
2026 PSAI Annual Conference
Parallel 2C - Statistical Methodologies

Small Area Estimation: Model Selection & Methodological Approaches

PRESENTER

John Rayven DC. Surio

20-21 October 2026

School of Statistics, University of the Philippines Diliman, Quezon City

INTRODUCTION

BACKGROUND



BACKGROUND

Poverty remains a major development concern in the Philippines, with implications extending beyond income to people's well-being and access to opportunities.



BACKGROUND

Poverty remains a major development concern in the Philippines, with implications extending beyond income to people's well-being and access to opportunities.

Reliable poverty estimates guide policymaking, resource allocation, and social protection programs. However, survey-based estimates may become less reliable as the geographic level of estimation becomes more disaggregated, creating a need for methods that can generate reliable estimates for smaller areas



BACKGROUND

Poverty remains a major development concern in the Philippines, with implications extending beyond income to people's well-being and access to opportunities.

Reliable poverty estimates guide policymaking, resource allocation, and social protection programs. However, **survey-based estimates may become less reliable as the geographic level of estimation becomes more disaggregated**, creating a need for methods that can generate reliable estimates for smaller areas



BACKGROUND

Poverty remains a major development concern in the Philippines, with implications extending beyond income to people's well-being and access to opportunities.

Reliable poverty estimates guide policymaking, resource allocation, and social protection programs. However, survey-based estimates may become less reliable as the geographic level of estimation becomes more disaggregated, creating a need for methods that can generate reliable estimates for smaller areas

This study provides empirical evidence on the applicability of selected statistical and machine-learning approaches to small area estimation for predicting poverty incidence at more disaggregated geographic levels, specifically the regional and provincial levels.



WHY SMALL AREA ESTIMATION FOR POVERTY INCIDENCE?



Small Area Estimation (SAE) provides estimates for geographic domains where direct survey estimates may be unavailable or insufficiently reliable.

Data Sources

Data Sources

2023 Family Income and Expenditure Survey (FIES)

Provided household income for constructing the modeling outcome and estimating the relationship between income and household characteristics.

Data Sources

2023 Family Income and Expenditure Survey (FIES)

Provided household income for constructing the modeling outcome and estimating the relationship between income and household characteristics.

2020 Census of Population and Housing (CPH) – Household

Provided household-level auxiliary variables for predicting income and poverty in smaller geographic domains.

Data Sources

2023 Family Income and Expenditure Survey (FIES)

Provided household income for constructing the modeling outcome and estimating the relationship between income and household characteristics.

2020 Census of Population and Housing (CPH) – Household

Provided household-level auxiliary variables for predicting income and poverty in smaller geographic domains.

2020 CPH – Barangay Characteristics, Infrastructure, and Local Facilities

Additionally used to incorporate barangay-level characteristics.

Data Sources

2023 Family Income and Expenditure Survey (FIES)

Provided household income for constructing the modeling outcome and estimating the relationship between income and household characteristics.

2020 Census of Population and Housing (CPH) – Household

Provided household-level auxiliary variables for predicting income and poverty in smaller geographic domains.

2020 CPH – Barangay Characteristics, Infrastructure, and Local Facilities

Additionally used to incorporate barangay-level characteristics.

2021 Official Poverty Threshold

Used to adjust and calibrate the model estimates.

Used as reference benchmarks when calculating the absolute difference between estimates and official results.

Modelling Approaches



OUTCOME

Natural Logarithm of Household Per Capita Income,
 $Y_i = \log(\text{PERCAPITA}_i)$

GEOGRAPHIC DOMAINS

Regional and Provincial Levels

Linear Regression

Two alternative specifications of the same linear modeling framework:

LM-AIC: backward stepAIC, followed by removal of predictors with $p > 0.05$

LM-CP: minimum Mallows' CP subset, followed by removal of predictors with $p > 0.05$

LASSO

Penalized linear regression; 10-fold cross-validation; λ_{1se}

XGBoost

Gradient-boosted regression trees; 200 rounds; depth = 6; learning rate = 0.1; subsample = 0.8; column subsampling = 0.8

Random Forest

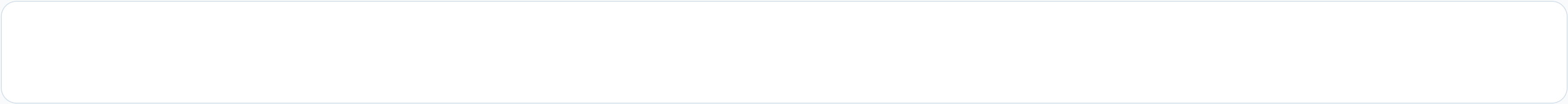
Ensemble of regression trees; 1,000 trees; mtry = $\lceil p/3 \rceil$; minimum node size = 10

Permutation-based variable importance and out-of-bag error retained

Empirical Bayes

Linear mixed-effects model; province-level random intercept; REML estimation. Statistically insignificant fixed effects iteratively removed

PROCESS



Modelling Approaches



OUTCOME

Natural Logarithm of Household Per Capita Income,
 $Y_i = \log(\text{PERCAPITA}_i)$

GEOGRAPHIC DOMAINS

Regional and Provincial Levels

Linear Regression

Two alternative specifications of the same linear modeling framework:

LM-AIC: backward stepAIC, followed by removal of predictors with $p > 0.05$

LM-CP: minimum Mallows' CP subset, followed by removal of predictors with $p > 0.05$

LASSO

Penalized linear regression; 10-fold cross-validation; λ_{1se}

XGBoost

Gradient-boosted regression trees; 200 rounds; depth = 6; learning rate = 0.1; subsample = 0.8; column subsampling = 0.8

Random Forest

Ensemble of regression trees; 1,000 trees; mtry = $\lceil p/3 \rceil$; minimum node size = 10

Permutation-based variable importance and out-of-bag error retained

Empirical Bayes

Linear mixed-effects model; province-level random intercept; REML estimation. Statistically insignificant fixed effects iteratively removed

PROCESS

Modelling in 2023 FIES



Predict PERCAPITA in 2020 CPH



Compute Poverty Incidence based on the Estimates



Adjust based on the 2021 Official Poverty Threshold



Abs Difference of predicted poverty incidence with official 2021 poverty estimates

Modelling



Car, AC, Computer, prop_OFW, and prop_Work_Private.

These variables were repeatedly selected or identified as important across modeling approaches and regions.

The variables shown below are the top 10 variables selected or identified as important for each method only; they are not a complete list of predictors.

 Linear Models (AIC and Mallows's CP)	 LASSO	 XGBoost	 Random Forest	 Empirical Bayes
total_members prop_OFW Car prop_Work_SelfEmp Ref AC Computer prop_Work_Private Water_CWS H_Ed_JHS_G	prop_OFW prop_Educ_Masteral_UG prop_Work_Gov prop_Non_Relative Car AC prop_Work_Private Computer prop_Educ_Elem_UG prop_Work_SelfEmp	AC total_members Ref Car H_Age_squared Computer prop_Educ_Elem_UG prop_Work_Private prop_OFW prop_Work_Gov	total_members Ref AC Car Computer prop_married prop_Work_Private prop_Educ_Elem_UG prop_Work_Gov prop_OFW	total_members prop_OFW Car prop_Work_SelfEmp AC Computer Water_CWS prop_Work_Private Ref H_Ed_JHS_G



Predictor recurrence reflects predictive usefulness within the fitted models; it should not be interpreted as evidence of causal effects.

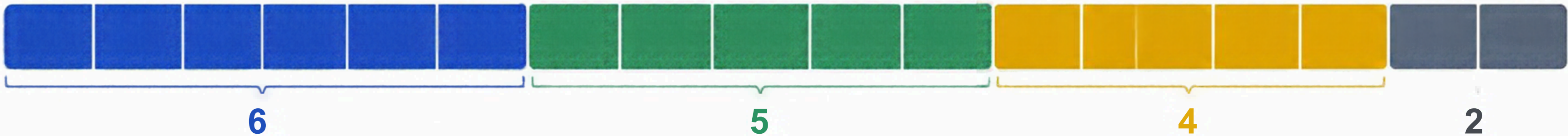
Regional-Level Estimates



Tree-Based Models Lead:
11 of 17 Regional Domains

XGBoost and Random Forest together account for 11 of 17 regional domains.

RESULTS ACROSS 17 REGIONS



excluding Negros Island Region



6
XGBOOST



5
RANDOM FOREST



4
LASSO



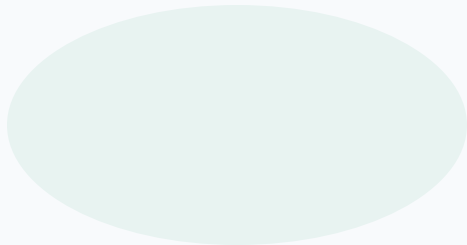
2
LINEAR REGRESSION
EMPIRICAL BAYES



TAKEAWAY

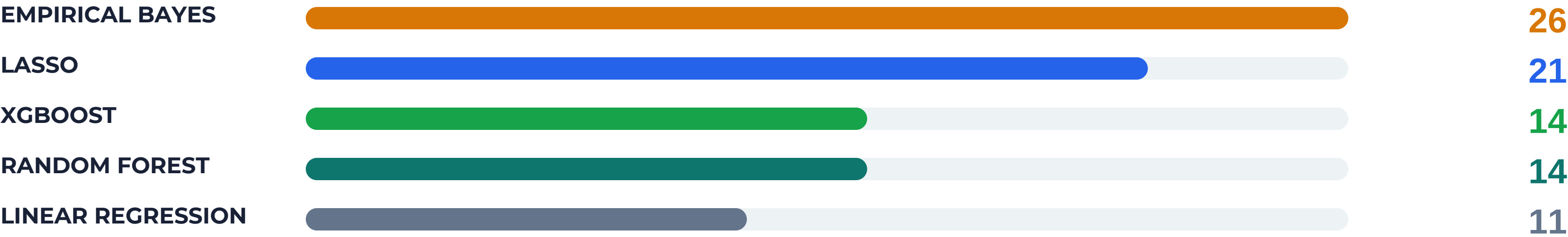
Nonlinear tree-based algorithms (XGBoost & Random Forest) captured complex macro-level interactions best, producing closest estimates in 11 out of 17 regional domains.

Provincial-Level Estimates



COUNT OF WITH CLOSEST PREDICTION BY METHOD

PROVINCES



TAKEAWAY

At fine spatial resolution, Empirical Bayes (province-level random intercepts) and LASSO shrinkage performed best by effectively controlling sampling variation in smaller samples.

No Single Superior SAE Model

Model performance varies fundamentally by geographic scale. The most appropriate approach depends on whether the target is regional or provincial estimation.

However, the results show which methods stand out in predicting poverty incidence at the regional and provincial levels using the five models.

REGIONAL SCALE

XGBoost and Random Forest together produced the closest estimates in 11 of 17 regional domains.

PROVINCIAL SCALE

Empirical Bayes performed best in 26 of 82 domains, followed by LASSO in 21 of 82 domains.

WAYS FORWARD

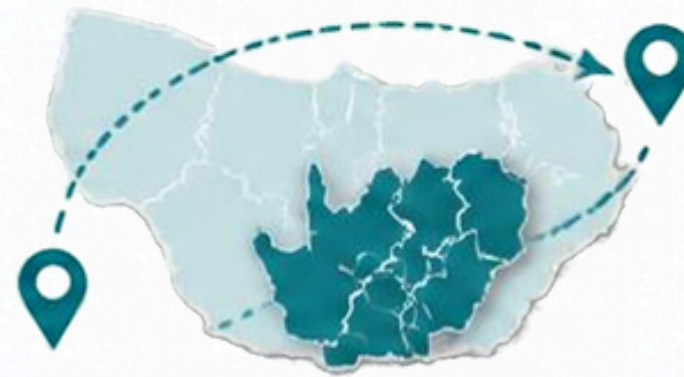
1



Extend modeling to the provincial level

Develop and evaluate the modeling framework at provincial level.

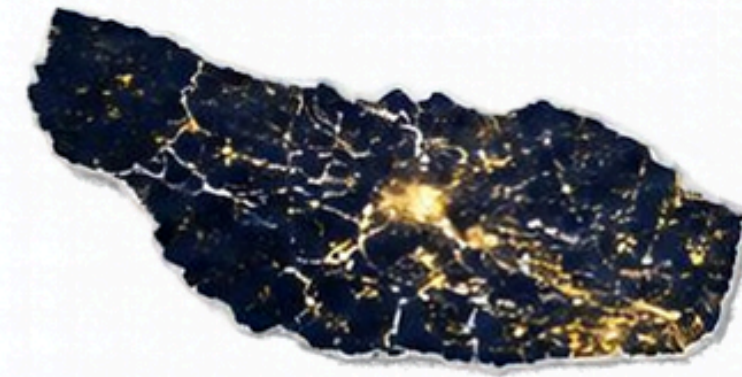
2



Continue prediction at both scales

Use the provincial-level modeling extension to produce predictions at regional and provincial levels.

3



Add night-time lights as a predictor

Incorporate night-time lights as an additional auxiliary predictor in the SAE models.



These extensions aim to strengthen small-area poverty estimation across geographic scales.

REFERENCES

1. Asian Development Bank, & United Nations Statistics Division. (2021). Guidelines for producing small area estimates. Asian Development Bank.
2. Battese, G. E., Harter, R. M., & Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83(401), 28–36. <https://doi.org/10.1080/01621459.1988.10478561>
3. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). Association for Computing Machinery. <https://doi.org/10.1145/2939672.2939785>
4. Das, S., & Haslett, S. (2019). A comparison of methods for poverty estimation in developing countries. *International Statistical Review*, 87(1), 1–25.
5. Edochie, I., Newhouse, D., Tzavidis, N., Schmid, T., Foster, E., Hernandez, A. L., Ouedraogo, A., Sanoh, A., & Savadogo, A. (2024). Small area estimation of poverty in four West African countries by integrating survey and geospatial data.
6. Elbers, C., Lanjouw, J. O., & Lanjouw, P. (2003). Micro-level estimation of poverty and inequality. *Econometrica*, 71(1), 355–364.
7. Frink, N. (2025). Generalized tree-based machine learning methods with applications to small area estimation (Doctoral dissertation, Otto-Friedrich-Universität Bamberg). Otto-Friedrich-Universität Bamberg. <https://doi.org/10.20378/irb-107660>
8. Krennmair, P., Würz, N., Schmid, T., & Tzavidis, N. (2024). Random forests and mixed effects random forests for small area estimation of general parameters: A poverty mapping case study in Mozambique.
9. Michal, V., Wakefield, J., Schmidt, A. M., Cavanaugh, A., Robinson, B., & Baumgartner, J. (2023). Small area estimation with random forests and the LASSO.
10. Molina, I. (2024). Small area estimation: Recent developments and future directions.
11. Molina, I., & Rao, J. N. K. (2010). Small area estimation of poverty indicators. *The Canadian Journal of Statistics*, 38(3), 369–385. <https://doi.org/10.1002/cjs.10051>
12. Nacion, J. A., & Pacificador, A. P. (2021). Small area estimation for poverty statistics in the Philippines.
13. Rao, J. N. K. (2017). Small area estimation. *Wiley Interdisciplinary Reviews: Computational Statistics*, 9(6), e1407.
14. Reyes, C. M., Tabuga, A. D., Mina, C. D., & Asis, R. D. (2004). Poverty mapping in the Philippines. Philippine Institute for Development Studies.
15. Velarde, R., & Fernandez, L. (2011). Measuring poverty in the Philippines. Philippine Institute for Development Studies.
16. Viljanen, M., Meijerink, L., Zwakhals, L., & van de Kasstele, J. (2022). A machine learning approach to small area estimation: Predicting the health, housing and well-being of the population of Netherlands. *International Journal of Health Geographics*, 21(4), 1–18.