New Project Proposal

Robust small-area estimation strategies for developing accurate stand-level diameter distributions

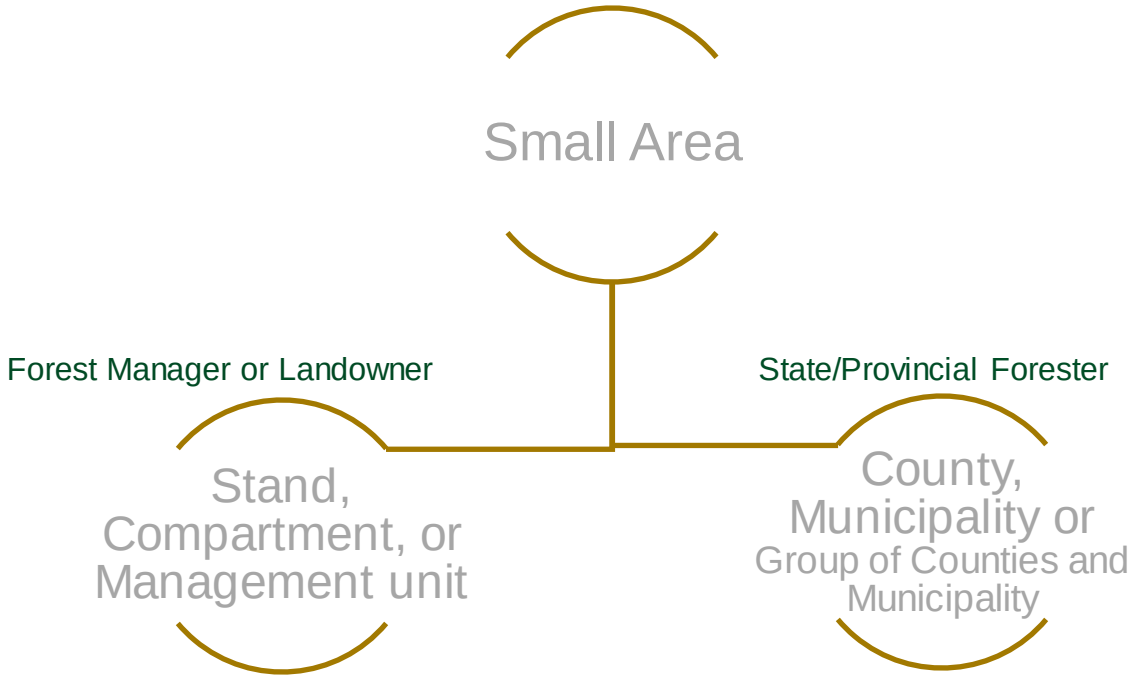
Jaslam Poolakkal, University of Idaho
Mark Kimsey, University of Idaho
David Affleck, University of Montana
Phil Radtke, Virginia Tech
Rachel Cook, North Carolina State University
Public-Private Research Collaborators

Presenter – Jaslam Poolakkal

Introduction

SAE in Forest Inventories

Small area estimation (SAE) refers generally to approaches for making population-level estimates within small domains for which sample sizes are deemed inadequate to produce estimates of acceptable precision using traditional design-based techniques.

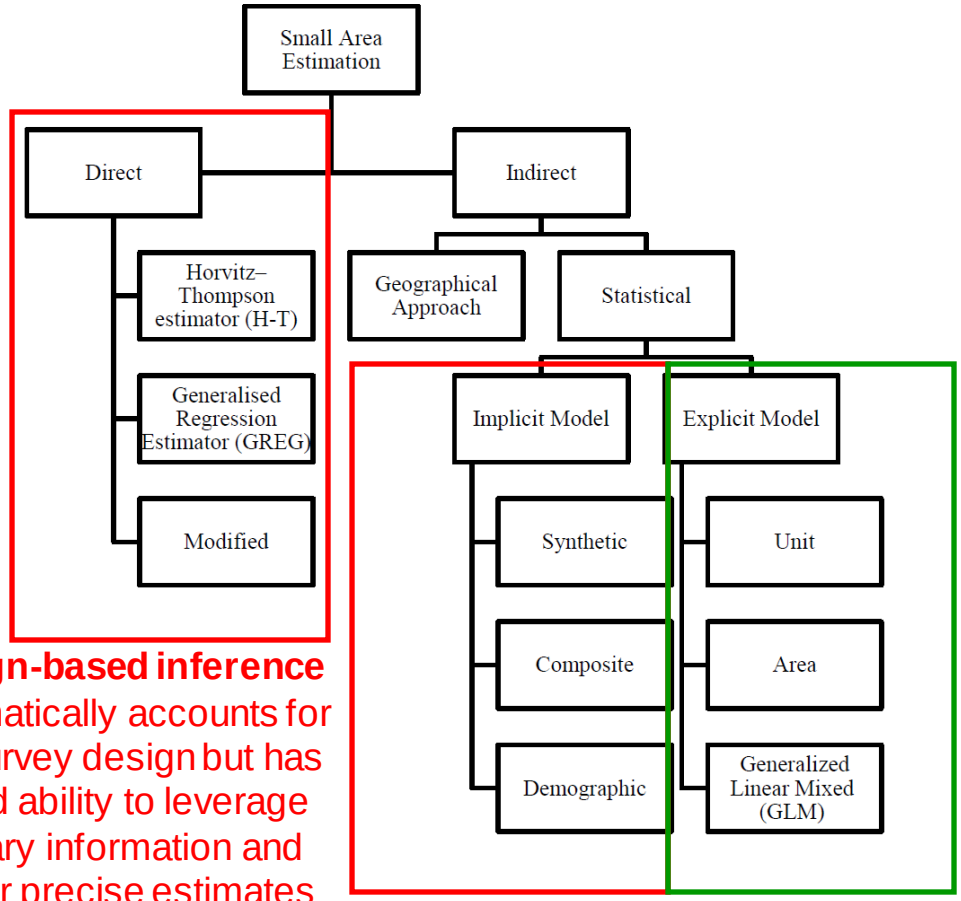


Ref: Guldin RW (2021)

"The term 'small area estimation' is somewhat confusing because it's the size of the sample from the area that causes estimation problems, not size of the area."
Pfeffermann (2013)

Introduction

The term Small Area Estimation refers to a number of methods that rely on ancillary data sources in order to "borrow" additional information, increasing the effective sample size – and consequently, the precision of the estimates – for the selected domain.



Design-based inference automatically accounts for the survey design but has limited ability to leverage ancillary information and deliver precise estimates for small sample sizes (i.e., small area estimates).

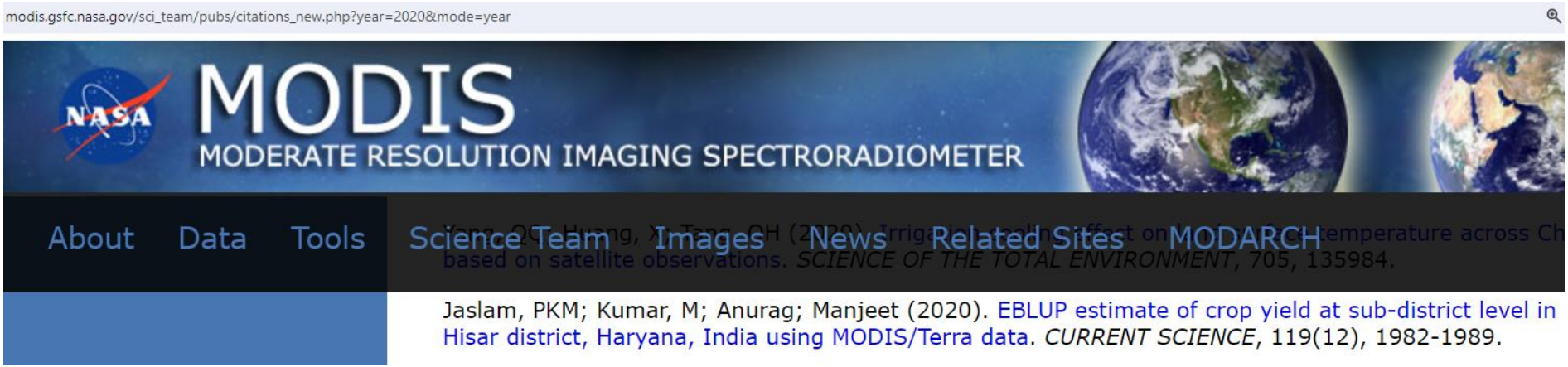
Model-based inference explicitly consider the design and data jointly but can use ancillary information and borrow from the rich modeling literature to deliver robust inference for small samples sizes.

Introduction

Model based small area estimation are broadly classified into two groups:

- **Unit level random effect models**, proposed originally by Battese *et al.(*1988). These models relate the unit values of a study variable to unit-specific covariates.

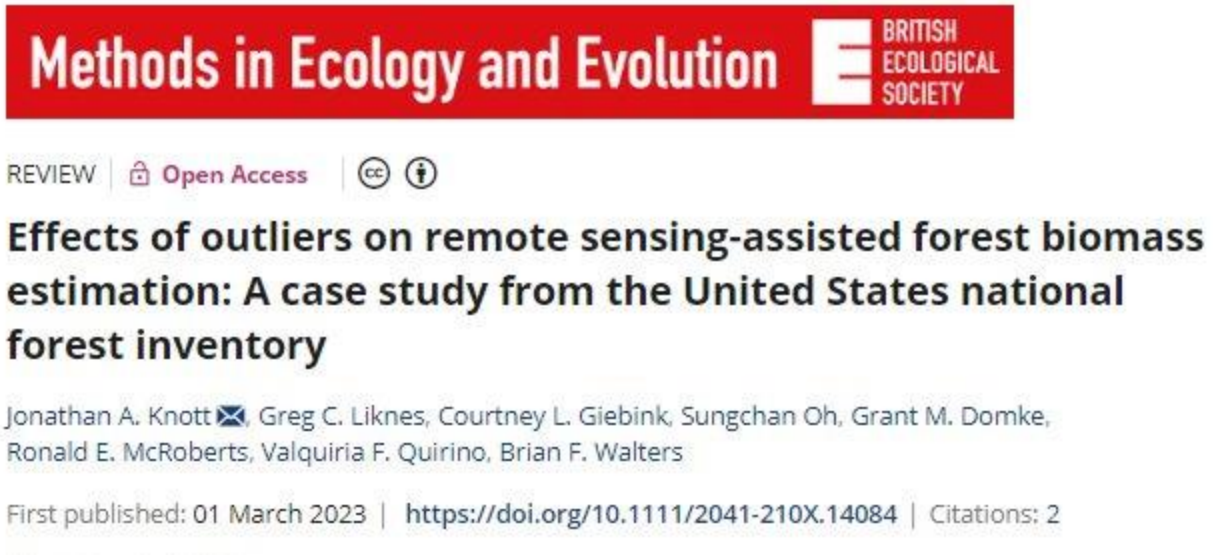
- **Area level random effect models**, which are used when auxiliary information is available only at area level. They relate small area direct survey estimates to area-specific covariates (Fay and Herriot, 1979).



NSF

Center for Advanced Forestry Systems

Justification



Methods in Ecology and Evolution | BRITISH ECOLOGICAL SOCIETY

REVIEW | Open Access

Effects of outliers on remote sensing-assisted forest biomass estimation: A case study from the United States national forest inventory

Jonathan A. Knott, Greg C. Liknes, Courtney L. Giebink, Sungchan Oh, Grant M. Domke, Ronald E. McRoberts, Valquiria F. Quirino, Brian F. Walters

First published: 01 March 2023 | https://doi.org/10.1111/2041-210X.14084 | Citations: 2

Inclusion or removal of potential outliers can have unintended consequences on population parameter estimates. When used to inform biomass mapping, carbon markets, greenhouse gas reporting and environmental policy, it is necessary to ensure the proper use of NFI and remotely sensed data in geospatial data harmonization.

Justification

AUSTRIAN JOURNAL OF STATISTICS
Volume 41 (2012), Number 4, 243–265

Robust Unit-Level Small Area Estimation: A Fast Algorithm for Large Datasets

Tobias Schoch

School of Business,
University of Applied Sciences Northwestern Switzerland

Abstract: Small area estimation is a topic of increasing importance in official statistics. Although the classical EBLUP method is useful for estimating the small area means efficiently under the normality assumptions, it can be highly influenced by the presence of outliers. Therefore, Sinha and Rao (2009; The Canadian Journal of Statistics) proposed robust estimators/predictors for a large class of unit- and area-level models. We confine attention to the basic unit-level model and discuss a related, but slightly different, robustification. In particular, we develop a fast algorithm that avoids inversion and multiplication of large matrices, and thus permits the user to apply the method to large datasets. In addition, we derive much simpler expressions of the bounded-influence predicting equations to robustly predict the small-area means than Sinha and Rao (2009) did.

Mixed linear models have, unlike location-scale or regression models, **no nice invariance structure**. Notably, this means that the parameters cannot be estimated consistently in the presence of contamination—there is an unavoidable asymptotic bias. In the presence of contamination, any method estimates the parameter at the core model plus an unknown bias. In the case of ML estimators, the bias can be arbitrarily large and renders these estimators extremely inefficient (Welsh and Richardson, 1997).

Justification

Model Assisted Statistics and Applications 18 (2023) 171–181
DOI 10.3233/MAS-221416
IOS Press

Analysis of unit level models for small area estimation in crop statistics assisted with satellite auxiliary information

Cite

Article type: Research Article

Authors: Jaslam, P.K. Muhammed[a,*] | Kumar, Manoj[a] | Bhardwaj, Nitin[a] | Salinder, [b] | Sumit, Vikash Kumar[c]

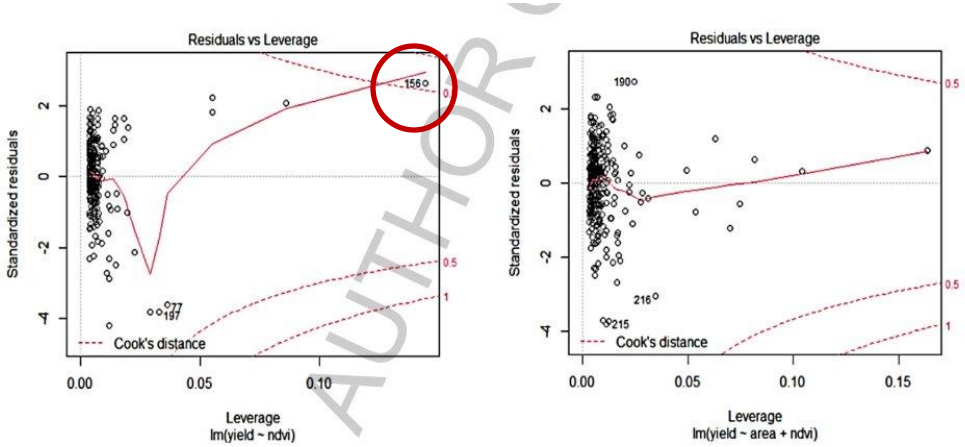


Fig. 1. Residual vs Leverage plot for Hisar and Sirsa districts.

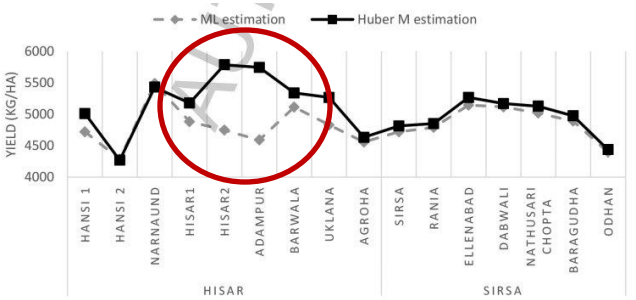


Fig. 3. REBLUP estimates of the wheat yield for Hisar and Sirsa districts.

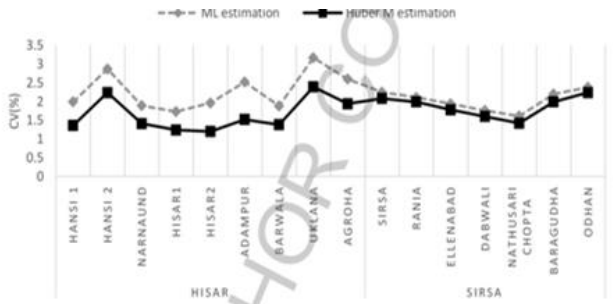


Fig. 2. Block wise CVs for the wheat yield estimate by ML and Huber M estimation.

Justification

Stand-level SAE

Model-based Small Area Estimation (SAE) combines direct and synthetic estimates, leveraging specific data while mitigating instability from small sample sizes or variability.

For sampled stands, the EBLUP is a weighted average of the direct estimator obtained using only the ground information and the synthetic estimator.

$$\hat{Y}_i^{EBLUP} = X_i^T \hat{\beta} + \hat{v}_i = \hat{\gamma}_i y_i + (1 - \hat{\gamma}_i) X_i^T \hat{\beta}$$

Stands with one or No Plots

Very few plots with Unreliable Direct Estimates

The model relies more or only on synthetic prediction

Hypotheses or Objectives

- Traditional SAE techniques often struggle with **statistical assumptions, non-normal data, outliers, and nonlinear relationships.** We propose *integrating robust estimation methods* and *machine learning algorithms* into the SAE framework to *estimate diameter distributions*, a key forest metric.

- A comprehensive baseline assessment of traditional SAE versus multiple robust and machine learning diameter distribution models will be assessed and validated against independent datasets. Model assisted techniques that strike a balance between precision and bias will be recommended.

- Our post-hoc analysis will ***quantify uncertainty, sensitivity****, and* ***reliability of estimates***, ***prioritizing the explainability*** of machine learning-based SAE models.

Study Area and Data Sources

- Complex, mixed conifer forests of the Pacific Northwest and Rocky Mountains, and more uniform Southeastern pine forests.

- Leverage national CAFS SDImax project dataset (w/permissions), additional sourced inventory data from project collaborators.

- Auxiliary data from the national CAFS SDImax database (climate, topography, geology, soil) + remote sensing data from 3D NAIP, Sentinel, and where available, free 3DEP LiDAR.

Methods

ANNUAL REVIEWS

Discussed SAE methods that are developed under weaker assumptions and SAE methods that are robust in certain ways, such as in terms of outliers or model failure. Also includes topics such as nonparametric SAE methods, Bayesian approaches, model selection and diagnostics, and missing data

Annual Review of Statistics and Its Application

Robust Small Area Estimation: An Overview

Jiming Jiang[1] and J. Sunil Rao[2]

[1]Department of Statistics, University of California, Davis, California 95616, USA; email: jimjiang@ucdavis.edu

[2]Department of Public Health Sciences, University of Miami, Miami, Florida 33136, USA

Machine Learning Based SAE

Methods

These techniques excel in leveraging intricate **nonlinear relationships and interactions** within data. They adeptly incorporate spatial and temporal information, enhancing the precision and resilience of Small Area Estimation (SAE) estimates.

A Forest for the Trees: Using Random Forests for Small Area Estimation on US Forest Inventory Data

Citation
Schmitt, Julian Francis. 2023. A Forest for the Trees: Using Random Forests for Small Area Estimation on US Forest Inventory Data. Bachelor's thesis, Harvard University Engineering and Applied Sciences.

The Survey Statistician, **2022, Vol. 86, 22–31.**

New and Emerging Methods

Tree-Based Machine Learning in Small Area Estimation

Patrick Krennmair [1], Nora Würz [2] and Timo Schmid[3]

[1]Freie Universität Berlin, Germany, patrick.krennmair@fu-berlin.de
[2]Freie Universität Berlin, Germany, nora.wuerz@fu-berlin.de
[3]Otto-Friedrich-Universität Bamberg, Germany, timo.schmid@uni-bamberg.de

Statistics > Methodology

[Submitted on 12 Feb 2024]

A step towards the integration of machine learning and small area estimation

Tomasz Żądło, Adam Chwila

The use of machine-learning techniques has grown in numerous research areas. Currently, it is also widely used in statistics, including the official statistics for data collection (e.g. satellite imagery, web scraping and text mining, data cleaning, integration and imputation) but also for data analysis. However, the usage of these methods in survey sampling including small area estimation is still very limited. Therefore, we propose a predictor supported by these algorithms which can be used to predict any population or subpopulation characteristics based on cross-sectional and longitudinal data. Machine learning methods have already been shown to be very powerful in identifying and modelling complex and nonlinear relationships between the variables, which means that they have very good properties in case of strong departures from the classic assumptions. Therefore, we analyse the performance of our proposal under a different set-up, in our opinion of greater importance in real-life surveys. We study only small departures from the assumed model, to show that our proposal is a good alternative in this case as well, even in comparison with optimal methods under the model. What is more, we propose the method of the accuracy estimation of machine learning predictors, giving the possibility of the accuracy comparison with classic methods, where the accuracy is measured as in survey sampling practice. The solution of this problem is indicated in the literature as one of the key issues in integration of these approaches. The simulation studies are based on a real, longitudinal dataset, freely available from the Polish Local Data Bank, where the prediction problem of subpopulation characteristics in the last period, with "borrowing strength" from other subpopulations and time periods, is considered.

Company Benefits/ Deliverables

Benefits

- ML-based SAE models can demonstrate the ability to develop robust estimates of stand characteristics in areas with limited or no sample data.

- SAE characterization of the diameter distribution of forest stands managed for timber production provides improved tree lists for growth and yield modeling and carbon estimation.

Deliverables

- Project model output provided through GitHub (model forms and their coded script)

- Methods and outcomes published in a peer-reviewed journal

Summary

Demonstration of robust, bias-controlled, explainable, and independently validated SAE models will illustrate their role for resource and time efficiency in the development of accurate stand level forest metrics (tree lists). Derived methods and models will further support current FIA's long term strategic plan priorities 6,7,8,10,11, and aid the industry's need for growth and yield modeling inputs across large landscapes.