Continuing Project

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

CAFS.24.105

Jaslam Poolakkal, University of Idaho
Mark Kimsey, University of Idaho
David Affleck, University of Montana
Paul Parker, University of California-Santa Cruz
Nathaniel Naumann, PotlatchDeltic Corp.

Presenter: Jaslam Poolakkal

The Problem: Justification

Forest decisions—like harvest planning, habitat protection, and carbon accounting—require **highly localized data** (e.g., tree sizes, species mix, biomass). But:

- ❑ Typical stands have only 1-2 measurement plots
- ❑ Many stands have no measurements at all
- ❑ Plots often miss stand variability (density changes, gaps, microsites)
- ❑ Exhaustive sampling is financially impossible at operational scales
- ❑ While remote sensing (free/low-density LiDAR) provides landscape-scale data, it cannot directly measure critical variables like diameter distributions

Justification

Traditional methods fail to produce precise estimates of diameter distributions at the stand level due to:

- ❑ Small or unbalanced sample sizes
- ❑ Complex stand structures (e.g., mixed species, irregular distributions)

Why Small Area Estimation (SAE)?

SAE enables reliable estimation in small domains where traditional approaches lack precision. SAE effectively “borrows strength” from auxiliary data (e.g., LiDAR, NAIP, Sentinel-2) to improve predictions even when sample sizes are small.

Stand-level forest metrics (e.g., diameter class distributions) require robust, interpretable, and flexible modeling approaches that can handle complex stand heterogeneity.

Common issues: skewed data, outliers, complex correlation

The SAE Advantage:

Justification

SAE fills the gaps by intelligently blending:

- ✓ Limited field measurements (e.g., 1-2 plots per stand)
- ✓ Auxiliary data (LiDAR, climate, soil maps)
- ✓ Statistical models to ensure reliability

Why Current Methods Fall Short:

- Inadequate Sample Sizes
- Rigid Distributional Assumptions
- Parametric Model Constraints
- Fails to capture nonlinear covariate interactions and hierarchical structures

This project advances SAE through robust model-based techniques integrated with machine learning

Hypotheses or Objectives

Objective: Demonstrate robust, interpretable machine learning framework for stand-level SAE estimates across multiple U.S. regions.

Specific Objectives:

- ❑ Characterize diameter distributions using univariate area-level SAE models (Traditional, Robust, ML).
- ❑ Develop multivariate linear SAE models to estimate diameter distribution indices and stem density.
- ❑ Generate synthetic estimates via multi-output machine learning, integrating data-driven approaches.
- ❑ Evaluate and benchmark model performance using datasets from National Forest Systems (NFS) and Industry.

Innovative Methods: Incorporate robust methods and machine learning to enhance SAE, addressing issues like model misspecification

Methods

Study Regions & Data Integration

Geographic Scope: Mixed conifer forests (PNW: FS Regions 1, 4, 6, Rockies) and even-aged southern pine plantations (FS Region 8).

SAE Domains: Stand-level (Area level SAE Models)

Ground Data Sources:

- ❑ Industrial CFI data (PotlatchDeltic, Green Diamond, Manulife)
- ❑ NFS stand exams
- ❑ FIA plots (benchmark)

Auxiliary Variables:

- ❑ 3D-NAIP PC & Ortho Imagery,
- ❑ ClimateNA – (Annual, Month, Season),
- ❑ Geology and Soil layer - (SGMC & gSSURGO geodatabase),
- ❑ Topography extraction from 30m DEM (Slope, Aspect, Topographic wetness index, Solar radiation)

Methods

Objective 1 – Univariate SAE Models

- Base: Fay-Herriot model with area-level covariates
- Robust variants: OBP, M-quantile, transformed response
- ML plug-in estimators integrated within SAE
- Parameter recovery: Weibull, Johnson SB, Finite Mixture Models (FMM)

Objective 2 – Multivariate SAE Models

- Multivariate Fay-Herriot for correlated diameter indices
- Joint estimation for efficiency and shared strength
- Transform responses counts by diameter class (Poisson/lognormal)

Objective 3 – ML-Based Synthetic SAE

- Multi-output RF, XGBoost, Neural Networks for stem density by diameter class
- Mixed Effects RF (MERF), Random Weight NNs
- Clustered ML for hierarchical stand structure

Objective 4 – Model Evaluation & Uncertainty

- Bootstrap and analytic MSE estimators
- Validation via stand exams and simulation
- Metrics: RMSE, MAE, MAPE, AIC, adj. R^2, residual diagnostics

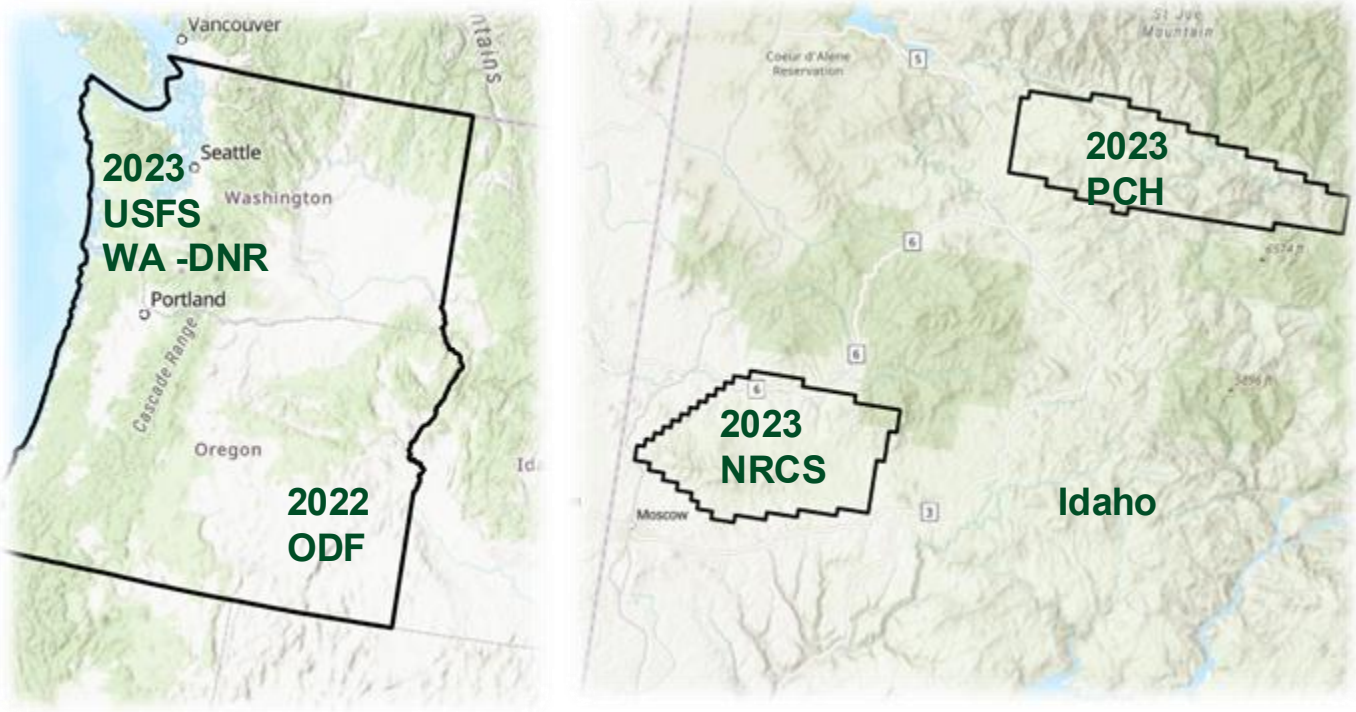
NSF

Center for Advanced Forestry Systems

Progress 2024-25

Forest Inventory Data: Engaged with industry partners, public land managers, and research networks to compile data across the Pacific Northwest and Southeast U.S.

Auxiliary Data: Leveraging publicly available datasets and initiating procurement of 3D NAIP products via project collaborations.



PotlatchDeltic

GREEN DIAMOND Resource Company

FOREST SERVICE U.S. DEPARTMENT OF AGRICULTURE

HEXAGON

OREGON DEPARTMENT OF FORESTRY

NRCS Natural Resources Conservation Service

WASHINGTON DEPARTMENT OF NATURAL RESOURCES

WASHINGTON STATE DEPT OF NATURAL RESOURCES

Southern U.S. Identified three AOIs (~600 km² each) in Arkansas, Mississippi, and South Carolina - high industry stand exam coverage facilitated by Green Diamond and PotlatchDeltic

Initial Implementation Area: Idaho AOI's

- **St. Joe National Forest**

Extensive coverage from **PotlatchDeltic** and **Idaho Department of Lands** (IDL) industrial forests.

- **Moscow Mountain**

Includes **UI Experimental Forest** for **validation and field calibration**.

LiDAR Preprocessing Workflow

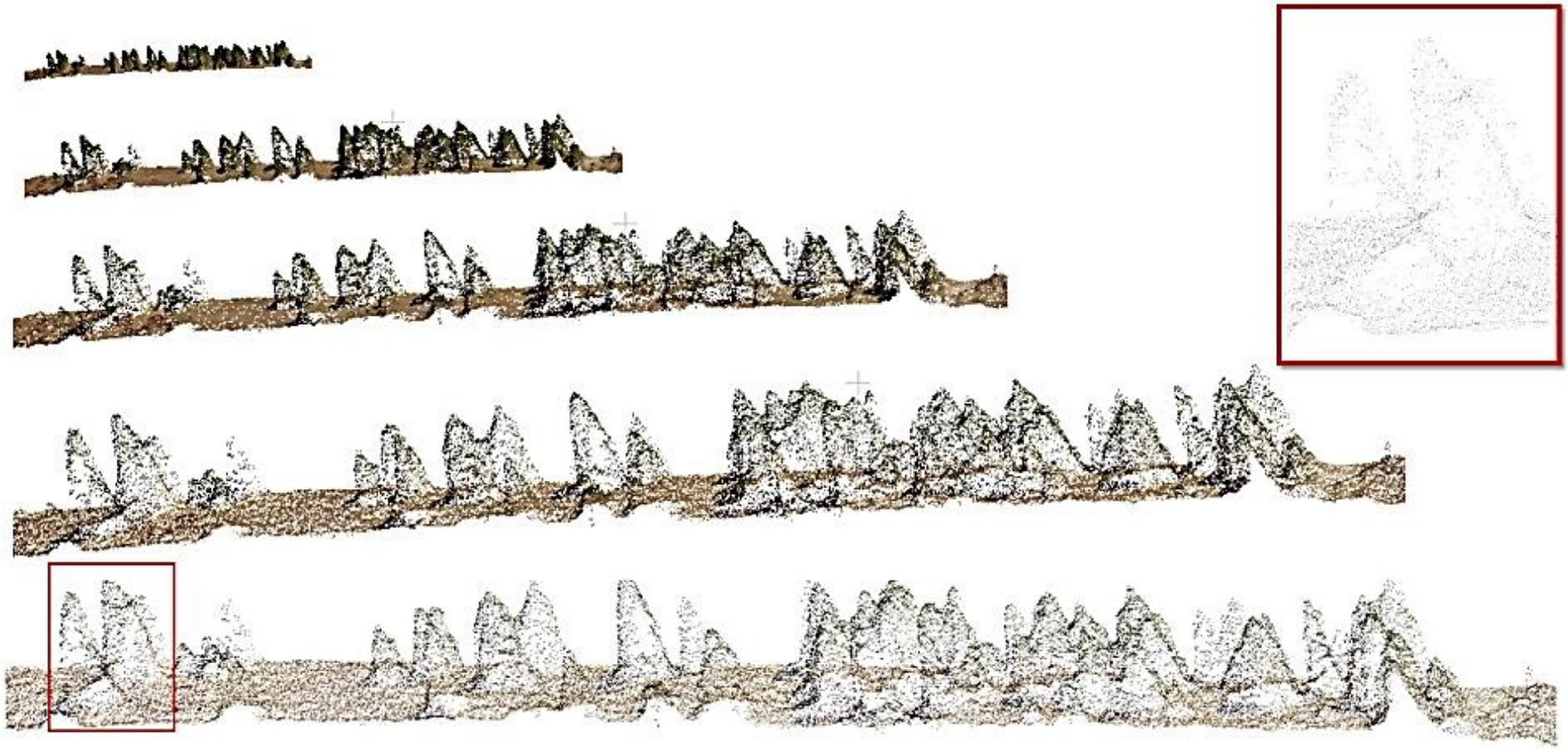
Steps: Noise filtering → Ground classification → Normalization → Metric gridding → Canopy surface modeling

Outputs: Stand-level auxiliary variables

Progress 2024-25

Colorized 3D NAIP Point Cloud (GSD: 30 cm, Frame Sensor)

Progress 2024-25



Progress 2024-25

Structural Metrics from LiDAR

Height (HAG/HABS):

Max, Mean, SD, CV, Skew, Kurtosis, Q05–Q95, Min/Max

Canopy Structure:

Canopy Cover (Total, Percentiles), Density ≥2–20 m, VCI, UCI, GFP, Canopy Relief Ratio, Foliage Height Diversity

Complexity & Stratification:

CCI, Height Evenness/Stratification, Taller Tree Dominance, Height-Weighted Density, Density–Height Ratio, Tall Stem Skew, Canopy Closure Proxy

Spectral & Environmental Covariates

Ortho imagery :

Vegetation: NDVI, GNDVI, NDWI, SAVI, EVI
Color: ExG, VARI, GRVI, NGRDI, TGI
Normalized Bands: R_norm, G_norm, B_norm

Terrain (30 m DEM):

Slope, Aspect, TWI, Solar Radiation

Climate (ClimateNA):

Annual, monthly, and seasonal

Soils & Geology:

gSSURGO & SGMC

Initial Target Variable: Quadratic Mean Diameter (QMD)

As an initial step, we focus on Quadratic Mean Diameter (QMD) — a widely used stand-level metric that effectively summarizes the central tendency of tree diameter while accounting for basal area.

QMD is especially valuable in forest inventory and modeling because it reflects both the size and distribution of trees within a stand, making it a representative and interpretable indicator of diameter structure.

Feature Selection & Modeling Pipeline

Step	Purpose
1. VarianceThreshold	Removes low-information numeric features
2. OneHotEncoding	Handles categorical variables
3. Correlation Filter	Removes highly correlated variables
4. Mutual Information	Selects best from each correlated pair
5. VIF Filter	Reduces multicollinearity
6. LassoCV	Final feature selection

Model	Method	Estimation Strategy	Proposed By / Citation	Purpose & Strengths
M1	reml	Restricted Maximum Likelihood	Rao & Molina (2015), Prasad & Rao (1990)	Common default. Reduces small-sample bias vs ML. Stable fixed effect estimation.
M2	ml	Maximum Likelihood	Fay & Herriot (1979)	Direct likelihood optimization. Slightly biased in small samples but efficient.
M3	amrl	Adjusted ML – Random Effects Level	Marhuenda, Molina & Morales (2013)	Reduces bias in random effect variance. Improves small-area accuracy.
M4	ampl	Adjusted ML – Prediction Level	Marhuenda et al. (2013)	Improves prediction MSE by adjusting variance at prediction level.
M5	amrl_yl	Adjusted ML with Y-link Transformation	Marhuenda et al. (2014)	Handles heteroscedasticity; transforms response for better normality.
M6	ampl_yl	Adjusted Prediction Level + Y-link	Marhuenda et al. (2014)	Balances transformed modeling and bias correction at prediction level.
M7	reblup	Robust EBLUP	Sinha & Rao (2009)	Incorporates robustness against outliers in small areas.
M8	reblupbc	Robust EBLUP with Bias Correction	Molina et al. (2017)	Adds bootstrap-based bias correction to improve inference.
M9	PCA + Stepwise	Principal Component Regression	Jolliffe (2002), adapted in SAE by Tzavidis et al. (2016)	Condenses correlated features into orthogonal PCs. Reduces dimensionality without loss of information.

Major Findings

Well-performing models:

- **REML (M1)** & **Adjusted ML (M3–M6)**: Stable estimates, high explanatory power (FH_R^2 ≈ 0.95), low MSE, good for general applications.

- **PCA + Stepwise ML (M9)**: Achieved best AIC (646.37), showing value of feature reduction.

- **REBLUP/REBLUPBC (M7–M8)**: Robust against outliers, delivering high-precision coefficients and stable inference under non-normality.

Major Findings

- Most influential predictors across methods: Cruise Design, Canopy Height Metrics, Soil and Drought Variables repeatedly significant and strong in effect size.

Residuals & Diagnostics: Good fit, but residual skewness/kurtosis and some non-normal random effects suggest opportunities for improvement.

Work in progress

Model Generalization: Validating current models on unseen domains or bootstrap samples will reveal if SAE methods overfit. ML-based models can be tuned for generalization with cross-validation or out-of-sample testing.

Hybrid Approaches: Emerging research shows combining SAE with ML (e.g., random forest residual correction, neural-net variance modeling) improves prediction in real-world small areas.

Deliverables & Company Benefits

📦 Deliverables

- SAE & ML models for diameter distribution
- Tools for parameter recovery & class estimation
- Validated results (FIA/CFI) + Jupyter notebooks
- Peer-reviewed publication & GitHub code

🏢 Benefits to Partners

- Accurate stand-level metrics from limited data
- Supports volume, biomass, and product planning
- Customizable across ownerships & regions
- Aligns with FIA goals & remote sensing use

Next Steps & Future Directions

Beyond QMD — Toward Full Diameter Distributions

- Current phase focused on Idaho & univariate QMD.
- Forest applications need full **diameter distributions**, not just summary stats.

Upcoming Objectives

- ❑ **Multivariate SAE** for joint modeling of moments/percentiles.
- ❑ **Diameter class estimation** using multi-output regression.
- ❑ **ML-based SAE** to handle nonlinearity, interactions, and hierarchical data.
- ❑ **Validation** on external datasets (FIA, CFI); empirical simulation testing.
- ❑ **Advanced feature selection** (e.g., mRMR, embedded ML techniques).

Summary

- **Objective:** Improve stand-level diameter estimates using robust SAE and machine learning.
- **Challenge:** Sparse field data; complex forest structure.
- **Solution:** Fuse LiDAR, NAIP, soils, and climate data with flexible, interpretable models.
- **Progress:** High-accuracy QMD predictions ($R^2 \approx 0.95$); key features include canopy metrics & site factors.
- **Impact:** Reliable, scalable tools for partners to support planning, inventory, and reporting.
- **Next:** Extend to full diameter distributions with multivariate and ML-based SAE.

ROBUST SMALL-AREA ESTIMATION STRATEGIES FOR DEVELOPING ACCURATE STAND-LEVEL DIAMETER DISTRIBUTIONS

Partnership for
Small Area Estimation

Principle Investigators

Jaslam Poolakkal, University of Idaho,
Mark Kimsey, University of Idaho,
David Affleck, University of Montana,
Paul Parker, University of California-Santa Cruz.,
Nathaniel Naumann, PotlatchDeltic Corp.,

Project Advisory Committee

Phil Radtke, Virginia Tech University,
Dale Hogg, Green Diamond Resource Company,
Rachel Cook, North Carolina State University,
Jacob Strunk, USDA Forest Service-FIA,
Karin Wolken, USDA Forest Service-NFS,
David Lindbo, USDA Natural Resources Conservation Service,
Peter Gould, Mason, Bruce & Girard,

NSF

Center for Advanced Forestry Systems