CHAPTER 7

Sampling Designs for SAR-Assisted Forest Biomass Surveys

Hans-Erik Andersen, Research Forester/Vegetation Monitoring and Remote Sensing (VMaRS) Team Leader, USDA Forest Service

ABSTRACT

Sampling designs that efficiently integrate information from plot data and a variety of remote sensing systems, including spaceborne SAR, are required to support cost-effective monitoring of forest biomass/carbon at regional and global scales. In particular, sampling designs and statistical modelling/estimation frameworks are desired that provide sound, statistically-rigorous assessments of uncertainty and make efficient use of expensive field plot data and more extensive use of less-expensive remotely-sensed information. In addition, these designs should also provide the flexibility to accommodate a variety of field plot configurations and remote sensing data acquisition strategies/resolutions. This chapter discusses several important considerations in quantifying uncertainty in multi-level sampling designs, including both the model-assisted and model-based inferential frameworks, and use simulation to illustrate the statistical properties of the estimators associated with these designs, with the goal of informing the design of forest inventory and monitoring programs in remote regions.

7.1 Background

International efforts to reduce carbon emissions from the forest sector have created increased demands on the capabilities of national and regional forest monitoring systems to provide timely, accurate information on forest carbon stocks and changes due to deforestation and degradation (GFOI 2016). At the same time, it is recognized that traditional forest inventory sampling designs, which typically rely heavily on large numbers of field plot measurements distributed over a region, are difficult or impossible to implement in many remote, underdeveloped regions of the world (e.g., high latitudes, tropics) due to logistical complexity and/or high costs. For this reason, there is increasing interest in the development of new sampling designs for the monitoring of forest biomass/carbon that can efficiently utilize the low-cost mapped information on forest structure (biomass/carbon), at the global scale, that is increasingly available with the recent and future launches of several satellite SAR missions, such as Advanced Land Observation Satellite (ALOS) Phased Array type L-band Synthetic Aperture Radar (PALSAR) (Hoekman et al. 2010) and ALOS-2 PALSAR-2 (JAXA 2014). For this reason, sampling designs and statistical modelling/estimation frameworks are increasingly sought with the following properties:

(1) Provide the basis for sound, statistically-rigorous assessment of uncertainty (e.g., Gregoire et al. 2016)

(2) Use a fewer number of expensive field plots and more extensive, efficient use of less-expensive remotely-sensed information (including airborne light detection and radar (lidar), satellite-based L-band SAR)

(3) Provide flexibility to accommodate a variety of field plot configurations and remote sensing data acquisition strategies/resolutions

This chapter discusses several important considerations in the assessment of uncertainty in forest biomass surveys and how these considerations should factor into the design and implementation of a sampling design for biomass inventory and monitoring using L-band spaceborne SAR in remote regions.

7.1.1 SOURCES OF UNCERTAINTY IN A CARBON INVENTORY AND MONITORING PROGRAM

There are three primary sources of variability in the context of a forest carbon inventory and monitoring system: (1) measurement error, (2) modelling error, and (3) sampling error. In making the choice of a field measurement protocol, sampling design, and inferential framework, all three types of errors should be considered. In the context of carbon monitoring programs, measurement error—or discrepancy between a recorded field measurement and the expected value of the measurement as defined by documented protocol—is often introduced through inadequate training or lack of adherence to protocol. In practice, measurement error is usually assessed and mitigated (if possible) through quality assurance/quality control (QA/QC) procedures (that can be quite costly to implement), and otherwise is assumed to be minimal in comparison to the measurement itself (Gregoire & Valentine 2008).

In the context of forest carbon monitoring using remote sensing, modelling error is introduced in two ways: (1) the use of allometric models to estimate tree-level biomass/carbon using various tree measurements (diameter at breast height, height, etc.), and (2) the use of models relating the remotely-sensed measurement (SAR backscatter, air photo-derived canopy height and cover, etc.) to the plot-level biomass/carbon.

7.1.2 ALLOMETRIC MODELS FOR BIOMASS

Given the difficulty of measuring aboveground tree biomass directly, virtually all carbon monitoring programs rely upon allometric models to convert tree measurements obtained in a forest inventory (e.g., height, stem diameter) to aboveground biomass (or carbon) estimates. Due to the relatively small samples used to develop these models, and the wide range of variability in wood density and height/diameter relationships across the geographic range of trees, it is widely acknowledged that lack-of-fit in the allometric models used to estimate biomass can contribute significantly to the true overall error budget for carbon monitoring—although national forest inventory programs often do not explicitly account for this error in official reports. While several recent efforts have made progress in improving the quality of allometric models used in national- or regional-scale carbon monitoring programs (Chojnacky et al. 2014, Chave et al. 2014) and the emergence of new technologies, such as terrestrial laser scanning (Calders et al. 2015) hold promise for improving the efficiency of field measurements, uncertainty due to allometric modelling remains the most difficult source of error to account for in large scale carbon monitoring programs (Duncanson et al. 2017).

7.1.3 ESTIMATION OF BIOMASS USING SAR

Due to its sensitivity to forest biomass, global coverage, and capability to penetrate cloud cover, L-band satellite radar has been used extensively as an auxiliary source of data to support forest monitoring programs across a range of biomes (Ryan et al. 2011, Hoekman et al. 2010). L-band dual-polarization (HH, HV) backscatter has been shown to be well-correlated with forest biomass up to approximately 150 Mg/ha, lending it particular utility in assessing forest biomass levels in low-biomass forests characteristic of high-latitude boreal forest biome as well as semi-arid, savanna forests of the tropics (Atwood et al. 2014, Tanase et al. 2014). However, it has been noted that generalizing relationships between L-band radar backscatter and biomass across forest types is inadvisable since radar backscatter from a forest scene is a function of numerous forest structural characteristics (stem density, height, stem diameter), as well as other scene properties (soil moisture, slope, etc.) with varying correlation to tree biomass (Woodhouse et al. 2012). Although the L-band backscatter signal saturates at higher biomass levels (>150 Mg/ha), limiting its usefulness as a stand-alone correlate for biomass in high-biomass forests, there is evidence to suggest that including additional forest structure information, perhaps obtained from lidar or repeat-pass interferometry (Treuhaft and Siqueira 2000), can help to decouple the complex relationships between backscatter and forest structural attributes that can obscure the biomass-backscatter signal at higher biomass levels (Joshi et al. 2017).

Once the measurement protocols and modelling frameworks have been established in a forest carbon monitoring system, the next step is determining the proper sampling design to obtain the required precision for carbon estimators within the limitations of the available resources. Although it is typical to only be able to directly measure trees (and estimate biomass via allometry) on a very limited portion of the landscape—leading the third source of variability in carbon estimates, sampling error—the use of remote sensing provides a means of obtaining a much more comprehensive picture of forest structure across an area of interest. This chapter explores sampling approaches that utilize a combination of field data and auxiliary information—including wall-to-wall satellite SAR imagery and sampled high-resolution (e.g., lidar) in multilevel inventory designs—to estimate support forest monitoring programs.

7.2 Use of Remote Sensing to Support Carbon Surveys

7.2.1 MODES OF INFERENCE

Traditionally, forest inventory and monitoring programs have been based on the principles of *design-based* inference, where field plots were distributed as a probability sample, and each unit in the population of interest has a positive probability of being selected in a sample. In design-based sampling, the population is considered fixed, and all uncertainty in the estimation of a population parameter (total biomass, volume, etc.) is due to variability between randomly drawn samples from the population. Depending on the objectives of the study or inventory, probabilities of selection can vary across the population to reduce costs or increase the statistical precision of the estimates. For example, in stratified sampling, the units of the population can be grouped into homogeneous strata and the population-level estimate is calculated as a weighted average of the stratum-level estimates with the weights based on stratum sizes. *Model-assisted* inference is a means of using lower-cost auxiliary data (e.g., maps, imagery, photo plots) and a model describing the relationship between auxiliary measurements and inventory parameters to improve precision of estimates within the design-based inferential paradigm. In model-assisted approaches, data at every level are still collected as probability samples, but the number of field plots required to achieve a given level of precision can be reduced significantly (compared to designs using only field plots) if there is a strong correlation between auxiliary data and inventory parameters.

In contrast, *model-based* inference is usually based on the so-called *superpopulation* model, where each value from an element in the population is considered a realization of a random variable with a specific probability distribution. Therefore, all population-level values (e.g., total or mean biomass, etc.) are also considered random variables. In the model-based inferential paradigm, the uncertainty in the estimation of a population parameter is due to randomness in the values observed for each population element. Because the validity of inferences in the model-based paradigm are not dependent upon a random (probability) sample, it can be applied in situations where collecting a sufficiently large probability sample of field plots is either too expensive or logistically difficult, such as estimation within small areas or remote regions lacking transportation infrastructure.

Due to very different underlying assumptions, the results from model-based and model-assisted approaches are difficult to compare directly. The advantage of model-based approaches is that there is no requirement that the field plots be a probability sample, while this is a requirement of model-assisted approaches. However, inferences in the model-based context are conditional on the model and may produce severely biased estimators in cases where the model is developed using an unrepresentative sample. In contrast, design-based (including model-assisted) estimators can, from a practical standpoint, be considered unbiased (for reasonably large sample sizes) regardless of the model that is used. Obviously, in a regional or national forest inventory and monitoring context—where estimates are often used to

INTERFERENCE TYPE	DESCRIPTION	STRENGTHS	WEAKNESSES	APPLICATION
Design-based	All data collected as a probability sample, Population is considered fixed; uncertainty is due to variability between randomly-drawn samples	Simple, well- documented designs and formulae for point and variance estimators; design-unbiased estimation; reliable confidence intervals	Requirement of probability sample may be logistically infeasible to cost-prohibitive in some cases; Less efficient if strongly-correlated auxiliary data is available	National forest inventories, National Greenhouse Gas inventories
Model-assisted	Uses lower-cost auxiliary data and models to improve precision of estimates within the design-based paradigm; data at every level are still collected as probability samples	Increased efficiency (fewer field plots for given level of precision) and lower cost if there is a strong correlation between auxiliary data and inventory parameters	Probability samples required at every level of the design; Form of estimators potentially very complex; Only design-unbiased for large samples; confidence intervals less reliable for small samples	REDD+ applications, NFI in remote regions
Model-based	Population values and parameters are random variables. Uncertainty due to randomness in the values observed for each population element.	Probability sampling not required; potentially much less expensive to implement than design-based approaches	Not design-unbiased; Estimators based on models developed with unrepresentative samples can be severely biased	Small area estimation; Tactical forest management; Inventory over large, remote regions lacking transportation infrastructure

Table 7.1 *Description of strengths, weaknesses and main applications of the models addressed in this chapter.*

support forest policy decisions and fulfill Reducing Emissions from Deforestation and forest Degradation (REDD+) and Net Green House Gas (NGHG) monitoring and reporting requirements—the quality of unbiased data is critical and the model-assisted approach may be more appropriate. Model-based approaches may be more appropriate for assessment of remote, or small, inadequately sampled areas, or to support tactical-level forest management decisions.

7.3 Exercise 1: Simulating an Artificial Population

Simulation can be a useful approach to gain insight into the statistical properties of various survey estimators, especially in the case of somewhat complex, multi-level sampling designs (Ene et al. 2016, Saarela et al. 2017). Here, simulation implemented in the R statistical software package is used to demonstrate the implementation of several SAR-assisted, multi-level sampling designs. Proficiency in R programming is not required to carry out the exercises, since the scripts can be run by simply copying and pasting the code at the R command line.

When generating a simulated population, it is desirable to include realistic correlations between the response variable (e.g., biomass) and the predictor variables used in the inventory. While a multivariate normal distribution can be used to model correlation between several variables, it may also be important for the purposes of gaining insight into the properties of the point and variance estimators, as well as implications for sample size and modelling effort, that these variables have more realistic marginal distributions (gamma, exponential, etc.). A copula function is a useful mathematical tool to simulate a population with specified multivariate correlation structure and marginal distributions (Ene et al. 2012, Nelsen 2006). While an in-depth discussion of copula models is outside the scope of this chapter, they essentially allow for expressing multivariate distributions in terms of their corresponding univariate marginal distributions and a copula function. In this exercise, a copula function is used to simulate a large population where each element has a value for forest/nonforest classification, biomass (Mg/ha), a lidar-based measurement (function of lidar-derived height and cover), and a SAR-based measurement (function of HH and HV backscatter). Realistic marginal distributions and correlation structure between remote sensing measurements and field-based biomass were developed based on an analysis of airborne lidar, SAR, and field biomass data from a site in interior Alaska (Andersen et al. 2013). In order to introduce realistic spatial heterogeneity across the simulated area, a binary random field (150 × 150 grid cells) generated an image with a realistic simulated spatial distribution of “forest” and “nonforest” areas. The grid cells within the simulated forest/nonforest image were then populated with elements from the simulated population generated using the copula function. In this way, each element in the image had a value for forest/nonforest, biomass, lidar, and SAR, and the simulated population had realistic marginal distributions, correlation structure, and long-range spatial heterogeneity (**Figs. 7.1** and **7.2**).

7.3.1 DESIGN-BASED ESTIMATION

7.3.1.1 Simple Random Sampling

Simple random sampling (SRS) represents the most fundamental type of design-based sampling and is often used as the basis of comparison for more complex sampling designs. Given a probability sample of elements of size *n* from a population of size *N*, where a forest attribute of interest (Y_i) is obtained for each element *i*, the SRS estimator of the population mean is given by the sample mean:

$$\hat{\mu}_{SRS} = \hat{\mu}_{ma,1} = \overline{Y} = \frac{1}{n}\sum\nolimits_{j=1}^{n} Y_i \qquad (7.1)$$

and the variance estimator is given by

$$\hat{V}\left(\hat{\mu}_{SRS}\right) = \frac{1}{n\left(n-1\right)}\sum\nolimits_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 \qquad (7.2)$$

7.3.3 POST-STRATIFICATION

The precision of an SRS estimator can be increased at the estimation stage if the population can be stratified in such a way that plots with similar values for an inventory parameter are grouped together in the same class or stratum, a technique called post-stratification. In post-stratification, the estimator of the population mean is given by

$$\hat{\mu}_{PS} = \sum\nolimits_{h} W_h \overline{y}_h \qquad (7.3)$$

with a variance estimator:

$$\hat{V}\left(\hat{\mu}_{PS}\right) = \frac{1}{n}(\sum\nolimits_{h} W_h n_h V(\overline{y}_h) + \frac{1}{n^2}\sum\nolimits_{h}\left(1 - W_h\right) n_h V(\overline{y}_h) \;, (7.4)$$

where W_h is the proportion of the population in stratum *h* (i.e., $W_h = \frac{N_h}{N}$) and $V(\overline{y}_h)$) is the variance of the mean of plots in stratum *h*.

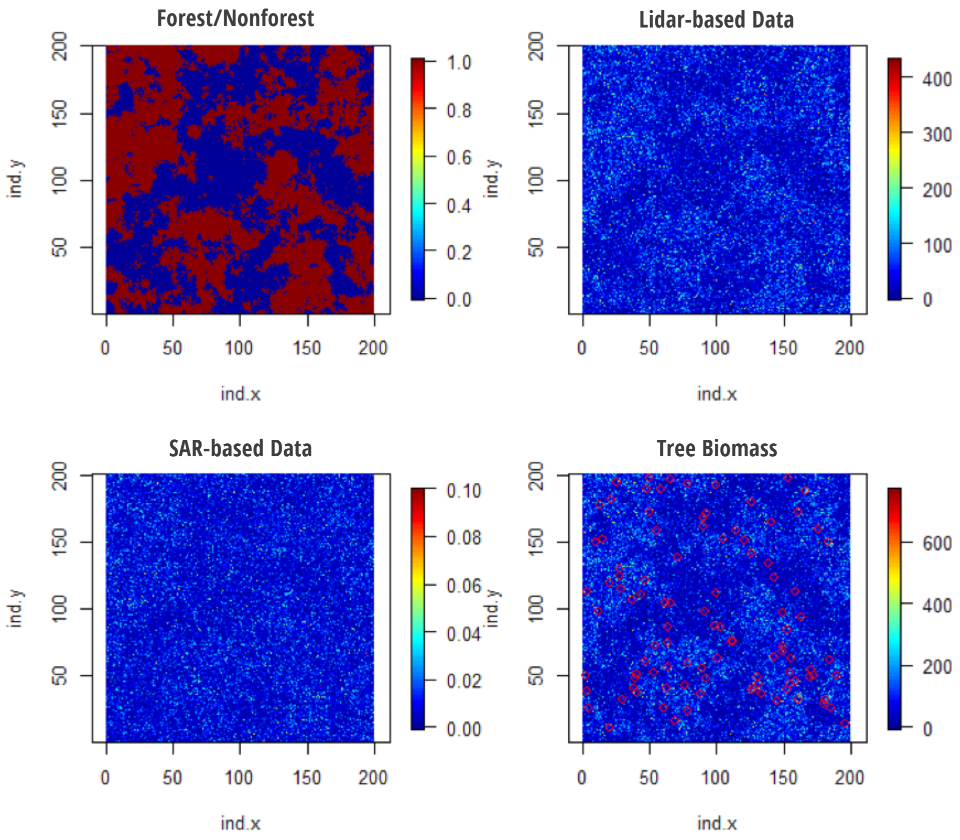

Figure 7.1 Simulated population with biomass, forest/nonforest, lidar-based measurements, and SAR-based measurements. Simulated plots (red) are shown overlaid on tree biomass image.

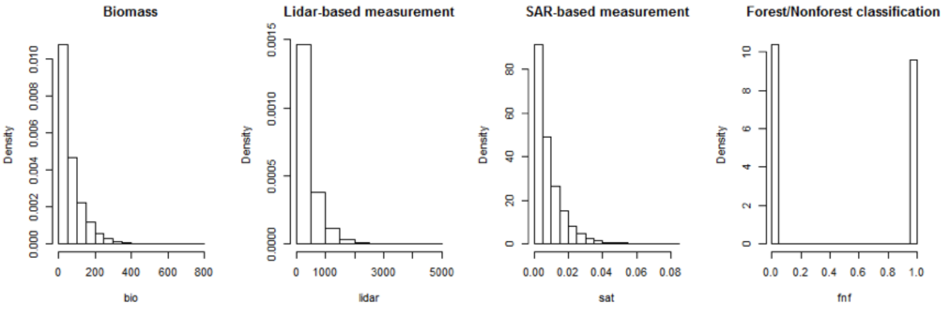

Figure 7.2 Simulated marginal distributions of biomass, lidar-based measurements, SAR-based measurements, and forest/nonforest classification. Exponential distributions used to model biomass, lidar, and SAR variables; Bernoulli distribution used to model forest/non-forest class.

	Biomass	Lidar-based	SAR-based	Forest/Nonforest
Biomass	1.00	0.88	0.66	0.36
Lidar-based	0.88	1.00	0.56	0.30
SAR-based	0.66	0.56	1.00	0.15
Forest/Nonforest	0.36	0.30	0.15	1.00

Table 7.2 Correlation matrix for a simulated population

7.4 Exercise 2: Properties of Estimators via Simulation

The statistical properties of the various estimators can be assessed using the simulated population developed previously. At each iteration, a simple random sample of n elements is drawn from the population, and the point estimator μ and the variance estimator $V(\mu)$ are calculated.

Given that we know the actual population mean μ, we can then calculate the mean percent bias of the point estimator

$$\left(\overline{\hat{\mu}}_{\text{iterations}} - \mu\right) \Big/ \mu \times 100\% \ , \qquad (7.5)$$

the relative standard error of the point estimator

$$SD\left(\hat{\mu}_{\text{iterations}}\right) \Big/ \overline{\hat{\mu}}_{\text{iterations}} \times 100\% \ , \qquad (7.6)$$

and the empirical coverage probability of the 95% confidence interval for the point estimator:

$$Prob\left(\hat{\mu} - t_{\left(0.025, n_2 - 2p\right)} \sqrt{\hat{V}\left(\hat{\mu}\right)} < \mu < \hat{\mu} + t_{\left(0.975, n_2 - 2p\right)} \sqrt{\hat{V}\left(\hat{\mu}\right)}\right) \times 100\% \ . \qquad (7.7)$$

The empirical coverage probability provides an indication of how reliable (i.e., unbiased) the variance estimator is for a parameter. An empirical coverage probability (95% CP) near 95% is an indicator that the 95% confidence intervals (CIs) calculated using this estimator are reliable. Empirical coverage probabilities of less than 95% indicate that the calculated 95% CIs are giving a falsely precise estimate of uncertainty, while coverage probabilities greater than 95% indicate that the 95% CIs obtained from this estimator are overly conservative.

When the SRS estimator is assessed via simulation, the results indicate the increase in precision due to increasing sample size, as well as the improvement in 95% coverage probability with increasing sample size (it is well-documented that variance estimators can be biased for small samples drawn from highly-skewed populations).

When using a forest/non-forest layer for post-stratification of the SRS sample, the precision is increased a small amount.

7.4.1 REGRESSION ESTIMATORS

Model-assisted estimators essentially provide a means to use models based on auxiliary data (e.g., remote sensing) to improve inferences within the

		Bias (%)	SE (%)	95% CP
n	25	0.4%	20.1%	92.1%
	50	-0.3%	15.1%	92.4%
	100	-0.3%	10.6%	93.2%
	200	0.4%	7.0%	95.2%

Table 7.3 *Statistical properties (bias, relative standard error, and 95% coverage probability) for SRS estimator (based on 1,000 iterations).*

		Bias (%)	SE (%)	95% CP
n	25	0.0%	19.2%	91.5%
	50	-0.5%	13.3%	94.7%
	100	0.2%	9.8%	93.1%
	200	0.3%	6.7%	94.2%

Table 7.4 *Statistical properties (bias, relative standard error, and 95% coverage probability) for post-stratified estimator (based on 1,000 iterations).*

design-based inferential framework (McRoberts et al. 2014). In other words, random (probability) sampling at all levels in the design is the basis for all inference. Model-assisted regression estimators are based on a model of the relationship between the forest attribute of interest (e.g., biomass/carbon), *Y*, and a vector *X*, of auxiliary variables, formulated as,

$$Y_i = f(X_i;\beta) + \varepsilon_i \ , \qquad (7.8)$$

where $f(X_i;\beta)$ expresses the mean of *Y* given observation of *X*, β are the parameters to be estimated, and ε_i is a random residual term. In practice, the entire population is not observed, but the parameters of the regression relationship $\hat{\beta}$ based on a sample of the population is estimated. Then this regression model and observed vector of auxiliary variables are used to predict the inventory attribute for a particular unit of the population:

$$\hat{Y}_i = f\left(X_i;\hat{\beta}\right) \ . \qquad (7.9)$$

A regression estimator for the population mean, $\hat{\mu}_{ma,1}$ when a single source of wall-to-wall auxiliary information (e.g., satellite SAR or spectral imagery) is given by the following expression:

$$\hat{\mu}_{ma,1} = \frac{1}{N}\sum\nolimits_{j=1}^{N}\hat{Y}_i + \frac{1}{n}\sum\nolimits_{i=1}^{n}\left(Y_i - \hat{Y}_i\right) \ , \qquad (7.10)$$

where *N* is the population size, *n* is the sample size, and $\hat{Y}_i$ is obtained from Eq. 7.9 (Särndal et al. 1992). The first right-hand term in this equation is the sum of the model predictions for the entire population, and the second right-hand term is a correction term which, when added to the first term, compensates for model bias. The regression estimator can be expressed in different forms, but the above formulation is the easiest form to interpret in our context, since the model predictions $\hat{Y}_i$ are based on remotely sensed imagery or measurements and the second term is the mean of the residuals observed at the field plots. For *n* much smaller than *N*, an approximately unbiased estimator of the corresponding variance is formulated as:

$$\hat{V}\left(\hat{\mu}_{ma,1}\right) = \frac{1}{n\left(n-1\right)}\sum\nolimits_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \ . \qquad (7.11)$$

The advantage of the regression estimator over the *SRS* estimator is that the variance estimator is based on residuals, $Y_i - \hat{Y}_i$ rather than differences, $Y_i - \overline{Y}$, between observations and their mean. Therefore, it can be seen that the degree to which the relationship with *X* explains variability in *Y* will determine the gain in precision from using the regression estimator as opposed to the *SRS* estimator. It should be noted that post-stratification—where population-level strata proportions are used to improve precision of an estimate in the estimation (rather than the design) stage—is a special case of regression estimation where the predictors are categorical variables (for example, satellite image-based landcover classes).

7.5 Exercise 3: Simulation-Based Assessment of Model-Assisted Estimator With a Single Source Of Auxiliary Data

The statistical properties of the model-assisted estimator with one source of auxiliary data (assumed to be collected wall-to-wall, such as SAR imagery) and various sample sizes for field plots (**Table 7.5**). It is evident from these results that there is a small reduction in the standard error (in comparison to the SRS estimator) through including a single auxiliary that is moderately correlated with biomass (Mandallaz et al. 2013). It is noted that the sampling distribution of this variance estimator is bell-shaped, but with heavier tails than a normal distribution. Therefore, this approach was followed and confidence intervals calculated using a student's t-distribution with $n_2 - 2p$ degrees of freedom.

		Bias (%)	SE (%)	95% CP
n_2	25	0.3%	16.1%	90.6%
	50	0.4%	11.5%	92.7%
	100	-0.1%	7.8%	93.5%
	200	0.1%	5.5%	94.4%

Table 7.5 *Statistical properties of a model-assisted regression estimator with single-auxiliary (bias, relative standard error, 95% coverage probability) for four different phase-1 sample sizes (250, 500, 1,000, 2,000) and four different phase-2 sample sizes (25, 50, 100, 200), based on 1,000 iterations.*

7.5.1 MODEL-BASED APPROACHES

Following McRoberts et al. (2010) and Saarela et al. (2016) if *Y* is the random variable (Above Ground Biomass (AGB)) with a mean μ and standard deviation σ, the observed AGB value at the i^{th} pixel (y_i) can be represented as

$$y_i = \mu_i + \epsilon_i \ , \qquad (7.12)$$

where $\epsilon_i \sim N(0,\sigma^2)$. The mean AGB at the i^{th} pixel is then given by

$$\mu_i = f(\mathbf{X}_i;\beta) \ , \qquad (7.13)$$

which is estimated by

$$\hat{\mu}_i = f\left(\mathbf{X_i};\hat{\beta}\right) \ , \qquad (7.14)$$

where X_i is the lidar-based predictor variable at the i^{th} pixel, and $\hat{\boldsymbol{\beta}}$ is the vector of *p* predicted regression coefficients. The model-based estimate of mean AGB over the entire areas is:

$$\mu_U = \iota'_{\mathbf{U}}\mathbf{X_U}\hat{\beta} \ , \qquad (7.15)$$

where $\iota'_{\mathbf{U}}$ is an N-length column vector where every element equals 1/N, X_U is an N × (p + 1) matrix of satellite auxiliary variables available for each element in the population U. The variance of the model-based mean

AGB estimate is given by

$$\mathbf{V}\left(\widehat{\mu_U}\right) = \iota_{\mathbf{U}}'\mathbf{X}_{\mathbf{U}}\mathbf{V}_{\hat{\beta}}\mathbf{X}_{\mathbf{U}}'\iota_{\mathbf{U}} \quad , \tag{7.16}$$

where $\mathbf{V}_{\hat{\beta}}$ is the variance-covariance matrix for the regression model parameter estimates $\hat{\boldsymbol{\beta}}$. For example, in the case of $p = 2$, $\mathbf{V}_{\hat{\beta}}$ is given by:

$$\begin{bmatrix} \hat{V}\left(\hat{\beta}_0\right) & \widehat{Cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) \\ \widehat{Cov}\left(\hat{\beta}_1, \hat{\beta}_0\right) & \hat{V}\left(\hat{\beta}_1\right) \end{bmatrix} . \tag{7.17}$$

It should be noted that when using internal models developed from an SRS sample at all levels of the sampling design, the model-based estimator will yield virtually the same point estimate and variance estimator as the model-assisted estimator. However, as noted above, the assumptions behind these estimators differ and provide more flexibility in the application of the model-based estimator (e.g., application to nonprobability samples). Care must be taken to ensure that models are based on a representative (if not random) sample to reduce bias in the point and variance estimators (see Exercise 4).

7.6 Exercise 4: Simulation-Based Assessment of Model-Based Estimator with One Source Of Auxiliary Data

In order to illustrate the perils of an incorrectly specified model in the context of model-based estimation, in this exercise, the model is developed from a sample selected only from the forested plots within the population, and then used to estimate biomass—using both model-assisted and model-based estimators—over the entire population. **Table 7.6** indicates that use of an incorrectly specific model (based on an unrepresentative sample) can lead to significant bias in the point estimates (28% in this case), while the model-assisted estimator remains virtually unbiased (0.5%).

		Bias (%)	SE (%)	95% CP
n_2 = 50	M-A, 1-aux	0.5%	11.3%	96.8%
	M-B, 1-aux	28.2%	9.2%	39.1%

Table 7.6 Statistical properties of model-assisted and model-based regression estimators with single-auxiliary (bias, relative standard error, 95% coverage probability) using a mid-range second-phase sample size of 50 (based on 1,000 iterations).

7.6.1 SAMPLING DESIGNS WITH MULTIPLE SOURCES OF AUXILIARY DATA

In some cases, two types of auxiliary information are available, where one (e.g., satellite SAR imagery) is collected wall-to-wall and another type of (more expensive and higher resolution) remotely-sensed data is collected in a sampling mode. For example, multi-level sampling design may consist of: (1) a large sample of relatively inexpensive photo-interpreted plots distributed over an area of interest, with (2) detailed, relatively expensive, field measurements of the attribute of interest (e.g., tree biomass/carbon) collected on a subsample of these photo plots, and (3) free, or very inexpensive, satellite image data (SAR) available over the entire area. Depending on application and how the data were collected, this type of multi-level sampling design can be approached from a model-based or model-assisted inferential standpoint.

7.6.2 MODEL-ASSISTED

Following Mandallaz et al. (2013), a model-assisted estimator of mean aboveground tree biomass can be developed using field plot data and two sources of auxiliary data in the following manner: as in the previous example, (1) a vector X_U of remote sensing-derived variables that are known for all N elements in the population (U), and (2) a vector X_1 of remote sensing-derived variables that are known only for the elements in the first phase sample of n_1 units, and the inventory attribute of interest, Y_2, is only measured on a relatively small second-phase subsample of n_2 photo plots. As a specific example, the X_{Ui} variables may represent satellite image data (e.g., SAR HV/HH backscatter, Landsat tasselled cap bands) available wall-to-wall over the entire study area, and the X_1 variables represent photo plot measurements (average tree height, cover, forest type) that are only available at a sample of locations distributed over the area of interest. Regression analysis is used to develop a linear model for predicting biomass from photo-based measurements:

$$Y_{2i} = f(X_{1i};\beta_1) + \varepsilon_{1i} \ , \tag{7.18}$$

while satellite-derived predictor variables are used to predict biomass using satellite-based measurements:

$$Y_{2i} = f(X_{Ui};\beta_U) + \varepsilon_{Ui} \ . \tag{7.19}$$

Again, following Mandallaz et al. (2013), this design yields the following estimator of mean biomass for the study area:

$$\hat{\mu}_{ma,2} = \frac{1}{N}\sum\nolimits_{j=1}^{N}\hat{Y}_{Ui} + \frac{1}{n_1}\sum\nolimits_{i=1}^{n_1}\left(\hat{Y}_{1i} - \hat{Y}_{Ui}\right) + \frac{1}{n_2}\sum\nolimits_{k=1}^{n_2}\left(Y_k - \hat{Y}_{1k}\right) \tag{7.20}$$

		n_1											
		250			500			1000			2000		
		Bias (%)	SE (%)	95% CP	Bias (%)	SE (%)	95% CP	Bias (%)	SE (%)	95% CP	Bias (%)	SE (%)	95% CP
n_2	25	0.5%	11.0%	91.3%	0.0%	10.8%	89.0%	0.5%	10.8%	89.0%	0.4%	10.3%	89.6%
	50	-0.2%	8.3%	91.5%	-0.1%	7.6%	92.8%	0.5%	7.4%	92.0%	0.0%	7.7%	89.9%
	100	0.2%	6.5%	93.3%	0.0%	5.7%	94.3%	0.4%	5.3%	94.0%	0.0%	5.1%	93.5%
	200	0.0%	5.4%	93.7%	0.0%	4.4%	95.2%	0.1%	4.0%	94.4%	-0.1%	3.6%	95.2%

Table 7.7 Statistical properties of model-assisted estimator with two auxiliaries (bias, relative standard error, 95% coverage probability) for four different phase 1 sample sizes (250, 500, 1,000, 2,000) and four different phase 2 sample sizes (25, 50, 100, 200). Satellite image-derived measurements were assumed to be available for every unit in the population (based on 1,000 iterations).

		n_1											
		250			500			1000			2000		
		Bias (%)	SE (%)	95% CP	Bias (%)	SE (%)	95% CP	Bias (%)	SE (%)	95% CP	Bias (%)	SE (%)	95% CP
n_2 = 50	**M-A, 2-aux**	-0.4%	8.2%	95.6%	-0.2%	7.8%	96.0%	-0.1%	7.6%	95.6%	0.1%	7.3%	95.7%
	M-B, 2-aux	10.8%	7.8%	85.7%	11.2%	7.1%	82.9%	10.7%	7.1%	82.8%	11.5%	6.5%	81.1%

Table 7.8 *Statistical properties of model-assisted and model-based regression estimators with two auxiliaries (bias, relative standard error, 95% coverage probability) using a mid-range second-phase sample size of 50 (based on 1,000 iterations).*

with variance estimator:

$$\hat{V}\left(\hat{\mu}_{ma,2}\right)=\frac{1}{n_1}\frac{1}{n_2}\sum_{i=1}^{n_2}\left(y_{2i}-\hat{y}_{Ui}\right)^2+\frac{1}{n^2}\left(1-\frac{n_2}{n_1}\right)\sum_{i=1}^{n_2}\left(y_{2i}-\hat{y}_{1i}\right)^2 \quad (7.21)$$

$\hat{y}_{1k}=\mathbf{x}'_{1k}\hat{\mathbf{B}}_{1s}$ (for k Є U) are the satellite image-based predictions at each satellite image pixel (obtained via linear regression).

7.7 Exercise 5: Statistical Properties of Model-Assisted Estimators with Two Sources of Auxiliary Data

7.7.1 MODEL-BASED

A model-based approach to utilizing auxiliary data collected at multiple levels was developed by Saarela et al. (2016). As in the previous example of model-based estimator, the relationship between the inventory attribute Y, which is the random variable (AGB) with a mean μ and standard deviation σ, the observed mean AGB value at the ith pixel (y_i) can be represented as:

$$\mu_i = f(X_i;\beta) + \epsilon_i \ , \quad (7.22)$$

where ϵ_i~N(0,σ^2). The mean AGB at the ith pixel is given by

$$\hat{\mu}_i=f\left(\mathbf{X_i};\beta\right) \quad (7.23)$$

which is estimated by

$$\hat{\mu}_i=f\left(\mathbf{X_i};\hat{\beta}\right) \quad (7.24)$$

where X_i is the set of lidar-based predictor variables available for the second phase sample n_2 of the population and $\hat{\boldsymbol{\beta}}$ is the vector of p predicted regression coefficients. This linear model is used to estimate the mean ABG at every pixel in the first phase sample n_1:

$$\hat{\mu}_{1i}=f\left(\mathbf{X_{1i}};\hat{\beta}\right) \ . \quad (7.25)$$

In this hierarchical modelling framework, a second model is developed relating the satellite-based predictor variables Z_U available over the entire population to the $\hat{\mu}_1$ predictions available within the first phase sample:

$$\hat{\mu}_{1i}=f\left(\mathbf{Z_{1i}};\boldsymbol{\alpha}_1\right)+\omega_{1i} \quad (7.26)$$

where ω_{1i}~N(0,σ^2) and α_1 is the vector of model coefficients linking lidar-estimated AGB values and the satellite predictor variables estimated by $\hat{\boldsymbol{\alpha}}_1$. The model-based estimate of mean AGB over the entire area is

$$\widehat{\mu_U}=\boldsymbol{\iota}'_{\mathbf{U}}\mathbf{Z_U}\hat{\boldsymbol{\alpha}}_1 \ . \quad (7.27)$$

The variance of the model-based mean AGB estimate is given by:

$$\mathbf{V}\left(\widehat{\mu_U}\right)=\boldsymbol{\iota}'_{\mathbf{U}}\mathbf{Z_U}\mathbf{V}_{\widehat{\alpha_1}}\mathbf{Z}'_{\mathbf{U}}\boldsymbol{\iota}_{\mathbf{U}} \ , \quad (7.28)$$

where $\mathbf{V}_{\widehat{\alpha_1}}$ is the variance-covariance matrix for the regression model parameter estimates given by:

$$\mathbf{V}_{\hat{\alpha}}=\frac{\hat{\boldsymbol{\omega}}'_1\hat{\boldsymbol{\omega}}_1}{\mathbf{M}-\mathbf{q}-\mathbf{1}}\left(\mathbf{Z}'_1\mathbf{Z}_1\right)^{-1}+\left(\mathbf{Z}'_1\mathbf{Z}_1\right)^{-1}\mathbf{Z}'_1\left[\mathbf{X}_1\mathbf{V}_{\hat{\beta}}\mathbf{X}'_1\right]\mathbf{Z}_1\left(\mathbf{Z}'_1\mathbf{Z}_1\right)^{-1} \ , \quad (7.29)$$

where $\mathbf{V}_{\hat{\beta}}$ is the variance-covariance matrix for the regression model parameter estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\omega}}_1=\mathbf{X}_1\hat{\boldsymbol{\beta}}-\mathbf{Z}_1\hat{\boldsymbol{\alpha}}_1$ is an n_1 length vector of model residuals.

7.8 Exercise 6: Statistical Properties of Model-Based Estimators with 2-Auxiliaries with Biased Model

Refer to **Table 7.8**.

7.9 Exercise 7: Estimation of Tree Biomass Using Field Plots, Lidar Plots, and SAR

In this exercise, the model-based estimators are applied with two sources of auxiliary data using an actual dataset collected for a region of interior Alaska (USA). The data consist of: (1) estimates of aboveground tree biomass (Mg/ha) collected over relatively sparse sample of field plots (*n_2* = 30 1/30th ha circular plots), (2) height-based metrics collected over a denser (systematic) sample of lidar plots (*n_1* = 325 1/30th ha circular plots), and (3) wall-to-wall L-band satellite SAR-derived imagery (see **Fig. 7.3**). Tree height, tree diameter, and species were collected for each tree on the plots, and allometric models were applied to these measurements to estimate tree—and aggregated plot-level biomass (Yarie et al. 2007).

Here, the model-based estimator with one source of auxiliary data (SAR imagery) developed in Exercise 4, and the estimator for two sources of auxiliary data (wall-to-wall L-band SAR backscatter, large sample of lidar plots) developed in Exercise 6 are compared in the estimation of total aboveground tree biomass (**Table 7.9**).

	Mean (Mg/ha)	SE (Mg/ha)
Model-based estimator: SAR only	49.35	9.43
Model-based estimator: SAR and Lidar plots	50.83	7.07

Table 7.9 *Mean biomass estimate for Tok study area (interior Alaska) using field data, lidar plots, and SAR imagery.*

7.10 Application in Monitoring, Reporting, and Verification (MRV) systems

7.10.1 REQUIREMENTS OF REDD+ MRV PROGRAMS

The Intergovernmental Panel on Climate Change (IPCCC) has specified good practice as it pertains to the concept of REDD+ forest monitoring as inventory design that "neither over- nor under-estimates so far as can be judged, and in which uncertainties are reduced as far as is practicable" (GFOI 2016). This guidance essentially promotes the implementation of monitoring programs that maximize precision of estimates, while minimizing bias, within the constraints of available resources. The multi-level estimators for forest biomass presented in this chapter provide a range of options for design of carbon monitoring programs, including model-assisted approaches requiring probability samples for all levels of the design that provide design-unbiased estimators and model-based approaches that may be less expensive to implement due to the lack of requirements for a probability sample, but at the cost of a possibly biased estimator if the model is incorrectly specified.

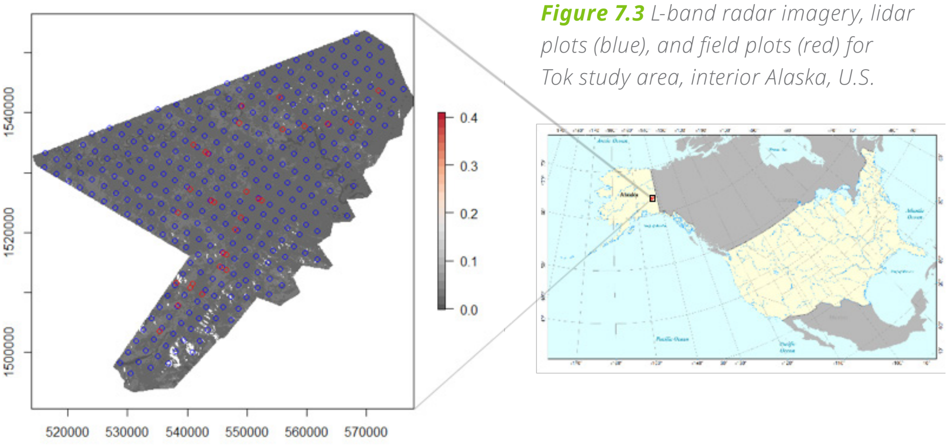

Figure 7.3 *L-band radar imagery, lidar plots (blue), and field plots (red) for Tok study area, interior Alaska, U.S.*

7.11 References

Andersen, H., J. Strunk, H. Temesgen, D. Atwood, and K. Winterberger. 2012. Using multilevel remote sensing and ground data to estimate forest biomass resources in remote regions: A case study in the boreal forests of interior Alaska. Canadian Journal of Remote Sensing 37: 596-611.

Atwood, D.K., H-E Andersen, B Matthiss, F. Holecz. 2014. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing Vol. 7(8): 3262 – 3273.

Calders, K. G Newnham, A. Burt, S. Murphy, P. Raumonen, M. Herold, M. Kaasalainene. Nondestructive estimates of above-ground biomass using terrestrial laser scanning. Methods Ecological Evolution 6(2):198-208.

Chave, J. et al. 2014. Improved allometric models to estimate the aboveground biomass of tropical trees. Global Change Biolory 20(10):3177-90

Chojnacky, DC, L. Heath, J. Jenkins. 2014. Updated generalized regression biomass equations for North American tree species, Forestry 87(1):129-51. .

Duncanson, L. W. Huang, K. Johnson, A. Swatantran, R. McRoberts, and R. Dubayah. 2017. Implications of allometric model selection for county-level biomass mapping. Carbon Balance and Management 12:18.

Ene, L. T., E. Naesset, T. Gobakken, T. G. Gregoire, G. Stahl, R. Nelson. 2012. Assessing the accuracy of regional lidar-based biomass estimation using a simulation approach. Remote Sensing of Environment volume 123: 579-592.

Ene, L., E. Naesset, and T. Gobakken. 2016. Simulation-based assessment of sampling strategies for large-area biomass estimation using wall-to-wall and partial coverage airborne laser scanning surveys. Remote Sensing of Environment 176:328-340.

GFOI 2016, Integration of remote-sensing and ground-based observations for estimation of emissions and removals of greenhouse gases in forests: Methods and Guidance from the Global Forest Observations Initiative, Edition 2.0, Food and Agriculture Organization, Rome

Gregoire, T.G., Næsset, E., McRoberts, R. E.; Ståhl, G. Andersen, H.-E.; Gobakken, T., Ene, L. Nelson, R. 2016. Statistical rigor in LiDAR-assisted estimation of aboveground forest biomass. Remote Sensing of Environment. 173: 98-108.

Gregoire, T. and H. Valentine. 2008. Sampling strategies for national resources and the environment. Chapman and Hall/CRC, Boca Raton.

Hoekman, D. 2010. PALSAR wide-area mapping of Borneo: methodology and map validation. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, Vol. 3, No. 4.

JAXA. 2014. ALOS-2/Calibration Result of JAXA Standard Products; Japan Aerospace Exploration Agency, Earth Observation Research Center: Tsukuba, Japan.

Joshi, N., E.T.A. Mitchard, M. Brolly, J. Schumacher, A. Fernández-Landa, V. K. Johannsen, M. Marchamalo and R. Fensholt. 2017. Understanding 'saturation' of radar signals over forests. Scientific Reports 7:3505.

McRoberts, R. 2010. Probability and model-based approaches to inference for proportion forest using satellite imagery as ancillary data. Remote Sensing of Environment 114:1017-1025.

McRoberts, R., H.-E. Andersen, and E. Naesset. 2014. Using airborne laser scanning data to support forest sample surveys. In Maltamo, M., E. Naesset, and J. Vauhkonen. 2014. Forest applications of airborne laser scanning. Springer: Dordrecht.

Mandallaz, D., J. Breschan, A. Hill. 2013. New regression estimators in forest inventories with two-phase sampling and partially exhaustive information: a design-based Monte Carlo approach with applications to small-area estimation. Canadian Journal of Forest Research, 2013, 43:1023-1031.

Nelsen, R. B. 2006. An Introduction to Copulas. Springer-Verlag New York.

Ryan, C., T. Hill, E. Woollen, C. Ghee, E. Mitchard, G. Cassells, J. Grace, I. Woodhouse, M. Williams. 2012. Quantifying small-scale deforestation and forest degradation in African woodlands using radar imagery. Global Change Biology 18:243-257.

Saarela, S., Holm, S., Grafström, A., S. Schnell, E. Naesset, T. Gregoire, R. Nelson, and G. Ståhl. 2016. Hierarchical model-based inference for forest inventory utilizing three sources of information. Annals of Forest Science 73: 895

Särndal, CE, Swensson, B., Wretman, J. 1992. Model-Assisted Survey Sampling. Springer-Verlag, New York.

Tanase, M., R. Panciera, K. Lowell, S. Tian, J. M. Hacker, J. P. Walker. 2014. Airborne multi-temporal L-band polarimetric SAR data for biomass estimation in semi-arid forests Remote Sensing of Environment 145:93-104.

Treuhaft, R., and P. Siqueira. 2000. Vertical structure of vegetated land surfaces from interferometric and polarimetric radar. Radio Science 35(1):141-177.

Woodhouse, I., E. Mitchard., M. Brolly, D. Maniatis, C. Ryan. 2012. Radar backscatter is not a "direct measure" of biomass. Nature Climate Change 2:556-557.

Yarie, J., E. Kane, and M. Mack. 2007. Aboveground biomass equations for the trees of interior Alaska. AFES Bulletin 115, University of Alaska-Fairbanks, Fairbanks, Alaska, USA (https://www.uaf.edu/files/snre/B115.pdf) last accessed 12/20/2018.

DR. HANS-ERIK ANDERSEN received the Ph.D. degree in quantitative resource management from the University of Washington, Seattle, WA, USA. He was a Research Scientist with the University of Washington Precision Forestry Cooperative from 2003 to 2006, where he developed applications of airborne lidar and interferometric SAR for forest inventory and wildfire fuels assessment. He joined the USDA Forest Service Pacific Northwest (PNW) Research Station as a Research Forester in 2006, based in Anchorage, AK, USA, where he worked on the development of multi-level forest inventory designs for remote regions utilizing both field and remote sensing data. Since 2011, he has been a Team Leader of the Vegetation Monitoring and Remote Sensing (VMaRS) team within the PNW Resource Monitoring and Assessment Program, based in Seattle, WA, USA

Funding for this project was provided by the NASA-SERVIR program. The author would also like to thank reviewers for the help in improving this chapter.

Andersen, Hans-Erik. “Sampling Designs for SAR-Assisted Forest Biomass Surveys.” SAR Handbook: Comprehensive Methodologies for Forest Monitoring and Biomass Estimation. Eds. Flores, A., Herndon, K., Thapa, R., Cherrington, E. NASA 2019. DOI: 10.25966/3k0h-8908