**PROJECT ID:** CAFS.23.104

**PROJECT TITLE:** The Interplay between Sampling Design and Small Area Estimation to Improve Forestland Inventory

**INVESTIGATOR(S):** Temesgen Hailemariam1, Aaron Weiskittel2, Mike Premer2, Rachel Cook3, Phil Radtke4, Corey Green4

1Oregon State University

2University of Maine

3North Carolina State University

4Virgina Tech University

|  |
| --- |
| **PROJECT DESCRIPTION:** One of the challenges often faced in forestland inventory/valuation is estimating gross and net merchantable volume for smaller areas of interest consisting of delineated stands within a larger forested population or ownership. In addition, in many cases, sampling designs commonly used in forest inventories lead to insufficient sample sizes in stands to meet adequate error levels, usually in 5% to 10% allowable error ranges. Small sample sizes for subpopulations of interest, defined geographically or by types (e.g., species composition, site productivity classes, or habitats), lead to unreliable estimates when using the sample data alone. Managers and analysts require reliable small-area estimates (SAE) for acquisitions, fiber supply analyses, silvicultural prescriptions, or habitat capability modeling despite the small sample sizes.  The Fay-Herriot (FH) models are widely used to obtain efficient estimates for small areas. The FH models are a particular case of linear mixed models. They incorporate small area-dependent random effects to explain random variation between small areas that the model’s fixed effects cannot explain. For remote sensing-based forest inventory, Temesgen et al. (2022) reported that FH models are a valuable alternative when there are no precise coordinates for the ground plots or data from fixed-area plots are unavailable. These make FH models useful for 1) operational forest inventories based on LiDAR or other remote-sensing auxiliary information sources, 2) quantifying uncertainty for those stand-level estimates; and 3) improving the allocation of ground samples over ownership.  Variable selection is a critical component of FH models. Many predictor variables are commonly acquired from climate, terrain, and remotely sensed databases. This poses an issue of selecting the optimal set of predictor variables to be included in a model. Variable selection has been addressed in several statistical methods, including linear regression, parametric regression, nonparametric regression, and additive models. Variable selection is not only an issue in linear regression but is also critical to nonparametric models and SAE models. Unfortunately, the procedures used for parametric methods that employ many observations cannot be used in SAE methods because the techniques are fundamentally different. For instance, model accuracy in training data does not automatically improve as more predictor variables are added, and the definition of model complexity is not straightforward. Hence, Packalen et al. (2012) asserted that variable selection is more important than the method used to select nearest neighbors in many cases. The same holds for selecting variables for SAE. For example, Rao and Molina (2015) emphasized that choosing auxiliary data that are strong predictors of the variables of interest should be a major focus when using SAE techniques.  One must guard against having two or more high (positively) correlated variables entering (or maintained in) a regression equation with opposite signs and, in so doing, each nullifying the effect of the other. Hence, this is also when selecting variables for SAE. In operational inventory applications, auxiliary data should be chosen with consideration given to 1) meaningful relationships with the parameters of interest, 2) reliable availability, 3) consistent format and collection protocol, and 4) cost of both acquisition and analysis. In this study, predictor variables will be chosen to reflect these guidelines. In selecting model forms and covariates, we avoided potential issues with multicollinearity by identifying and omitting the most highly correlated covariates. We used the literature and our experience in choosing model forms and covariates.  The FH model requires a solid variable selection and sampling error variance estimation, which involves specifying a sampling protocol and sample size allocation for the direct estimator and then for composite and empirical Bayes estimators. Efficient sampling strategies will facilitate cruise design that effectively captures variation within and between stands. The varying ratio: model error (,)/ (model error ()+ sampling error variance (σ2v)) is the shrinkage estimator. In addition, these errors, along with random effect standard deviation and the residual variation, are essentially a proxy for different types of response variables in the presence of covariates of varying quality. Hence, a thorough analysis is required to examine:   1. Variable selection methods for stand exam-oriented SAE, 2. Efficient sampling design and sample size for stand exam-oriented SAE 3. Allocation of samples in small domains |
| **HYPOTHESES and/or OBJECTIVES:** This proposal has three sets of objectives.   1. Examine variable selection methods for developing small-area estimation models that link inventory plots and remotely sensed data for timberland inventory. 2. Examine the performance of selected sampling designs and sample sizes for applying SAE models. In that, we seek to examine the use of small-area estimators to either reduce sample size when precision is given or improve precision when the sample size is fixed; and 3. Allocate sample size to subpopulation, including optimal allocation of samples in small domains. |
| **METHODS:** After thoroughly reviewing the literature, we will identify variable selection/reduction methods for estimating stand volume and site index from climate, terrain, and remotely sensed data. The list of variables extracted from these auxiliary variables is voluminous. We must select a few variables/covariates that improve prediction at the stand and ownership levels. We seek to choose a reasonable number of predictors from many climates, terrain indices, and remotely sensed attributes. Few studies address this issue, but there has been some research in machine learning for selecting variables for classification models and prediction. A typical example of a classification task is where variable selection plays a vital role in gene selection from microarray data (Guyon and Elise, 2003).  We will first use data collected in western Oregon and then expand to other key regions in CAFS to achieve the study’s objectives. The data include:   1. Individual inventory plot summarizations, stand summarizations and polygons 2. For every stand, Total or operable area (stand acre), number of ground plots, Compiled merchantable volume/acre for each plot (variable radius or ground plot), and measured site index 3. LiDAR plot data, LiDAR layers, ownership layer. Date of collection and system used to collect data, average altitude, speed, scanning angle, and the nominal pulse density per square meter 4. Sentinel data, including mean reflectance values from selected Sentinel-2 bands and additional variables derived from sentinel data sets   Use advanced statistical models, evaluate the selected approaches to allocate samples to small domains, and document the results. |
| **PROJECT TIMELINE:** This is a two-year project starting September 2023 and will be completed in June 2025. Key milestones include   1. Recruiting a MS graduate student – Fall 2023 2. Modeling and variable selection – Fall 2024 3. Sampling design and simulations – Winter 2025 4. Evaluation of sample allocation to small domains – Spring 2025 5. Write reports/manuscripts – Summer 2025 |
| **EXPECTED DELIVERABLES – ONE YEAR:** Key deliverables would be a compiled dataset for SAE comparisons and an annual progress report. |
| **EXPECTED DELIVERABLES – LONG-TERM:** A key long-term deliverable would be a model to aid in accurate small-area estimation for timberland inventory. This would include protocols that use remotely sensed data and thinning status to reduce uncertainty in predicting the number of trees, basal area, and volume estimate. Additional long-term deliverables would be a MS thesis, publications/presentations, and peer-reviewed journal articles. |
| **POTENTIAL MEMBER COMPANY BENEFITS:**   * Member companies borrow strength from the freely available remotely sensed and ground data and reduce the costs of data acquisitions. They will benefit by reducing the cost of establishing ground plots and improving the estimation and prediction of selected stand variables, including gross and net merchantable volume. * The project demonstrates ways for incorporating FH models into operational forest inventory in plantations. * Members will acquire protocols that use remotely sensed data and thinning status to reduce uncertainty in predicting the number of trees, basal area, and volume estimate. * Translate some of the theories in small-area estimation to practice. * Operationalize small area estimation methods by linking remotely sensed and ground data and developing cost-effective algorithms solutions for improved stand-and ownership-level estimation. |
| **NEXT YEAR’S PROJECT BUDGET – NSF/CAFS PORTION:** The project budget for next year would be $35,000 and would support a PhD graduate student stipend, health insurance, and tuition. The position would be cost-shared between NSF and CAFS funding. |
| **NEXT YEAR’S PROJECT BUDGET - OTHER SOURCES, INCLUDING SITE-SPECIFIC:** Additional funding for the PhD graduate student’s travel, required equipment, and additional expenses would be covered for additional sources including the US Forest Service Joint Venture Agreements and other site-specific resources. |