Improving Estimates of Mean Welfare and Uncertainty in Developing Countries*

Joshua D. Merfeld† Hai-Anh Dang‡ David Newhouse§

2025-05-16

Abstract

Reliable small-area estimates of economic welfare significantly inform the design and evaluation of development policies. This paper compares the accuracy of wealth estimates obtained from the empirical best predictor (EBP) of a linear nested error model, Cubist regression, extreme gradient boosting, and boosted regression forests. The evaluation draws two-stage samples from unit-level household census data in seven developing countries, combines them with publicly available geospatial indicators to generate small area estimates of assets for all seven countries and poverty for two, and evaluates these estimates against census-derived benchmarks. Extreme gradient boosting and Cubist regression generally produce more accurate predictions than traditional EBP models. A proposed two-stage residual bootstrap procedure slightly underestimates confidence intervals, but leads to higher coverage rates than the parametric bootstrap approach used for EBP predictions. These results demonstrate that, given a sufficiently large sample of enumeration areas, predictions from extreme gradient boosting or Cubist regression with a two-stage residual block bootstrap generally provide more accurate point and uncertainty estimates for generating small-area welfare estimates.

Keywords: poverty, welfare, prediction, machine learning, geospatial
JEL Codes: C53, C80, I32, O10

*We thank the editor Tavneet Suri, three anonymous reviewers, Jed Friedman, Haishan Fu, Keith Garrett, Dean Jolliffe, Talip Kilic, and Bo Pieter Johannes Andree for comments and support. We thank Lina Cardona, Dilhanie Deepawansa, Carlos De Maia, Francis Mulangu, and Mario Negre for help obtaining data. This project was partially funded by the Knowledge for Change Program's Phase IV-funded programmatic research project "Understanding Trends in Sub-National Differences in Economic Well-Being in Low- and Middle-Income Countries".

†KDI School of Public Policy and Management and IZA; merfeld@kdis.ac.kr
‡World Bank, IZA, Indiana University, and University of Economics, Ho Chi Minh city; hdang@worldbank.org
§World Bank and IZA; dnewhouse@worldbank.org

1 Introduction

Accurate measures of welfare for spatially disaggregated areas are valuable inputs into the design and evaluation of effective development policies (Atkinson, 2019; Blumenstock, 2016; Ravallion, 2015; Merfeld and Morduch, 2023; McBride et al., 2022). Yet, most estimates of welfare are derived from household surveys that can only produce reliable statistics at higher levels of aggregation, mostly because of the high cost of data collection (Fujii and van der Weide, 2020; Kilic et al., 2017). Small area estimation combines survey data with more comprehensive auxiliary data to obtain estimates for small areas, in this case defined as one administrative level below that which survey estimates are considered to be sufficient reliable to publish. These more granular estimates can improve geographic targeting (Elbers et al., 2007) as well as program and policy evaluation (Ratledge et al., 2021). While some countries can draw on rich administrative data such as income tax records to serve as auxiliary data, developing countries do not typically maintain accurate and up-to-date administrative data sources. It is quite common to predict welfare with census data and a contemporaneous survey – following Elbers et al. (2003) or Molina and Rao (2010). But in poorer countries, census data are usually collected infrequently. As a result, official statistics on welfare in small areas tend to be dated.

Against this backdrop, recent advances in machine learning and the growing availability of non-traditional data sources have led to the proliferation of new options for small area estimation. For example, Blumenstock et al. (2015) use mobile phone records to infer the socioeconomic status of phone owners in Rwanda and Aiken et al. (2023) use mobile phone call data records to predict targeting performance of programs in Togo. However, one drawback of mobile phone data is that – like banking records – the population of mobile-phone owners may be systematically different from the population of those without phones. Satellite-derived geospatial data do not suffer from this selection bias and have become increasingly popular in economics (Donaldson and Storeygard, 2016). Previous research has demonstrated that geospatial data is a promising source of data to estimate economic growth (Henderson et al., 2012), labor force participation (Merfeld et al., 2022), and welfare more generally (Jean et al., 2016; Yeh et al., 2020; Chi et al., 2022; Engstrom et al., 2022; Van der Weide et al., 2024).[1]

In this paper, we evaluate four main methods to estimate welfare at low levels of aggregation in developing countries. Importantly, we propose and implement a two-stage weighted residual bootstrap procedure to estimate uncertainty for the set of machine learning methods; estimates of uncertainty are an ongoing, but current unsolved, issue with machine learning predictions (Chi et al., 2022) and our results are encouraging.

[1] Newhouse (2024) provides a recent review of the literature of applications of geospatial small area estimation to wealth and poverty.

We evaluate the performance of these methods using unit-level census data across seven developing countries in Africa and Asia: Burkina Faso, Madagascar, Malawi, Mozambique, Sri Lanka, Tanzania, and Vietnam. These countries were selected due to the availability of georeferenced census data. In Malawi and Tanzania we are able to extend the evaluation to include a headcount poverty measure in addition to the asset index, while in the other five countries we evaluate prediction of an asset index, similar to Chi et al. (2022) and Masaki et al. (2022).

The first main method we evaluate is empirical best predictions based on a nested error linear model, which we refer to as EBP. This estimation method comes from a long history of small area estimation in statistics. (Battese et al., 1988; Jiang and Lahiri, 2006; Molina and Rao, 2010). We also evaluate three newer machine learning methods: Cubist regression models (Quinlan et al., 1992; Wang and Witten, 1997), extreme gradient boosting (Chen and Guestrin, 2016) – more commonly known as XGBoost – and boosted regression forests, or BRF (Friedberg et al., 2020; Tibshirani et al., 2018). As a robustness check, we also evaluate another traditional method commonly used in small area estimation, the ELL method (Elbers et al., 2003).[2] These methods differ in their level of parsimony and transparency on the one hand, and their predictive accuracy on the other, and a key goal of this exercise is to better understand the terms of this trade-off in the context of welfare prediction. In each case, we specify models at the sub-area level, which in these contexts refers to highly disaggregated administrative areas akin to groups of villages. We then aggregate predictions to obtain estimates at the target area for each country.

For predictors, we use satellite-derived geospatial indicators that are available across much of the globe, meaning that the methods and data evaluated here are widely applicable in cases where geolocated survey data are available. We use shapefiles from the seven countries to pull geospatial data from multiple sources, which is then combined with samples drawn from the unit-level census data. In each country, we simulate 100 two-stage samples – first randomly selecting enumeration areas and then randomly selecting households based on survey designs of actual household surveys in each country – and compare the overall performance across simulations, ensuring that the results are derived from one hundred possible samples rather than a single sample. Under these conditions, XGBoost and Cubist regression tend to outperform EBP and BRF in terms of accuracy, as measured both by Pearson correlations and Mean Squared Deviation. This can be seen clearly in the first two rows of Table 1.

A key contribution of the paper is the evaluation of a two-stage residual block bootstrap to estimate uncertainty

[2]See Das and Haslett (2019) for a comparative analysis of several poverty mapping methods including the ELL method, EBP, and M-quantile. Pratesi and Spagnolo (2023) offer a recent overview of small area estimation methods for measuring poverty. Another potential SAE method is ESPREE (Isidro et al., 2016). However, we know of no readily available software packages that implement either M-quantile or ESPREE.

Table 1: Summary of results

	EBP	Cubist	XGBoost	BRF
Correlation (pearson)	0.829	0.868	0.885	0.822
Squared deviation	0.135	0.110	0.096	0.159
Width of CI	0.915	0.883	0.841	1.149
Coverage	0.784	0.842	0.867	0.885
Area under the curve (AUC)	0.897	0.917	0.925	0.909
Poverty targeting (FGT1)				
Malawi				
10-percent target	0.058	0.054	0.056	0.067
20-percent target	0.052	0.050	0.048	0.058
Tanzania				
10-percent target	0.014	0.016	0.015	0.015
20-percent target	0.012	0.012	0.011	0.013

Note: The table shows the average of the statistics listed down the rows, separately for each of the four methods and across all seven countries. We calculate AUC only for the asset indices. Higher values equate to better performance. The poverty targeting results are post-transfer values for only Malawi and Tanzania and are calculated based on what percentage of the total population is targeted (either 10 or 20 percent). Lower values indicate better targeting.

for the three machine learning method, similar to proposals by Chambers and Chandra (2013) and Luo and Lai (2021). The presentation of uncertainty statistics has traditionally been less common for machine learning methods (Chi et al., 2022) but remains an important component of official statistics. The weighted residual block bootstrap accounts for the hierarchical nature of the data and proceeds in two steps, sampling residuals – accounting for informative sampling using sampling weights – separately at both the target area level and the sub-area level, which is the unit of analysis. This procedure slightly underestimates uncertainty, with average coverage rates of 88.5 percent for BRF, 86.7 percent for XGBoost, and 84.7 percent for Cubist. These coverage rates, however, exceed the average for EBP (78.4 percent), which are derived from the parametric bootstrap procedure typically used to generate uncertainty estimates for EBP predictions (Butar and Lahiri, 2003; González-Manteiga et al., 2008). Because XGBoost and Cubist are more accurate on average than EBP, their confidence intervals are smaller (rows three and four of Table 1). Alternative accuracy statistics lead to similar conclusions. XGBoost is the most accurate and EBP is the least accurate when we compare "area under the curve" (AUC) estimates across simulations (Hanna and Olken, 2018). Finally, when we simulate a hypothetical cash transfer covering a fixed percentage of the population, targeting areas based on XGBoost or Cubist predictions reduces poverty by more than EBP predictions in Malawi. However, EBP performs similarly to XGBoost in Tanzania.

We also document differences in performance across in-sample and out-of-sample areas, given recent evidence that out-of-sample predictions generated by EBP models can be significantly less accurate than in-sample predictions when using geospatial data (Newhouse et al., 2025). On average, all four estimators generate

more accurate predictions in sample than out of sample. EBP shows the largest drop in performance out of sample, as shown in subsection 3.1. The more flexible nature of the machine learning models appears to be better suited for predicting into out-of-sample areas than EBP. In addition, out-of-sample EBP predictions suffer from the unavailability of sample data to condition on when estimating the random area effect.

There are also noticeable differences in precision across in-sample and out-of-sample areas. Since the machine learning models do not include a random effect conditioned on the sample, the bootstrap procedure estimates confidence intervals that are roughly the same size for in-sample and out-of-sample areas. Out-of-sample accuracy is lower than in-sample accuracy, meaning that out-of-sample coverage rates tend to be lower as well; for example, average coverage rates for wealth estimates using XGBoost fall from 91.8 percent to 83.3 percent. This does not result from overfitting the model to the sample, given that LASSO is used for model selection in the EBP model and that regularization methods are built into the machine learning algorithms. Instead, because villages are sampled proportional to their population size, out-of-sample areas tend to be less populated and may therefore systematically different from in-sample areas, even after taking into account sampling probabilities. Although sample weights are included, out-of-sample predictions suffer from the relative paucity of training data from rural, less populated areas.

To better understand which geospatial predictors are important in the models, we examine the prevalence of different types of variables selected by LASSO for the EBP models. In addition, we calculate Shapely decompositions of the impact of predictors on the predictors generated by extreme gradient boosting models (Lundberg et al., 2019). Overall, land cover classification variables, including vegetation indices, consistently emerge as important predictors. Population estimates, pollution variables, and night time lights also make important contributions in different contexts. This is consistent with geospatial predictors acting as proxies for population density and urbanity, which are in turn systematically correlated with wealth and poverty. (Newhouse, 2024)

Finally, we examine the impact of altering the model specification by expanding the set of candidate geospatial variables and including additional interactions. In particular, we include MOSAIKS variables (Rolf et al., 2021) and separately experiment with interacting all predictors (including the MOSAIKS variables) with a measure of urbanity. Neither adding geospatial predictors nor adding these interactions leads to a meaningful improvement in prediction accuracy on average. However, both adding additional predictors and interactions moderately increases estimated uncertainty and improves coverage rates with the machine learning methods. Interestingly, the inclusion of more candidate predictors leads to worse results for EBP. These results suggest limited benefits from supplementing a relatively small set of publicly available features with MOSAIKS indicators. This result is useful because MOSAIKS variables, which are derived from a

convolutional neural network, are less transparent and interpretable than the geospatial features used in the main specification. Furthermore, since machine learning methods already have the flexibility to incorporate interactions, interacting features with a measure of urbanity has little systematic positive impacts on accuracy.

These findings primarily contribute to a newer literature using new types of data to estimate economic statistics of interest, especially welfare. In the past decade, there has been a proliferation in the use of satellite imagery to estimate poverty and welfare (Jean et al., 2016; Yeh et al., 2020; Engstrom et al., 2022; Newhouse et al., 2025; Chi et al., 2022). However, processing raw imagery is typically computationally intensive and requires specialized skills. In comparison, other types of data are easier to use, like mobile phone call data records (Aiken et al., 2023; Blumenstock et al., 2015), but these can be more difficult to access due to privacy concerns and also raise issues related to representativeness. The satellite indicators we use can be obtained from publicly available sources relatively easily and are much smaller in size.[3]

We also contribute to a related literature on small area estimation, which grew out of the statistics literature in the 1970s (Efron and Morris, 1973; Carter and Rolph, 1974; Fay III and Herriot, 1979; Battese et al., 1988). Earlier work proposed the use of census data for prediction (Elbers et al., 2003) and the empirical best predictor (Jiang and Lahiri, 2006; Molina and Rao, 2010; Tzavidis et al., 2018) is now one of the most common implementations of small area estimation. One reason the EBP model is preferred in many applications is its transparency; a nested-error regression model allows for a straightforward estimation of linear coefficients with random effects specified at the target area level. A simple table of coefficients indicates exactly how each variable is related to the measure of household welfare. EBP estimates are also "design consistent," in the sense that the estimates converge to the population value as the sample becomes large. On the other hand, machine learning methods, while possibly generating more accurate predictions, suffer from a lack of parsimony and transparency (Efron, 2020). While we employ Shapley decompositions to shed light on which predictors have large impacts on predicted outcomes, it is not straightforward to understand the relationship between the set of predictors and the prediction. In addition, much of the formal statistical theory related to measuring the uncertainty associated with predictions from tree-based machine learning is new (Athey et al., 2019). To the best of our knowledge, this is the first paper to show rigorously that a two-stage, residual bootstrap can estimate confidence intervals at least as accurate as those currently used for EBP predictions, when evaluated against unit-level census data as ground truth. In cases where economists and statisticians are willing to sacrifice parsimony and transparency to achieve more accurate predictions and the sample data are sufficiently rich to train accurate machine learning models, the availability of a simple and accurate bootstrap method for estimating uncertainty surmounts a crucial barrier to the use of tree-based machine

[3]For example, most of the predictors we use in this paper are (freely) available through Google Earth Engine.

learning algorithms.

These findings are also policy relevant, particularly regarding providing more accurate inputs into designing cost-effective poverty targeting and budget transfers. For example, recent World Bank guidelines strongly encourage the adoption of EBP models where census microdata is available (Corral et al. 2022). These results offer more nuanced insights into the performance of EBP and other methods when applied to geospatial data, which presents policy makers with different modeling options for the best decision-making outcomes. If additional information is available for within-community targeting (Elbers et al., 2007), more accurate poverty estimates would translate into significant poverty reduction. Furthermore, since small area estimation methods are generally less costly than additional data collection, they offer appealing policy choices in a poorer country context.[4]

The rest of the paper is organized as follows. In section 2, we provide a brief overview of the data, the estimation methods, and the method utilized to validate estimates for accuracy and uncertainty. Then, in section 3, we review detailed results of simulations, including distinguishing between in and out of sample performance, performance in poorer or richer areas, using alternative metrics, and different specifications. Section 4 concludes.

2 Methods

This paper evaluates four different methods for generating predictions of district-level poverty rates: linear empirical best predictor (EBP) models (Battese et al., 1988; Jiang and Lahiri, 2006; Molina and Rao, 2010); cubist regression models (Quinlan et al., 1992; Wang and Witten, 1997), which we refer to as Cubist throughout the results section; extreme gradient boosting (Chen and Guestrin, 2016), more commonly known as XGBoost; and boosted regression forests, or BRF (Friedberg et al., 2020; Tibshirani et al., 2018). Table 2 summarizes key features of these methods, which we describe in greater detail below. Importantly, we evaluate these methods in the context of developing countries, where census data is often unavailable or dated. One of our goals is to improve the estimation of key development outcomes in such contexts. As such, we propose using data that is widely available across the globe: remote sensing and geospatial data. In addition, we adapt and apply a two-stage residual bootstrap procedure to estimate uncertainty for the machine-learning models.

These techniques may improve on existing methods but require rigorous evaluation of their accuracy and

[4]Data gaps are a well-known challenge affecting poverty measurement in poorer countries. While census data are less available in these poorer countries, fewer household surveys are also infrequently collected. Even where household surveys are available, they likely display inconsistent quality over longer time periods. Data imputation methods have been increasingly used to help fill these data gaps (Beegle et al. (2016); Dang et al. (2019); Dang and Lanjouw (2023)).

precision in multiple contexts before they can be applied. To do that, we compare estimates from these models to ground-based "truth" derived from unit-level census data in seven countries: Burkina Faso, Madagascar, Malawi, Mozambique, Sri Lanka, Tanzania, and Vietnam. While these countries were selected because of the availabiltiy of census data with either enumeration area geocoordinates or sub-area identifiers with corresponding shapefiles, they also cover a wide variety of contexts, varying in incomes as well as geography.[5] The official administrative boundaries available in shapefiles differ across countries. In general, we pull geospatial data for the lowest administrative level possible. For example, this is the Fokontany in Madagascar, while in Malawi it is the Enumeration Area. We refer to these levels as "subareas" throughout this paper.

Table 3 presents the number of areas and subareas for each country. Sub-areas, in particular, vary greatly across countries, from as few as 8,763 in Burkina Faso to as many as 67,239 in Mozambique.[6] Areas, on the other hand, show less variation, ranging from 169 (Tanzania) to 1,515 (Madagsacar). The table also shows the number of households in the unit-level census data. We have the full census for Madagascar, Mozambique, and Sri Lanka, but a sub-sample of the full census for the other countries.[7]

2.1 Outcomes

We focus on two separate measures of welfare: an asset index and poverty rates. We calculate an asset index for all countries and estimated poverty rates for Malawi and Tanzania, both of which have a near-concurrent household survey that allows us to impute poverty into the census. Using survey data, we predict household per capita expenditures for all households in the census using assets. We then classify households as poor (or not) based on the quantile of their imputed expenditures per capita and the poverty rate in the household survey. We set a poverty rate of approximately 50 percent in Malawi and 21 percent in Tanzania, which matches the national poverty rate in the household surveys. Appendix B provides more details on the imputation procedure, as well as the calculation of the asset indices across countries.

The use of unit-record census data remains the preferred gold standard option when recent census data are available. However, census data tend to be collected infrequently in most developing countries, and small area estimates based on satellite indicators are a preferred alternative to reporting direct survey estimates when census data are old or there have been rapid changes in spatial welfare patterns. We focus on satellite and other remotely sensed data because they are widely available and predictive of spatial variation in welfare. Importantly, their wide coverage means that these indicators cover all areas of a given country, which is a prerequisite for unbiased small area estimates.

[5] For example, only four of the countries are classified as lower income, while the geography varies from the Sahel to Southeast Asia.
[6] The sub-area in Mozambique is below the admin4 level, while the sub-area in all other countries is the admin4.
[7] The size of the sub-sample varies from 10 to 20 percent.

Table 2: Comparison of key features across methods

		ML method		
Feature	EBP	Cubist	XGBoost	BRF
Prediction target	Census data without assets	Census data without assets	Census data without assets	Census data without assets
Functional form	Single linear function	Multiple linear functions (one at each decision tree node)	Decision Trees	Decision Trees
Assumed data generation process	Specified linear model with nested error structure	Average linear prediction based on specified predictors at each decision tree node	Sum of decision trees based on specified predictors and residuals	Sum of decision trees based on specified predictors and residuals
Objective (Loss) function	Minimize mean squared error of predictions	Minimize standard error of dependent variable when determining each split. Minimize mean squared error at each node	Minimize mean squared error of predictions	Maximize penalized heterogeneity in the dependent variable when determining each split. The penalty favors balanced splits.
Random effect conditioned on survey data	Yes	No	No	No
Predictors used	All selected by LASSO	Selected through growing decision trees	Selected through growing decision trees	Selected through growing decision trees from randomly selected subset of candidate predictors at each node
Use different subsamples of the data for growing trees and making predictions	N/A	No	No	Yes
Transparency/ Ease of use	More	Less	Less	Less

Note: See Appendix C for more details.

Table 3: Census summary statistics

	Year	Areas	Subareas	Households	Land area ('000s km^2)	Outcomes calculated: Assets	Outcomes calculated: Poverty
Burkina Faso	2019	346	8,763	748,961	274	X	
Madagascar	2017	1,515	14,412	5,007,602	597	X	
Malawi	2018	420	18,700	796,925	118	X	X
Mozambique	2017	1,282	67,239	6,133,769	802	X	
Sri Lanka	2012	331	13,984	4,842,300	66	X	
Tanzania	2012	169	13,857	2,612,518	947	X	X
Vietnam	2019	711	11,159	2,322,464	332	X	

Note: The table shows the number of areas and subareas, the number of households, the total land area in thousands of square kilometers, and the outcomes calculated for different countries. Assets are available in all countries, while we impute poverty into the census using concurrent household surveys in Malawi and Tanzania.

2.2 Geospatial features

We pull geospatial features from Google Earth Engine using Python and the earth engine api library. Table A1 in the appendix lists the geospatial features used. Importantly, we often derive several additional statistics from different indicators. For example, data on temperature is used to construct average temperature, maximum temperature, and minimum temperature during the year, while data on pollution is used to generate many distinct indicators.[8] In addition, we also create features by aggregating to higher levels by taking means. For example, we calculate mean urbanity at the admin3 level for all countries, and create predictors with other variables in similar ways. By combining these features in different ways and across different levels of aggregation, we end up with several hundred different predictive features. While we include all of these features in the machine learning methods, we use lasso to select features for the EBP model, following others (Engstrom et al., 2022; Masaki et al., 2022; Newhouse et al., 2025).[9]

We pull the most recent data available for all countries. Different geospatial features have different time coverage, meaning we sometimes pull data from after the census. For example, our preferred land classification data only goes back to 2019, so we extract 2019 data for the 2012 Tanzania census. We recognize that this is not ideal and can result in additional stress-testing for the proposed methods (e.g., by introducing noise into the data).

[8] More concretely, we construct weather variables separately by month throughout the year of the census, maximum and minimum values throughout the same year, as well as climate-related variables that are long-run means and standard deviations from the 10 years prior to the census. We construct similar variables (except the long-run averages) for pollution.

[9] All three machine learning methods have their own regularization methods which make lasso redundant.

2.3 Linear Empirical Best Predictor Models

We utilize the *povmap* package in R to generate the Empirical Best Predictor (EBP) estimates.[10] This is an updated version of the *emdi* package (Kreutzmann et al., 2019), which implements the models described in Molina and Rao (2010) with additional features. We implement a "sub-area level model," in which we estimate the model at the sub-area level before aggregating to the area level for the final predictions. While we include descriptions of the other estimators in Appendix C, we discuss some details of the sub-area model, since it is one of the most common methods for small area estimation of welfare.

The subarea-level model is a model of the form:

$$G(y_{sar}) = \beta_1 X_{sar} + \beta_2 X_{ar} + \eta_{ar} + \varepsilon_{sar}, \tag{1}$$

where $G(y_{sar})$ is a transformation of outcome y for sub-area s in area a in region r,[11] X_{sar} is a vector of sub-area-specific geospatial features, X_{ar} is a vector of area-specific geospatial features – which may include region dummies – η_{ar} is an area-level random effect, assumed to be normally distributed and conditioned on the sample data, and ε_{sar} is a classical error term assumed to be normally distributed. When predicting assets, no transformation is used, implying that $G\left(y_{sar}\right) = y_{sar}$. When predicting poverty rates, the arcsin transformation is used, implying that $G\left(y_{sar}\right) = \arcsin\left({y_{sar}}^{0.5}\right)$.

Because the random effects are conditioned on the survey data, they are shrunk towards zero, with the shrinkage factor depending on the relative estimated variances of the random effect η and the ideosyncratic error term ϵ. After estimating the model's coefficients, estimates for areas are generated by calculating $E\left[G^{-1}\left(y_{sar}\right)\right]$, conditional on estimated parameters $\hat{\beta}, \hat{\sigma}^2_\eta$, and $\hat{\sigma}^2_\epsilon$. We include both survey weights and population weights, with the latter taken from WorldPop estimates rather than the censuses. The software calculates measures of uncertainty using 100 parametric bootstrap replications. For more details on the model, we refer readers to Tzavidis et al. (2018) and Molina and Rao (2010).

As with any regression model whose main goal is prediction, EBP models can be prone to overfitting. This is especially true in our case, where we have several hundred possible features from which to choose. To help prevent overfitting, we select features using lasso, implemented using the R package *glmnet* (Friedman et al., 2010). We select the optimal lambda using cross-validation.

[10] The package is a spin-off of the EMDI package developed by Ifeanyi Edochie and colleagues and available for download at: https://github.com/SSA-Statistical-Team-Projects/SAEplus.

[11] The level at which we are interested in predicting outcomes is the area, described above.

2.4 Cubist

The second prediction method that we evaluate is Cubist regression, which is closely related to M5 regression model trees and is derived from the work of Kuhn and Johnson (2013), Wang and Witten (1997), Quinlan (2014), and Quinlan (1992). We implement it in R with the *Cubist* package (Kuhn et al. 2022), using a procedure described in detail in Kuhn and Johnson (2013) and the publicly available source code. The input is a set of training data with a dependent variable and set of candidate independent variables. The output is a set of piecewise linear models. The procedure uses tree-based prediction methods to develop "rules", which correspond to leaves of the tree, and linear models are estimated for every rule. The user can set the number of rules or allow the algorithm to determine the optimal number of rules based on cross-validation. In short, the procedure estimates a set of linear models that are estimated on various subsets of the data, which are selected to maximize the accuracy of the predictions. Further details on the Cubist algorithm can be found in Appendix C.

2.5 XGBoost

The third estimator – Extreme Gradient Boosting – is a popular implementation of gradient boosted trees, commonly called XGBoost (Chen and Guestrin 2016). XGBoost develops a set of regression forests, which like the committees in the Cubist model sequentially predict residuals from the past regression. Appendix C contains further details on the estimation of the algorithm; we just summarize the material found in the online XGBoost documentation[12] as well as in the original paper by Chen and Guestrin (2016).

2.6 Boosted Regression Forests

The last method we compare is Boosted Regression Forests (BRF), implemented in the *GRF* package for R and described in the online documentation to that package as well as in Athey, Tibshirani, and Wager (2019). Boosted Regression Forests are very similar to XGBoost, in that both estimate a series of regression forests that successively predict the residuals from the previous round. However, BRF differs from XGBoost by using one subsample of the data to grow trees and another to generate predictions at the leaves of the tree, a procedure which is more theoretically sound. Each regression forest consists of a set of decision trees that the algorithm grows on randomly selected subsets of the data. Further details on BRF are in Appendix C.

[12] https://XGBoost.readthedocs.io/en/stable/tutorials/model.html

2.7 Uncertainty Estimates for ML Estimators

For EBP, we model welfare at the sub-area level, which is then aggregated up to the target area level. Because the model specification contains a random effect at the target area level, this procedure accounts for the hierarchical nature of the data. The procedure uses a parametric bootstrap to estimate uncertainty, drawing from the estimated distributions for the random effects and the error terms, and accounting for the arcsin transformation implemented for EBP. For the three ML estimators, we also estimate the models at the subarea level.

While Friedberg et al. (2020) prove a central limit theorem for local linear forests that allows for the construction of uncertainty estimates, variance estimation has not yet been implemented for Boosted Regression Forests. As an alternative, we propose a non-parametric bootstrap procedure that draws from previous work on residual bootstraps with hierarchical data (e.g. Luo and Lai (2021)).[13]

For subarea sa, consider the sample direct estimate of the outcome: $\hat{y}_{sa}^{direct}$. In addition, there is the prediction from the machine learning algorithm, $\hat{y}_{sa}^{ML}$. With these two estimates, we calculate subarea-specific "residuals" as:

$$\hat{R}_{sa} = \hat{y}_{sa}^{ML} - \hat{y}_{sa}^{direct}. \tag{2}$$

We can likewise calculate residuals at the area level, by aggregating $\hat{y}_{sa}^{ML}$ and $\hat{y}_{sa}^{direct}$ to the area, weighting by estimated population from WorldPop:

$$\hat{R}_{a} = \hat{y}_{a}^{ML} - \hat{y}_{a}^{direct}. \tag{3}$$

The proposed bootstrap continues in two steps. First, note that we can only calculate this residual for in-sample subareas and in-sample areas. After estimating predictions, we first calculate the residuals in Equation 2. Then, we randomly draw one residual from the vector $\hat{R}_{sa}$ for each in-sample subarea and add that residual to the prediction: $\hat{y}_{sa}^{ML} + \hat{R}_{sa}$. We do this sampling with replacement, for in- and out-of-sample subareas.

We now have adjusted subarea predictions for all subareas. We aggregate all of these predictions to the area level, using estimated population from WorldPop as weights. At the area level, we pursue a similar strategy, but this time we draw area-level residuals, with replacement, for all areas, regardless of sample status. We

[13] We estimate the EBP model using REML as the default. We apply weights using the weights option in the *lme* function of the *nlme* package and we rescale the weights to sum to the number of observations within group. When estimating BRF, the R *GRF* package does not allow for aggregate variance estimates.

repeat this residual bootstrap 1,000 times and the standard deviation of the estimate, which functions as our estimate of the standard error.

Importantly, the drawing of observations from the vectors of residuals is done by weighting based on aggregate sampling weights at the subarea and area level. As such, we are taking into account the fact that sampling is informative and some (sub) areas have a higher probability of inclusion than others. Since only in-sample (sub)areas are included in the calculation of the residual bootstrap vector, this method works when only a household survey is available.

The proposed bootstrap consists of the following steps:

1. Predict an outcome (asset index or poverty) using XGBoost, BRF, or Cubist.
2. Calculate subarea residuals for in-sample subareas by differencing the prediction and the direct estimate from the survey. Call this vector of residuals $\hat{R}_{sa} = \hat{y}_{sa}^{ML} - \hat{y}_{sa}^{direct}$.
3. Aggregate predictions to the area level, using estimated population from WorldPop as weights.
4. Calculate area residuals for in-sample subareas by differencing the prediction and the direct estimate from the survey. Call this vector of residuals $\hat{R}_{a} = \hat{y}_{a}^{ML} - \hat{y}_{a}^{direct}$.
5. With original subarea predictions, bootstrap with replacement from $\hat{R}_{sa}$ for all subareas, with sampling probability determined through aggregated sampling weights.
6. Aggregate these new predictions to the area level.
7. Bootstrap with replacement from $\hat{R}_{a}$ for all areas, with sampling probability determined through aggregated samplign weights.
8. Repeat steps five through seven 1,000 times, saving new area estimates after each replication.
9. Calculate percentiles across the 1,000 replications.

Although residuals can only be calculated for sampled sub-areas and areas, these are used to generate uncertainty estimates for both samples and non-sampled areas meaning that, unlike EBP – which conditions the random effect on the sample for in-sample areas – estimated confidence intervals are approximately equal in size for in-sample and out-of-sample areas

Importantly, the poverty rate is a variable bounded by zero below and by one above. In order to respect these restrictions throughout the process, we do not estimate the poverty rate in levels. Instead, we estimate an arcsin (square root) transformed poverty rate: $p_{sa}^{transformed} = sin^{-1}(\sqrt{p}_{sa})$. We carry over this transformation

throughout the entirety of the bootstrap procedure, only back transforming it at the end, in step 9.[14]

2.8 Evaluating Performance

We have access to unit-level census data for seven countries. This allows us to calculate true sampling distributions with the unit-level census data by simulating separate surveys and saving the results from each iteration. In all countries, we treat subareas as enumeration areas. For each country, we design a sampling scheme that mimics sampling of an actual household survey from that country. For example, we mimic the sampling scheme used in the fifth Integrated Household Survey (IHS5) for Malawi. In all countries, we design the sampling process in a two-stage manner, first stratifying, then drawing an "enumeration area" (which in our case is always the lowest level of aggregation available in the census), before finally drawing households.

We independently draw 100 separate surveys, predict our outcomes of interest with each method, and then evaluate the performance of the methods against the ground truth derived from the full unit-level census data. We calculate the following statistics, where i indexes areas, $\hat{y}$ refers to the predicted outcome for area i, and y_i^{truth}: refers to the true value for area i:

- Correlation: We calculate both the Pearson correlation coefficient, r, and the Spearman (rank) correlation coefficient, ρ. We present the means across all 100 independent samples:

$$\frac{1}{100}\sum_{s=1}^{100} r_s \text{ and } \frac{1}{100}\sum_{s=1}^{100} \rho_s \tag{4}$$

- Absolute deviation: This is defined for each area as $\left|\hat{y}_{si} - y_{si}^{truth}\right|$. We present the average across all areas and all simulations:

$$\frac{1}{100N}\sum_{s=1}^{100}\sum_{i=1}^{N} \left\|\hat{y}_{si} - y_{si}^{truth}\right\| \tag{5}$$

- Squared deviation: This is defined for each area as $\left(\hat{y}_{si} - y_{si}^{truth}\right)^2$. Similarly, we present the average across all areas and simulations:

$$\frac{1}{100N}\sum_{s=1}^{100}\sum_{i=1}^{N} \left(\hat{y}_{si} - y_{si}^{truth}\right)^2 \tag{6}$$

- Width of confidence interval: For the EBP estimates, this is derived from the estimated MSE $\hat{E}_{si}$ as

[14]We of course also back transform for the original point estimate.

follows:

$$CIW_{ebp} = \frac{3.92}{100N} \sum_{s=1}^{100} \sum_{i=1}^{N} \sqrt{\hat{E}_{si}} \tag{7}$$

- For the three ML estimators, the width is defined as the average, across areas, of the difference between the 95th and 5th percentiles of the bootstrap replications.

$$CIW_m = \frac{1}{100N} \sum_{s=1}^{100} \sum_{i=1}^{N} \left(\hat{y}_{si}^{p95} - \hat{y}_{si}^{p5} \right) \tag{8}$$

- Coverage rate: This is defined as $I\left(y_i^{truth} \in [CI_i^{lower}, CI_i^{upper}]\right)$, where $I(\cdot)$ is the indicator function and CI refers to the confidence interval for a given area. We calculate the proportion of areas with true values that fall within the confidence interval:

$$\frac{1}{100N} \sum_{s=1}^{100} \sum_{i=1}^{N} I\left(y_i^{truth} \in [CI_i^{lower}, CI_i^{upper}]\right) \tag{9}$$

- Area Under the Curve (AUC). For each country, we construct an average Receiver Operating Characteristic (ROC) curve for predicted area asset index values across the one hundred samples. To plot this curve, we calculated true positive rates (TPR) and false positive rates (FPR) for fifty quantiles q of the asset index distribution across areas, and take the average across samples. For method m and quantile q,

$$TPR_{m,q} = \sum_{s=1}^{100} \frac{\sum_{i=1}^{N} I\left(\hat{y}_{msi} < T_q^{ms}\right) * I\left(y_i^{truth} < T_q^{truth}\right)}{100 * \sum_{i=1}^{N} I\left(y_i^{truth} < T_q^{truth}\right)} \tag{10}$$

and

$$FPR_{m,q} = \sum_{s=1}^{100} \frac{\sum_{i=1}^{N} I\left(\hat{y}_{msi} < T_q^{sm}\right) * I\left(y_i^{truth} \geq T_q^{truth}\right)}{100 * \sum_{i=1}^{N} I\left(y_i^{truth} \geq T_q^{truth}\right)} \tag{11}$$

Where $\hat{y}_{msi}$ represents the predicted asset index associated with method m, simulation s, and area i. $I()$ is an indicator function T_q^{ms} is a "area poverty threshold" for quantile q, method m, and simulation s, defined as the q^{th} percentile of the predicted asset index distribution for method m and simulation s across areas. T_q^{truth} is the qth percentile of the true distribution of average asset indices across areas.

Thus, the TPR represents the average (across simulations) share of "poor" areas in the census that are correctly classified as poor by the predictions, while the FPR represents the average share of "non-poor" areas that were incorrectly predicted to be poor. After plotting 50 points, one for each quantile, we calculate the area under the curve (AUC), which is a summary measure of targeting accuracy for each prediction. AUC values typically range from a low of 0.5, in which case the prediction is no more accurate than a random guess, to a perfect score of 1.

- Poverty Targeting Simulations. Finally, to check the implications of using different estimates for targeting purposes, we perform targeting simulations based on predicted poverty rates in Malawi and Tanzania. The simulations are conducted for ten different population coverage thresholds ranging from 5 to 50 percent. At each threshold, predictions from each method are used to identify beneficiary areas that will be given simulated transfers – Traditional authorities in Malawi and Districts in Tanzania. Beneficiary areas are selected based on the predicted poverty rates obtained from each method, starting with the area predicted to be poorest, until the covered population equals the population threshold. Per capita expenditure is then increased by a constant amount, equal to ten percent of the poverty line, for all households in beneficiary areas. This simulates each household obtaining a transfer proportional to the size of their household.

  We then report the poverty gap (P1) for each method m and population coverage rate q, defined as:

$$PG_{mq} = \sum_{s=1}^{100} \frac{\sum_{h=1}^{H} I\left(y'_{qmsh} < Z\right) * \left(Z - y'_{qmsh}\right) * n_h}{100 * Z * \sum_{h=1}^{H} n_h} \tag{12}$$

  where H is the number of households in the census, n_h is the size of household h, Z is the poverty line, and y'_{qmsh} is the post-transfer per capita consumption for household h (in the census) for simulation s, method m, and population coverage rate q. Therefore,

$$y'_{qmsh} = y_h^{truth} + 0.1 * Z * I\left(\hat{y}_{msh} < T_{msq}\right) \tag{13}$$

  T_{msq} is the eligibility threshold for method m associated with population coverage rate q in simulation s. This is equal to the q^{th} percentile of the population-weighted distribution of $\hat{y}_{msi}$, the area-level predictions of the asset index generated by method m in simulation s. This ensure that q percent of the population is covered by the hypothetical transfer program, regardless of the prediction method used to identify beneficiary areas.

For the poverty targeting simulations, we examine impacts on the poverty gap rather than the headcount poverty rate (P0) because targeting areas based on their poverty headcount rates maximizes the impact of a transfer program on the poverty gap (Besley and Kanbur, 1991; Kanbur, 1986). Reporting the headcount poverty rate, on the other hand, could give a misleading picture of targeting effectiveness by rewarding methods that successfully identify households whose predicted consumption is just under the poverty line and would therefore exit poverty as a result of the transfer.

A key difference between the poverty targeting measure and all others previously considered is that it is averaged across individuals rather than areas. This implicitly weights areas according to their population, rewarding methods that more accurately estimate poverty rates for areas that have larger population. Weighting areas according to their population during model estimation could yield estimates better suited for the purpose of minimizing this indicator, and would be a useful topic for further research.

A final caveat is useful. The final performance of a specific ML method may depend on a combination of different factors related to data sources (including data-specific features), prediction targets, the prediction method, and performance metrics. While one method may work better for a specific application or according to a specific evaluation metric, it may perform worse for another. To better focus the comparison, we present in Table 2 a general comparison of the key features of the four methods that we use in this paper. To compare their performance, we use the same prediction targets (i.e., imputing into census data without any assets or poverty data) and performance metrics (as discussed above). However, they differ by design regarding the functional form (including whether random effects are employed), (assumed) data generation process, objective function, and other features such as how they select the predictor variables and whether they use different subsamples of the data in growing trees and making predictions. Finally, these ML methods offer various degrees of transparency and ease of use.

3 Results

We first examine the accuracy of the four candidate estimators, starting with basic average accuracy statistics.[15] Table 4 presents two separate indicators of accuracy: pearson and spearman (rank) correlations. The values are averages across all areas and the 100 samples.

Although the relative performance of different methods varies across cases, consistent patterns emerge. When looking at average Pearson correlations, XGBoost and Cubist nearly always outperform traditional EBP, which has been the workhorse of small area prediction for decades. In all nine cases (seven for assets and two

[15]Section 3.3 considers alternative measures of accuracy

for poverty), XGBoost improves upon EBP. In only one case, assets in Burkina Faso, does EBP improve upon cubist regression. The performance of BRF relative to EBP is more mixed, with EBP notably outperforming BRF when predicting the asset index in Burkina and poverty in Tanzania and giving comparably accurate estimates for the asset index in Malawi and Sri Lanka. On the other hand, BRF estimates are more accurate than EBP estimates for assets in Mozambique and Tanzania, as well as for poverty in Malawi. When comparing XGBoost and Cubist regression, XGBoost generally produces the most accurate estimates, although Cubist regression is more accurate when predicting assets in Tanzania.

When looking at averages across countries, the rankings across methods are the same for both assets and poverty: XGBoost is the most accurate on average, followed by Cubist, BRF, and EBP. The magnitude of the average differences are notable, as XGBoost is approximately six percentage points more accurate than EBP for assets on average and around four percentage points more accurate than EBP for poverty when averaging across the two countries with poverty estimates. These general patterns remain when looking at Spearman rank correlations instead of Pearson correlations.

Rank correlations are an important measure of accuracy because they reflect targeting accuracy, or the ability to discern the poorest areas. However, correlations do not typically capture bias, in the sense that correlations are unchanged when a constant is added to all predictions. Because of this, Table 5 examines deviations from truth, which captures bias. When looking at absolute and squared deviations, XGBoost and Cubist again substantially outperform EBP and BRF, at least on average. XGBoost has lowest mean absolute error in five of the nine cases and the lowest mean squared error in another six cases. EBP has the lowest errors in one case (Madagascar assets) and Cubist has the lowest errors in several cases. When comparing XGBoost and EBP, some of the differences are quite large. For example, in Malawi (assets), the MAE for XGBoost is 20 percent lower than that for EBP, similar to the results for poverty in the same country.

Likewise, comparing the values for MSE, XGBoost is 30 percent lower for assets and around 22 percent lower for poverty in Malawi. Looking at the averages, both Cubist and XGBoost outperform EBP for assets and poverty. This difference is particularly marked for poverty, with Cubist, in particular, greatly reducing the MSE relative to EBP and XGBoost also providing a substantial reduction. These large differences are driven by the fact that, when EBP performs better, it does so only slightly, whereas the two ML methods sometimes perform much better than EBP.

Table 4: Correlations across countries and methods

	Pearson				Spearman			
	EBP	Cubist	XGB	BRF	EBP	Cubist	XGB	BRF
Panel A: Assets								
Burkina Faso	0.743	0.734	0.850	0.694	0.715	0.698	0.812	0.683
Madagascar	0.875	0.907	0.910	0.906	0.795	0.847	0.851	0.849
Malawi	0.664	0.768	0.844	0.664	0.658	0.794	0.848	0.741
Mozambique	0.897	0.921	0.917	0.922	0.781	0.812	0.803	0.827
Sri Lanka	0.919	0.937	0.935	0.916	0.884	0.911	0.911	0.908
Tanzania	0.868	0.911	0.899	0.895	0.804	0.862	0.842	0.838
Vietnam	0.868	0.890	0.904	0.866	0.873	0.887	0.902	0.863
Average	0.833	0.867	0.894	0.838	0.787	0.830	0.853	0.816
Panel B: Poverty								
Malawi	0.786	0.865	0.851	0.737	0.786	0.841	0.839	0.710
Tanzania	0.836	0.874	0.858	0.794	0.852	0.889	0.872	0.832
Average	0.811	0.870	0.854	0.766	0.819	0.865	0.855	0.771

Note: The first four columns present the simple mean of pearson correlation across simulations for each country. The last four columns present the simple mean of spearman correlation across simulations for each country. The results are based on 100 simulations for each country and method.

Table 5: Accuracy across countries and methods

	Mean absolute error				Mean squared error			
	EBP	Cubist	XGB	BRF	EBP	Cubist	XGB	BRF
Panel A: Assets								
Burkina Faso	0.303	0.285	0.237	0.466	0.143	0.119	0.081	0.273
Madagascar	0.320	0.395	0.336	0.355	0.152	0.197	0.154	0.163
Malawi	0.440	0.369	0.348	0.462	0.373	0.294	0.254	0.532
Mozambique	0.278	0.198	0.200	0.191	0.115	0.065	0.067	0.066
Sri Lanka	0.159	0.134	0.135	0.164	0.043	0.031	0.033	0.058
Tanzania	0.413	0.345	0.336	0.366	0.224	0.156	0.154	0.179
Vietnam	0.236	0.215	0.204	0.228	0.106	0.092	0.076	0.102
Average	0.307	0.277	0.257	0.319	0.165	0.136	0.117	0.196
Panel B: Poverty								
Malawi	0.180	0.123	0.142	0.180	0.047	0.028	0.037	0.049
Tanzania	0.085	0.069	0.073	0.090	0.015	0.009	0.010	0.013
Average	0.132	0.096	0.108	0.135	0.031	0.019	0.024	0.031

Note: The first four columns present the simple mean of the mean absolute error (MAE) across simulations for each country. The last four columns present the simple mean of the mean squared error (MSE) across simulations for each country. The results are based on 100 simulations for each country and method.

So far, EBP is the only type of linear model we have evaluated. A popular alternative method was developed in Elbers et al. (2003) and is widely known as ELL. A key difference between EBP and ELL is that ELL does not condition the estimated random effect on the sample data. Partly because of that, ELL does not assume that the error terms are distributed normally. Appendix Table A3 compares Pearson and Spearman correlations for EBP and ELL. In all cases except assets in Tanzania, EBP outperforms ELL. On average, the benefit from using EBP rather than ELL is moderate, in terms of correlation with the true estimates, equal to 2.8 percentage points for assets and one percentage point for poverty. In two cases, Burkina Faso and Vietnam, the differences are larger, at 9.2 and 6.3 percentage points, respectively. Overall, in these contexts, EBP tends to generate more accurate estimates than ELL.

Accuracy of the prediction, however, is not the only measure of concern. It is also important to accurately estimate measures of uncertainty, since these are often used to determine whether the estimates are sufficiently reliable to publish publicly. Machine learning methods have been consistently shown to generate accurate point estimates, but are less amenable to estimating uncertainty. Table 6 presents two key statistics related to uncertainty, with the uncertainty measures calculated using a weighted two-stage residual block bootstrap, as described in section 2.7. The first statistic is the coverage rate, which shows how often the true value (from the census) lies within the estimated confidence intervals for a given prediction. We calculate coverage rates for 95% confidence intervals, meaning that if the point and uncertainty estimates are accurate, the coverage rate should be approximately 0.95. The second statistic is the total width of the confidence interval, which is an indicator of the extent to which estimated uncertainty varies across estimation methods.

Coverage rates for the four estimators vary quite a bit depending on the context and the method. On average, BRF is associated with the highest coverage rates, averaging 88.4 percent for assets and 88.8 percent for poverty. This is partly because BRF is less accurate than the other ML methods on average, leading to wider confidence intervals due to the use of a residual bootstrap. For assets, XGBoost has the second-highest coverage rates at 87.2 percent, followed by Cubist at 84.7 percent. For poverty, the ranking is the same, as XGBoost has a colverage rate of 85 percent with Cubist following at 82.3 percent. Coverage rates vary by context, however. Uncertainty estimates when predicting assets in Tanzania, for example, appear to be systematically underestimated for all estimators, especially for EBP at 39.1 percent. Figure A2 in the Appendix shows that part of the reason for this may be a systematic underestimation of poverty across the entire distribution, leading to decent correlation but poor coverage rates. We also note that the Tanzania census is from 2012, while some of the geospatial features such as land classification are from 2019. As such, we hesitate to put too much weight on the results from Tanzania. The coverage rate for Sri Lanka is also low for EBP (68.4 percent). We encourage further research to better understand why coverage rates are

particularly low in these cases and why all four methods seem to follow similar patterns in this regard.

In addition to coverage rates, Table 6 also presents the average width of confidence intervals ("CI width"). As expected, the results demonstrate the trade-off between estimated precision, accuracy, and coverage rates. BRF, despite being less accurate than XGBoost and Cubist on average, consistently yields the largest confidence intervals, which explains its higher coverage rates. When estimating assets, average CI widths are 1.33 for BRF, as opposed to 0.97, 1.04, and 1.06 for XGBoost, Cubist, and EBP, respectively, though the EBP mean is pulled down by the misleadingly low width in Malawi, Madagascar, and Sri Lanka. We note that the higher coverage rates for XGBoost and Cubist come with lower estimated confidence intervals, in general. In other words, at least in the current context, these two ML methods combined with our proposed residual bootstrap procedure calculate smaller and more accurate confidence intervals for assets than EBP. We see generally similar patterns of the ML methods when looking at poverty (Panel B). Both XGBoost and Cubist consistently outperform EBP in terms of both coverage rates and the width of the confidence interval.

3.1 In-Sample and Out-of-Sample Estimates

We next break out measures of accuracy and precision separately for sampled and non-sampled areas. This is important because prediction into non-sampled areas can be less accurate than prediction into sampled areas, due to bias in the estimated model parameters.[16] Since the share of target areas that are sampled can vary in practical applications, in-sample and out-of-sample performance may be more relevant in particular contexts. In each of the 100 samples per country, we randomly selected subareas, with probability proportional to size, and then randomly selected households from within each subarea. We define an area as being "in sample" if at least one subarea is sampled from within that area.

In general, there are three reasons why out-of-sample predictions could be less accurate than in-sample predictions. The first is that the model is overfit, although this seems unlikely given the use of regularization methods. For example, when implementing EBP, models are selected using LASSO to avoid overfitting. Meanwhile, the ML estimators employ different methods to help avoid overfitting, such as estimating regularized objective functions in the case of XGBoost and Cubist regression, and using a random subset of data and predictors across trees and splits, in the case of BRF. Another more likely explanation is that smaller areas – which are much less likely to appear in the sample – differ systematically from sampled areas, leading to bias when extrapolating predictions. Third, for EBP estimates in particular, out-of-sample estimates are less accurate due to the absence of sample data to condition on (Tzavidis et al., 2018).

[16] Pfeffermann and Sverchkov (2007) and Newhouse et al. (2025) demonstrate this for EBP models. We know of no evidence on out-of-sample predictions for small area estimates when using BRF, Cubist, or XGBoost.

Table 6: Uncertainty statistics across simulations

	Coverage rate				Width of CI			
	EBP	Cubist	XGB	BRF	EBP	Cubist	XGB	BRF
Panel A: Assets								
Burkina Faso	0.928	0.730	0.917	0.899	1.331	0.811	0.951	1.552
Madagascar	0.691	0.953	0.947	0.974	0.799	1.453	1.301	1.596
Mozambique	0.863	0.711	0.730	0.756	1.778	0.858	0.828	1.414
Sri Lanka	0.684	0.948	0.923	0.954	0.731	1.017	0.923	1.109
Vietnam	0.913	0.904	0.873	0.915	0.667	0.566	0.521	0.834
Malawi	0.391	0.699	0.735	0.710	0.701	0.912	0.941	0.985
Tanzania	0.973	0.982	0.976	0.983	1.444	1.645	1.340	1.809
Average	0.777	0.847	0.872	0.884	1.064	1.037	0.972	1.329
Panel B: Poverty								
Malawi	0.743	0.778	0.862	0.855	0.477	0.395	0.492	0.614
Tanzania	0.874	0.868	0.837	0.920	0.311	0.283	0.273	0.426
Average	0.809	0.823	0.850	0.888	0.394	0.339	0.382	0.520

Note: The first four columns present the coverage rate across simulations for each country. The coverage rate is defined as the proportion of confidence intervals that contain the true value, derived from the census. The last four columns present the average width of the confidence interval across simulations for each country. The results are based on 100 simulations for each country and method.

We start with accuracy statistics in Table 7. The first four columns of Table 7 include in-sample areas only, while the last four columns include out-of-sample areas. Statistics are means based on all 100 independent samples, similar to previous results. Looking at in-sample areas, on average XGBoost is the most accurate for assets on average, followed by Cubist, and finally by BRF and EBP. For poverty, Cubist is slightly more accurate than XGBoost on average, followed by BRF and EBP. When considering both assets and poverty, Cubist is more accurate than XGBoost in five of the nine cases, but the differences are never greater than 5.2 percentage points. On the other hand, XGBoost produces much more accurate estimates of assets than Cubist in two cases: Burkina Faso (by 9.8 pp) and Mozambique (by 5.7 pp). EBP and BRF are always less accurate than cubist or XGBoost with the exception of Madagascar, where BRF is slightly more accurate than Cubist.

The right portion of Table 9 provides accuracy estimates for non-sampled areas. As expected, out-of-sample estimates are consistently less accurate than in-sample estimates. On average, XGBoost estimates are most accurate for assets, while Cubist is the most accurate for poverty. The relatively poor performance of EBP in out-of-sample areas is consistent both with the linear nature of the model, given that the sampled areas are systematically different than the non-sampled areas, as well as the lack of sample data on which to condition the random effects.

Overall, a consistent pattern emerges both in and out-of sample. XGBoost and Cubist are generally the most accurate, except for Sri Lanka and Mozambique where out-of-sample BRF predictions are most accurate by a small margin. Meanwhile, either EBP or BRF is always the least accurate – sometimes by large margins. The relative performance of XGBoost and Cubist varies across countries and outcomes. For example, XGBoost estimates are much more accurate in Burkina Faso and Mozambique, but are slightly less accurate than Cubist for assets and poverty in Malawi. For out of sample areas, XGBoost and BRF suffer the smallest drop off in accuracy; out-of-sample correlations for assets are less than 8 percentage points lower, while that number is 12 pp for EBP and 10 pp for Cubist.

Finally, to further investigate the differences between in and out-of-sample estimates, we examine how the out-of-sample accuracy penalty changes for different methods when area fixed effects are included. Including area fixed effects controls for fixed characteristics of areas, meaning that the remaining difference in accuracy is attributable solely to these areas being excluded from the sample. Table 8 presents a set of regressions where we estimate the difference in absolute deviation between areas that were included or excluded from the sample. In each pair of columns, the first column includes only simulation fixed effects, while the second column we also include area fixed effects that restrict identification to within-area changes in sample status across simulations.

Table 7: Correlations across countries and methods

	In sample				Out of sample			
	EBP	Cubist	XGB	BRF	EBP	Cubist	XGB	BRF
Panel A: Assets								
Burkina Faso	0.799	0.778	0.876	0.698	0.638	0.649	0.771	0.692
Madagascar	0.895	0.917	0.921	0.919	0.862	0.900	0.902	0.898
Malawi	0.712	0.846	0.903	0.818	0.649	0.747	0.830	0.624
Mozambique	0.950	0.960	0.958	0.955	0.846	0.883	0.876	0.888
Sri Lanka	0.931	0.944	0.942	0.921	0.832	0.880	0.875	0.883
Tanzania	0.899	0.926	0.917	0.913	0.762	0.867	0.837	0.829
Vietnam	0.878	0.893	0.906	0.868	0.614	0.595	0.741	0.698
Average	0.866	0.895	0.918	0.870	0.743	0.789	0.833	0.787
Panel B: Poverty								
Malawi	0.817	0.924	0.872	0.883	0.782	0.840	0.843	0.668
Tanzania	0.899	0.921	0.907	0.840	0.669	0.727	0.697	0.663
Average	0.858	0.922	0.890	0.861	0.725	0.783	0.770	0.665

Note: The first four columns present the simple mean of pearson correlations for in-sample areas across simulations for each country. The last four columns present the simple mean of pearson correlations for out-of-sample areas across simulations for each country. An area is defined as in-sample if at least one of that area's enumeration areas is in the sample on a given simulation.

Table 8: Difference in accuracy within areas: in-sample vs. out-of-sample

	EBP		Cubist		XGB		BRF	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Assets								
Burkina Faso	-0.088***	-0.106***	-0.074***	-0.051***	-0.055***	-0.016***	-0.089***	-0.018***
	(0.013)	(0.009)	(0.011)	(0.003)	(0.009)	(0.002)	(0.014)	(0.002)
Madagascar	-0.017***	-0.010***	-0.021***	-0.003**	-0.020***	-0.004***	-0.024***	-0.005***
	(0.005)	(0.001)	(0.005)	(0.001)	(0.005)	(0.001)	(0.004)	(0.001)
Malawi	-0.164***	-0.058**	-0.139***	-0.057**	-0.118***	-0.040*	-0.239***	-0.016**
	(0.027)	(0.028)	(0.028)	(0.023)	(0.026)	(0.021)	(0.044)	(0.006)
Mozambique	0.016***	-0.010***	0.020***	-0.010***	0.031***	-0.008***	0.039***	-0.008***
	(0.005)	(0.001)	(0.005)	(0.001)	(0.005)	(0.001)	(0.006)	(0.0009)
Sri Lanka	-0.083***	-0.041***	-0.053***	-0.026***	-0.060***	-0.024***	-0.067***	-0.012***
	(0.010)	(0.005)	(0.007)	(0.004)	(0.008)	(0.003)	(0.013)	(0.002)
Tanzania	-0.064***	-0.011	-0.052***	-0.004	-0.040*	0.002	-0.035	0.002
	(0.018)	(0.008)	(0.017)	(0.006)	(0.022)	(0.004)	(0.024)	(0.004)
Vietnam	-0.298***	-0.129***	-0.145**	-0.056***	-0.150***	-0.071***	-0.189***	-0.055***
	(0.112)	(0.021)	(0.059)	(0.012)	(0.030)	(0.011)	(0.029)	(0.006)
All	-0.040***	-0.024***	-0.032***	-0.013***	-0.024***	-0.010***	-0.042***	-0.008***
	(0.005)	(0.002)	(0.004)	(0.001)	(0.004)	(0.001)	(0.006)	(0.0006)
Panel B: Poverty								
Malawi	-0.066***	-0.013*	-0.046***	-0.019**	-0.042***	-0.006*	-0.084***	-0.011***
	(0.008)	(0.008)	(0.008)	(0.009)	(0.006)	(0.003)	(0.009)	(0.003)
Tanzania	-0.050***	-0.014***	-0.040***	-0.015***	-0.042***	-0.019***	-0.019**	-0.009***
	(0.012)	(0.003)	(0.010)	(0.003)	(0.010)	(0.003)	(0.007)	(0.002)
All	-0.062***	-0.014***	-0.044***	-0.017***	-0.042***	-0.012***	-0.068***	-0.010***
	(0.007)	(0.004)	(0.006)	(0.005)	(0.005)	(0.002)	(0.007)	(0.002)
Fixed effects:								
Simulation	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Area	No	Yes	No	Yes	No	Yes	No	Yes

Note: Standard errors are clustered at the simulation and area level. Each cell is a separate regression, where the dependent variable is absolute deviation from truth for a given area/estimator and the independent variable is an indicator for whether the area appears in the sample in a given simulation. The all rows estimate regressions with all countries simultaneously (separately for assets and poverty in Panel A and Panel B, respectively).
* p<0.1 ** p<0.05 *** p<0.01

The first column in each pair shows that the in-sample areas have higher accuracy (lower absolute deviations) than the out-of-sample areas. The differences for EBP range from approximately 0.017 to 0.298. There are several important differences in the second column when area fixed effects are included. First, the inclusion of area fixed effects generally decreases the difference in accuracy based on sample status. In other words, controlling for differences in characteristics across areas substantially shrinks the out-of-sample penalty for accuracy. Some of these decreases are quite remarkable; for example, the penalty in Tanzania for assets almost completely disappears for all four methods after taking into account area fixed effects.

A second notable finding from Table 8 is that, without taking into account area fixed effects, the out-of-sample penalty on average for assets is largest for BRF and EBP (0.042 and 0.040, respectively), followed by Cubist (0.032) and XGBoost (0.024). When controlling for area effects, however, differences in average accuracy across methods are small, ranging from 0.008 for BRF to 0.024 for EBP. This suggests that, among the methods considered, the ML methods are the best at predicting into out of sample areas that are systematically different than those included in the sample. EBP may be hampered in this respect by the assumption of a linear functional form, in addition to conditioning on the sample data. Overall patterns are similar for poverty with respect to the comparison of the ML methods to EBP, though the penalty is lower for EBP than for Cubist after taking into account area fixed effects.

Table 9 turns to uncertainty estimates and examines how coverage rates vary for in and out of sample areas. For the three machine learning methods, coverage rates are higher for in-sample than out-of-sample areas. For sampled areas, the three ML methods tend to yield higher coverage rates than EBP, with BRF generally exhibiting the highest coverage rates, followed closely by EBP and then Cubist. EBP performs particularly poorly when estimating the uncertainty of asset estimates in Tanzania. Low coverage rates can reflect either inaccurate point estimates or consistent underestimates of uncertainty. However, in these cases, the low coverage rates appear to result from the underestimation of uncertainty rather than inaccuracy, as Table 7 shows that EBP estimates of assets in Tanzania are only slightly less accurate than the other three methods. For non-sampled areas, the average coverage rates for asset estimates are again highest for BRF at 85.2 percent, followed closely by XGBoost at 83.3 percent, with EBP and Cubist showing similar results of 80.3 and 80.6 percent. Average coverage for out of sample poverty rates has a wider variance across methods, however, ranging from 89.8 percent for BRF to 73.1 percent for Cubist.

One disadvantage of the random effects block bootstrap is that it can only utilize data from sampled subareas and areas, and therefore cannot distinguish between sampled and non-sampled areas when estimating uncertainty. Despite the presence of sample weights, the sample systematically under-represents less populous areas and extrapolations into non-sampled areas are less accurate than estimates for sampled areas. Yet this

Table 9: Coverage across countries and methods

	In sample				Out of sample			
	EBP	Cubist	XGB	BRF	EBP	Cubist	XGB	BRF
Panel A: Assets								
Burkina Faso	0.903	0.768	0.938	0.916	0.968	0.670	0.884	0.872
Madagascar	0.673	0.959	0.953	0.976	0.702	0.949	0.944	0.972
Malawi	0.777	0.766	0.775	0.833	0.917	0.676	0.702	0.707
Mozambique	0.577	0.937	0.901	0.938	0.721	0.952	0.930	0.960
Sri Lanka	0.919	0.920	0.892	0.928	0.876	0.806	0.762	0.839
Tanzania	0.343	0.723	0.754	0.728	0.519	0.637	0.685	0.662
Vietnam	0.974	0.983	0.976	0.983	0.920	0.955	0.926	0.951
Average	0.738	0.865	0.884	0.900	0.803	0.806	0.833	0.852
Panel B: Poverty								
Malawi	0.757	0.857	0.899	0.952	0.734	0.727	0.839	0.794
Tanzania	0.878	0.910	0.883	0.933	0.865	0.753	0.715	0.886
Average	0.818	0.884	0.891	0.942	0.799	0.740	0.777	0.840

Note: The first four columns present the average coverage rate (using 95-percent confidence intervals) for in-sample areas across simulations for each country. The last four columns present the the average coverage rate for out-of-sample areas across simulations for each country. An area is defined as in-sample if at least one of that area's enumeration areas is in the sample on a given simulation.

added source of model error is not reflected in uncertainty estimates. This partly contributes to the pattern of lower coverage rates out of sample than in-sample observed for the machine learning methods in Table 9. There are possible alternatives, such as using a parametric bootstrap – which is how EBP calculates uncertainty – or explicitly modeling heteroscedasticity, as in Elbers et al. (2003). However, the random effect block bootstrap does reasonably well in the simulations reported above, with out-of-sample coverage rates averaging 85 percent for BRF and 84 percnet for XGBoost for assets, which is higher than the 80 percent average coverage rate for EBP. For poverty, however, the average for EBP exceeds that for XGboost by a couple of percentage points.

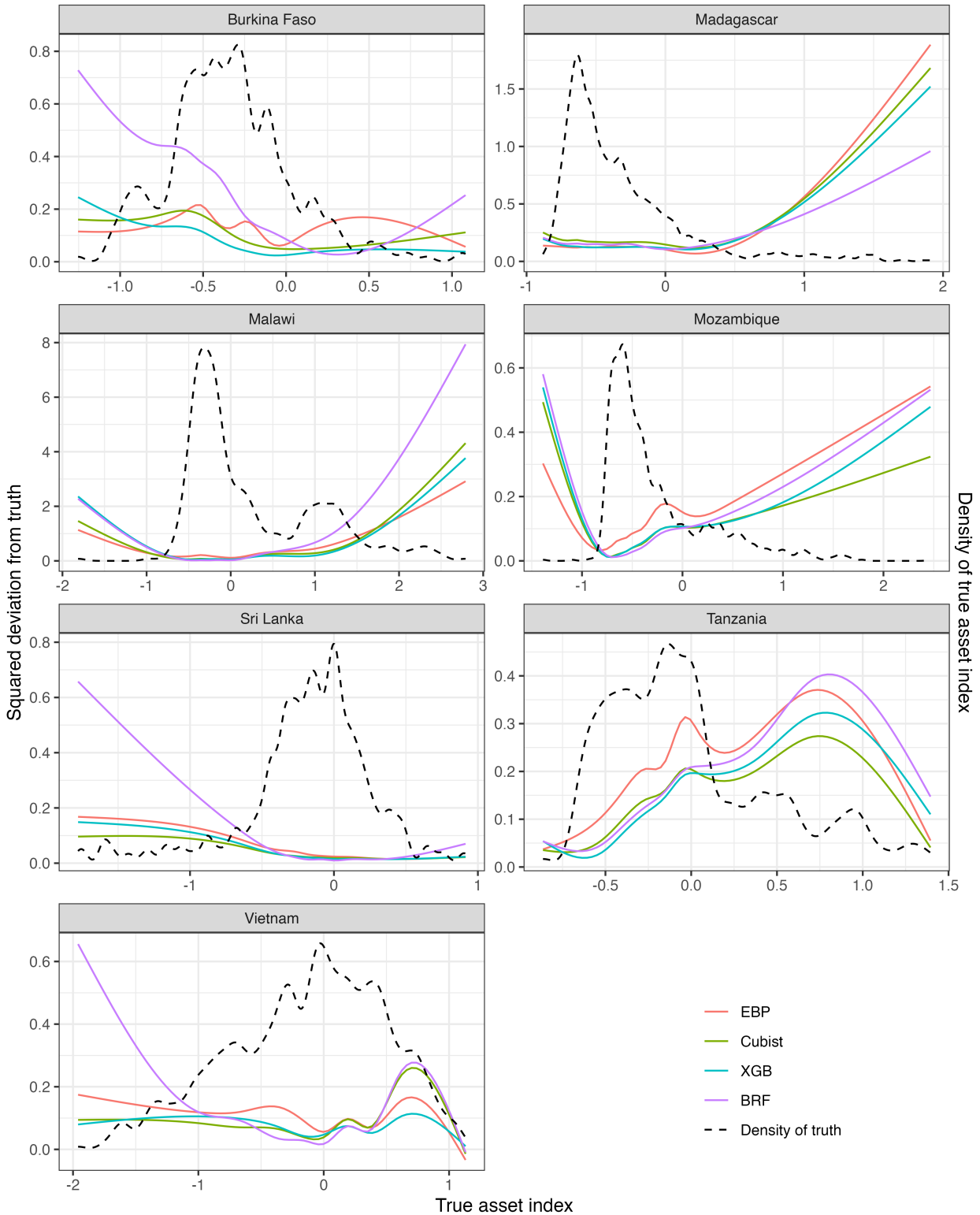
3.2 Accuracy for Poorer and Richer Areas

The previous sections examined accuracy both overall and separately for in- and out-of-sample areas. Figure 1 shows how accuracy varies for poorer and wealthier areas. The left y-axis represents mean squared deviation across samples for each area. The dotted black line shows the kernel density estimate of the true value, represented on the right y-axis. Several patterns emerge. First, predictions are most accurate in areas where the sample density is highest, where the dotted black lines are higher.

Second, there are notable differences in how accuracy varies across the distribution by method. BRF, for example, gives particularly inaccurate predictions at the tails in several cases, such as Burkina Faso, Malawi, Sri Lanka and Vietnam. EBP also predicts poorly in the right tail in Madagascar and Mozambique, but is also notably more accurate than several of the ML methods in many of the left tails. On the other hand, EBP is less accurate in the denser parts of the distribution in Mozambique, Tanzania, and Vietnam. Thus, the relatively poor performance of BRF appears to reflect poor predictive performance at the tails where there is little data, perhaps due to the use of different subsets of the data for tree building and leaf estimation. Meanwhile, the linear functional form associated with EBP, while also problematic at the tails in some cases, also appears to reduce accuracy through much of the distribution in cases such as Mozambique and Tanzania.

Third, areas of low density in the sample do not always correspond to smaller areas. In Madagascar and Mozambique, for example, the fewest observations are in the upper tail of the distribution, where the asset index is highest. These areas correspond to larger, more urban areas, with high populations. This also means that these areas are much more likely to appear in the sample (since probability of inclusion is related to size). Nevertheless, we see the least accurate predictions in these areas because they are systematically different from most areas in the sample. This raises the possibility of revising sampling strategies to oversample the tails of the distribution, in terms of the outcome of interest, in addition to more populous areas. More generally, we encourage future research to explore alternative sampling strategies as data fusion techniques grow in popularity and ease of use.

Figure 1: Deviations from truth across the distribution (assets)



Note: Figures show lowess-smoothed plots of squared deviation relative to truth, across all 100 simulations, separately for all seven countries' asset indices. The dashed line shows the density of the true asset index for each country.

Figure A2 in the appendix shows predicted-true plots for all nine cases. In many cases, as expected, predictions are overestimated towards the left and underestimated towards the right, with the exception of Tanzania assets where predictions are consistently underestimated except at the smallest values. Overall, it is difficult to make out consistent patterns, except in a few cases. In Burkina Faso, XGBoost and Cubist predict much more accurately than BRF, and the same is true towards the upper tail for assets in Malawi and the lower tail for assets in Vietnam. For poverty in Malawi, XGBoost and Cubist give much more accurate predictions than EBP and BRF both at the lower and upper tails.

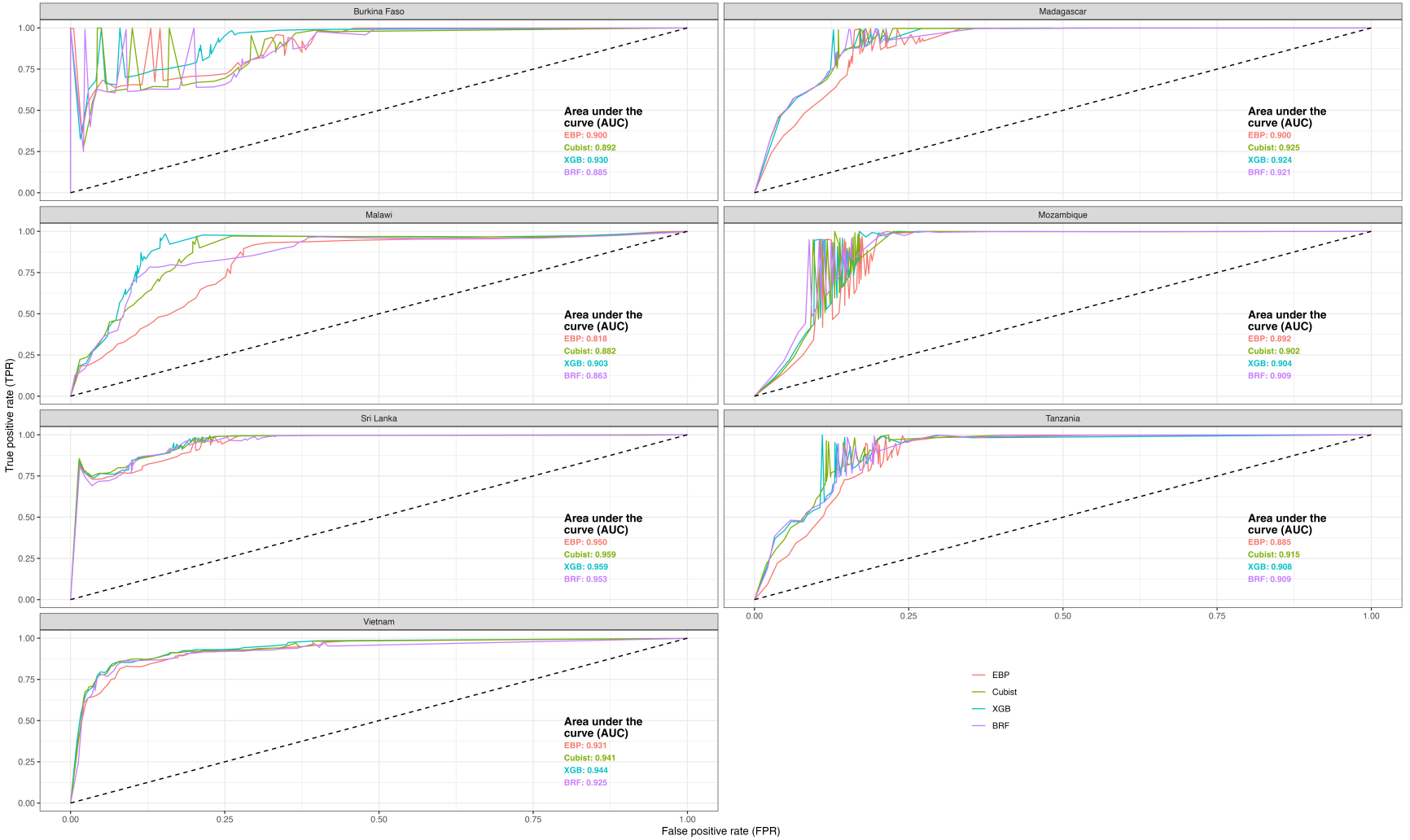
3.3 Alternative measures of accuracy

The previous section only considered a limited set of accuracy measures: pearson and rank correlations, mean absolute error, and mean squared error. But it is often unclear how these translate into targeting outcomes, which are a common application of small area estimation. To investigate this further, this section examines an alternative measure of targeting accuracy for asset predictions, the area under the curve (Hanna and Olken, 2018). In addition, we simulate the poverty impacts of a hypothetical transfer program based on the poverty estimates for Malawi and Tanzania obtained by different methods.

Figure 2 shows the estimates for area under the curve (AUC) for assets in all seven countries. The curve in question is a receiver operator characteristic (ROC) curve, applied to target areas, as described in section 2. Figure 2 indicates that either XGBoost or Cubist has the highest area under the curve in all cases. EBP and BRF are typically associated with the lowest score, though which is lower depends on the country. In summary, the AUC results confirm that XGBoost and Cubist deliver better targeting outcomes than EBP and BRF.

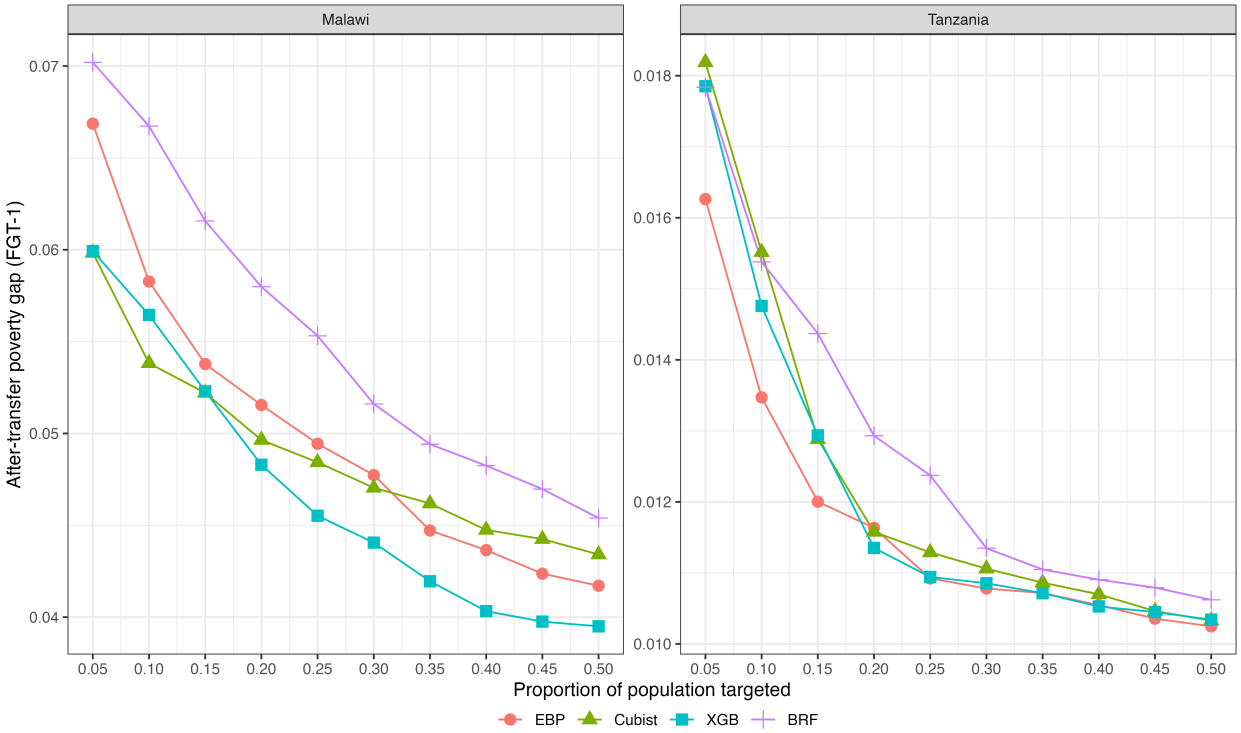
Finally, we show the results of a poverty targeting simulation in Figure 3. The y-axis represents the poverty gap (P1) associated with basing targeted geographic transfers on different prediction methods. In Malawi, basing transfers on the XGBoost rankings reduces poverty the most irrespective of the share of the population targeted, with the only exception being 20 percent, where Cubist performs slightly better. This is followed by Cubist, until 35 percent of the population is targeted, at which point EBP does slightly better at reducing the poverty gap. Ranking areas by BRF is consistently least effective. For Tanzania, the results are different: Ranking areas using EBP estimates reduces poverty severity the most until 20 percent of the population is covered, at which point XGBoost does slightly better, though EBP and XGBoost then flip spots repeatedly over the higher values.

Figure 2: Asset index targeting



Note: The figure plots the True Positive Rate (Y-axis) against the False Positive Rate (x-axis) for every percentile of the mean asset score distribution . The values are simple means across 100 simulations. For example, a value of 0.05 on the x-axis indicates that the false positive rate is 5 percent for that particular percentile threshold, and the value of the y-axis is true positive rate at that percentile threshold. The dashed lined indicates expected results if ordering were completely randomized.

Figure 3: Poverty targeting results



Note: For Malawi and Tanzania, we rank areas based on their estimated poverty rate. We then make transfers, equal to 10-percent of the per-capita poverty line, to X% of the population, shown on the x-axis. We rank areas by poverty rates and make transfers until X% of the population is reached. The y-axis presents the resulting, country-level poverty gap rates after the transfers, as a simple average over all 100 simulations.

There are important caveats to these results. First, this targeting exercise transfers 10 percent of the per-capita poverty line. Higher (or lower) transfers could lead to different results as the decrease in the poverty gap hinges on the distribution of households just below the poverty line. Second, we assume that this transfer is given to *all* households in a targeted area, meaning non-poor households also receive the transfers. An alternative option would be to target based on sub-area estimates instead of area estimates (which is feasible with geospatial data) or means testing households in targeted areas (which is not feasible without large expenditures or recent household data on all households). Finally, this ranking is done using the official poverty rates from each country. Another alternative would be to target a different poverty rate (e.g. 25-percent headcount poverty in Malawi instead of the 50-percent current rate).

3.4 Which predictors are important for prediction?

This section turns to briefly examining the "importance" of different types of geospatial predictors in predicting asset wealth. Doing so can shed light on which geospatial indicators are most important to include as candidate predictors in models. We assess the importance of predictors in two ways. First, we examine the share of

samples for which different types of predictors are selected when implementing LASSO to select variables for the EBP model. Predictors are grouped into eight categories: Weather and climate variables, Vegetation (specifically the Normalized Difference Vegetation Index), Land Cover Classification (LC), modeled population estimates, pollution indicators, distance to different locations, and latitude and longitude (GPS). The results are shown in Table 10.

When predicting assets, Pollution, Distance, and Land Cover are selected most frequently, more than 96 percent of the time. Weather and vegetation are also quite common, at 90 and 88 percent, respectively. GPS and population are less common, as they are only selected 58 and 52 percent of the time, respectively. Patterns are fairly similar for poverty prediction. Vegetation, Land cover, and distance to major cities are also frequently selected. Weather-related variables, however, are only selected in 65 percent of the samples and the GPS indicators (latitude and longitude) are almost never selected.

Selection by LASSO is a coarse measure of the importance of predictor variables. We therefore perform Shapley decompositions of the predictors in one sample per country and indicator using the XGBoost results, following Lundberg et al. (2019). For each sub-area, this procedure calculates the average of the absolute value of changes in the prediction of assets or poverty due to each predictor variable, taken over all possible orderings of the predictors in the tree. These average Shapley values are then averaged across all sub-areas in the sample. The variables with the largest Shapley values for each country and indicator are listed in Table 11 .

In nearly all cases, the landcover shares (coverfraction) stand out as key variables in the XGBoost models. These include the fraction of the sub-areas classified as urban, shrub, grass, bare, trees, crops, and water. On average these variables and their higher-level aggregates account for almost half of the top 10 variables for asset prediction and 60 percent of the top 10 variables for poverty prediction. Pollution variables such as ozone (o3) and nitrogen dixode (no2) also appear in several cases, as do vegetation (ndvi), population (pop), weather (starting with weather) and distance to cities (variables starting with distto). In general, the listed variables tend to be correlated with population density, which is in turn correlated with asset ownership and poverty.

Table 10: Variable selection through lasso across simulations

	Weather	NDVI	LC	Population	Pollution	Distances	GPS	Regions
Panel A: Assets								
Burkina Faso	0.99	0.54	0.99	0.00	1.00	0.97	0.29	1.00
Madagascar	0.77	0.79	1.00	0.77	1.00	0.99	0.68	1.00
Malawi	0.99	0.99	0.74	0.53	0.99	0.96	0.14	0.99
Mozambique	0.81	1.00	1.00	0.72	0.98	0.99	0.97	0.99
Sri Lanka	0.74	1.00	1.00	0.23	0.88	1.00	0.98	1.00
Tanzania	0.99	0.97	1.00	0.56	1.00	0.99	0.02	1.00
Vietnam	1.00	0.90	1.00	0.81	1.00	0.90	0.99	1.00
Average	0.90	0.88	0.96	0.52	0.98	0.97	0.58	1.00
Panel B: Poverty								
Malawi	0.37	1.00	0.74	0.52	1.00	0.71	0.05	1.00
Tanzania	0.93	0.79	1.00	0.07	1.00	0.98	0.00	1.00
Average	0.65	0.90	0.87	0.30	1.00	0.84	0.03	1.00

Note: The columns list the different types of candidate variables used in estimation. For each country, we calculate the proportions of simulations in which lasso selects at least one of each variable type for use in EBP. GPS variables include the longitude and latitude of each subarea centroid, while region is defined separately by country, but is generally dummy variables for the strata used in sampling (usually admin2, sometimes also stratified by urban/rural).

Table 11: Feature importance in XGBoost: Shapley values

Burkina Faso	Madagascar	Malawi	Mozambique	Sri Lanka	Tanzania	Vietnam
			Panel A: Assets			
pop (levels)	urbancoverfraction	urbancoverfractiondist	urbancoverfractiondist	urbancoverfraction	shrubcoverfraction	urbancoverfractionprov
urbancoverfraction	urbancoverfractiondist	area	urbancoverfraction	shrubcoverfractionloc	grasscoverfraction	cropscoverfraction
pop (log)	o3prov	urbancoverfraction	area	ntl	ndviaveragemin	grasscoverfraction
shrubcoverfraction	grasscoverfractiondist	ndvi2	ndvi2	ndviyearmaxloc	disttogampaha	waterseasonalcover-fractiondist
popdist	no2	pop (log)	shrubcoverfraction	shrubcoverfractiondist	ndviaveragemindsd	barecoverfraction
disttogaoua	barecoverfractionprov	grasscoverfraction	grasscoverfraction	shrubcoverfraction	ndviyearmin	ndvi2prov
disttobanfora	disttosambava	shrubcoverfractiondist	pop (log)	o3loc	ntldist	treecoverfractionprov
no2	urbancoverfractionprov	shrubcoverfraction	cropscoverfraction	treecoverfraction	ndviyearmindsd	lon
so2	shrubcoverfractiondist	cropscoverfraction	grasscoverfractiondist	disttopemba	hchodsd	grasscoverfractiondist
barecoverfraction	lon	ndvi11	weathermeandist	so2	barecoverfraction	weather9
			Panel B: Poverty			
		urbancoverfraction			urbancoverfractiondist	
		urbancoverfractiondist			urbancoverfraction	
		urbancoverfractionprov			weather5dist	
		ndvi3prov			disttododoma	
		grasscoverfractiondist			weather11prov	
		ndvi3dist			grasscoverfractiondist	
		grasscoverfraction			weather5prov	
		disttoilala			grasscoverfraction	
		weather1prov			grasscoverfractionprov	
		shrubcoverfractiondist			shrubcoverfractiondist	

Note: The table shows the top ten variables for each country, based on the mean SHAP value across one simulation. The variables are not necessarily the same across countries, but the variables are listed in order of importance, with the most important variable listed first. Variable names followed by an integer represent the month of the year (e.g. ndvi2 is NDVI in February). Variable names with suffixes (e.g. dist, prov, or loc) indicate variables aggregated to a higher level (in the examples given, district, province, and localidade, respectively). Area refers to the geographic area of the subarea. Finally, suffixes of min and max refer to the minimum and maximum values of the variable in the location throughout the year. All of the coverfraction variables indicate land classifications, defined as the proportion of the given area covered by the given land type.

3.5 Additional geospatial predictors

Previous sections reported results on the accuracy of point and uncertainty estimates generated using an identical set of publicly available geospatial indicators, as described in section 2.2. However, these are not the only freely available geospatial features. In addition, the predictor variables used were never interacted with any other variables; this may be particularly problematic for EBP, which assumes a linear functional form, while the ML methods are flexible enough to implicitly take into account possible interactions. This section therefore extends the previous analysis by examining the extent to which adding additional features and interactions improves the accuracy of predictions. For additional features, we use the MOSAIKS features (Rolf et al., 2021), which are derived from raw satellite imagery. In addition, we experimented by interacting all variables with the share of land classified as urban (at the sub-area level).

Table 12 shows the results comparing EBP (which is most likely to change when including interactions) and XGBoost, our overall preferred method in terms of accuracy.[17] Somewhat surprisingly, adding MOSAIKS features to the set of candidate predictors reduces the accuracy of EBP predictions on average, for both assets and poverty. For XGBoost, adding MOSAIKS features reduces correlation for assets on average by 1.6 percentage points and increases correlations for poverty by 0.3 percentage points. Adding urbanity interactions slightly increases the accuracy of EBP estimates for assets and slightly decreases it for poverty. For XGBoost, adding interactions with urbanity as candidate variables increases the accuracy of the average asset estimates by 0.9 pp and decreases the accuracy of the average poverty estimates by 0.8 pp. In short, neither adding MOSAIKS features nor interactions has a consistently positive impact on accuracy. In general, impacts on accuracy are small, although adding MOSAIKS leads to a sizeable decline of approximately 5 percent points in accuracy when predicting wealth in Burkina Faso, and a sizable increase of about the same amount in Vietnam when using EBP.

Finally, Table 13 reports uncertainty measures when MOSAIKS features plus interactions are added. Compared with Table 6, average coverage rates for assets fall 3 pp for EBP, increase 3 pp for Cubist regression, and 1 pp for XGBoost and BRF. For poverty, average coverage rates hardly change for EBP and BRF, decline about 1 pp for Cubist, and increase nearly 3 pp for XGBoost. For XGBoost and BRF, including additional candidate variables appears to have modestly positive impacts on coverage rates, despite the limited impacts on the accuracy of the point estimates. However, this increase seems to come from larger confidence intervals; the width of the confidence interval in the last four columns is markedly larger than those in Table 6.

[17] We choose to present only these two methods for parsimony in the table.

Table 12: Comparing candidate feature choices

	(1) Basic variables		(2) MOSAIKS variables		(3) MOSAIKS and interactions	
	EBP	XGB	EBP	XGB	EBP	XGB
Panel A: Assets						
Burkina Faso	0.743	0.850	0.699	0.809	0.689	0.840
Madagascar	0.875	0.910	0.872	0.905	0.866	0.907
Malawi	0.664	0.844	0.541	0.811	0.544	0.833
Mozambique	0.897	0.917	0.893	0.917	0.902	0.919
Sri Lanka	0.919	0.935	0.923	0.934	0.930	0.937
Tanzania	0.868	0.899	0.879	0.906	0.891	0.912
Vietnam	0.868	0.904	0.911	0.916	0.911	0.930
Average	0.833	0.894	0.817	0.886	0.819	0.897
Panel B: Poverty						
Malawi	0.654	0.868	0.631	0.878	0.613	0.858
Tanzania	0.836	0.858	0.845	0.858	0.855	0.861
Average	0.745	0.863	0.738	0.868	0.734	0.860

Note: The first two columns present the simple mean of pearson correlation across simulations for each country using all candidate features as well as interactions between features and urbanity. The last two columns present the same correlations but restricting the candidate feature set by excluding all mosaiks features and interactions. The results are based on 100 simulations for each country and method.

Table 13: Uncertainty statistics across simulations, all features and interactions

	Coverage rate				Width of CI			
	EBP	Cubist	XGB	BRF	EBP	Cubist	XGB	BRF
Panel A: Assets								
Burkina Faso	0.913	0.806	0.929	0.902	2.275	1.492	1.793	2.818
Madagascar	0.626	0.940	0.962	0.991	0.722	1.298	1.442	1.984
Malawi	0.809	0.769	0.723	0.738	1.633	1.080	0.919	1.278
Mozambique	0.641	0.953	0.913	0.971	0.687	1.060	0.887	1.348
Sri Lanka	0.909	0.930	0.930	0.919	0.607	0.718	0.686	1.001
Tanzania	0.344	0.713	0.721	0.771	0.627	0.940	0.951	1.181
Vietnam	0.980	0.994	0.978	0.988	1.281	1.572	1.201	1.723
Average	0.746	0.872	0.880	0.897	1.119	1.166	1.126	1.619
Panel B: Poverty								
Malawi	0.744	0.763	0.872	0.831	0.464	0.452	0.495	0.575
Tanzania	0.871	0.865	0.884	0.943	0.297	0.297	0.295	0.515
Average	0.808	0.814	0.878	0.887	0.380	0.374	0.395	0.545

Note: The first four columns present the coverage rate across simulations for each country. The coverage rate is defined as the proportion of confidence intervals that contain the true value, derived from the census. The last four columns present the average width of the confidence interval across simulations for each country. The results are based on 100 simulations for each country and method. Candidate features include all variables used in the main specifications, along with mosaiks and interactions between features and urbanity.

4 Conclusion

This paper evaluates the use of three tree-based machine learning techniques and linear empirical best predictor (EBP) models for the purposes of contemporaneous small area estimation of wealth and poverty using household census data and publicly available geospatial indicators from seven countries. In addition to evaluating accuracy, we propose and implement a bootstrap algorithm to estimate the uncertainty associated with machine learning predictions and evaluate its performance using real-world data. Two of the three tree-based machine learning methods evaluated – Cubist Regression and Extreme Gradient Boosting (XGBoost) significantly outperformed the empirical best predictor model traditionally used for small area estimation (Molina and Rao 2010; Tzavidis et al. 2018). A third machine learning method, BRF, generated estimates comparable to EBP. Both point estimates and uncertainty estimates generated using Cubist regression models are generally a bit less accurate than those generated using XGBoost.

The results make a strong case for the use of XGBoost or Cubist regression in cases where the sample includes a sufficiently large number of subareas, those subareas comprise a small share of subareas in the population, and transparency and parsimony are not first-order concerns. In general, in these contexts, it does not appear that the benefit of EBP, in terms of conditioning on the sample data, outweighs the restrictions of its linear function form. In addition, the acccuracy of BRF estimates in this context likely suffers from using random subsets of predictors and different subsets of the data for training and prediction. Nonetheless, appropriate diagnostics, such as estimating linear models, interpreting the results of machine learning models, and evaluating results using cross-validation techniques, remain crucial in practice. Furthermore, accuracy is not always the most important factor when selecting methods. In cases where transparency and parsimony are important, or the number of sampled subareas is small, linear mixed EBP models or Cubist regression with a small number of rules are viable options. We do not claim that these results hold for all types of samples and settings. While recent research has made some progress towards combining tree-based machine learning with conditional random effects (Krennmair and Schmid, 2022; Messer and Schmid, 2024), these have yet to gain widespread acceptance and use. Our results highlight potential benefits from this line of research.

One notable finding, which has also been observed in other contexts, is the significantly greater accuracy of estimates in sampled areas than non-sampled areas. However, XGBoost appears to be markedly more accurate than other methods in predicting out-of-sample estimates. These differences diminish when comparing accuracy estimates within the same area across simulations. This shows how the tendency for samples to under-represent less densely populated areas can lead to less accurate predictions out of sample, even when taking sampling probabilities into account. This finding underscores the benefits of including all target areas

in the sample. When this is not possible, however, XGBoost appears to offer a notable advantage when predicting into non-sampled areas. Further research could explore whether and how re-weighting the sample can improve the accuracy of out-of-sample predictions. Related to this is the difficulty of predicting accurately at the tails of the welfare distribution, suggesting potential benefits from oversampling tails. Further research can also explore how the relative accuracy of different methods depends on the size and structure of the sample, as this analysis only considers one type of sample.

A key contribution of the paper is to evaluate the accuracy of the uncertainty estimates generated by a two-stage residual block bootstrap. These estimates perform reasonably well on average, as coverage rates average approximately 89 percent for BRF, 87 percent for XGBoost, and 84 percent for Cubist, all of which exceed the 78 percent average for EBP. Overall, the results indicate that the combination of Extreme Gradient Boosting or Cubist Regression, the random effect block residual bootstrap, and publicly available geospatial data offer a practical way to improve on EBP estimates when geolocated survey data are available.

While our results offer useful input for poverty estimates at a more disaggregated level that can result in more cost-effective budget planning, particularly in contexts where recent census data are unavailable, it should be noted that these results may not extend to other welfare outcomes. In addition, while we include seven different countries, it is not possible to claim generalizability in all contexts. The relative performance of different machine learning methods may well depend on the specific outcomes and model features under consideration. For example, one method may work well for poverty but may work less well for food insecurity or children's malnutrition status. The performance of different methods will also depend on the nature of the sample data. We encourage further research on comparing the performance of different methods, contexts, and outcomes. Further research could also shed light on whether models are sufficiently stable to use for intertemporal prediction. It is also useful to evaluate the performance of these methods in richer data contexts where we can compare models using satellite-based variables versus those using variables coming from household surveys and censuses.

Last but not least, a practical challenge with implementing machine learning methods involves limitations with analytical capacity of national statistical offices. We note that in these contexts, using existing tools to investigate modeling issues is good practice and should be done as part of the standard data checking and model exploration process before employing more advanced ML methods. For these methods to be employed more widely in a responsible way to better address data gaps, we call for more efforts by various stakeholders to improve local analytical capacity. For example, donors can support leading experts to jointly collaborate with national statistical staff to produce statistical training and better results. Providing ungated, public access to the latest research results and well-documented software packages (e.g., the *EMDI* and *povmap*

R packages) are critical for supporting these efforts. This is consistent with the Sustainable Development Goals' (Goal 17) recent focus on enhancing capacity-building support to developing countries to significantly increase the availability of high-quality, timely, and reliable data.

References

Aiken, E. L., Bedoya, G., Blumenstock, J. E., and Coville, A. (2023). Program targeting with machine learning and mobile phone data: Evidence from an anti-poverty intervention in Afghanistan. *Journal of Development Economics*, 161:103016.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests.

Atkinson, A. B. (2019). *Measuring Poverty around the World*. Princeton University Press.

Battese, G. E., Harter, R. M., and 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.

Beegle, K., Christiaensen, L., Dabalen, A., and Gaddis, I. (2016). *Poverty in rising Africa*. World Bank.

Besley, T. and Kanbur, R. (1991). *The principles of targeting*. Springer.

Blumenstock, J., Cadamuro, G., and On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264):1073–1076.

Blumenstock, J. E. (2016). Fighting poverty with data. *Science*, 353(6301):753–754.

Butar, F. B. and Lahiri, P. (2003). On measures of uncertainty of empirical bayes small-area estimators. *Journal of Statistical Planning and Inference*, 112(1-2):63–76.

Carter, G. M. and Rolph, J. E. (1974). Empirical bayes methods applied to estimating fire alarm probabilities. *Journal of the American Statistical Association*, 69(348):880–885.

Chambers, R. and Chandra, H. (2013). A random effect block bootstrap for clustered data. *Journal of Computational and Graphical Statistics*, 22(2):452–470.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Chi, G., Fang, H., Chatterjee, S., and Blumenstock, J. E. (2022). Microestimates of wealth for all low-and middle-income countries. *Proceedings of the National Academy of Sciences*, 119(3):e2113658119.

Dang, H., Jolliffe, D., and Carletto, C. (2019). Data gaps, data incomparability, and data imputation: A review of poverty measurement methods for data-scarce environments. *Journal of Economic Surveys*, 33(3):757–797.

Dang, H.-A. and Lanjouw, P. (2023). Regression-based imputation for poverty measurement in data scarce

settings. In *Jacques Silber. (Eds). Handbook of Research on Measuring Poverty and Deprivation*, pages 141–150. Edward Elgar Publishing.

Das, S. and Haslett, D. (2019). A Comparison of Methods for Poverty Estimation in Developing Countries. *International Statistical Review*, 87(2):368–392.

Donaldson, D. and Storeygard, A. (2016). The view from above: Applications of satellite data in economics. *Journal of Economic Perspectives*, 30(4):171–198.

Efron, B. (2020). Prediction, estimation, and attribution. *International Statistical Review*, 88:S28–S59.

Efron, B. and Morris, C. (1973). Stein's estimation rule and its competitors —- an empirical Bayes approach. *Journal of the American Statistical Association*, 68(341):117–130.

Elbers, C., Fujii, T., Lanjouw, P., Özler, B., and Yin, W. (2007). Poverty alleviation through geographic targeting: How much does disaggregation help? *Journal of Development Economics*, 83(1):198–213.

Elbers, C., Lanjouw, J. O., and Lanjouw, P. (2003). Micro-level estimation of poverty and inequality. *Econometrica*, 71(1):355–364.

Engstrom, R., Hersh, J., and Newhouse, D. (2022). Poverty from space: Using high resolution satellite imagery for estimating economic well-being. *The World Bank Economic Review*, 36(2):382–412.

Fay III, R. E. and Herriot, R. A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74(366a):269–277.

Friedberg, R., Tibshirani, J., Athey, S., and Wager, S. (2020). Local linear forests. *Journal of Computational and Graphical Statistics*, 30(2):503–517.

Friedman, J. H., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33:1–22.

Fujii, T. and van der Weide, R. (2020). Is predicted data a viable alternative to real data? *The World Bank Economic Review*, 34(2):485–508.

González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D., and Santamaría, L. (2008). Bootstrap mean squared error of a small-area eblup. *Journal of Statistical Computation and Simulation*, 78(5):443–462.

Hanna, R. and Olken, B. A. (2018). Universal basic incomes versus targeted transfers: Anti-poverty programs in developing countries. *Journal of Economic Perspectives*, 32(4):201–226.

Henderson, J. V., Storeygard, A., and Weil, D. N. (2012). Measuring economic growth from outer space. *American Economic Review*, 102(2):994–1028.

Isidro, M., Haslett, S., and Jones, G. (2016). Extended structure preserving estimation (espree) for updating small area estimates of poverty. *Annals of Applied Statistics*, 10(1):451–476.

Jean, N., Burke, M., Xie, M., Alampay Davis, W. M., Lobell, D. B., and Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794.

Jiang, J. and Lahiri, P. (2006). Mixed model prediction and small area estimation. *Test*, 15:1–96.

Kanbur, S. (1986). *Budgetary rules for poverty alleviation*. IIES.

Kilic, T., Serajuddin, U., Uematsu, H., and Yoshida, N. (2017). Costing household surveys for monitoring progress toward ending extreme poverty and boosting shared prosperity. *World Bank Policy Research Working Paper*, (7951).

Krennmair, P. and Schmid, T. (2022). Flexible domain prediction using mixed effects random forests. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 71(5):1865–1894.

Kreutzmann, A.-K., Pannier, S., Rojas-Perilla, N., Schmid, T., Templ, M., and Tzavidis, N. (2019). The R package emdi for estimating and mapping regionally disaggregated indicators. *Journal of Statistical Software*, 91:1–33.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. (2019). Explainable ai for trees: From local explanations to global understanding. *arXiv preprint arXiv:1905.04610*.

Luo, W. and Lai, H. C. (2021). A weighted residual bootstrap method for multilevel modeling with sampling weights. *Journal of Behavioral Data Science*, 1(2):89–118.

Masaki, T., Newhouse, D., Silwal, A. R., Bedada, A., and Engstrom, R. (2022). Small area estimation of non-monetary poverty with geospatial data. *Statistical Journal of the IAOS*, 38(3):1035–1051.

McBride, L., Barrett, C. B., Browne, C., Hu, L., Liu, Y., Matteson, D. S., Sun, Y., and Wen, J. (2022). Predicting poverty and malnutrition for targeting, mapping, monitoring, and early warning. *Applied Economic Perspectives and Policy*, 44(2):879–892.

Merfeld, J. and Morduch, J. (2023). Poverty at higher frequency.

Merfeld, J. D., Newhouse, D. L., Weber, M., and Lahiri, P. (2022). Combining survey and geospatial data can significantly improve gender-disaggregated estimates of labor market outcomes.

Messer, P. and Schmid, T. (2024). Gradient boosting for hierarchical data in small area estimation. *arXiv preprint arXiv:2406.04256*.

Molina, I. and Rao, J. N. (2010). Small area estimation of poverty indicators. *Canadian Journal of statistics*, 38(3):369–385.

Newhouse, D. (2024). Small area estimation of poverty and wealth using geospatial data: What have we learned so far? *Calcutta Statistical Association Bulletin*, 76(1):7–32.

Newhouse, D., Ramakrishnan, A., Swartz, T., Merfeld, J., and Lahiri, P. (2025). Small area estimation of monetary poverty in Mexico using satellite imagery and machine learning. *Oxford Bulletin of Economics and Statistics*.

Pfeffermann, D. and Sverchkov, M. (2007). Small-area estimation under informative probability sampling of areas and within the selected areas. *Journal of the American Statistical Association*, 102(480):1427–1439.

Pratesi, M. and Spagnolo, F. S. (2023). Small area methodology for measuring poverty at a local level. In *Jacques Silber. (Eds). Research Handbook on Measuring Poverty and Deprivation*, pages 129–140. Edward Elgar Publishing.

Quinlan, J. R. et al. (1992). Learning with continuous classes. In *5th Australian Joint Conference on Artificial Intelligence*, volume 92, pages 343–348. World Scientific.

Ratledge, N., Cadamuro, G., De la Cuesta, B., Stigler, M., and Burke, M. (2021). Using satellite imagery and machine learning to estimate the livelihood impact of electricity access. Technical report, National Bureau of Economic Research.

Ravallion, M. (2015). *The economics of poverty: History, measurement, and policy*. Oxford University Press.

Rolf, E., Proctor, J., Carleton, T., Bolliger, I., Shankar, V., Ishihara, M., Recht, B., and Hsiang, S. (2021). A generalizable and accessible approach to machine learning with global satellite imagery. *Nature Communications*, 12(1):4392.

Tibshirani, J., Athey, S., Friedberg, R., Hadad, V., Hirshberg, D., Miner, L., Sverdrup, E., Wager, S., Wright, M., and Tibshirani, M. J. (2018). Package 'grf'. *Comprehensive R Archive Network*.

Tzavidis, N., Zhang, L.-C., Luna, A., Schmid, T., and Rojas-Perilla, N. (2018). From start to finish: a framework for the production of small area official statistics. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 181(4):927–979.

Van der Weide, R., Blankespoor, B., Elbers, C., and Lanjouw, P. (2024). How accurate is a poverty map based on remote sensing data? an application to malawi. *Journal of Development Economics*, 171:103352.

Wang, Y. and Witten, I. H. (1997). Inducing model trees for continuous classes. In *Proceedings of the Ninth European Conference on Machine Learning*, volume 9, pages 128–137. Citeseer.

Yeh, C., Perez, A., Driscoll, A., Azzari, G., Tang, Z., Lobell, D., Ermon, S., and Burke, M. (2020). Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature Communications*, 11(1):2583.

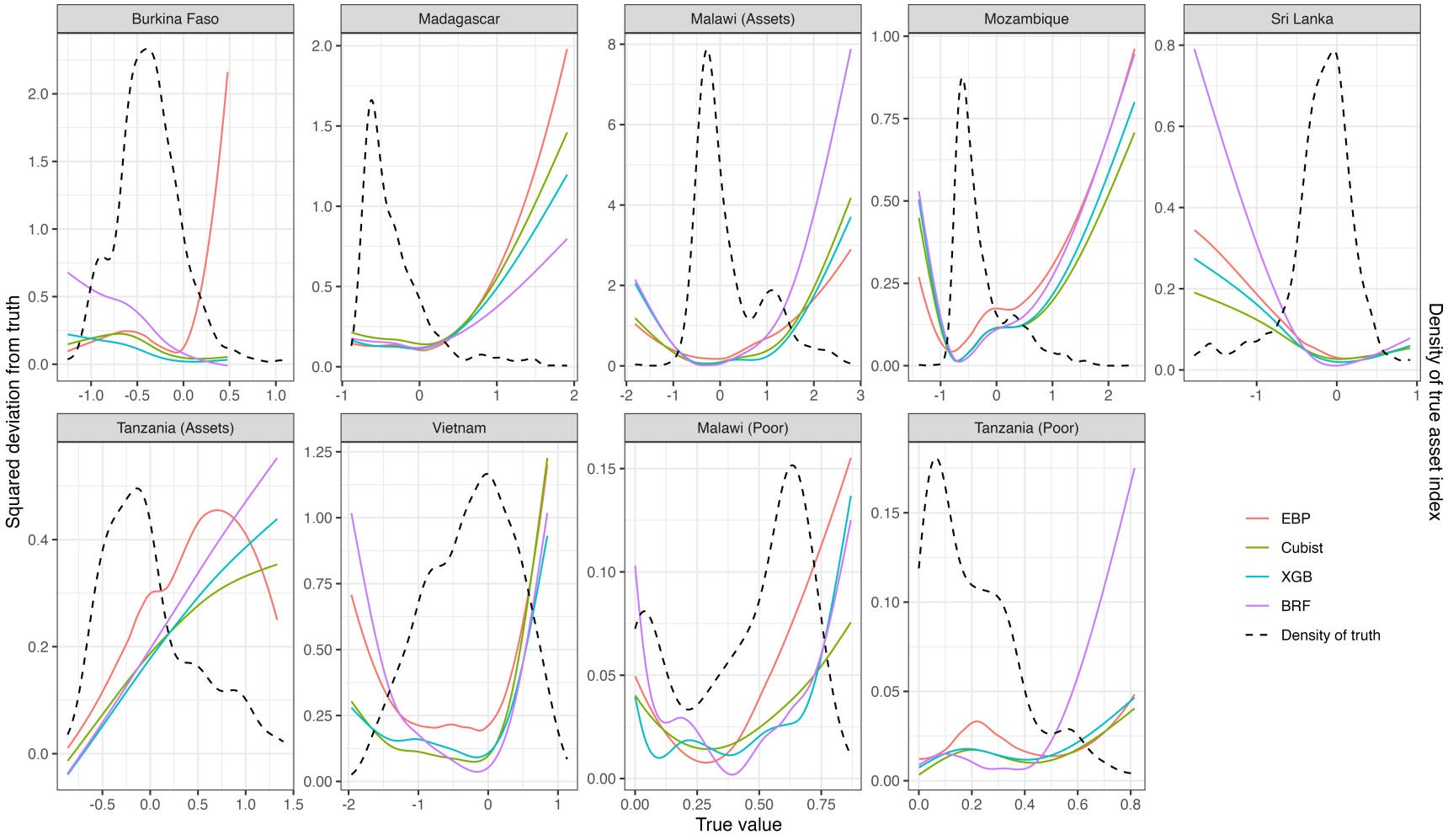
Appendix A - Additional tables and figures

Table A1: Geospatial features

Indicator	Source
Population	WorldPop
Precipitation	TerraClimate
Temperature	TerraClimate
Nightlights	NOAA VIIRS
Land cover	EU Copernicus
Elevation	Conservation Science Partners
NDVI	MODIS
Pollution measures	EU Copernicus
Distance to key cities[1]	Collected by authors
Mosaiks[2]	Rolf et al. (2021)

[1] We collect the (approximate) location of all of the Admin 2 or Admin 1 capitals (depending on the country and number of admin areas) and calculate the distance from each enumeration area in the population.
[2] Mosaiks is only used in an alternative specification. The main results do not include the Mosaiks features.

Figure A1: Deviations from truth vs true values, out-of-sample areas only



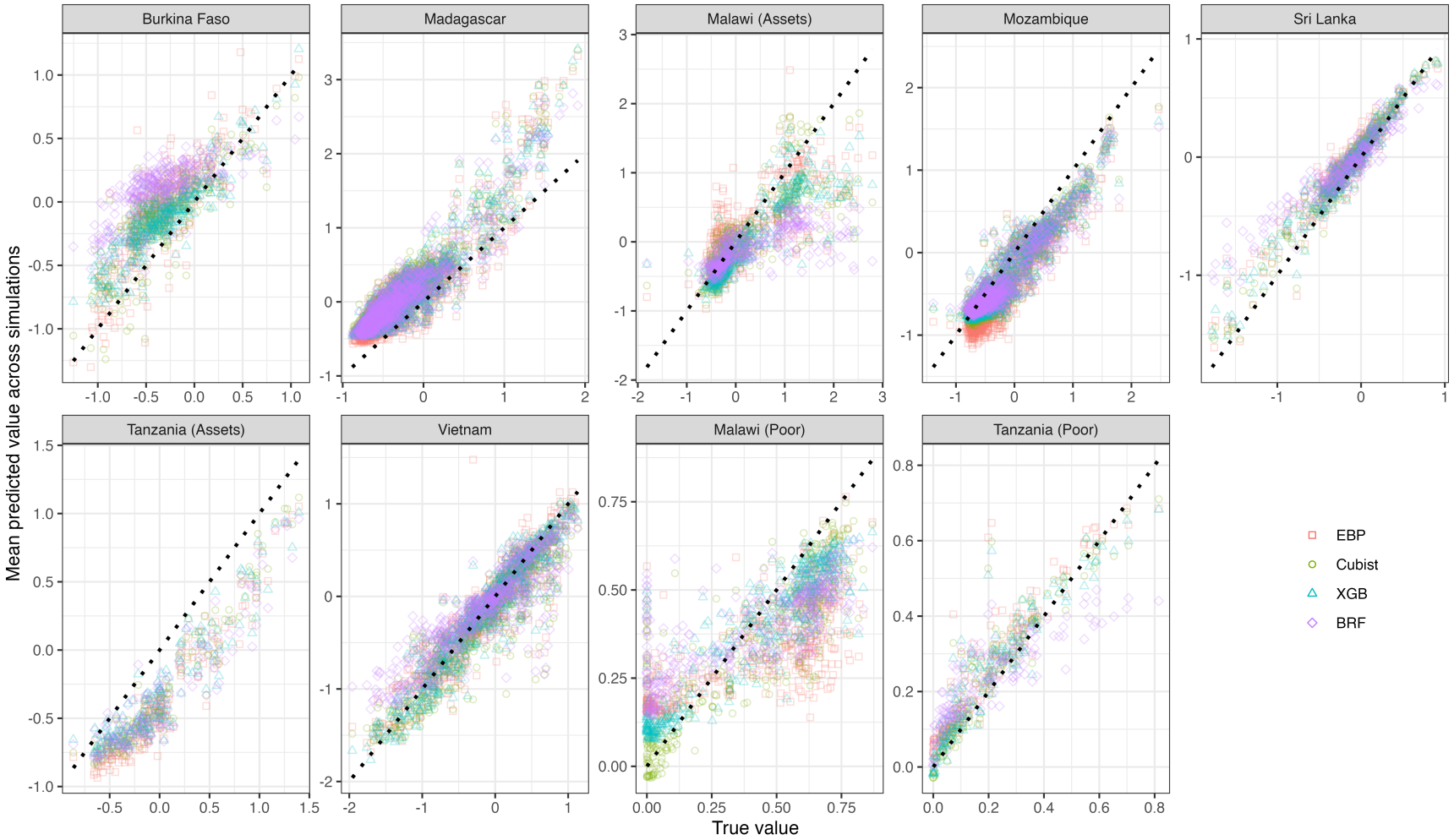
Note: All figures are smoothed conditional means of the mean squared deviation across 100 independent samples (first y-axis) on truth (x-axis), with means restricted to only samples in which an area does not appear (out-of-sample areas). The kernel density estimate refers to the density of truth, which is on the x-axis.

Table A2: Optimal hyperparameters from cross validation

	Assets							Poverty	
	BFA	MDG	MWI	MOZ	LKA	TZA	VNM	MWI	TZA
Panel A: Cubist									
Committees	100	100	100	50	100	100	100	100	50
Neighbors	0	0	9	0	9	9	0	9	9
Panel B: XGBoost									
Rounds	100	200	200	200	100	200	100	200	100
Max. depth	4	4	6	6	4	6	4	6	4
η	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.3	0.15
γ	1	1	1	1	0	1	0	1	0
Min. child weight	0	0	0	0	0	0	0	0	0
Col sample by tree	0.8	0.9	0.9	0.8	0.8	0.8	0.8	0.9	0.8
Subsample	0.9	0.9	0.8	0.8	0.8	0.9	0.9	0.8	0.8

Note: The table shows the optimal hyperparameters from a single random sample for each country. Due to computing time, we use the same hyperparameters across all simulations for a given country. The column names are the country's three-digit ISO code. The first two rows show the hyperparameters for Cubist, the next seven for XGBoost. For XGBoost, η is the learning rate and γ is the minimum loss reduction required to make a further partition. Due to the implementation of BRF, the hyperparameters differ for each simulation so are not shown here.

Figure A2: Predicted-true plots across areas, scatter plot



Note: In all figures, areas are ordered by true value, which is displayed on the x-axis. The y-axis presents the estimated value for each given area.

Table A3: Correlations across countries and methods, EBP vs. ELL

	Pearson		Spearman	
	EBP	ELL	EBP	ELL
Panel A: Assets				
Burkina Faso	0.743	0.651	0.715	0.607
Madagascar	0.875	0.870	0.795	0.787
Malawi	0.664	0.641	0.658	0.623
Mozambique	0.897	0.892	0.781	0.773
Sri Lanka	0.919	0.900	0.884	0.866
Tanzania	0.868	0.877	0.804	0.812
Vietnam	0.868	0.805	0.873	0.810
Average	0.833	0.805	0.787	0.754
Panel B: Poverty				
Malawi	0.786	0.672	0.786	0.696
Tanzania	0.836	0.818	0.852	0.842
Average	0.811	0.745	0.819	0.769

Note: The first four columns present the simple mean of pearson correlation across simulations for each country. The last four columns present the simple mean of spearman correlation across simulations for each country. The results are based on 100 simulations for each country and method.

Appendix B - Calculating the asset index and imputing poverty

B1 Assets

For all countries, we compute an asset index using the assets available in the unit-level census data. Importantly, while we calculate the asset index for the census to calculate ground truth, we also calculate the asset index independent on each random sample. In other words, we do not use the asset index values from the entire census when computing the sub-area asset index during the simulations. In all cases, we use principal components analysis and allow for just one factor, which we use as the asset index.

In the following sections, we discuss which assets we use in each country.

B1.1 Burkina Faso

We use the following assets to calculate the asset index in Burkina Faso:

- Radio ownership
- TV ownership
- Landline ownership
- Mobile phone ownership
- Fridge ownership
- Computer ownership
- Stove ownership
- Electric stove ownership
- Bike ownership
- Motorcycle ownership
- Tricycle ownership
- Car ownership
- Canoe ownership
- Cart ownership
- Camel ownership
- Horse ownership
- Donkey ownership
- Lighting type (dummies)
- Energy type (dummies)
- Toilet type (dummies)
- Walls type (dummies)
- Roof type (dummies)
- Floor type (dummies)

B1.2 Madagascar

We use the following assets to calculate the asset index in Madagascar:

- Radio ownership
- TV ownership
- VCR ownership
- Oven ownership
- Fridge ownership
- Washing machine ownership
- Dryer ownership
- Computer ownership
- Internet access
- Landline ownership
- Mobile phone ownership
- Car ownership
- Air conditioning ownership
- Scooter/motorcycle ownership
- Bike ownership
- Energy type (dummies)
- Toilet type (dummies)
- Walls type (dummies)
- Roof type (dummies)
- Floor type (dummies)
- Housing type

B1.3 Malawi

We use the following assets to calculate the asset index in Malawi:

- Radio ownership
- TV ownership
- VCR/DVD ownership
- Mobile phone ownership
- Computer ownership
- Fridge ownership
- Bike ownership
- Table ownership
- Bed ownership
- Iron ownership
- Solar panel ownership
- Lamp/torch ownership
- Car ownership
- Walls type (dummies)
- Roof type (dummies)
- Floor type (dummies)
- Dwelling type (dummies)
- Housing ownership

B1.4 Mozambique

We use the following assets to calculate the asset index in Mozambique:

- Radio ownership
- TV ownership
- Landline ownership
- Mobile phone ownership
- Computer ownership
- Internet access
- Iron ownership
- Stove ownership
- Electric or gas stove ownership
- Fridge ownership
- Car ownership
- Motorcycle ownership
- Bike ownership
- Walls type (dummies)
- Roof type (dummies)
- Floor type (dummies)
- Dwelling type (dummies)

B1.5 Sri Lanka

We use the following assets to calculate the asset index in Sri Lanka:

- Radio ownership
- TV ownership
- Landline ownership
- Mobile phone ownership
- Desktop computer ownership
- Laptop computer ownership
- Internet access
- Fax machine ownership
- Electricity access
- Water type (dummies)
- Toilet type (dummies)
- Walls type (dummies)
- Roof type (dummies)
- Floor type (dummies)
- Dwelling type (dummies)
- Number of rooms in dwelling (different rooms)

B1.6 Tanzania

We use the following assets to calculate the asset index in Tanzania:

- Radio ownership
- TV ownership
- Landline ownership
- Mobile phone ownership
- Car ownership
- Bike ownership

- Motorcycle ownership
- Computer ownership
- Fridge ownership
- Iron ownership
- House ownership
- Power tiller ownership
- Wheelbarrow ownership
- Plough ownership
- Walls type (dummies)
- Roof type (dummies)
- Floor type (dummies)

B1.7 Vietnam

We use the following assets to calculate the asset index in Vietnam:

- TV ownership
- Radio ownership
- Computer ownership
- Mobile ownership
- Fridge ownership
- Washing machine ownership
- Water heater ownership
- Air conditioning ownership
- Motorbike ownership
- Bike ownership
- Boat ownership
- Car ownership
- Dwelling type (dummies)
- Year dwelling built (dummies)
- Bedrooms in dwelling
- Walls type (dummies)
- Roof type (dummies)
- Floor type (dummies)
- Toilet type (dummies)
- Source of lighting (dummies)
- Source of cooking energy (dummies)
- Source of water (dummies)

B2 Poverty

We impute welfare into the 2018 Malawi census using the 2019 Integrated Household Survey (IHS) and into the 2012 Tanzania census using the third wave of the National Panel Survey (2012-2013). Both the census and the household surveys include information on household assets and key household demographics; in addition, the surveys have information on expenditures/consumption. We select the variables that are common to both datasets (separately by country) and then use lasso to select the most predictive variables

to use in the imputation procedure. The assets available in the household survey are more numerous than those available in the census, meaning the matched sets of assets are those listed above.

The post-lasso regression results for both countries are in Table B1. We remove all selected region/district dummies from the table for parsimony. The r-squared is quite high considering these estimates are at the household level; in Malawi the regression explains around 43% of total variation in per capita expenditures while in Tanzania the regression explains around 46% of the total variation.

We predict welfare directly into the census using the results in Table B1. We then use the census-derived expenditures to calculate a poverty line using the same quantile as the poverty rate in the household survey. In Malawi, this quantile is approximately 0.5. In Tanzania, this quantile is approximately 0.21. In other words, by construction, the poverty rate in the census is equal to the poverty rate in the household survey.

Table B1: Welfare imputation regressions

	Malawi	Tanzania
Head male	-0.068	
	(0.012)	
Head educ primary	0.088	
	(0.013)	
Head educ secondary	0.14	
	(0.018)	
Head educ university	0.224	
	(0.037)	
Dwelling owned	-0.116	-0.229
	(0.014)	(0.021)
Dwelling permanent	0.05	
	(0.028)	
Dwelling semi-permanent	0.022	
	(0.02)	
Rooms in dwelling	-0.066	
	(0.005)	
Roof grass	-0.087	0.064
	(0.017)	(0.055)
Room cement	0.605	
	(0.534)	
Roof metal		0.109
		(0.055)
Roof other		0.23
		(0.083)
Wall mud	-0.028	
	(0.027)	
Wall concrete	0.015	0.025
	(0.028)	(0.03)
Wall bricks	-0.026	-0.053
	(0.018)	(0.023)
Floor cement	0.126	
	(0.018)	
Floor wood	0.503	
	(0.377)	
Floor tile	0.378	
	(0.082)	
Floor easrth		-0.241
		(0.025)
Floor other		0.167
		(0.132)
Mobile ownership	0.095	0.229
	(0.012)	(0.022)
Radio ownership	0.071	0.105
	(0.012)	(0.018)
TV ownership	0.069	0.176
	(0.024)	(0.029)
Computer ownership	0.348	0.345
	(0.038)	(0.05)
Fridge ownership	0.199	0.086
	(0.03)	(0.034)
Bike ownership	0.027	-0.07
	(0.012)	(0.019)
Table ownership	0.052	
	(0.013)	
Bed ownership	0.181	
	(0.014)	
Iro ownership	0.11	0.131
	(0.017)	(0.023)
Solar panel ownership	0.033	
	(0.014)	
CD/DVD ownership	0.043	
	(0.025)	
Car ownership	0.366	0.446
	(0.041)	(0.053)
Cycle ownership		0.159
		(0.035)
Land ownership		-0.153
		(0.022)
Wheelbarrow ownership		0.012
		(0.063)
Plough ownership		-0.047
		(0.037)
Intercept	12.425	13.528
	(0.046)	(0.064)
District/region dummies	31 total	24 total
r-squared	0.431	0.461

Robust standard errors are in parentheses.

Appendix C - Methods

This section describes the details of the different machine learning methods. Due to computing time, we tune hyperparameters once and use these in all simulations for a given country and outcome. The optimal hyperparameters are in Table A2.

C1 Cubist

The Cubist algorithm proceeds as follows:

1. **Form a decision tree by conducting an exhaustive search over the predictor space and training set samples**. Splits are determined by minimizing the standard error of the dependent variables within groups. In mathematical terms, splits are chosen recursively to maximize the reduction in a measure of error. Defining S as the entire set of data and $S_1, ..., S_p$ as the P subsets of the data after splitting, the algorithm maximizes

$$\text{reduction} = SD(S) - \sum_{i=1}^{P} \frac{n_i}{n} SD(S_i), \tag{14}$$

   where SD is the standard deviation, n is the number of sample observations considered, and n_i is the number of sample observations in partition i. In other words, the algorithm identifies the set of splits that maximizes the reduction in the weighted average, across child nodes, of the standard deviations within the nodes. Splitting ceases, and the node becomes a leaf, when the maximum residual falls below a minimum tolerance level or when the number of training cases falls below a minimum threshold.[18]

2. **Estimate and simplify linear models at each node**. At each node of the tree, a linear model is estimated using only the variable attributes used to split the sub-tree above the node. In other words, the model for the first split from the top will be a bivariate regression with a single predictor. At subsequent nodes further down the tree, the set of candidate variables expands to include the set of all variables used for splitting to that point.

   Not all candidate variables are actually used in the models. In particular, the resulting linear models are simplified to avoid overfitting, by greedily dropping variables to minimize "adjusted error rate". The adjusted error rate is the mean absolute error multiplied by a term to penalize models with many

[18] The minimum tolerance level is set at five percent of the standard deviation of the dependent variable in the full training data (Wang and Witten, 1997). The minimum number of observations is set to 10 percent of the sample if the sample is less than 2000 observations, or 20 if the sample is more than 2000 observations.

variables, defined as:

$$\text{adjusted error rate} = \frac{n^* + p}{n^* - p} \sum_{i=1}^{n^*} |y_i - \hat{y}_i(X_i)| \,, \tag{15}$$

where n^* is the number of observations in the training data at the node used to build the model; p is the number of parameters, equal to the number of independent variables plus one; and $\hat{y}_i(X_i)$ is the predicted value from the model given a set of predictor variables X_i. The variable that leads to the largest reduction in the adjusted error rate is removed, sequentially, until the adjusted error rate increases when removing any of the remaining predictors. Removing attributes inevitably increases mean absolute error but also reduces the multiplication factor $\frac{n^*+p}{n^*-p}$, which may reduce the adjusted error rate.

Finally, the procedure performs an outlier check, defining outliers as cases where residuals are greater than five times the average absolute value of the model residuals for that node. At each node, before finalizing the model, outliers are eliminated from the estimation sample and the model is re-estimated and re-simplified.

3. **Prune the rules**. Each leaf of the tree is translated into a set of "rules" based on the sequence of splitting conditions that lead to the leaf. For example, a rule based on a leaf with two branches above it would consist of three conditions, for example $X_1 > 10$, $X_2 < 2$ and $X_3 = 1$. These rules are then "pruned", a process that eliminates conditions that are harmful or not useful for predicting the full set of training data. To measure prediction accuracy, the algorithm uses the adjusted error rate defined in Equation 15, applied to the full set of training data.

   As a first step, the algorithm calculates smoothed predictions across the various conditions of a rule, which corresponds to particular nodes along the tree that lead to a leaf, using the following formula [@hastie1990shrinking]:

$$\hat{Y}_{par} = a\hat{Y}_{kid} + (1 - a)\hat{Y}_{par}, \tag{16}$$

   where $\hat{Y}_{par}$ is the prediction of the model estimated at the parent node and $\hat{Y}_{kid}$ is the prediction of the model estimated at the current node. a is equal to

$$a = \frac{var(e_{par}) - cov(e_{kid}, e_{par})}{var(e_{par} - e_{kid})}, \tag{17}$$

where e_{par} are the model residuals from the parent node and e_{kid} are the model residuals from the current node, for training cases under consideration at the current node. All rule pruning is based on these smoothed predictions.

The second step is to eliminate all conditions (nodes) that increase in the adjusted rate, defined as in Equation 16 except taken over the full training sample. The program identifies the condition (node) that, when removed, leads to the largest decline in the adjusted rate. If removing that condition does not increase the adjusted error rate, it is removed. This proceeds sequentially until no condition can be removed without increasing the adjusted error rate.

The third step repeats step 2, except that conditions are removed as long as they do not raise the adjusted error rate by more than 0.5 percent. This additional step is implemented to further simplify the tree structure. Finally, if necessary, conditions are further pruned until the number of remaining rules is equal to the maximum number of rules specified by the user.

4. **Generate smoothed models for each rule**. For each rule, a model is created by coefficients at the leaves with all the models above it on the path to the initial split, similar to Equation 16. The model coefficients for each rule (leaf) are averaged according to the following formula:

$$\hat{\beta}_{par} = a\hat{\beta}_{kid} + (1 - a)\hat{\beta}_{par}. \tag{18}$$

5. This procedure smooths the model coefficients by collecting the sequence of linear models at each node into a single, smoothed representation of the models. The algorithm adjusts the final model so that all continuous cutpoints match those present in the data, by changing the cutpoint to equal the closest value in the data.

   The Cubist software also allows an option to estimate "committees," which are sets of Cubist models that successively correct errors in the previous estimates, similar to boosting. In other words, each committee produces a series of rules and associated models that iteratively predict the errors from the previous committeeś prediction. The package uses cross-validation to determine the optimal number of

rules and committees.

The Cubist method, like the other machine learning methods, is associated with several hyperparameters. We opt to tune the hyperparameters just once, using a random survey sample drawn in the exact same way as described above in order to cut down on total computation time. We then keep this set of hyperparameters constant through all survey draws. In addition, we tune the hyperparameters via cross validation, hand coding the folds to draw all subareas in a given area to help preserve the hierarchical nature of the data and prevent some information leakage across folds. Table A1 in the appendix shows the optimal hyperparameters for cubist that we use in all sample iterations.

C2 XGBoost

Extreme gradient boosting estimates a function that predicts the dependent variable y_i as a function of the set of independent variables x_i. This function is defined as the sum of a series of individual decision tree functions. In mathematical terms, for a single observation i of a set of predictors x_i,

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), f_k \in F, \tag{19}$$

where K is the number of trees estimated in the model and f_k is a decision tree function mapping x_i to $\hat{y}_i$ in the functional space F, which is the set of all possible decision trees. $f_1, ..., f_K$ is defined as the minimum of the following objective function of $\phi(x_i)$:

$$obj(\phi) = \sum_{i=1}^{n} l\left(y_i, \phi(x_0)\right) + \sum_{k=1}^{K} \omega(f_k), \tag{20}$$

where l is a differential convex loss function that measures the distance between the predicted value and the training value and $\omega(f_k)$ is a regularization term that penalizes model complexity, defined below.

Because the algorithm is optimizing over a set of feasible functions f_k, instead of parameters, it is not possible to use standard optimization tools. Instead, the algorithm proceeds by estimating each individual $f_k(x_i)$ tree function in a "greedy" manner (Friedman 2001). Specifically, the algorithm identifies a tree $f_t(x_i)$ at step t to minimize the following objective function:

$$obj(t) = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \omega(f_t). \tag{21}$$

This sequentially adds the f_t that provides the largest improvement in performance according to the objective function, Equation 20, given the previous round's prediction, $\hat{y}_i^{(t-1)}$. $\hat{y}_i^0$ is set to zero so the first iteration generates the tree $f_1(x_i)$ that minimizes $\sum_{i=1}^{n} l\left(y_i, f_t(x_i)\right) + \omega(f_t)$.

The mean value of the asset index is continuous when aggregated to the subarea level – the level at which we estimate welfare – we use mean-squared error as the loss function, thus:

$$l\left(y_i, \hat{y}_i + f_t(x_i)\right) = \left(\hat{y}_i^{(t-1)} + f_t(x_t) - y_i\right) \tag{22}$$

The resulting objective function at step t, after removing constants, becomes

$$obj(t) = \sum_{i=1}^{n} \left[2\left(\hat{y}_i^{(t-1)} - y_i\right) f_t(x_i) + f_t(x_i)^2\right] + \omega(f_t), \tag{23}$$

which the algorithm minimizes at each step by choosing $f_t(x_i)$.

Regularization prevents overfitting and, in a general case, is defined as

$$\omega(f_i) = \gamma T + \frac{\lambda}{2} \sum_{j=1}^{T} w_j^2, \tag{24}$$

where T represents the number of leaves on tree f_k and w_j is the score assigned to leaf j. λ and γ are tuning parameters controlling the extent of regularization. We follow the default and set $\lambda = 1$ and $\gamma = 0$ for estimation. Table A1 in the appendix shows the optimal hyperparameters for XGBoost that we use in all sample iterations; we use default values for any unlisted hyperparameters.

C3 Boosted Regression Forests (BRF)

To grow a tree, the algorithm first takes a random sample of the data. The share of the data selected is a parameter determined by cross-validation. The algorithm then begins the tree at the root node with this random sample, and recursively splits the data to create child nodes. At each split, the algorithm randomly selects a subset of the predictor variables as splitting candidates. For each splitting candidate, the algorithm considers all the possible values these variables take on in the data. For all values taken on by all the splitting candidates, the algorithm first considers whether the split would meet three basic eligibility criteria:

1. That the resulting children have a minimum number of observations that exceeds a minimum absolute node size parameter
2. That each child contains more than a minimum threshold fraction of the parent observations, to prevent splits that are too imbalanced.
3. That the split improves heterogeneity in outcomes as defined in equation (19) below

The minimum node size and balance thresholds are parameters estimated through cross-validation, as described below. Of the remaining candidate splits, the algorithm selects the threshold that maximizes heterogeneity in the average outcome across the child nodes. All observations with variables below that threshold are assigned to child 1 and all observations with variables above that threshold are assigned to child 2. For boosted regression forests, heterogeneity in the split, denoted H, is defined as:

$$H = \frac{N_{C1}N_{C2}}{N_P^2(\bar{y}_{C1} - \bar{y}_{C2})^2} - \left(\frac{IP}{N_{C1}} + IPN_{C2}\right), \tag{25}$$

where N_{C1}, N_{C2}, and N_P^2 are the number of observations in child 1, child 2, and the parent node, respectively, and $\bar{y}_{C1}$ and $\bar{y}_{C2}$ are the average values of the predicted outcome in the children. IP is an imbalance penalty parameter, selected through cross validation, that favors more balanced splits.

To simplify the process, BRF automates the tuning of hyperparameters using cross-validation. However, for computational reasons, the hyperparameters are estimated once in the first boosting step.