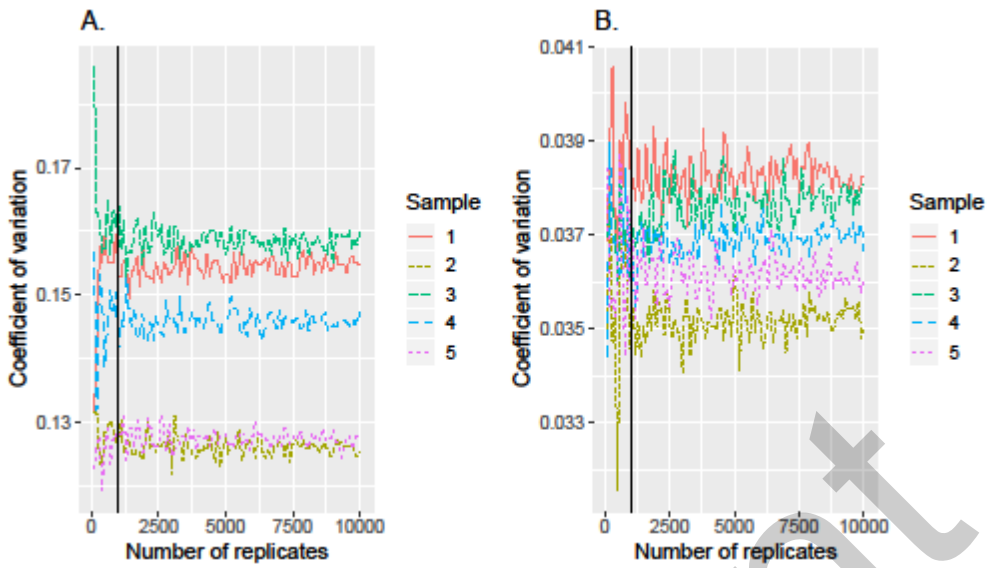


Figure 4.—Coefficients of variation for estimates of total numbers of (A) ownerships and (B) acres of family forests (1+ acres) in the empirical dataset by number of bootstrap replicates. The vertical black lines indicate the critical value discussed in the text.

Estimates of Totals

The actual number of family forest ownerships with 1+ acres of forest land in The empirical dataset , given the methods described above, is 549,856 (Table 2). The median value of the estimated number of family forest ownerships in the state is 548,011 with a 95 percent HDI of (-0.02, 0.04) (Fig. 5)

Table 2.—Actual values, weighted estimates, and unweighted estimates for number and area (in acres) of family forest ownerships (1+ acres) in The empirical dataset . Numbers in parentheses are sampling errors.

		Actual		Weighted Estimates		Unweighted Estimates	
Variable	Statistic	Ownership	Acres	Ownership	Acres	Ownership	Acres
–	Total	549,856	11,112,619	570,629	10,621,603	226	20,697
		(–)	(–)	(89,823)	(407,852)	(0)	(121)
Size	Mean	20.2	107.5	18.6	91.6	91.6	251.1
		(–)	(–)	(2.9)	(8.3)	(121.2)	(243.4)
Size	Q_0	1.0	–	1.1	–	1.1	–
Size	Q_1	2.1	–	1.9	–	27.8	–

Postprint

Size	Q_2	5.4	–	4.4	–	54.5	–
Size	Q_3	22.4	–	22.1	–	103.7	–
Size	Q_4	4,978.0	–	1,041.7	–	1,041.7	–
y_1	Proportion	0.50	0.50	0.48	0.49	0.49	0.46
		(–)	(–)	(0.08)	(0.03)	(0.01)	(0.01)
y_2	Proportion	0.09	0.35	0.08	0.36	0.36	0.65
		(–)	(–)	(0.02)	(0.03)	(0.01)	(0.01)
y_3	Mean	60.9	60.9	62.6	61.1	61.1	62.0
		(–)	(–)	(1.0)	(0.5)	(7.6)	(8.0)

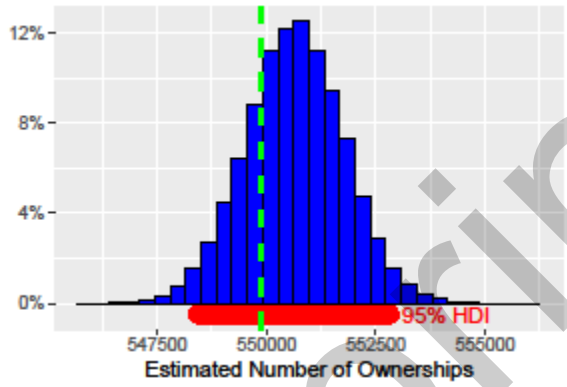


Figure 5.—Bayesian inference of estimates of total numbers of ownerships of family forests (1+ acres) in The empirical dataset . The red bar is the 95 percent high density interval (HDI). The vertical, green, dashed line is the true population value.

The acreage of family forest ownerships with 1+ acres of forest land in The empirical dataset is 11,112,619 (Table 2). The 95 percent HDI for the estimate is (11,112,777, 11,135,234) and the median effect size is 0.03 (Fig. 6).

Postprint

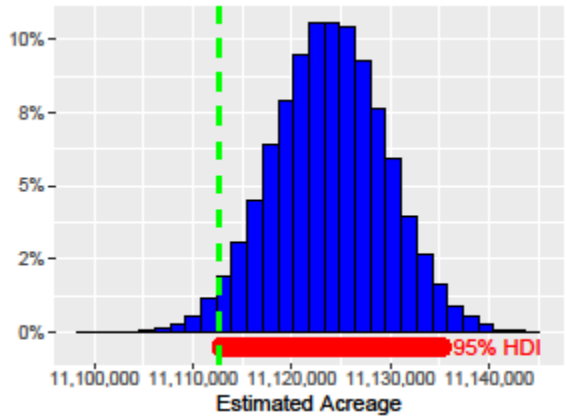


Figure 6.—Bayesian inference of estimates of total acreage of family forests (1+ acres) in The empirical dataset . The red bar is the 95 percent high density interval (HDI). The vertical, green, dashed line is the true population value.

Estimates of Proportions

Factor Variable [y_1 ~ Bern(0.5)]

The proportion of family forest ownerships (1+ acres) in The empirical dataset with $y_1 = 1$ is 0.50 (Table 2). The 95 percent HDI for the estimate is (0.50, 0.50) with a median effect size of -0.04 (Fig. 7).

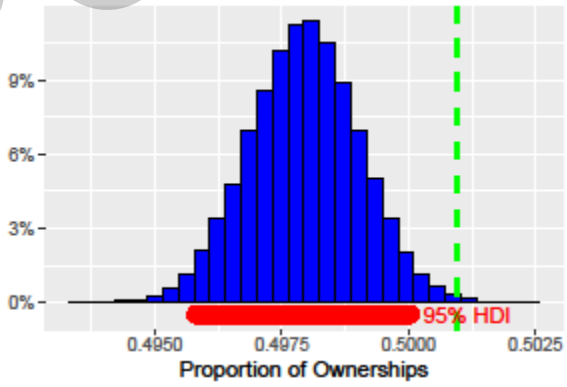


Figure 7.—Bayesian inference of estimates of proportion of ownerships of family forests (1+ acres) in The empirical dataset with $y_1 = 1$. The red bar is the 95 percent high density interval

Postprint

(HDI). The vertical, green, dashed line is the true population value.

The proportion of family forest land (1+ acres) in The empirical dataset with $y_1 = 1$ is 0.50 (Table 2). The 95 percent HDI for the estimate is (0.50, 0.50) with a median effect size of -0.07 (Fig. 8).

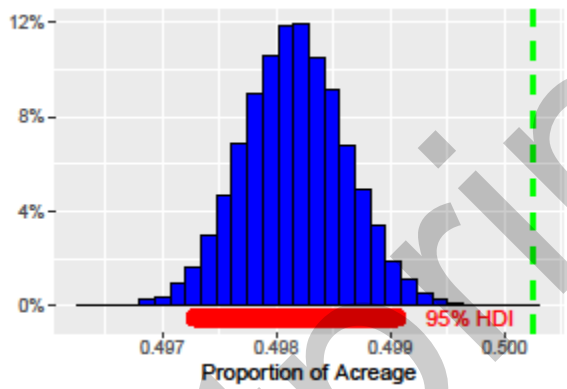


Figure 8.—Bayesian inference of estimates of proportion of family forest acreage (1+ acres) in The empirical dataset with $y_1 = 1$. The red bar is the 95 percent high density interval (HDI). The vertical, green, dashed line is the true population value.

Factor Variable

$$\left(y_2 \sim Bern\left(\frac{e^{-4.0+(0.8ln(a_{si}))}}{1+e^{-4.0+(0.8ln(a_{si}))}}\right)\right)$$

The proportion of family forest ownerships (1+ acres) in The empirical dataset with $y_2 = 1$ is 0.09 (Table 2). The 95 percent HDI for the estimate is (0.09, 0.09) with a median effect size of 0.02 (Fig. 9).

Postprint

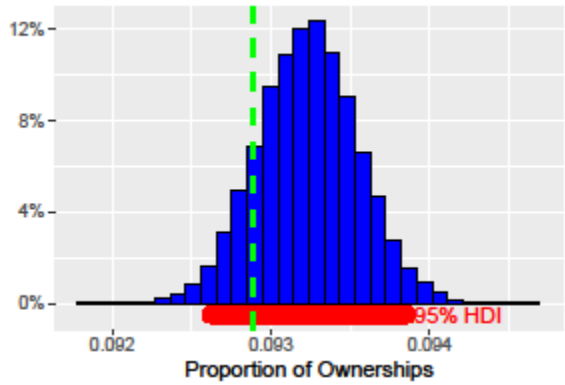


Figure 9.—Estimates of proportion of ownerships of family forests (1+ acres) in The empirical dataset with $y_2 = 1$. The red bar is the 95 percent high density interval (HDI). The vertical, green, dashed line is the true population value.

The proportion of family forest land (1+ acres) in The empirical dataset owned by people where $y_2 = 1$ is 0.35 (Table 2). The 95 percent HDI for the estimate of this parameter is (0.35, 0.35) with a median effect size of -0.01 (Fig. 10).

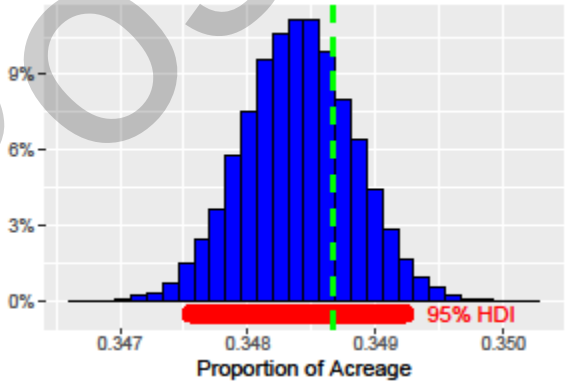


Figure 10.—Estimates of proportion of family forest acreage (1+ acres) in The empirical dataset with $y_2 = 1$. The red bar is the 95 percent high density interval (HDI). The vertical, green, dashed line is the true population value.

Postprint

Estimates of Means

Size of Holdlings

The mean size of a family forest holding (1+ acres) in The empirical dataset is 20.2 acres (Table 2). The 95 percent HDI for the estimate is (20.4, 20.6) with a median effect size of 0.1 (Fig. 11).

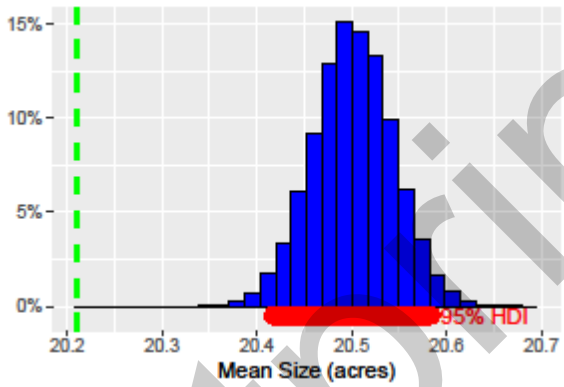


Figure 11.—Bayesian inference of the estimated mean size of family forest holdings (1+ acres) per ownership in The empirical dataset . The red bar is the 95 percent high density interval (HDI). The vertical, green, dashed line is the true population value.

The mean size of family forest holdings (1+ acres) on a per acre basis in The empirical dataset is 107.5 acres (Table 2). The 95 percent HDI for the estimate of this parameter is (104.5, 105.3) with a median effect size of -0.2 (Fig. 12).

Postprint

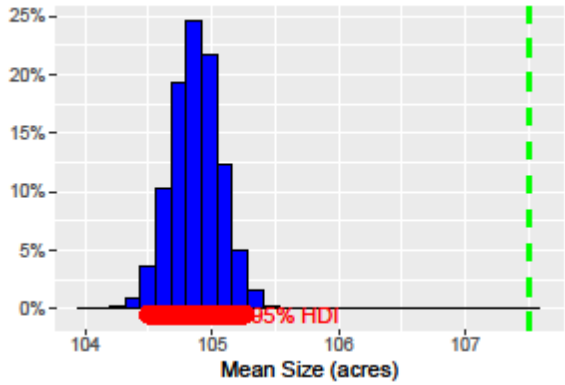


Figure 12.—Bayesian inference of the estimated mean size of family forest holdings (1+ acres) per acre in The empirical dataset . The red bar is the 95 percent high density interval (HDI). The vertical, green, dashed line is the true population value.

y_3 ~ N (60.9, 7.5)

The mean value of y_3 for family forest ownerships (1+ acres) in The empirical dataset is 60.9 (Table 2). The 95 percent HDI for the estimate is (60.9, 61.0) with a median effect size of 0.03 (Fig. 13).

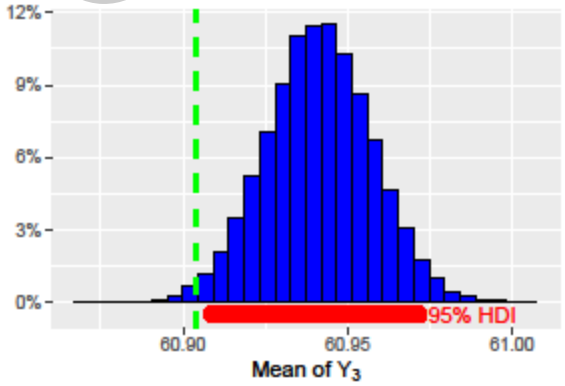


Figure 13.—Bayesian inference of the estimated mean of y_3 for family forest ownerships (1+ acres) in The empirical dataset . The red bar is the 95 percent high density interval (HDI). The

Postprint

vertical, green, dashed line is the true population value.

The mean value of y_3 for family forest ownerships (1+ acres) in The empirical dataset on a per acre basis is 60.9 (Table 2). The 95 percent HDI for the estimate of this parameter is (60.9, 60.9) with a median effect size of -0.1 (Fig. 14).

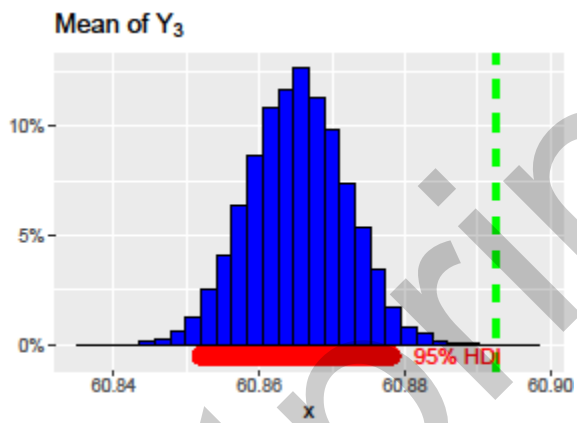


Figure 14.—Bayesian inference of the estimated mean of y_3 per acre for family forest ownerships (1+ acres) in The empirical dataset . The red bar is the 95 percent high density interval (HDI). The vertical, green, dashed line is the true population value.

Postprint

Estimates of Quartiles

The quantiles [i.e., quartile probability = (0.00, 0.25, 0.50, 0.75, 1.00)] of the size of holdings for family forest ownerships (1+ acres) in The empirical dataset are 1.0, 2.1, 5.4, 22.4, and 4,978.0 for Q_0 , Q_1, Q_2, Q_3 , and Q_4, respectively (Table 2). The 95 percent HDI for the estimates of these parameters are (1.2, 1.2), (2.2, 2.2), (5.8, 5.9), (23.2, 23.6), and (1,795.9, 1,848.4), respectively (Fig. 15). With the exception of the fourth quartile, the estimates are slightly higher than the actual values, but the relative and absolute effect sizes are small, with the absolute differences all being within 1 acre. The estimate for the fourth quartile is

substantially smaller than the actual value.

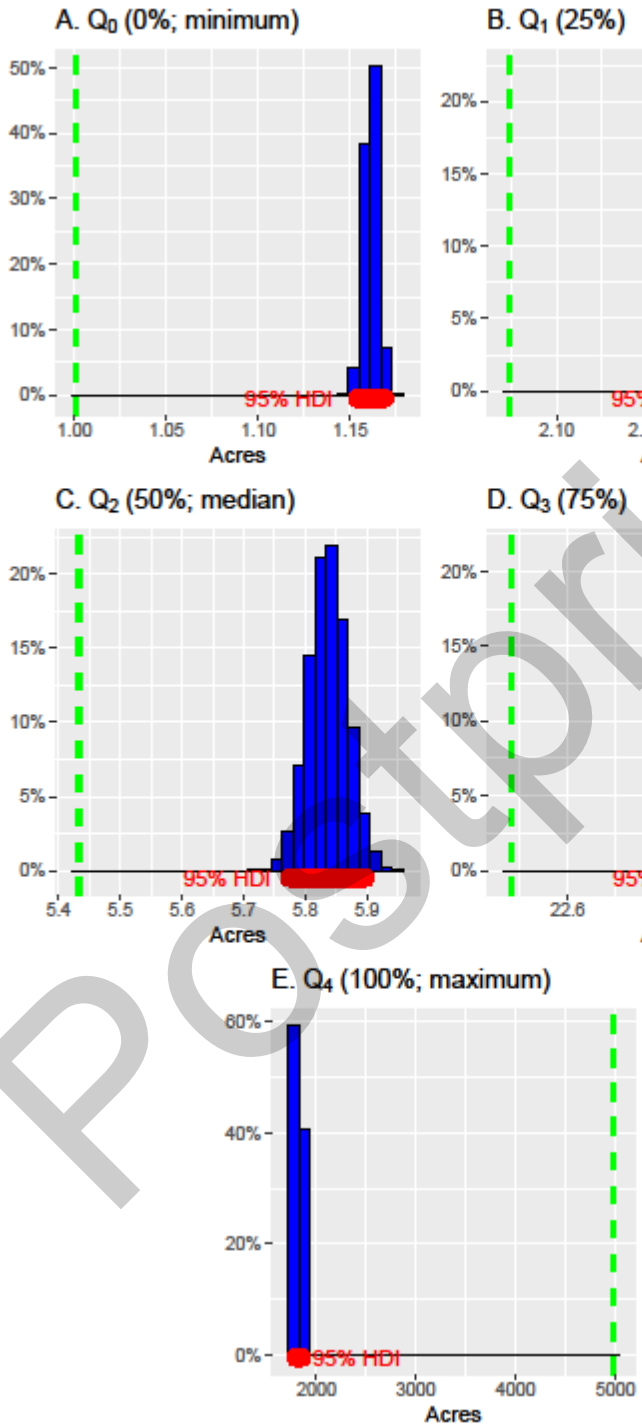


Figure 15—Bayesian inference of the estimated quartiles, of size of family forest holdings (1+ acres) in The empirical dataset . The red bars are the 95 percent high density intervals (HDI).

Postprint

The vertical, green, dashed lines are the true population value

DISCUSSION

The weighting approach provides a relatively straightforward method to produce estimates from the NWOS, and the estimators behaved well. However, a few issues need to be considered and are further discussed here:

- Robustness of estimates
- Weighting versus not weighting
- Effects of small ownerships on estimates
- Nonresponse bias

Robustness of Estimates

Overall, the estimators are robust. The 95 percent HDIs overlap with or are close to the estimates for most parameters, and for most estimates, the effect sizes are small. In some cases, nonoverlapping intervals with small effect sizes are due to the large number of replicates. Two instances where there is poor agreement between the estimates and the actual values are the zeroth and fourth quantiles (minimum and maximum). This is logical given that the sampling procedure does not, by definition, include all ownerships, and hence the most extreme values are unlikely to be included in the sample. Due to this finding, the estimates of the zeroth and fourth quantiles should not be considered reliable.

The variance estimation procedure also appears robust and avoids issues related to closed form estimation approaches. One concern with this bootstrapping approach is the computational time

Postprint

required, but this can be solved by faster computers (e.g., cloud computing) or reducing the number of bootstrap iterations.

It is interesting to note the variation in estimates across replicates due to the nature of random sampling. This is a known issue, but the results presented further illustrate this. For example, the proportion of family forest ownerships where $y_1 = 1$ is 0.50. The median estimate of this parameter is very close to the true value, as are the vast majority of the estimates (i.e., $Q_1 = 0.45$ and $Q_3 = 0.55$), but there is the chance, albeit unlikely, of extreme values that differ substantially (*min.* = 0.25 and *max.* = 0.76). This is the nature of random sampling, but it is something that should be considered when interpreting the results of any estimates that rely on random sampling. Reporting sampling errors and taking them into consideration helps to mitigate this issue, but does not completely address it.

There are some differences in the definitions used to generate the empirical data presented in this report and the definitions used by the NWOS. The analysis of the empirical data presented here uses a land cover definition, while the NWOS uses a land use definition. In addition, due to the very large amount of data in the full empirical dataset, the categorization of ownerships and the identification of ownerships with multiple parcels uses an automated procedure, which is less precise process compared to the method used by the NWOS. These differences hamper the direct comparisons between the estimates from these empirical data and the NWOS results.

To Weight or Not to Weight?

Weight! Using weights for all estimates is necessary to fully account for the sample design. Weighting allows something meaningful to be said about the population rather than just the

Postprint

sample. Another way of thinking about this is if weights are not used, this is equivalent to assigning all observations equal weights (i.e., $\omega_{fsi} = 1$). The implications can be seen by comparing the statistics for the estimates with and without weights (Table 2). Without weights, the totals are basically meaningless. The minimum and maximum values are the same as the weighted statistics, but the means and quartiles are substantially different. For the three ancillary variables, the values for y_1 and y_3 are within one percentage point of each other, but the unweighted estimate of y_2 is off by 23 percentage points for ownerships and is off by 10 percentage points for acres. The greater discrepancies in values for y_2 are related to this variable being correlated with size of holdings and the other variables being randomly generated.

Responses are roughly equivalent to acreage estimates, at least for statistics that summarize central tendencies. The estimates are the same for means of *Size* and *y3* and proportions for y_1 and y_2 due to all ownerships in the sample having the same point counts (i.e., $n_s = 1$). These findings would not be identical if the point counts were different, as can be the case in the NWOS, or if estimates were made across multiple states with different sampling intensities.

While the use of sample weights is relatively straightforward and required to get unbiased estimates for univariate population level estimates, the procedures for multivariate analyses (e.g., regression) are less clear and there is some disagreement within the statistical community as to the best approach. While there have been techniques developed to incorporating sample weights (Lumley and Scott 2017), there are fundamental questions related to the statistical assumptions that are still unresolved and unweighted approaches are often recommended (Winship and Radbill 1994). Given the conclusions of Winship and Radbill (1994) and others, the complex sample design of the NWOS, and the fact that area of forest land owned is part of the NWOS weights (Eq. 1) and this variable is commonly a predictor or independent

Postprint

variable in models, an unweighted approach to modeling using the NWOS data is recommended.

Effects of Small Ownerships on Estimates

Due to inclusion probabilities that proportional to size of holdings, ownerships with small holdings can have very large weights. These extreme values have big impacts on the estimates and their associated variances. The sampling errors associated with estimates of number of ownerships is substantially reduced if the domain of interest excludes ownerships with smaller acreage (e.g., examination of family forest ownerships with 10+ acres) as was done with the primary results from the 2013 and 2018 iterations of the NWOS (Butler et al. In review, 2016).

Nonresponse Bias

The detection and mitigation of nonresponse bias, in terms of both unit and item nonresponse, is an important topic for all surveys (Groves et al. 2002). Potential methods for testing for unit nonresponse bias include early/late responses, mail/phone responses, and auxiliary data (e.g., parcel size). Potential methods for correcting for unit nonresponse bias include post-stratification, response propensity modeling, generalized regression, and raking (Valliant et al. 2013). The missingness pattern for item nonresponse needs to be assessed and then appropriate techniques can be used to mitigate it, imputation (Rubin 2004) is the most common mitigation method. For the NWOS, response propensity score modeling (Brick 2013) is used to adjust the weights for unit nonresponse biases and multiple imputation by chained equations (van Buuren 2018) is used to address item nonresponse (Butler et al. In review).

R package

While the primary purpose of this report was to provide a theoretical justification and empirical

Postprint

validation of the overall weighting estimation approach, a secondary objective was to develop the computer code for implementing it. The R package created to implemented the procedures is freely available through GitHub (https://github.com/familyforestresearchcenter/nwos).

CONCLUSIONS

This report outlines a weighting method for producing estimates for the Forest Service's National Woodland Owner Survey. This methodology has advantages over previous approaches in that it is more transparent and uses a more robust variance estimation method. Comparing the values generated by the estimators to known population values provides evidence that the estimators are robust. Although the results presented here are specific for the NWOS, the basic principles are applicable to other surveys that have similar sample designs.

Acknowledgments

Thanks go to Paul Patterson, Stephanie Snyder, and John Stanovick for helpful comments provided on earlier versions of this report.

Postprint

LITERATURE CITED

Bååth, R. 2014. **Bayesian first aid: A package that implements Bayesian alternatives to the classical *. test functions in R**. Proceedings of UseR. Available at http://www.sumsar.net/papers/baath_user14_abstract.pdf (accessed on May 11, 2020).

Bechtold, W.A.; Patterson, P.L., eds. 2005. **The enhanced Forest Inventory and Analysis program-national sampling design and estimation procedures**. Gen. Tech. Rep. SRS-80. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station. 85 p. https://doi.org/10.2737/srs-gtr-80.

Brick, J.M. 2013. **Unit nonresponse and weighting adjustments: A critical review**. Journal of Official Statistics. 29(3): 329–353. https://doi.org/10.2478/jos-2013-0026.

Butler, B.J.; Butler, S.M.; Caputo, J.; Dias, J.; Robillard, A.; Sass, E.M. In review. **Family forest ownerships of the United States, 2018: Results from the USDA Forest Service, National Woodland Owner Survey**. Madison, WI: USDA Forest Service, Northern Research Station.

Butler, B.J.; Hewes, J.H.; Dickinson, B.; Andrejczyk, K.; Butler, S.M.; Markowski-Lindsay, M. 2016. **USDA Forest Service National Woodland Owner Survey: National, regional, and state statistics for family forest and woodland ownerships with 10+ acres, 2011-2013**. Res. Bull. NRS-99. Newtown Square, PA: U.S. Department of Agriculture, Forest Service, Northern Research Station. 39 p. http://dx.doi.org/10.2737/NRS-RB-99.

van Buuren, S. 2018. **Flexible imputation of missing data**. Boca Raton, FL: CRC Press. 415 p. ISBN: 978-1-138-58831-8.

Cohen, J. 1988. **Statistical power analysis for the behavioral sciences**. 2nd ed. New York, NY: Academic Press. 567 p. ISBN: 978-0-8058-0283-2.

Dickinson, B.J.; Butler, B.J. 2013. **Methods for estimating private forest ownership statistics: Revised methods for the USDA Forest Service's National Woodland Owner Survey**. Journal of Forestry. 111(5): 319–325. https://doi.org/10.5849/jof.12-088.

Efron, B.; Tibshirani, R. 1986. **Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy**. Statistical Science. 1(1): 54–75. https://doi.org/10.1214/ss/1177013815.

Groves, R.M.; Dillman, D.A.; Eltinge, J.L.; Little, R.J.A., eds. 2002. **Survey nonresponse**. New York, NY: Wiley. 500 p. ISBN: 0-471-39627-3.

Kruschke, J.K. 2011. **Doing Bayesian data analysis: A tutorial with R and BUGS**. San Diego, CA: Elsevier Academic Press. 653 p. ISBN: 978-0-12-381485-2.

Levenshtein, V.I. 1966. **Binary codes capable of correcting deletions, insertions, and reversals**. Soviet Physics Doklady. 10(8): 707–710.

Lohr, S.L. 1999. **Sampling: Design and analysis**. Pacific Grove, CA: Duxbury Press. 494 p. ISBN: 978-0-534-35361-2.

Lumley, T.; Scott, A. 2017. **Fitting regression models to survey data**. Statistical Science. 32(2): 265–278. https://doi.org/10.1214/16-STS605.

Metcalf, A.L.; Finley, J.C.; Luloff, A.E.; Shumway, D.; Stedman, R.C. 2012. **Private forest landowners: Estimating population parameters**. Journal of Forestry. 110(7): 362–370. https://doi.org/10.5849/jof.11-089.

O'Connell, B.M.; Conkling, B.L.; Wilson, A.M.; Burrill, E.A.; Turner, J.A.; Pugh, S.A.; Christiansen, G.; Ridley, T.; Menlove, J. 2016. **The Forest Inventory and Analysis Database: Database description and user guide version 6.1 for phase 2**. Washington, DC: U.S. Department of Agriculture, Forest Service. 892 p. www.fia.fs.fed.us/library/database-documentation/.

R Core Team. 2019. **R: A language and environment for statistical computing**. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org.

Rubin, D.B. 2004. **Multiple imputation for nonresponse in surveys**. Hoboken, NJ: Wiley-Interscience. 287 p. ISBN: 978-0-471-65574-9.

U.S. Geological Survey (USGS). 2014. **NLCD2011 USFS percent tree canopy (analytical version)**. Sioux Falls, SD: USGS/EROS. https://www.mrlc.gov/nlcd11_data.php (accessed on July 18, 2017).

USDA Forest Service. 2016. **Forest Inventory and Analysis glossary**. Newtown Square, PA: U.S. Department of Agriculture, Forest Service, Northern Research Station. www.nrs.fs.fed.us/fia/data-tools/state-reports/glossary/default.asp (accessed on December 4, 2018).

USDA Forest Service. 2019. **Forest Inventory and Analysis national core field guide:**

Postprint

Volume I - field data collection procedures for phase 2 plots, version 9.0. Washington, DC: U.S. Department of Agriculture, Forest Service. https://www.fia.fs.fed.us/library/field-guides-methods-proc/ (accessed on May 12, 2020).

Valliant, R.; Dever, J.A.; Kreuter, F. 2013. **Practical tools for designing and weighting survey samples**. New York, NY: Springer. 670 p. ISBN: 978-1-4614-6448-8.

Winship, C.; Radbill, L. 1994. **Sampling weights and regression analysis**. Sociological Methods & Research. 23(2): 230–257. https://doi.org/10.1177/0049124194023002004.

Wisconsin State Cartographer's Office; Wisconsin Land Information Program. 2016. **V2 statewide parcel data (v2.0.5)**. Available at www.sco.wisc.edu/images/stories/publications/V2/data/ (accessed on July 11, 2017).

Postprint