Improving estimates of mean welfare and uncertainty in developing countries*

Joshua D. Merfeld† David Newhouse‡

2023-01-18

Abstract

Reliable estimates of economic welfare for small areas are valuable inputs into the design and evaluation of development policies. This paper compares the accuracy of point estimates and confidence intervals for small area estimates of wealth and poverty derived from four different prediction methods: linear mixed models, Cubist regression, extreme gradient boosting, and boosted regression forests. The evaluation draws samples from unit-level household census data from four separate developing countries, combines them with publicly and globally available geospatial indicators to generate small area estimates, and evaluates these estimates against aggregates calculated using the full census. Predictions of wealth are evaluated in four countries and poverty in one. All three machine learning methods outperform the traditional linear mixed model, with extreme gradient boosting and boosted regression forests generally outperforming the other alternatives. Our proposed residual bootstrap procedure reliably estimates confidence intervals for the machine learning estimators, with estimated coverage rates across simulations falling between 94 and 97 percent. These results demonstrate that predictions obtained using tree-based gradient boosting with a random effect block bootstrap generate more accurate point and uncertainty estimates than prevailing methods for generating small area welfare estimates.

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

*thanks to...
†KDI School of Public Policy and Management and IZA; merfeld@kdis.ac.kr
‡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 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 Weide 2020; Kilic et al. 2017). 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 an accompanying survey – following Elbers, Lanjouw, and Lanjouw (2003) or Molina and Rao (2010) – but in poorer countries, recent census data are not always available. 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, Cadamuro, and On (2015) use mobile phone records to infer the socioeconomic status of phone owners in Rwanda and Aiken et al. (2022) to use mobile phone call data records to predict targeting performance of programs in Togo. However, one drawback to 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, Storeygard, and Weil 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, Hersh, and Newhouse 2022; Newhouse et al. 2022; Masaki et al. 2022).

In this paper, we evaluate four different methods to estimate welfare at low levels of aggregation in developing countries. We validate the performance of these methods using unit-level census data across four separate developing countries: Madagascar, Malawi, Mozambique, and Sri Lanka. In Malawi we are able to extend the evaluation to include a headcount poverty measure in addition to the asset index, while in the other three countries we evaluate prediction of an asset index, similar to Chi et al. (2022) and Masaki et al. (2022).

The four methods we evaluate are linear empirical best predictor (EBP) models (Battese, Harter, and Fuller 1988; Jiang and Lahiri 2006; Molina and Rao 2010; Masaki et al. 2022), which come from a long history of small area estimation in statistics, and three newer machine learning methods: cubist regression models (Wang

and Witten 1997; Quinlan 1992), extreme gradient boosting (Chen and Guestrin 2016) – more commonly known as XGBoost – and boosted regression forests, or BRF (Friedberg et al. 2020; J. Tibshirani et al. 2022). 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. While we specify the EBP model at the household level, the others are specified at the sub-area level, which in these contexts are 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 data that is available across much of the globe, meaning that the methods and data we evaluate here are widely applicable in cases where geolocated survey data is available. We use shapefiles from all four countries to pull geospatial data from multiple sources, which is then combined with the unit-level census data. In each country, we simulate 100 two-stage samples – first randomly selecting enumeration areas and then randomly selecting households – and compare the overall performance across simulations, meaning our results are representative of one hundred possible samples rather than a single sample.

All three machine learning models substantially outperform EBP in terms of accuracy, as seen clearly in Table 1. XGBoost and BRF perform best, with XGBoost yielding slightly more accurate predictions on average. XGBoost is notably more accurate than BRF in Sri Lanka, especially among the poorest in the left tail of the distribution. Cubist regression, which estimates a set of linear models on different subsets of the data, performs slightly less well than XGboost and BRF, even when a boosting procedure is utilized. All three machine learning methods greatly outperform EBP. The average (pearson) correlation is 7.5 percent higher for XGBoost than EBP, which is the current workhorse for small area estimation for practitioners who require accurate estimates of uncertainty. We also examine additional accuracy measures related to deviation from truth. XGBoost and BRF also outperform EBP in terms of squared deviation, and in this case the difference is even larger: squared deviation for XGBoost is 37.1 percent lower than for EBP, indicating large improvements in accuracy. The other two ML estimators also outperform EBP by large amounts in terms of both correlation and squared deviation.

A key contribution of the paper is the evaluation of the random effect block bootstrap procedure, proposed by Chambers and Chandra (2013) in a different context, to estimate uncertainty for the three machine learning methods. The presentation of uncertainty statistics has traditionally been less common for machine learning methods (Chi et al. 2022). The random effect block bootstrap accounts for the hierarchical nature of the data and proceeds in two steps, sampling residuals separately at both the target area level and the sub-area level, which is the unit of analysis in the machine learning estimation. This procedure estimates accurate

Table 1: Summary of results

	EBP	cubist	XGBoost	BRF
correlation (pearson)	0.841	0.883	0.904	0.895
squared deviation	0.114	0.080	0.072	0.082
width of CI	1.212	1.316	1.006	1.381
coverage	0.950	0.966	0.940	0.968

Note: Measures of accuracy and uncertainty are simple averages for the asset index predictions across the four hundred independent samples (100 samples for each country).

confidence intervals, with average coverage rates of 96.6 percent for cubist, 94.0 percent for XGBoost, and 96.8 percent for BRF. These coverage rates are in line with the coverage rates for EBP, which are derived from the parametric bootstrap procedure typically used to generate uncertainty estimates for EBP predictions. XGBoost achieves approximately correct coverage rates but is more likely than BRF or Cubist to understate coverage, as coverage rates range from 0.894 to 0.979 for XGBoost, from 0.934 to 0.987 for BRF, and from 0.925 to 0.983 for Cubist. It appears that uncertainty estimates using the random effects block bootstrap slightly underestimate uncertainty for XGboost, and slightly overstate uncertainty when using Cubist and BRF.

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. 2022). On average, all four estimators predict in sample better than out of sample. EBP shows the biggest drop in performance out of sample, at least in terms of correlation. While the three ML estimators also show drops in accuracy, these drops tend to be relatively smaller. For example, both XGBoost and BRF have pearson and rank correlations above 0.8 in out-of-sample areas across the five cases that we examine. The more flexible nature of XGBoost and BRF models are better suited for predicting out-of-sample. In addition, out-of-sample EBP predictions suffer from the unavailability of sample data to serve as a prior when estimating the conditional random area effect.

There are also noticeable differences in precision across in-sample and out-of-sample areas. We do not include a conditional random effect in the machine learning models, and the bootstrap procedure therefore calculates confidence intervals that are roughly the same size for in- and out-of-sample areas. Since out-of-sample accuracy is lower than in-sample accuracy, out-of-sample coverage rates tend to be lower as well; coverage drops to just 83 percent for XGBoost in Malawi assets, for example. 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 are therefore systematically different from in-sample areas. Although sample weights are included, out-of-sample predictions suffer from the relative paucity of training data from rural, less populated areas.

Our final set of results takes the four estimators and applies them in the context of an actual publicly-available household survey with offset GPS coordinates: the 2019 Integrated Household Survey (IHS) in Malawi. We show that the main findings are robust when using the IHS instead of one hundred simulated surveys drawn from the census. Accuracy measures show that the machine learning methods continue to outperform the EBP model. Confidence intervals for the machine learning models are all slightly conservative, leading to higher coverage rates than the 95-percent target. Compared with EBP, XGBoost achieves this with smaller confidence intervals for poverty both in and out of sample, similar size confidence intervals for the asset index in sample, and confidence intervals just half the size of EBP's out of sample (the latter of which is still underestimating uncertainty). These findings indicate that, in these settings, the machine learning methods generate estimates that are more accurate and more precise than EBP, regardless of whether we are using simulated surveys drawn from the censuses or the actual IHS in thee case of Malawi.

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, Hersh, and Newhouse 2022; Newhouse et al. 2022; Chi et al. 2022). However, image processing is computational intensive and unwieldy. In comparison, other types of data are easier to use, like mobile phone call data records (Aiken et al. 2022; Blumenstock, Cadamuro, and On 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.

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 and Herriot 1979; Battese, Harter, and Fuller 1988; Rao and Molina 2015). Earlier work proposed the use of census data for prediction (Elbers, Lanjouw, and Lanjouw 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 (conditional) random effects specified at the target area level. A simple table with coefficients indicates exactly how each variable is related to the measure of household welfare. On the other hand, the machine learning methods, while generating more accurate predictions, suffer from a lack of parsimony and transparency (Efron 2020). While analytical techniques

can determine which features are most predictive in a machine learning model, 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, Tibshirani, and Wager 2019). To the best of our knowledge, this is the first paper to show rigorously that the Random Effect Block bootstrap estimates accurate confidence intervals for predictions from tree-based machine learning methods in this context, when validated against unit-level census data as ground truth. In contexts where economists and statisticians are willing to sacrifice parsimony and transparency to achieve more accurate predictions, the availability of a simple and accurate bootstrap method for estimating uncertainty surmounts a crucial barrier to the use of tree-based machine learning algorithms.

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, including the simulation results, both in and out-of sample. We discuss estimates from the 2019 Malawi IHS in section 4 before concluding in section 5.

2 Methods

This paper evaluates four different methods for generating predictions of district-level poverty rates: Linear Empirical Best Predictor (EBP) models (Battese, Harter, and Fuller 1988; Jiang and Lahiri 2006; Molina and Rao 2010), Cubist regression models (Wang and Witten 1997; Quinlan 1992), extreme gradient boosting (Chen and Guestrin 2016), and Boosted Regression Forest (BRF) models (Friedberg et al. 2020). Importantly, we evaluate these methods in the context of developing countries, where census data is typically collected rarely. 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 random effects block bootstrap procedure (Chambers and Chandra 2013) to estimate uncertainty for the machine-learning models.

These techniques improve on existing methods but require rigorous evaluation of their accuracy and 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 four countries: Madagascar, Malawi, Mozambique, and Sri Lanka. These countries were selected because of the availabilty of census data with either enumeration area geocoordinates or sub-area identifiers with corresponding shapefiles. We present information related to the censuses in Table 2. The official administrative boundaries available in shapefiles differ quite substantially

Table 2: Census data statistics

	Madagascar	Malawi	Mozambique	Sri Lanka
Area	Commune	Traditional Authority	Locality	DS Division
(count)	1,515	420	1,258	331
Subarea	Fokontany	Enumeration area	Bairro	GN Division
(count)	14,412	18,700	65,707	13,984
Sample or full?	Full	20%	Full	Full
Households (count)	5,007,602	796,925	5,992,349	4,842,300
Year	2017	2018	2017	2012

across countries. In general, we pull geospatial data for the lowest administrative level possible. In Madagascar, this is the Fokontany, while in Malawi it is the Enumeration Area, the Bairro in Mozambique, and GN Divisions[1] in Sri Lanka. We refer to these levels as "subareas" throughout this paper.

While we pull geospatial data at the subarea level, the target areas for prediction are one level above this in all cases. In Madagascar, this corresponds to the Commune; in Malawi it is the Traditional Authority; in Mozambique it is Localities (Localidades); and in Sri Lanka it is Divisional Secretary's Divisions (DS Divisions). These are below the levels at which the household survey is considered to be representative. There are large differences across countries in how administrative units are allocated across space, with some countries having relatively larger units and others relatively smaller ones.

For Madagascar, Mozambique, and Sri Lanka, we have access to the full set of unit-level census data for all households. In Malawi, we have a 20-percent extract, which is a random sample of the entire population.

2.1 Outcomes

We focus on two separate measures of welfare, an asset index and poverty rates, although we only estimate poverty rates in Malawi. In Malawi, in addition to the census data, we utilize a near-contemporaneous household survey – the 2019 IHS – that collected expenditure data and is made publicly available by the World Bank's Living Standards Measurement Survey Program.[2] Using that data, we predict household per capita expenditures for all the households in the census, and classify the bottom half of the distribution as poor. Thus poverty in based on predicted per capita consumption. Appendix B provides more data on the imputation procedure as well as the calculation of the asset index.

[1] "GN Divisions" stands for "Grama Niladhari" Divisions.

[2] The IHS is part of the Living Standards Measurement Survey project at the World Bank. Surveys under this project are generally implemented by country-specific national statistics offices but with support from the World Bank. More information on the 2019 IHS and the LSMS Program more generally is available on the World Bank website: https://microdata.worldbank.org/index.php/catalog/3818

2.2 Geospatial features

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. We pull geospatial features from Google Earth Engine using the rgee package in R. 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 during the year, and minimum temperature during the year, while data on pollution is used to generate many distinct indicators.[3] In addition, we also create features by aggregating to higher levels by taking means, medians, maximums, or minimums. By combining these features in different ways and across different levels of aggregation, we end up with more than 130 different predictive features. While we include all of these features in some methods, we use lasso to select features for the EBP model, following others (Engstrom, Hersh, and Newhouse 2022; Newhouse et al. 2022; Masaki et al. 2022).[4] We return to these points below.

2.3 Linear Empirical Best Predictor Models

We utilize the EMDIplus package in R to estimate the linear Empirical Best Predictor model.[5] 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 use a household-level model, which models household-level expenditures as a function of the chosen covariates.[6] While we include descriptions of the other estimators in the appendix, we spend some time here on EBP since it is the current workhorse method for small area estimation of welfare.

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

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

[3] For example, pollution data includes carbon monoxide, carbon dioxide, ozone, etc.

[4] We generally pull this data from the same year as the census. This is not possible for all indicators in Sri Lanka – which was several years before the other censuses – so we instead pull the most recent year wherever necessary.

[5] The package is a spin-off of the EMDI package developed by Ifeanyi Edochie and available for download from his GitHub page: https://github.com/SSA-Statistical-Team-Projects/SAEplus.

[6] The use of aggregate predictors/features in a household-level model is sometimes referred to as a "unit-context model." Unit-context models tend to generate more accurate and precise predictions than area-level models due to their ability to use sub-area level predictors Newhouse et al. (2022).

where $G(y_{hsa})$ is some transformation of outcome y for household h in sub-area s in area a in region r,[7] X_{sar} is a vector of sub-area-specific geospatial features, X_{ar} is a vector of area-specific geospatial features, γ_r is a set of region fixed effects, η_{ar} is an area-level random effect, and ε_{hsar} is a classical error term.

When estimating Malawian poverty rates, we use the rank order transformation proposed by Peterson and Cavanaugh (2019) and implemented in (Masaki et al. 2022), though the results are robust to using the log shift transformation recommended in Tzavidis et al. (2018). We set a specific threshold for the poverty line to match the poverty rate in the survey data., which is around 50 percent. We estimate the mean of the asset index directly. In both cases, the software estimates Equation 1 at the household level, with the conditional random effect effectively using the survey data as a prior estimate that is updated using predictions from the model. Once the model is estimated, poverty estimates for areas are generated by repeatedly drawing random area effects and idiosyncratic error terms from their estimated normal distributions, generating simulated poverty estimates one hundred times for each sub-area, and aggregating across sub-areas to the target area level. 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 estimation process, 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. To avoid this, we select features using LASSO (R. Tibshirani 1996), implemented using the R package glmnet (Friedman, Hastie, and Tibshirani 2010). We select the optimal lambda using cross validation. However, we are particularly concerned about the hierarchical structure of the data leading to information leakage across folds in the cross validation routine. Therefore, we modify the LASSO algorithm to assign areas instead of individual households to cross validation folds. We also allow the regional fixed effects to enter unpenalized; in other words, we force LASSO to select all regional dummies. Because the surveys are considered to be representative at the regional level, this improves model fit by prioritizing variables that explain within-region variation for selection. We implement this routine on each of the one hundred survey draws.

2.4 Cubist

The second prediction methods 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

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

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 cubist 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[8] 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.

2.7 Uncertainty estimates for ML estimators

For EBP, we model assets and monetary welfare at the household level, which is 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

[8] https://xgboost.readthedocs.io/en/stable/tutorials/model.html

uncertainty and we use the canned implementation of this procedure from the EMDIplus package, following González-Manteiga et al. (2008).

For the three ML estimators, we estimate the models at the subarea level.[9] In other words, we aggregate each sample to the subarea level, estimate the model at the subarea level, and then manually aggregate the predictions to the area level using estimated population from WorldPop as aggregation weights.

While Friedberg et al. (2020) prove a central limit theorem for local linear forests that allows for the construction of uncertainty estimates, their theory and the implementation of it in the accompanying R package only allow for uncertainty estimates at the same level as that of the estimation itself. As an alternative, we propose a non-parametric bootstrap procedure that draws from Chambers and Chandra (2013) and Krennmair and Schmid (2022). 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:

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

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 repeat this residual bootstrap 1,000 times and calculate the 2.5th percentile (25th value), the 97.5th percentile (975th value), and the standard deviation. We use the 2.5th and 97.5th percentile to calculate confidence intervals, not the standard deviation of the estimate.

[9] While predicting at the area level is also an option, this would discard important variation in the predictor variables. As such, we opt to estimate at the subarea level, instead.

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.
6. Aggregate these new predictions to the area level.
7. Bootstrap with replacement from $\hat{R}_{a}$ for all areas.
8. Repeat steps 5 through 7 1,000 times, saving new area estimates each replication.
9. Calculate percentiles and standard deviation 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.

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.[10]

2.8 Evaluating performance

We are fortunate to have access to unit-record census data to evaluate the performance of each data. 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. We draw 500 separate subareas with probability proportional to population size, using the census to define population sizes. We then randomly draw eight households from within each selected subareas. This results in a sample of approximately 4,000 households in each iteration of the survey.[11]

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

[11] In the rare cases in which a subarea contains less than eight households, we draw all households. This can lead to some sampling iterations to have slightly less than 4,000 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 as $\left|\hat{y}_i - y_i^{truth}\right|$. We present the average across all areas and all simulations:

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

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

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

- Width of confidence interval: For the EBP estimates, this is derived from the estimated MSE. [12] For the upper and lower bounds of the CI for the three ML estimators, it is derived using percentiles from a bootstrap distribution.
- 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:[ˆ13]

$$\frac{1}{100N}\sum_{s=1}^{100}\sum_{i=1}^{N} 1 \cdot I\left(y_i^{truth} \in [CI_i^{lower}, CI_i^{upper}]\right) \tag{7}$$

3 Results

We first examine the accuracy of the four candidate estimators. We present average accuracy statistics across all one hundred samples in Table 3. The table presents four separate indicators of accuracy – Pearson

[12]12

Table 3: Accuracy statistics across simulations

	EBP	cubist	XGBoost	BRF
Madagascar (assets)				
corr. (pearson)	0.830	0.881	0.897	0.893
corr. (spearman)	0.750	0.808	0.824	0.836
absolute dev.	0.259	0.209	0.231	0.230
squared dev.	0.109	0.070	0.078	0.072
Malawi (assets)				
corr. (pearson)	0.801	0.870	0.889	0.881
corr. (spearman)	0.810	0.863	0.879	0.882
absolute dev.	0.289	0.230	0.215	0.237
squared dev.	0.220	0.147	0.125	0.152
Malawi (poverty)				
corr. (pearson)	0.834	0.902	0.915	0.923
corr. (spearman)	0.797	0.862	0.877	0.887
absolute dev.	0.147	0.079	0.076	0.103
squared dev.	0.036	0.015	0.013	0.018
Mozambique (assets)				
corr. (pearson)	0.867	0.886	0.920	0.922
corr. (spearman)	0.758	0.786	0.816	0.841
absolute dev.	0.204	0.174	0.168	0.168
squared dev.	0.074	0.061	0.047	0.046
Sri Lanka (assets)				
corr. (pearson)	0.867	0.895	0.912	0.884
corr. (spearman)	0.850	0.880	0.903	0.896
absolute dev.	0.167	0.146	0.135	0.162
squared dev.	0.053	0.042	0.036	0.058

Note: Measures of accuracy are averages across all 100 independent samples drawn from the repsective country census. These are simple, unweighted averages across areas. EBP refers to small area estimation and BRF refers to local linear forests.

correlation, Spearman (rank) correlation, absolute deviation, and squared deviation. The values are averages across all areas and the one hundred samples.

There are large differences across estimators. When looking at correlations, there is a noticeable difference between traditional EBP – which has been the workhorse of small area prediction for decades – and the three machine-learning estimators; the latter consistently outperform traditional EBP across all countries and outcomes. In fact, there is not a single case where traditional EBP is more highly correlated with the reference measure than any of the ML methods, for any of the outcomes. Among the three ML methods, BRF and XGBoost perform slightly better in terms of correlations, though the differences are not large and the ranking of the two is not consistent.

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, we also examine deviations from truth, which are more sensitive to bias. When looking at absolute and squared deviations, the three ML estimators again substantially outperform EBP, at least on average. XGBoost is generally most accurate across the five outcomes, although cubist and BRF perform better in Madagascar. Some of the differences are quite large. Relative to traditional EBP, for example, XGBoost's absolute deviation is 10.9 percent lower in Madagascar and 18.9 percent lower in Sri Lanka. The largest differences are in Malawi, particularly for poverty; traditional EBP's absolute deviation is 94.9 percent larger than XGBoost's. Squared deviation is even starker: traditional EBP's absolute deviation is 183.6 percent larger than XGBoost's. All three ML methods vastly outperform traditional EBP wehn predicting poverty in Malawi. For assets, the average squared deviation across all four countries is 0.114 for EBP, 0.080 for cubist, 0.072 for XGBoost, and 0.082 for BRF. Overall, XGBoost appears to be the most accurate method in terms of deviations, and traditional EBP the least accurate.

However, accuracy is of course not the only measure of concern. In particular, it is important to accurately estimate measures of uncertainty, since these are often used to determine whether the estimates are sufficiently precise to be published. While machine learning methods have been consistently shown to be good predictors of a number of outcomes, they generally have had less success when it comes to measures of uncertainty. In Table 4, we present two key statistics of uncertainty, with the uncertainty measures calculated using a random-effect block residual bootstrap, as described in the methodology section. The first statistic is the coverage rate, which shows how often the true value (from the census) is within the confidence intervals for a given prediction. Since we calculate 95% confidence intervals, we expect the coverage rates to be around 0.95.

Overall, coverage rates for all four estimators are quite good, with all rates around 95 percent. Average coverage rates for assets are 0.950 for EBP, 0.966 for cubist, 0.940 for XGBoost, and 0.968 for BRF, while overall coverage rates including poverty in Malawi are 0.916 for EBP, 0.959 for cubist, 0.938 for XGBoost, and 0.961 for BRF. The large decrease in coverage rates for EBP comes from a coverage rate of just 0.780 for poverty, which may be partly due to the less accurate predictions for this particular outcome.

There is, however, variation in coverage rates. For example, coverage rates for XGBoost are closer to 0.9 for assets in Madagascar while BRF goes as high as almost 0.99 for assets in Mozambique. Some of the lower coverage rates are due to out-of-sample estimates, a point to which we return below. Nonetheless, average coverage rates are quite good for all estimators. This is particularly notable for the three ML estimators, since unlike the older, more established estimation methods for EBP, no widespread consensus exists on how best to measure uncertainty.

Table 4: Uncertainty statistics across simulations

	EBP	cubist	XGBoost	BRF
Madagascar (assets)				
coverage	0.953	0.974	0.894	0.978
CI width	1.334	1.224	0.761	1.147
Malawi (assets)				
coverage	0.925	0.925	0.915	0.936
CI width	1.312	1.365	1.122	1.646
Malawi (poverty)				
coverage	0.780	0.931	0.928	0.934
CI width	0.516	0.591	0.518	0.641
Mozambique (assets)				
coverage	0.957	0.983	0.979	0.987
CI width	1.259	1.496	1.101	1.345
Sri Lanka (assets)				
coverage	0.964	0.980	0.973	0.971
CI width	0.941	1.177	1.039	1.387

Note: Measures of uncertainty are averages across all 100 independent samples drawn from the repsective country census. These are simple, unweighted averages across areas. EBP refers to small area estimation and BRF refers to local linear forests. We do not use actual standard errors when calculating coverage rates for Cubist, XGBoost, and BRF. Instead, we use the appropriate percentiles of the bootstrapped distribution. As such, we present the width of the 95-percent confidence interval instead of the standard errors.

In addition to coverage rates, we also present the average width of confidence intervals ("CI width") as an additional metric. For EBP, the upper and lower confidence intervals are estimated by adding or subtracting 1.96 times the square root of the estimated mean squared error to the point estimate. For the machine learning methods, the CIs are calcuated using bootstrap percentiles for the three other estimators; as such, we present CI widths instead of standard errors in the table. In line with the (generally) improved accuracy of XGBoost over other methods, XGBoost also tends to have the narrowest confidence intervals.[13] XGBoost has the smallest intervals except for Malawi poverty and Sri Lanka assets. However, the smallest for Malawi poverty is traditional EBP, which achieves a coverage rate of just 0.780. Of the other cases, only for Sri Lanka assets is XGBoost more uncertain than the others. As noted, though, XGBoost, does slightly underestimate uncertainty for Madagascar.

Not surprisingly, the results show a trade-off between estimated precision and coverage rates. While XGBoost achieves accurate coverage on average with the smallest confidence intervals, it also clearly underestimates uncertainty slightly in at least one situation. Nonetheless, the coverage rates are still quite respectable in Madagascar, at around 0.9, nowhere near the worst performance of EBP (0.780 for poverty in Malawi). On

[13]Since we use a residual bootstrap, higher accuracy leads to smaller confidence intervals by construction. However, it is not a one-to-one improvement, since the bootstrap consists of two steps – using residuals of both areas and subareas – while we only present results for areas.

the other hand, BRF and cubist never underestimate uncertainty to the same extent as XGBoost does in Madagascar, but they do so with much larger confidence intervals.

3.1 In and out-of-sample estimates

We next look at accuracy and precision based on whether an area is included in the sample or not. In each of the 100 samples per country, we randomly select subareas, with probability proportional to size, and then randomly select households from within each subarea. Not all areas appear in every sample, with less populous areas more likely to excluded from the sample on a given draw. For this section, we define an area as "in sample" if at least one subarea from within that area is selected. We start with accuracy statistics in Table 5. The first five columns include in-sample areas only, while the last four columns include out-of-sample areas. The extra column for in-sample areas is due to direct estimates, which we also include as a way to gauge the performance of the other indicators. Statistics are means based on all 100 independent samples.

The direct estimates are in the first column. Across all outcomes, the direct estimate is actually quite accurate for four of the five outcomes, although it is just 0.794 for assets in Sri Lanka (and 0.735 for the rank correlation). Direct estimates consistently have the lowest correlation, followed by EBP, while the three machine learning methods tend to perform better. The main pattern in the table is that out-of-sample estimates are worse than in-sample estimates, which is true for all outcomes and all estimators. This is a result of calibrating the model on a sample that under-represents sparsely populated sub-areas. Since these areas are systematically different from larger areas, the accuracy the out-of-sample estimates suffers.

The average difference in performance in terms of correlations across the five outcomes is sometimes substantial. For example, (pearson) correlation for Malawi assets drops by 17.7 percent for EBP, by 15.2 percent for Cubist, by 11.9 percent for XGBoost, and by 11.3 percent for BRF. The smallest drop in performance is in Madagascar, where correlation drops by 13.7 percent for EBP, by 7.3 percent for Cubist, by 5.9 percent for XGBoost, and by 6.5 percent for BRF.

We also see variation in the accuracy as measured by deviations from truth. In Sri Lanka, squared deviation increases by 369.1 percent for EBP, by 676.2 percent for Cubist, by 601.2 percent for XGBoost, and by 541.9 percent for BRF. Malawi assets again shows the largest increases in deviation, across all estimators and both absolute and squared deviation.

Despite these differences, a consistent pattern emerges both in and out of sample. Of the four candidate methods, EBP is consistently the least accurate – sometimes by large margins – both in and out of sample. The three ML estimators vary across countries and outcomes. For example, BRF is slightly better than

Table 5: Accuracy statistics across simulations by sample status

	In sample					Out of sample			
	direct	EBP	cubist	XGB	BRF	EBP	cubist	XGB	BRF
Madagascar (assets)									
corr. (pearson)	0.874	0.911	0.926	0.933	0.934	0.786	0.858	0.878	0.873
corr. (spearman)	0.784	0.818	0.846	0.865	0.864	0.727	0.794	0.811	0.827
absolute dev.	0.252	0.243	0.199	0.217	0.218	0.265	0.212	0.236	0.234
squared dev.	0.126	0.100	0.062	0.073	0.065	0.112	0.072	0.080	0.075
Malawi (assets)									
corr. (pearson)	0.870	0.891	0.953	0.955	0.945	0.734	0.808	0.841	0.838
corr. (spearman)	0.794	0.841	0.908	0.909	0.915	0.767	0.816	0.846	0.846
absolute dev.	0.255	0.194	0.137	0.133	0.143	0.410	0.346	0.319	0.357
squared dev.	0.124	0.083	0.037	0.034	0.045	0.391	0.284	0.241	0.288
Malawi (poverty)									
corr. (pearson)	0.805	0.867	0.947	0.945	0.955	0.795	0.852	0.879	0.886
corr. (spearman)	0.766	0.796	0.885	0.887	0.904	0.779	0.830	0.851	0.861
absolute dev.	0.129	0.150	0.063	0.066	0.095	0.143	0.098	0.087	0.113
squared dev.	0.030	0.034	0.007	0.008	0.014	0.038	0.024	0.019	0.024
Mozambique (assets)									
corr. (pearson)	0.870	0.919	0.942	0.956	0.957	0.838	0.853	0.900	0.903
corr. (spearman)	0.740	0.782	0.831	0.858	0.881	0.753	0.768	0.802	0.827
absolute dev.	0.237	0.192	0.149	0.144	0.145	0.208	0.183	0.177	0.177
squared dev.	0.115	0.065	0.042	0.035	0.034	0.078	0.069	0.052	0.051
Sri Lanka (assets)									
corr. (pearson)	0.794	0.895	0.915	0.913	0.897	0.799	0.843	0.901	0.863
corr. (spearman)	0.735	0.862	0.893	0.901	0.899	0.796	0.826	0.896	0.863
absolute dev.	0.233	0.142	0.125	0.125	0.137	0.221	0.192	0.158	0.216
squared dev.	0.095	0.034	0.027	0.030	0.038	0.092	0.073	0.050	0.102

Note: Measures of accuracy are averages across all 100 independent samples drawn from the repsective country census. These are simple, unweighted averages across areas. EBP refers to small area estimation and BRF refers to local linear forests. Sample status is defined separately in each independent sample, such that a single area can be in sample in one draw and out of sample in another. In-sample areas refers to areas in which at least one subarea is randomly selected to be included in the sample. Out-of-sample areas refers to areas in which not a single subarea is randomly selected.

Table 6: Uncertainty statistics across simulations

	In sample					Out of sample			
	direct	EBP	cubist	XGB	BRF	EBP	cubist	XGB	BRF
Madagascar (assets)									
coverage	0.818	0.885	0.979	0.909	0.986	0.976	0.972	0.890	0.976
CI width	0.859	1.009	1.216	0.760	1.169	1.444	1.227	0.762	1.140
Malawi (assets)									
coverage	0.895	0.941	0.987	0.983	0.987	0.905	0.850	0.829	0.871
CI width	1.068	0.982	1.328	1.106	1.605	1.727	1.410	1.143	1.698
Malawi (poverty)									
coverage	0.756	0.700	0.987	0.984	0.988	0.881	0.862	0.856	0.866
CI width	0.595	0.423	0.605	0.532	0.670	0.632	0.573	0.500	0.603
Mozambique (assets)									
coverage	0.850	0.865	0.990	0.988	0.994	0.991	0.980	0.975	0.984
CI width	0.950	0.772	1.477	1.092	1.349	1.439	1.504	1.105	1.343
Sri Lanka (assets)									
coverage	0.884	0.972	0.994	0.982	0.988	0.946	0.952	0.955	0.935
CI width	0.939	0.833	1.174	1.037	1.383	1.175	1.186	1.041	1.396

Note: Measures of uncertainty are averages across all 100 independent samples drawn from the repsective country census. These are simple, unweighted averages across areas. EBP refers to small area estimation and BRF refers to local linear forests. We do not use actual standard errors when calculating coverage rates for Cubist, XGBoost, and BRF. Instead, we use the appropriate percentiles of the bootstrapped distribution. As such, we present the width of the 95-percent confidence interval instead of the standard errors. Sample status is defined separately in each independent sample, such that a single area can be in sample in one draw and out of sample in another. In-sample areas refers to areas in which at least one subarea is randomly selected to be included in the sample. Out-of-sample areas refers to areas in which not a single subarea is randomly selected.

XGBoost in Madagascar, the two are approximately the same in Mozambique, and XGBoost is slightly better in Malawi and Sri Lanka.

For the three machine learning models with the random effects block bootstrap, in-sample coverage rates tend to be higher than 0.95 across all outcomes and the estimators. Madagascar, where coverage rates are 0.909 for XGBoost, is the only exception. EBP, on the other hand, underestimates uncertainty for three of the five outcomes out of the five, with coverage rates of 0.700 for Malawi poverty, which is slightly below results from other contexts. The direct estimates also consistently underestimate uncertainty, with an average coverage rate of 0.841

As noted above, out-of-sample estimates tend to be less accurate than in-sample estimates, because the sample under-represents sparsely populated areas. The lower out-of-sample accuracy in turn leads to lower coverage rates out of sample. This is particularly noticeable in Malawi, where coverage rates for assets are 0.905 for EBP, 0.850 for cubist, 0.829 for XGBoost, and 0.871 for BRF. Uncertainty is similarly underestimated for poverty in Malawi, with coverage rates of 0.881 for EBP, 0.862 for cubist, 0.856 for XGBoost, and 0.866

for BRF. Meanwhile, in Madagascar, XGBoost slightly underestimates uncertainty while the other three estimators slightly overestimate uncertainty.

One disadvantage of the random effects block bootstrap is that it can only utilize data from sampled sub-areas and areas, and therefore cannot effectively distinguish between sampled and un-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 additional source of model error is not reflected in uncertainty estimates. There are possible alternatives, such as using a parametric bootstrap – which is how EBP calculates uncertainty – or modeling heteroscedasticity, as in Elbers, Lanjouw, and Lanjouw (2003). However, the random effect block bootstrap does quite well in the simulations reported above, with coverage rates always exceeding 85 percent out of sample. In these cases, the tendency for the bootstrap procedure to underestimate coverage out of sample may not be a major issue.

There are two possible reasons why out-of-sample predictions are less accurate than in-sample predictions. The first is that the model is overfit. However, this seems very unlikely given that regularization methods are used to avoid overfitting to the particular sample. For example, when implementing EBP, we select a model using LASSO to avoid overfitting. Meanwhile, The ML estimators employ different different methods to help avoid overfitting, such as using a random subset of data and/or variables across trees and splits, and estimating regularized objective functions. Another more likely explanation is that that smaller areas – which are much less likely to appear in the sample – differ systematically from sampled areas and are therefore related to the predictors in a different way.

To shed light on this, we examine differences in accuracy between estimates for sampled and non-sampled areas, and how these change 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 7 presents a set of regressions where we estimate the difference in (log) 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 in the second column we also include area fixed effects that restrict identification to within-area changes in sample status across simulations.

The first column in each pair shows that in-sample areas have much higher accuracy (lower absolute deviations) than out-of-sample areas. The differences ranges from approximately 12 to 39 percent. There are several important differences in the second column when area fixed effects are included. First, focusing on XGBoost, the inclusion of fixed effects substantially decreases the difference in accuracy based on sample status. For example, the fixed-effect coefficient is only 61.9 percent as large as the first coefficient in Madagascar, only

Table 7: Difference in accuracy within areas: in-sample vs. out-of-sample

	EBP		cubist		XGB		BRF	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
madagascar	-0.122***	-0.102***	-0.051***	-0.041***	-0.105***	-0.065***	-0.067***	-0.038***
	(0.020)	(0.016)	(0.018)	(0.006)	(0.018)	(0.011)	(0.019)	(0.006)
malawi (assets)	-0.624***	-0.005	-0.776***	-0.209***	-0.743***	-0.162***	-0.792***	-0.096***
	(0.053)	(0.015)	(0.054)	(0.016)	(0.058)	(0.024)	(0.065)	(0.011)
malawi (poor)	0.247***	0.020**	-0.020	-0.142***	-0.009	0.002	0.040	-0.057***
	(0.062)	(0.009)	(0.085)	(0.019)	(0.065)	(0.025)	(0.069)	(0.011)
mozambique	-0.123***	-0.028	-0.182***	-0.082***	-0.216***	-0.095***	-0.212***	-0.092***
	(0.022)	(0.021)	(0.020)	(0.008)	(0.020)	(0.012)	(0.024)	(0.007)
sri lanka	-0.387***	-0.172***	-0.379***	-0.178***	-0.270***	-0.002	-0.366***	-0.112***
	(0.039)	(0.023)	(0.038)	(0.016)	(0.037)	(0.015)	(0.055)	(0.013)
Fixed effects:								
simulation	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
area	No	Yes	No	Yes	No	Yes	No	Yes

Standard errors are clustered at the simulation and area level. Each row and column 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.
*p<0.1 **p<0.05 ***p<0.01

21.8 percent as large for Malawi assets, and only 44.0 percent as large in Mozambique. In Sri Lanka, the coefficient basically decreases to zero, while both coefficients are zero for poverty in Malawi. In other words, once we take into account differences in characteristics across areas, the difference in accuracy based on sample status shrinks substantially. For two of the outcomes, in fact, sample status does not predict any differences in accuracy. This suggests that systematic differences between sampled and non-sampled areas explain a large share of the difference in accuracy based on sample status.

3.2 Accuracy for poorer and richer areas

Figure 1 looks at how accuracy varies for poorer and wealthier areas. On the first y-axis is mean squared deviation across samples and on the x-axis is the true value for each area. On the second y-axis, we include the kernel density estimate of the true value. Several patterns emerge. First, accuracy tends to be lowest where density is lowest. In other words, predictions are most accurate in areas of common support. All prediction methods are far more accurate for interpolating within the sample than for extrapolating where there is little common support.

Second, areas of low density in the sample do not always correspond to smaller areas. In Madagascar, 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). Nonetheless, we see the worst predictions in these areas, because they are systematically different from most areas in the sample. In other words, the problem is more complex than simply whether an area is included 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. The results also higlights the importance of accurately reporting large confidence intervals at the tails of the distribution.

Figure A2 in the appendix shows predicted-true plots for all five outcomes, using local polynomial smoothing in order to better trace out average predictions across the range of truth. In general, average predictions hew relatively closely to truth, represented by the 45-degree dashed line. The most notable exception to this is assets in Malawi, where the tails are rather poorly estimated by all methods. However, Figure A3 in the appendix presents the accompanying scatterplot. The scatterplot makes clear that the Malawi result is really driven by just a few areas in the tails. In the lower tail, this is just a single area. While the upper tail has more areas laying further from the 45-degree line, the majority of these are not as extreme as the polynomial smoothing suggests. Finally, Figure **??** shows predicted-true plots for all five outcomes. The general patterns continue to enforce the same conclusions.

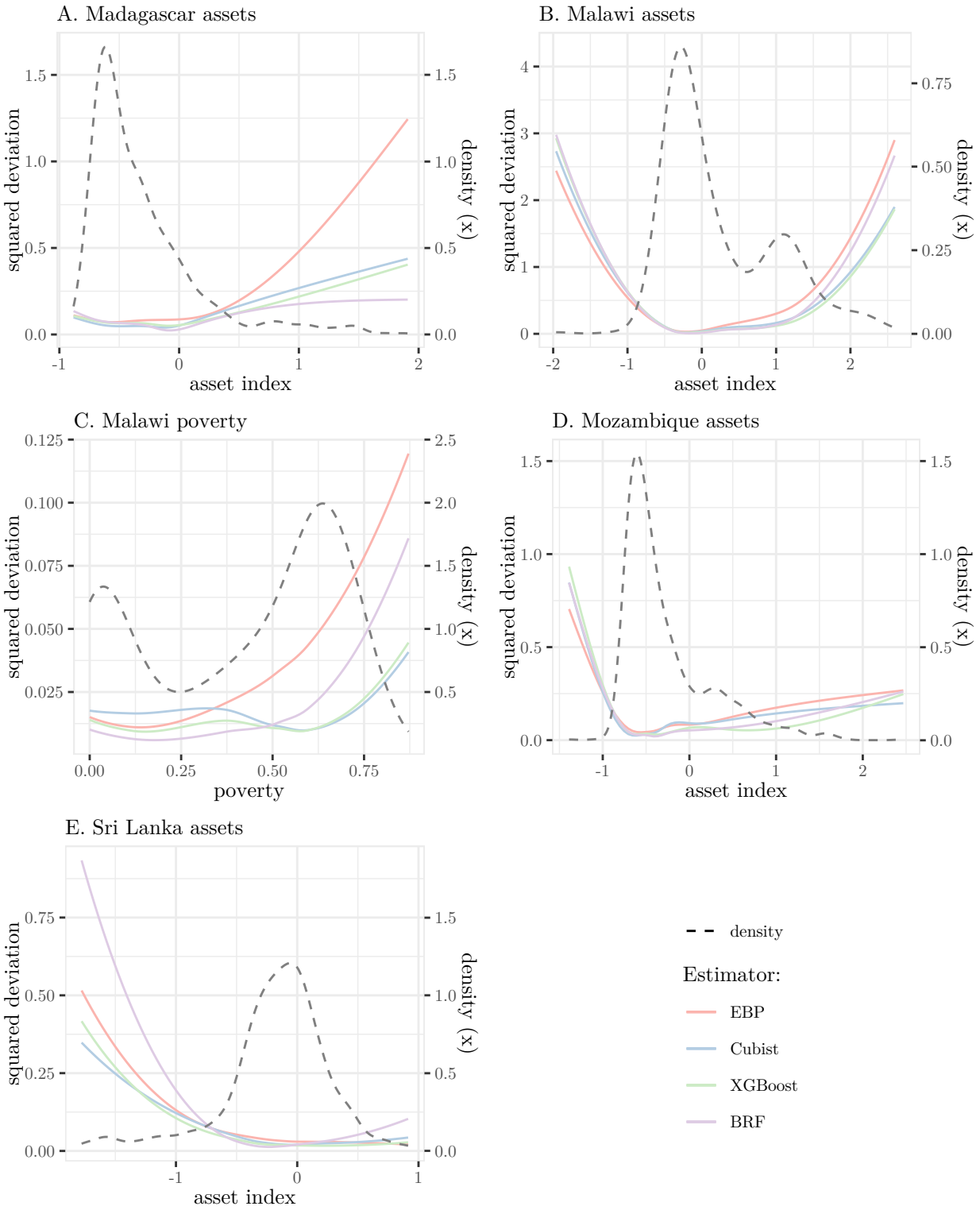
Consistent with previous results, we again see both XGBoost and BRF usually being closest to the 45-degree line. Importantly, EBP shows poor performance in the upper tail for poverty in Malawi, performing obviously worse than the other estimators. These areas are arguably the areas that are most important to predict accurately to guide interventions to reduce poverty.

3.3 Simple Cubist Regression

Two big advantages of traditional EBP are its parsimony and transparency. Models can be summarized in a simple regression table that indicates how each predictor is related to the outcome. The ML methods considered here, on the other hand, are opaque. This is because they aggregate large numbers of "weak learners" that are estimated using different subsets of candidate variables and observations. Because it is difficult to show a large number of weak learners in a tractable way, the predictions methods appear to be a black box.

Cubist Regression, when the estimation options are purposefully restricted, offers a middle ground between the transparency and parsimony of a linear model and the black box nature of more sophisticated machine learning methods. In the main results above, we tune hyperparameters for cubist in the same way as the others, to maximize prediction accuracy. However, we can set the parameters to make a more transparent

Figure 1: Deviations from truth vs. Welfare



Note: Both figures are smoothed conditional means of the mean squared deviation across 100 independent samples (first y-axis) on truth (x-axis). The kernel density estimate refers to the density of truth, which is on the x-axis.

model by allowing for a small number of rules and discarding "committees" (essentially multiple Cubist regression models). This option is not possible for XGBoost and BRF, given how they are designed. We therefore consider a particular simple cubist model, with three rules and no committee. Compared with EBP, this simple Cubist model benefits from being able to estimate three distinct linear models on different subsets of the data, which are themselves selected to maximize predictive accuracy. The model can be reported quiet simply as three regression tables with an additional row defining the relevant subsets of the data. Cubist regression, however, unlike EBP does not include a mixed effect that treats the direct survey estimate as a prior, and therefore discards some useful information in sampled areas.

We report the results from a simplified 3-rule Cubist model in Table 8. In column one, we present the EBP results (presented earlier) in order to directly compare those to a more transparent cubist model. We include all four accuracy statistics and both statistics related to uncertainty for each outcome. The average pearson correlation for EBP and the simplified cubist, across all five outcomes, is 0.840 and 0.848, respectively, while the respective statistics for rank correlation are 0.793 and 0.810. On average, the correlations of the simple cubist model are higher than those for EBP.

We can similarly compare deviations for the two estimators. In terms of absolute deviation, the average across the four outcomes with assets is 0.230 for EBP and 0.213 for cubist, while the squared deviations are 0.114 and 0.107, respectively. For poverty in Malawi, the simple cubist model vastly outperforms EBP in terms of deviations; the absolute deviation for the former is 25.5 percent smaller while the squared deviation is 30.0 percent smaller.

Overall, simple Cubist regression tends to outperform EBP in more cases than not, and thus may be a preferred option to EBP models, although the difference is not large. Furthermore, the simple Cubist model gives less accurate predictions in Sri Lanka. Another option, which we leave for further research, would be estimating a version of the simple Cubist models that includes a conditional random effect, bulding on Krennmair and Schmid (2022). This would incorporate the additional model flexibility that Cubist offers with the mixed effect framework offered by EBP. As for now, there appears to be a clear trade-off between accuracy (XGBoost and BRF) and transparency (EBP and simplified Cubist).

4 An application: The 2019 Malawi IHS

The results in the previous sections demonstrated that the three ML methods, particularly XGBoost and BRF, predict more accurately than EBP and that the random effects block bootstrap generates accurate

Table 8: Accuracy statistics across simulations

	EBP	Cubist (simple)
Madagascar (assets)		
corr. (pearson)	0.830	0.856
corr. (spearman)	0.750	0.793
absolute dev.	0.259	0.217
squared dev.	0.109	0.085
coverage	0.953	0.969
CI width	1.334	1.244
Malawi (assets)		
corr. (pearson)	0.801	0.801
corr. (spearman)	0.810	0.810
absolute dev.	0.289	0.289
squared dev.	0.220	0.220
coverage	0.925	0.914
CI width	1.312	1.467
Malawi (poverty)		
corr. (pearson)	0.834	0.848
corr. (spearman)	0.797	0.821
absolute dev.	0.147	0.110
squared dev.	0.036	0.025
coverage	0.780	0.909
CI width	0.516	0.584
Mozambique (assets)		
corr. (pearson)	0.867	0.882
corr. (spearman)	0.758	0.784
absolute dev.	0.204	0.175
squared dev.	0.074	0.063
coverage	0.957	0.979
CI width	1.259	1.429
Sri Lanka (assets)		
corr. (pearson)	0.867	0.853
corr. (spearman)	0.850	0.843
absolute dev.	0.167	0.172
squared dev.	0.053	0.061
coverage	0.964	0.971
CI width	0.941	1.256

Note: Measures of accuracy are averages across all 100 independent samples drawn from the repsective country census. These are simple, unweighted averages across areas. The first column is for EBP, while the second is for a simple version of cubist with just three rules and no committees.

estimates of uncertainty. However, the evidence presented so far relies on simulated samples. Though the samples come from real data, the sampling structure is somewhat simplified relative to most household surveys. To see whether these results carry over to an actual survey, we turn to the 2019 Malawi Fifth Integrated Household Survey (IHS).

The IHS took place the year following the census used above and is therefore well-timed to check whether the main results hold when using a real-world survey. The IHS collected information on both assets and expenditures. However, there are two minor concerns. First, the IHS took place the year after the census. While a year might not lead to large changes in a developed countries, this is not necessarily true in a country like Malawi, where rain-fed agriculture predominates and slightly different weather can lead to large changes in poverty. Second, we use the IHS in order to impute expenditures and poverty into the census. This could lead to information leakage for poverty which could lead to inflated measures of accuracy. This is not a concern with assets, however.

The 2019 IHS is a large survey. It consists of 11,249 households spread across 697 subareas (enumeration areas) and 321 areas (Traditional Authorities – TAs). In the country as a whole, there are 420 TAs, meaning that there are 99 out-of-sample areas in the survey. We can compare results in areas that are in the survey (in-sample areas) to these out-of-sample areas in order to see whether the prediction methods differ greatly in their predictive power based on whether an area is in the sample or not.

Table 9 presents accuracy statistics for in-sample areas (the first five columns) and out-of-sample areas (the last four columns). As before, we include direct estimates as a point of comparison. All four alternative estimators perform better than the direct estimates, across all accuracy statistics and for both assets and poverty. Since the IHS is a relatively large survey, the direct estimates are quite accurate; the in-sample pearson correlation is 0.906 for the mean asset index. However, as before, direct estimates are less accurate for poverty, at 0.804, because poverty rates are the mean of a discrete rather than continuous variable.

The overall patterns of accuracy remain consistent with the previous results. All four candidate small area estimation methods outperform the direct estimate, sometimes substantially. For both in-sample and out-of-sample areas, XGB and BRF perform noticeably better than EBP, while cubist performs better but less so. For assets in in-sample areas, the correlation is 2.8 percent higher for XGB than EBP, while the absolute deviation and squared deviation are 25.0 percent and 40.4 percent lower, respectively. For poverty, the differences are less pronounced.

We see relatively larger differences in out-of-sample areas, though the overall rank ordering is the same. For poverty, the correlation is 12.7 percent higher for XGBoost than EBP, while the absolute and squared

Table 9: Accuracy and uncertainty statistics for the 2019 Malawi IHS

	In sample					Out of sample			
	Direct	EBP	cubist	XGB	BRF	EBP	cubist	XGB	BRF
Assets									
corr. (pearson)	0.906	0.927	0.952	0.953	0.957	0.705	0.765	0.795	0.806
corr. (spearman)	0.859	0.871	0.900	0.904	0.906	0.725	0.777	0.813	0.817
abs. deviation	0.227	0.202	0.172	0.151	0.145	0.550	0.502	0.481	0.476
squared dev.	0.083	0.063	0.047	0.038	0.034	0.615	0.488	0.469	0.452
coverage	0.776	0.869	1.000	1.000	1.000	0.828	1.000	1.000	1.000
CI width	0.736	0.765	0.991	0.796	0.985	1.947	1.162	0.905	1.124
Poverty									
corr. (pearson)	0.804	0.886	0.895	0.895	0.913	0.799	0.815	0.832	0.828
corr. (spearman)	0.771	0.828	0.826	0.838	0.854	0.799	0.792	0.804	0.801
abs. deviation	0.119	0.089	0.086	0.086	0.080	0.140	0.128	0.129	0.134
squared dev.	0.023	0.013	0.012	0.012	0.010	0.036	0.032	0.029	0.031
coverage	0.882	0.882	1.000	1.000	1.000	0.990	1.000	1.000	1.000
CI width	0.574	0.393	0.432	0.335	0.427	0.861	0.462	0.351	0.455

Note: Measures of accuracy and uncertainty are averages across all areas.

deviations are 12.6 percent lower and 23.8 percent lower, respectively. The accuracy decreases markedly for all estimators when looking at assets, although there is a much less noticeable absolute drop off for poverty, especially for XGBoost. XGBoost sees a drop of 16.6 percent for pearson correlations and an increase of percent for absolute deviations and 68.6 percent for squared deviations for in-sample relative to out-of-sample areas, leading to XGBoost outperforming all other estimators for out-of-sample areas for poverty, though BRF appears very slightly more accurate for assets. Given that we only have one sample and how close some of the statistics are, however, it is difficult to determine conclusively whether XGBoost or BRF is most accurate.

Table 9 also presents the two statistics related to uncertainty: the coverage rate and the mean width of confidence intervals. Both the direct estimates and EBP underestimate standard errors, with the latter doing so for three of the four different subsamples, with only out-of-sample poverty being above 0.95 while the rest are all below 0.9. On the other hand, the three ML estimators overestimate uncertainty; not a single area falls outside the confidence intervals. Importantly, XGBoost accomplishes this with the smallest CIs, on average, among the estimators, with only EBP for in-sample assets being smaller. Out of sample, the widths of the CIs for XGBoost are 53.5 percent lower for assets and 59.3 percent lower for poverty than EBP, despite XGBoost obtaining higher coverage rates in both instances.

Comparing the out-of-sample accuracy penalty for assets and poverty gives some insight into the importance of different reference measures. In general, the out-of-sample penalty tends to be higher for assets than for poverty. This is surprising given the difference in how the reference benchmark is constructed. For assets, the

reference benchmark is measured directly in the census. Meanwhile, for poverty, the reference benchmark is derived from a model estimated using the 2019 IHS. By construction, the model is estimated using data that excludes non-sampled areas, which are systematically different from sampled areas. One might expect this to lead to particularly large discrepancies for poverty out of sample, compared both with in-sample estimates and with assets. The fact that the opposite is true suggests that in this context, that the non-representative nature of the IFS has limited impact on the comparison of accuracy across methods in out-of-sample areas.

Finally, Figure A6 in the appendix shows the predicted-true plot for assets (left panel) and poverty (right panel), again using conditional mean smoothed lines. The patterns are generally similare to the previous section for Malawi. The assets predictions are accurate in the middle of the distribution but poorly predicted in the upper and lower tails, but Figure A5 in the appendix shows that the lower range, in particular, is driven by just one or two areas.

Although these results are estimated using an independent survey, they generally echo the findings in the previous section. XGBoost and BRF, on average, perform best on measures of both accuracy and precision. EBP is least accurate of the four prediction methods, though EBP still offers a significant improvement on direct estimates in-sample. In all cases, in-sample estimates are more accurate than out-of-sample estimates by a substantial margin.

5 Conclusion

This paper evaluates the use of tree-based machine learning techniques for the purposes of small area estimation of wealth and poverty using household census data from four countries. In addition to evaluating accuracy, it suggests and implement a bootstrap routine for the machine learning estimators and validates its performance using real-world data. The three tree-based machine learning methods evaluated – Cubist Regression, Extreme Gradient Boosting (XGboost), and Boosted Regression Forests (BRF) – all significantly outperformed the linear mixed model traditionally used for small area estimation (Molina and Rao 2010; Tzavidis et al. 2018). Of these three methods, XGboost tended to perform equal to BRF except in the left tail of the Sri Lanka distribution wealth distribution where it was far more accurate. However, although the rank-ordering across statistics is not consistent and BRF is on more solid ground theoretically. Meanwhile, predictions generated using Cubist regression models are generally a bit less accurate.

The results make a strong case for the use of XGBoost or BRF in cases where transparency and parsimony are not first-order concerns. In other cases where transparency and parsimony are important, linear mixed

EBP models or Cubist regression with a small number of rules are viable options. The Cubist regression option allows for more flexible models while the EBP approach contains a mixed effect that allows the survey data to serve as a prior. Because predicted values are estimated fairly accurately and precisely relative to direct estimates, it does not appear that the inclusion of the mixed effect is a crucial element in generating accurate estimates in this context, as its beneficial effect is generally outweighed by the more flexible set of linear models offered by a simple Cubist regression with three rules. There are also methods that combine machine learning with mixed effects (Krennmair and Schmid 2022), though these cannot yet be applied to gradient boosting using readily available software.

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. This difference occurs despite the use of sampling weights and regularization methods to prevent overfitting. These differences are much diminished when comparing accuracy estimates within the same area across simulations. Nonetheless, the tendency for samples to under-represent less densely populated areas can lead to less accurate predictions out of sample. This finding underlines the benefits of including all target areas in the sample. 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, which implies that there are potential benefits of oversampling tails. Further research could 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 random effect block residual bootstrap, as proposed by Chambers and Chandra (2013). These estimates perform well across the board, as coverage rates never fall below 98 percent in-sample and 85 percent out of sample. Overall, the results indicate that the combination of Extreme Gradient Boosting or Boosted Regression Forests, the random effect block residual bootstrap, and publicly available geospatial data offer a practical way to significantly improve on both direct survey estimates and EBP estimates when geolocated survey data are available.

References

Aiken, Emily, Guadalupe Bedoya, Joshua Blumenstock, and Aidan Coville. 2022. "Program targeting with machine learning and mobile phone data: Evidence from an anti-poverty intervention in Afghanistan." *arXiv Preprint arXiv:2206.11400*.

Athey, Susan, Julie Tibshirani, and Stefan Wager. 2019. "Generalized Random Forests." *The Annals of Statistics* 47 (2): 1148–78.

Atkinson, Anthony B. 2019. *Measuring Poverty Around the World*. Princeton University Press.

Battese, George E, Rachel M Harter, and Wayne A Fuller. 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.

Blumenstock, Joshua. 2016. "Fighting Poverty with Data." *Science* 353 (6301): 753–54.

Blumenstock, Joshua, Gabriel Cadamuro, and Robert On. 2015. "Predicting Poverty and Wealth from Mobile Phone Metadata." *Science* 350 (6264): 1073–76.

Carter, Grace M, and John E Rolph. 1974. "Empirical Bayes Methods Applied to Estimating Fire Alarm Probabilities." *Journal of the American Statistical Association* 69 (348): 880–85.

Chambers, Raymond, and Hukum Chandra. 2013. "A random effect block bootstrap for clustered data." *Journal of Computational and Graphical Statistics* 22 (2): 452–70.

Chen, Tianqi, and Carlos Guestrin. 2016. "XGBoost: A scalable tree boosting system." In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 785–94.

Chi, Guanghua, Han Fang, Sourav Chatterjee, and Joshua Blumenstock. 2022. "Microestimates of Wealth for All Low-and Middle-Income Countries." *Proceedings of the National Academy of Sciences* 119 (3): e2113658119.

Donaldson, Dave, and Adam Storeygard. 2016. "The view from above: Applications of satellite data in economics." *Journal of Economic Perspectives* 30 (4): 171–98.

Efron, Bradley. 2020. "Prediction, Estimation, and Attribution." *International Statistical Review* 88: S28–59.

Efron, Bradley, and Carl Morris. 1973. "Stein's estimation rule and its competitors–an empirical Bayes approach." *Journal of the American Statistical Association* 68 (341): 117–30.

Elbers, Chris, Tomoki Fujii, Peter Lanjouw, Berk Özler, and Wesley Yin. 2007. "Poverty alleviation through geographic targeting: How much does disaggregation help?" *Journal of Development Economics* 83 (1): 198–213.

Elbers, Chris, Jean O Lanjouw, and Peter Lanjouw. 2003. "Micro-Level Estimation of Poverty and Inequality." *Econometrica* 71 (1): 355–64.

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

Fay, Robert E, and Roger A Herriot. 1979. "Estimates of income for small places: an application of James-Stein procedures to census data." *Journal of the American Statistical Association* 74 (366a): 269–77.

Friedberg, Rina, Julie Tibshirani, Susan Athey, and Stefan Wager. 2020. "Local Linear Forests." *Journal of Computational and Graphical Statistics* 30 (2): 503–17.

Friedman, Jerome. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics*, 1189–1232.

Friedman, Jerome, Trevor Hastie, and Rob Tibshirani. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33 (1): 1.

Fujii, Tomoki, and Roy van der Weide. 2020. "Is Predicted Data a Viable Alternative to Real Data?" *The World Bank Economic Review* 34 (2): 485–508.

González-Manteiga, Wenceslao, Maria J Lombardía, Isabel Molina, Domingo Morales, and Laureano Santamaría. 2008. "Bootstrap Mean Squared Error of a Small-Area EBLUP." *Journal of Statistical Computation and Simulation* 78 (5): 443–62.

Henderson, J Vernon, Adam Storeygard, and David N Weil. 2012. "Measuring Economic Growth from Outer Space." *American Economic Review* 102 (2): 994–1028.

Jean, Neal, Marshall Burke, Michael Xie, W Matthew Davis, David B Lobell, and Stefano Ermon. 2016. "Combining satellite imagery and machine learning to predict poverty." *Science* 353 (6301): 790–94.

Jiang, Jiming, and Partha Lahiri. 2006. "Mixed Model Prediction and Small Area Estimation." *TEST* 15 (1): 1–96.

Kilic, Talip, Umar Serajuddin, Hiroki Uematsu, and Nobuo Yoshida. 2017. "Costing Household Surveys for Monitoring Progress Toward Ending Extreme Poverty and Boosting Shared Prosperity." *World Bank Policy Research Working Paper*, no. 7951.

Krennmair, Patrick, and Timo Schmid. 2022. "Flexible domain prediction using mixed effects random forests." *arXiv Preprint arXiv:2201.10933*.

Kreutzmann, Ann-Kristin, Sören Pannier, Natalia Rojas-Perilla, Timo Schmid, Matthias Templ, and Nikos Tzavidis. 2019. "The R package emdi for estimating and mapping regionally disaggregated indicators." *Journal of Statistical Software* 91.

Kuhn, Max, and Kjell Johnson. 2013. *Applied Predictive Modeling*. Vol. 26. Springer.

Kuhn, Max, Steve Weston, Chris Keefer, and Maintainer Max Kuhn. 2022. "Package Cubist." *Rule-and Instance-Based Regression Modeling: Version 0.2* 3.

Masaki, Takaaki, David Newhouse, Ani Rudra Silwal, Adane Bedada, and Ryan Engstrom. 2022. "Small area estimation of non-monetary poverty with geospatial data." *Statistical Journal of the IAOS* 38 (3): 1035–51.

Merfeld, Joshua D, and Jonathan Morduch. 2022. "Poverty at Higher Frequency." *Working Paper*.

Merfeld, Joshua D, David Locke Newhouse, Michael Weber, and Partha Lahiri. 2022. "Combining Survey and Geospatial Data Can Significantly Improve Gender-Disaggregated Estimates of Labor Market Outcomes." *World Bank Working Paper 10077*.

Molina, Isabel, and JNK Rao. 2010. "Small Area Estimation of Poverty Indicators." *Canadian Journal of Statistics* 38 (3): 369–85.

Newhouse, David Locke, Joshua D Merfeld, Anusha Ramakrishnan, Tom Swartz, and Partha Lahiri. 2022. "Small Area Estimation of Monetary Poverty in Mexico Using Satellite Imagery and Machine Learning." *World Bank Working Paper 10175*.

Peterson, Ryan A, and Joseph E Cavanaugh. 2019. "Ordered Quantile Normalization: A Semiparametric Transformation Built for the Cross-Validation Era." *Journal of Applied Statistics*.

Quinlan, John R. 1992. "Learning with Continuous Classes." In *5th Australian joint conference on artificial intelligence*, 92:343–48. World Scientific.

———. 2014. *C4. 5: Programs for Machine Learning*. Elsevier.

Rao, John NK, and Isabel Molina. 2015. *Small Area Estimation*. John Wiley & Sons.

Ratledge, Nathan, Gabriel Cadamuro, Brandon De la Cuesta, Matthieu Stigler, and Marshall Burke. 2021. "Using Satellite Imagery and Machine Learning to Estimate the Livelihood Impact of Electricity Access." National Bureau of Economic Research.

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

Tibshirani, Julie, Susan Athey, Rina Friedberg, Vitor Hadad, David Hirshberg, Luke Miner, Erik Sverdrup, Stefan Wager, Marvin Wright, and Maintainer Julie Tibshirani. 2022. "Package Grf."

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society: Series B (Methodological)* 58 (1): 267–88.

Tzavidis, Nikos, Li-Chun Zhang, Angela Luna, Timo Schmid, and Natalia Rojas-Perilla. 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): 927–79.

Wang, Yong, and Ian H Witten. 1997. "Inducing Model Trees for Continuous Classes." In *Proceedings of the ninth European conference on machine learning*, 9:128–37. 1. Citeseer.

Yeh, Christopher, Anthony Perez, Anne Driscoll, George Azzari, Zhongyi Tang, David Lobell, Stefano Ermon,

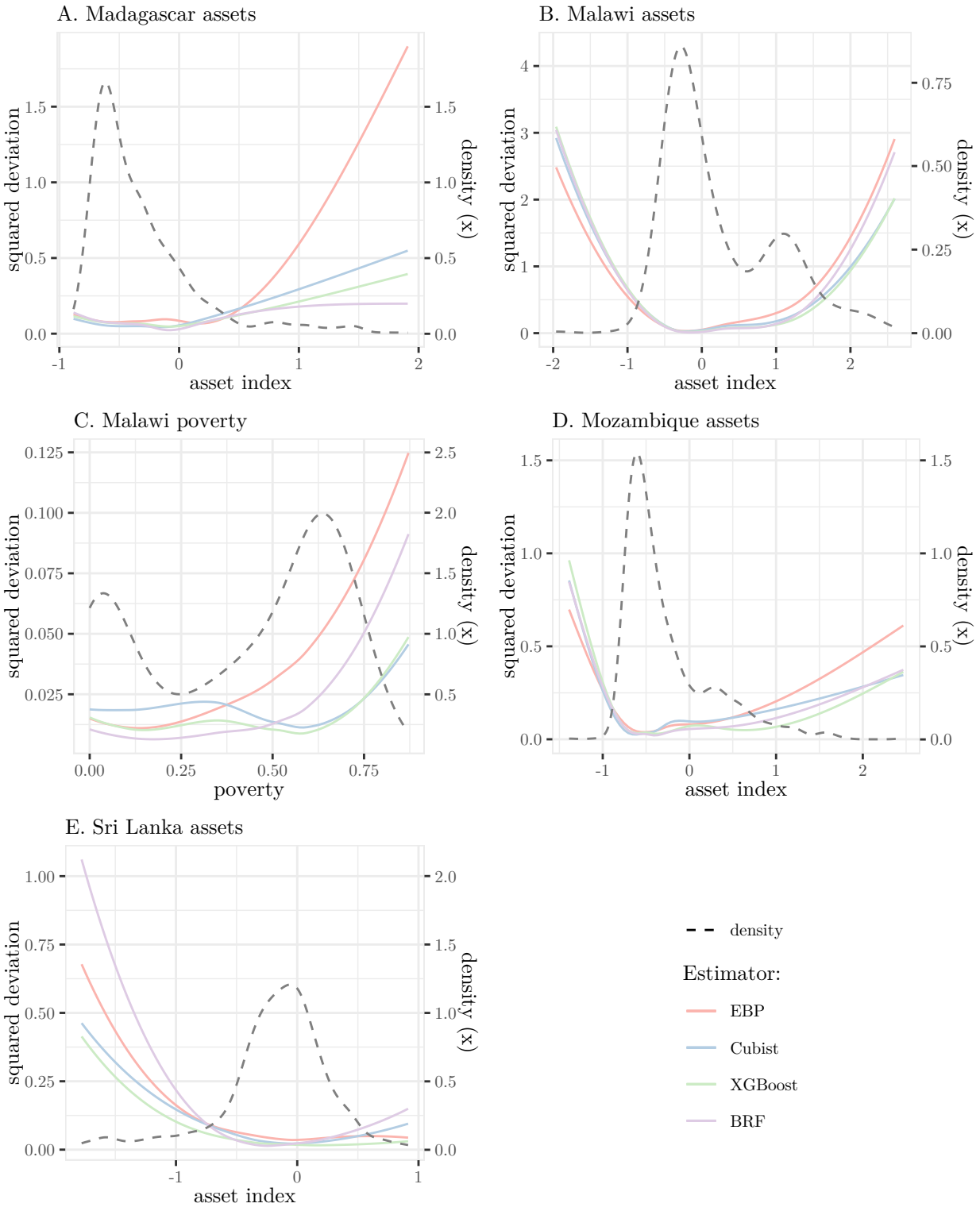
and Marshall Burke. 2020. "Using publicly available satellite imagery and deep learning to understand economic well-being in Africa." *Nature Communications* 11 (1): 1–11.

Appendix A

Table A1: Geospatial features

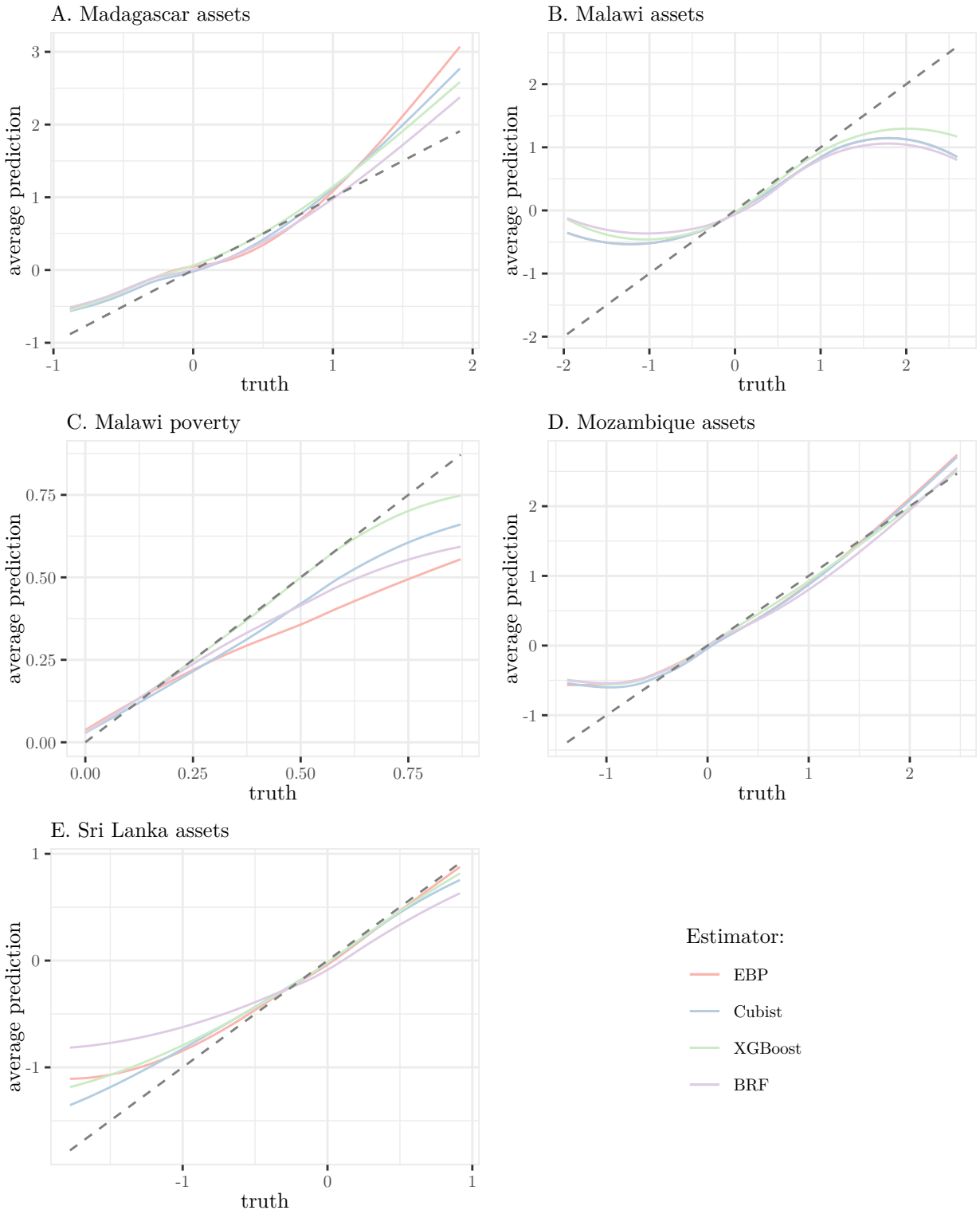
Indicator	Source
Population	WorldPop
Precipitation	TerraClimate
Temperature	TerraClimate
Nightlights	NOAA VIIRS
Land cover	EU Copernicus
Elevation	Conservation Science Partners
NDVI	MODIS
Pollution measures	EU Copernicus
Distance to key cities	Collected by authors

Figure A1: Deviations from truth and truth, out-of-sample areas only



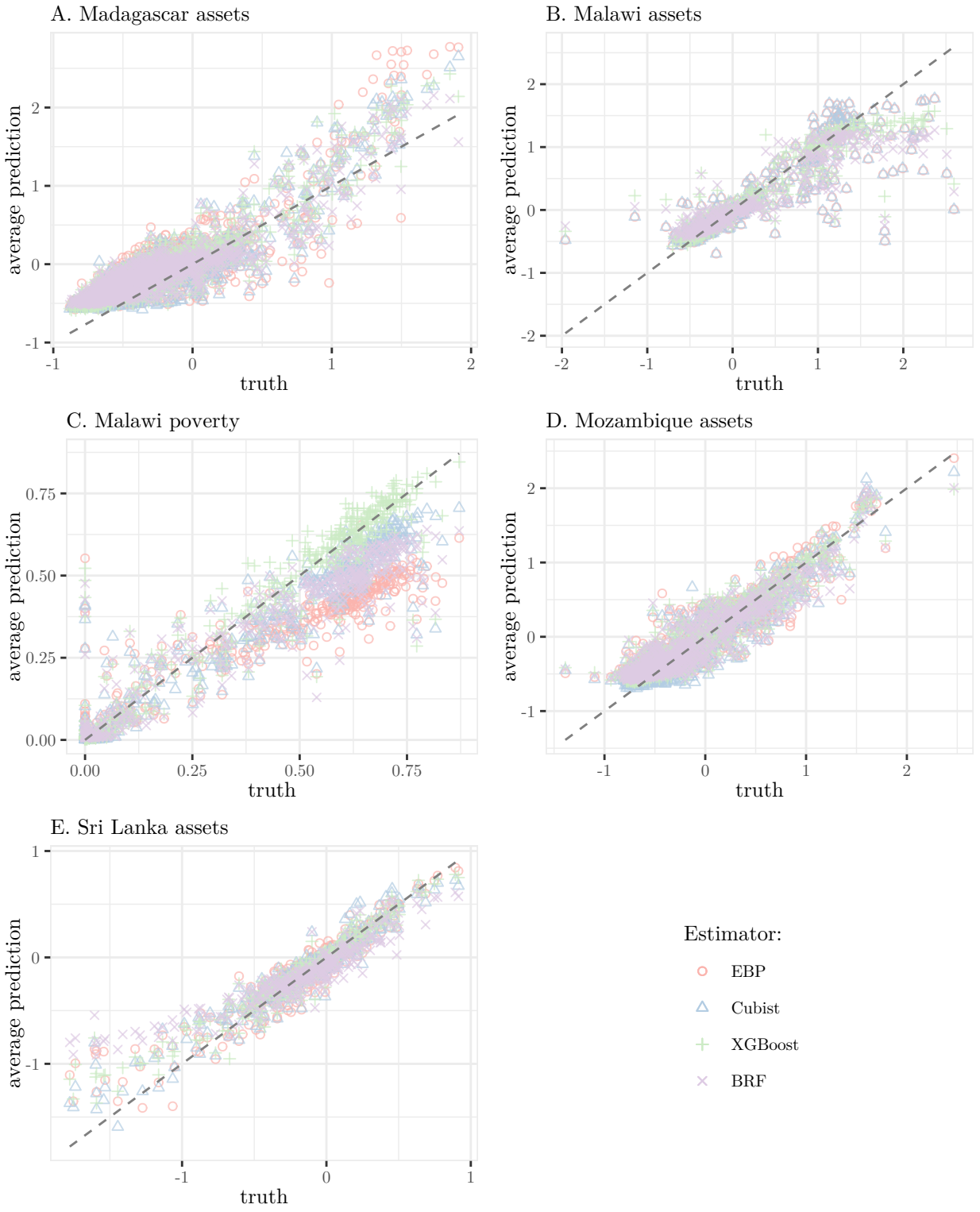
Note: All figures are smoothed conditional means of the mean squared deviation across 100 independent samples (first y-axis) on truth (x-axis), with means restricted to only samples in which an area appears (in-sample areas). The kernel density estimate refers to the density of truth, which is on the x-axis.

Figure A2: Predicted-true plots across areas



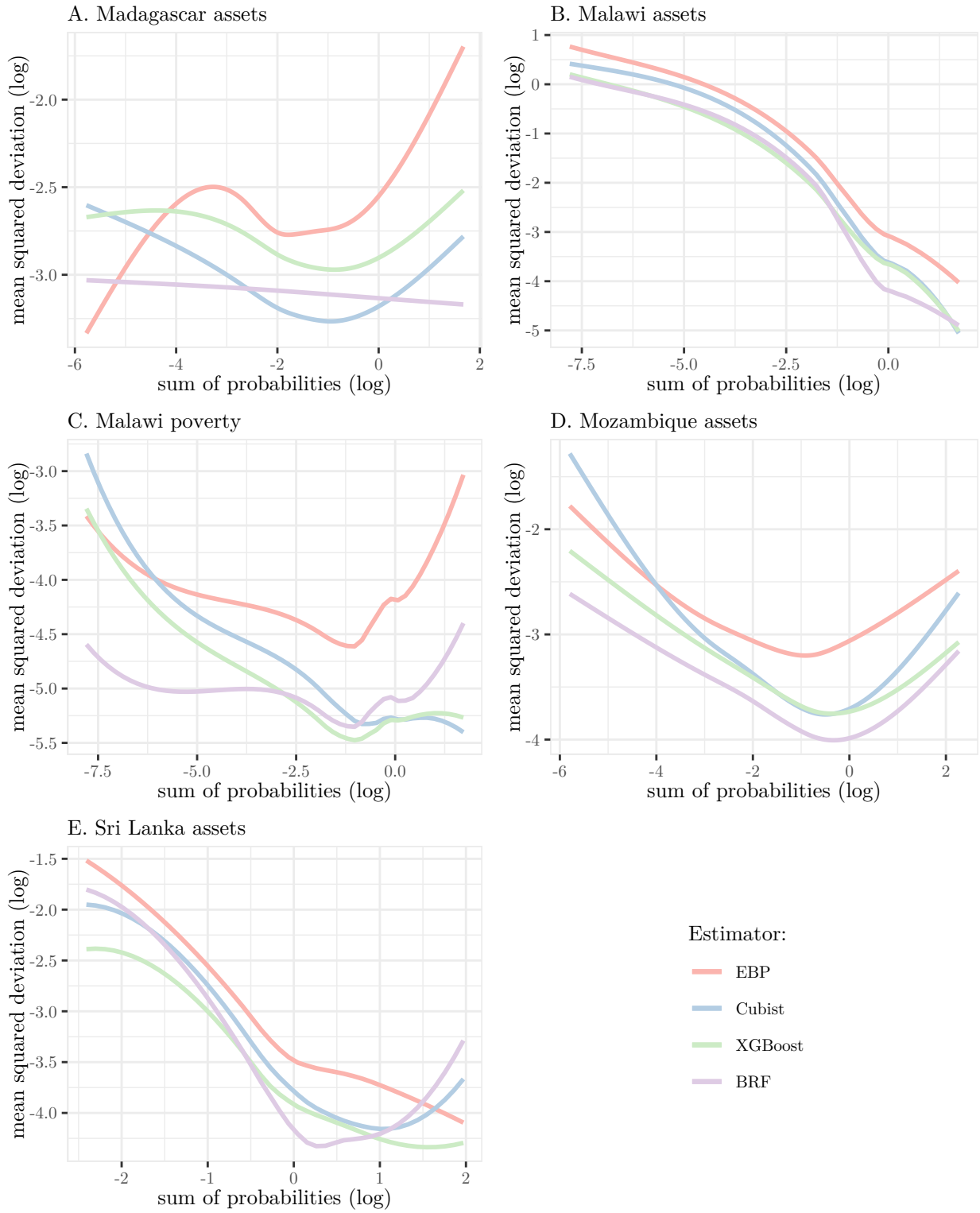
Note: All figures are smoothed conditional means of the estimated value (y-axis) on truth (x-axis).

Figure A3: Predicted-true plots across areas, scatter



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.

Figure A4: MSE and probability of appearing in the sample



Note:

Figure A5: Predicted-true plot, 2019 Malawi IHS

Note: In both figures, the true. The y-axis presents the log of the mean squared deviation for each area.

Table A2: Accuracy statistics across simulations

	EBP	cubist	XGBoost	BRF
Madagascar (assets)				
corr. (pearson)	0.830	0.881	0.897	0.893
corr. (spearman)	0.750	0.808	0.824	0.836
absolute dev.	0.259	0.209	0.231	0.230
squared dev.	0.109	0.070	0.078	0.072
Malawi (assets)				
corr. (pearson)	0.801	0.870	0.889	0.881
corr. (spearman)	0.810	0.863	0.879	0.882
absolute dev.	0.289	0.230	0.215	0.237
squared dev.	0.220	0.147	0.125	0.152
Malawi (poverty)				
corr. (pearson)	0.834	0.902	0.915	0.923
corr. (spearman)	0.797	0.862	0.877	0.887
absolute dev.	0.147	0.079	0.076	0.103
squared dev.	0.036	0.015	0.013	0.018
Mozambique (assets)				
corr. (pearson)	0.867	0.886	0.920	0.922
corr. (spearman)	0.758	0.786	0.816	0.841
absolute dev.	0.204	0.174	0.168	0.168
squared dev.	0.074	0.061	0.047	0.046
Sri Lanka (assets)				
corr. (pearson)	0.867	0.895	0.912	0.884
corr. (spearman)	0.850	0.880	0.903	0.896
absolute dev.	0.167	0.146	0.135	0.162
squared dev.	0.053	0.042	0.036	0.058

Note: Measures of accuracy are averages across all 100 independent samples drawn from the repsective country census. These are simple, unweighted averages across areas. EBP refers to small area estimation and BRF refers to local linear forests.

Table A3: Precision statistics across simulations

	EBP	cubist	XGBoost	BRF
Madagascar (assets)				
CI width	1.147	1.224	0.761	1.147
coverage	0.953	0.974	0.894	0.978
Malawi (assets)				
CI width	1.646	1.365	1.122	1.646
coverage	0.925	0.925	0.915	0.936
Malawi (poverty)				
CI width	0.641	0.591	0.518	0.641
coverage	0.780	0.931	0.928	0.934
Mozambique (assets)				
CI width	1.345	1.496	1.101	1.345
coverage	0.957	0.983	0.979	0.987
Sri Lanka (assets)				
CI width	1.387	1.177	1.039	1.387
coverage	0.964	0.980	0.973	0.971

Note: Measures of precision are averages across all 100 independent samples drawn from the repsective country census. These are simple, unweighted averages across areas. EBP refers to small area estimation and BRF refers to local linear forests. We do not use actual standard errors when calculating coverage rates for Cubist, XGBoost, and BRF. Instead, we use the appropriate percentiles of the bootstrapped distribution. As such, we present the width of the 95-percent confidence interval instead of the standard errors.

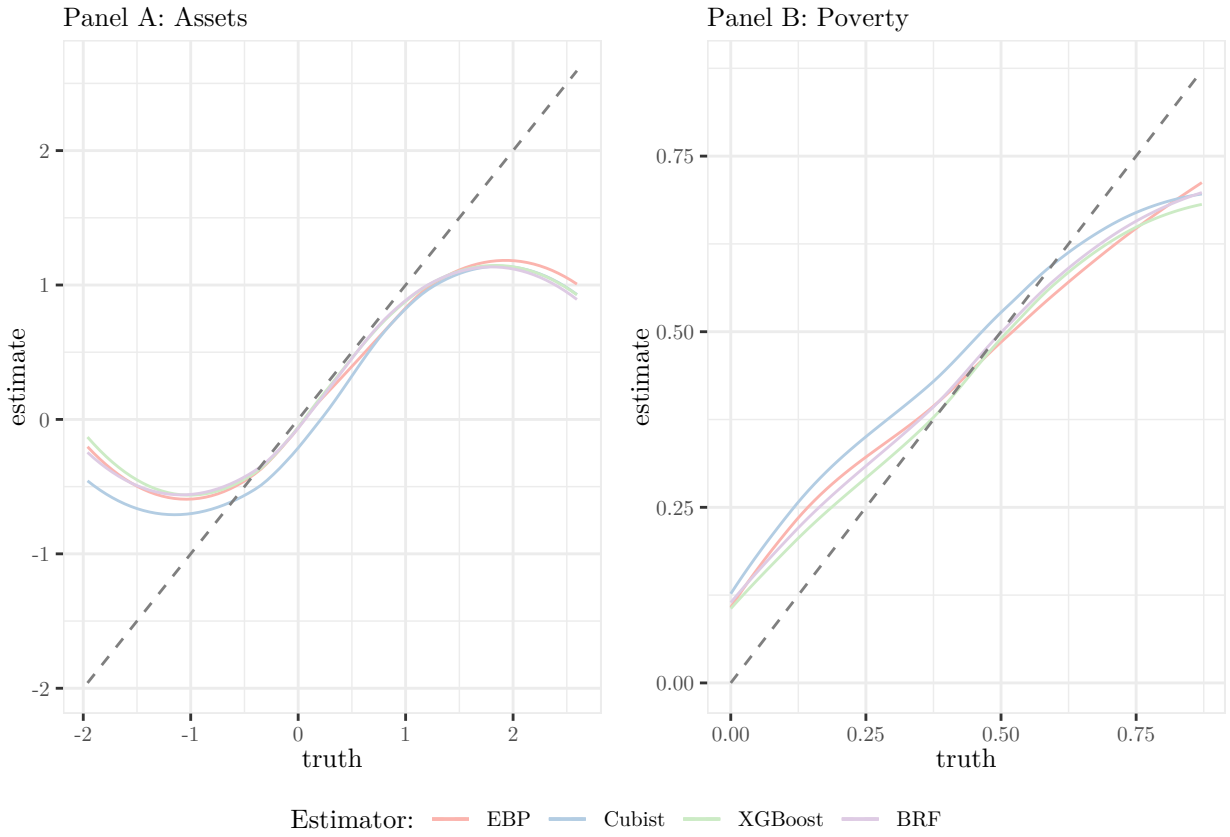
Table A4: Accuracy statistics across simulations, LASSO variables

	XGBoost	BRF
Madagascar (assets)		
corr. (pearson)	0.892	NA
corr. (spearman)	0.808	NA
absolute dev.	0.235	NA
squared dev.	0.077	NA
Malawi (assets)		
corr. (pearson)	0.881	NA
corr. (spearman)	0.862	NA
absolute dev.	0.227	NA
squared dev.	0.137	NA
Malawi (poverty)		
corr. (pearson)	0.900	
		NA
corr. (spearman)	0.858	NA
absolute dev.	0.101	NA
squared dev.	0.019	NA
Mozambique (assets)		
corr. (pearson)	0.900	NA
corr. (spearman)	0.800	NA
absolute dev.	0.181	NA
squared dev.	0.055	NA
Sri Lanka (assets)		
corr. (pearson)	0.904	NA
corr. (spearman)	0.896	NA
absolute dev.	0.139	NA
squared dev.	0.039	NA

Table A5: Optimal hyperparametercs from cross-validation

	Madagascar	Malawi assets	Malawi poverty	Mozambique	Sri Lanka
cubist					
rules	3	3	3	3	3
committees	3	5	5	1	5
sample	0.4	0.4	0.4	0.4	0.4
neighbors	0	0	0	0	0
XGB					
max depth	8	6	6	6	6
eta	0.02	0.01	0.01	0.01	0.01
col sample	0.6	0.6	0.4	0.6	0.6
subsample	0.6	0.6	0.4	0.4	0.4
BRF					
sample fraction	0.5	0.5	0.5	0.5	0.5
mtry	30	30	25	30	30
min. node size	4	5	5	5	4
honesty fraction	0.6	0.6	0.6	0.6	0.6
alpha	0.05	0.025	0.025	0.05	0.05

Figure A6: Predicted-true plot, 2019 Malawi IHS



Note: Both figures are smoothed conditional means of the estimated value (y-axis) on truth (x-axis).

Table A6: Accuracy statistics across simulations

	orderNorm	logShift
Madagascar (assets)		
corr. (pearson)	0.830	0.848
corr. (spearman)	0.750	0.771
absolute dev.	0.259	0.242
squared dev.	0.109	0.087
Malawi (assets)		
corr. (pearson)	0.801	0.800
corr. (spearman)	0.810	0.811
absolute dev.	0.289	0.304
squared dev.	0.220	0.253
Malawi (poverty)		
corr. (pearson)	0.834	0.834
corr. (spearman)	0.797	0.797
absolute dev.	0.147	0.147
squared dev.	0.036	0.036
Mozambique (assets)		
corr. (pearson)	0.867	0.876
corr. (spearman)	0.758	0.763
absolute dev.	0.204	0.195
squared dev.	0.074	0.069
Sri Lanka (assets)		
corr. (pearson)	0.867	0.854
corr. (spearman)	0.850	0.835
absolute dev.	0.167	0.182
squared dev.	0.053	0.059

Note: Measures of accuracy are averages across all 100 independent samples drawn from the repsective country census. These are simple, unweighted averages across areas. orderNorm refers to an ordered quantile normalization and logShift refers to a log-shift transformation.

Appendix B - Imputing poverty in the Malawi census

We impute welfare into the 2018 Malawi census using the 2019 Integrated Household Survey (IHS). Both the census and the IHS include information on household assets and key household demographics, while the survey has information on expenditures/consumption. We select the variables that are common to both datasets and then use lasso to select the most predictive variables to use in the imputation procedure. The post-lasso regression results are in Table B1.

We predict welfare directly into the census using the results in Table B1. Figure B1 shows the resulting densities of welfare (expenditures) in the survey and the census. The census-imputed results show a lot less variation – as we would expect. The estimated poverty rate in the IHS is 0.507. Given the differences in the variation of assets, we set the poverty threshold at median expenditures in the census, leading to a poverty rate of 0.500, almost equal to that in the survey.

Figure B1: Distribution of welfare in 2019 IHS and 2018 census

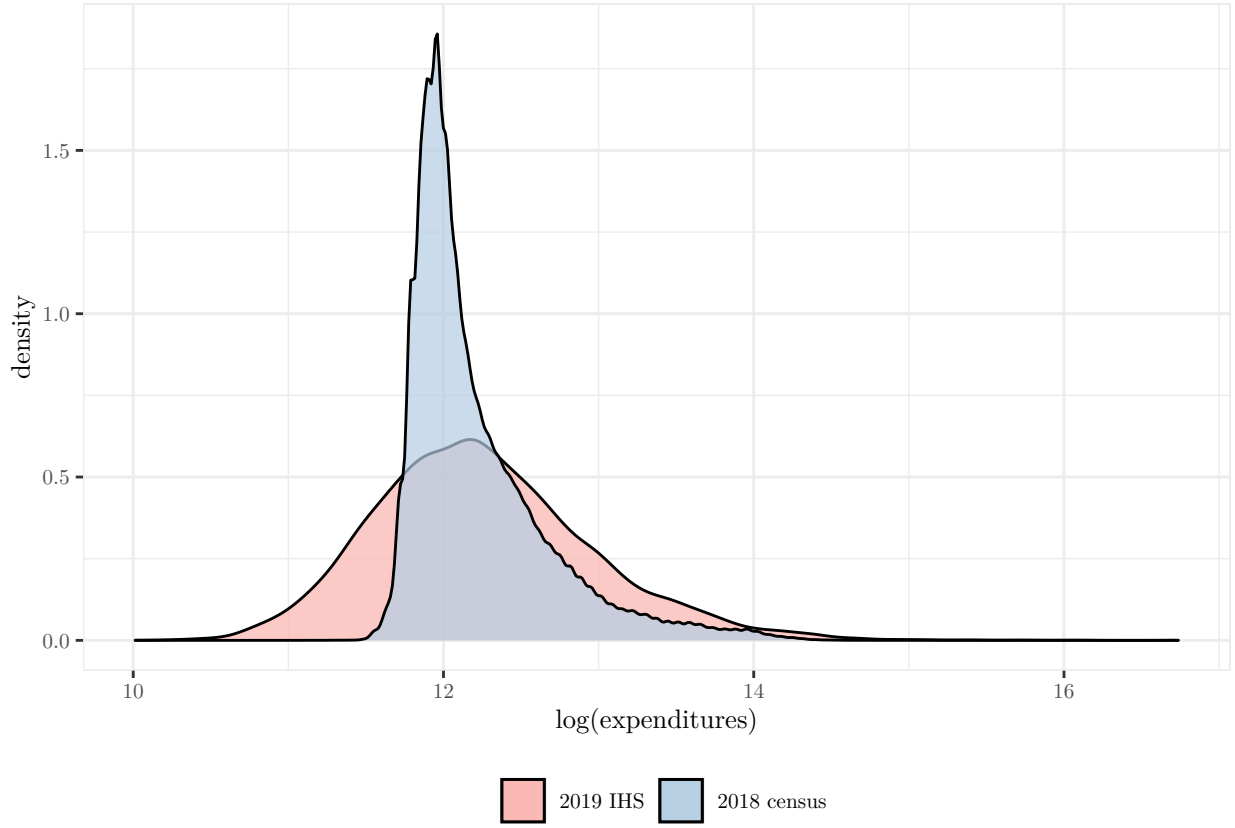


Table B1: Welfare imputation results

	Model 1
headAge	−0.0002 (0.0004)
headMale	−0.069 (0.012)***
headEducPrimary	0.103 (0.013)***
headEducSecondary	0.136 (0.018)***
headEducUniversity	0.217 (0.038)***
hhOwned	−0.137 (0.014)***
hhRoofGrass	−0.095 (0.097)
hhRoofIron	0.004 (0.097)
hhRoofCement	0.719 (0.553)
hhWallMud	−0.050 (0.059)
hhWallConcrete	−0.063 (0.063)
hhWallBricksBurnt	−0.069 (0.057)
hhWallBricksUnburnt	−0.051 (0.056)
hhFloorEarth	0.166 (0.223)
hhFloorCement	0.308 (0.222)
hhFloorWood	0.843 (0.444)*
hhFloorTile	0.608 (0.236)**
hhDwellingTypePerm	0.105 (0.028)***
hhDwellingTypeSemiperm	0.077 (0.020)***
hhDwellingRooms	−0.065 (0.005)***
hhAssetsCell	0.125 (0.012)***
hhAssetsRadio	0.063 (0.012)***
hhAssetsTV	0.078 (0.025)***
hhAssetsComputer	0.364 (0.039)***
hhAssetsFridge	0.197 (0.031)***
hhAssetsBike	−0.010 (0.012)
hhAssetsTable	0.050 (0.013)***
hhAssetsBed	0.219 (0.014)***
hhAssetsIron	0.122 (0.017)***
hhAssetsSolarPanel	0.028 (0.014)**
hhAssetsCDDVD	0.070 (0.026)***
hhAssetsCar	0.375 (0.042)***
r-squared	0.407
Observations	11.425

* = p<0.1, ** = p<0.05, *** = p<0.01

Appendix C - Methods

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

$$\text{reduction} = SD(S) - \sum_{i=1}^{P} \frac{n_i}{n} SD(S_i), \tag{8}$$

   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.[14]

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 variables, defined as:

[14] 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.

$$\text{adjusted error rate} = \frac{n^* + p}{n^* - p} \sum_{i=1}^{n^*} |y_i - \hat{y}_i(X_i)| \,, \tag{9}$$

where n^* is the number of observations int he 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 Equation 9, 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]:

$$\hat{Y}_{par} = a\hat{Y}_{kid} + (1 - a)\hat{Y}_{par}, \tag{10}$$

   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

$$a = \frac{var(e_{par}) - cov(e_{kid}, e_{par})}{var(e_{par} - e_{kid})}, \tag{11}$$

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 Equation 10 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 Equation 10. The model coefficients for each rule (leaf) are averaged according to the following formula:

$$\hat{\beta}_{par} = a\hat{\beta}_{kid} + (1 - a)\hat{\beta}_{par}. \tag{12}$$

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 reftab:featuresStats in the appendix shows the optimal hyperparameters for cubist that we use in all sample iterations. There are three rules in each country and the number of committees varies from 1 in Mozambique to 5 in Sri Lanka and Malawi; we use default values for any unlisted hyperparameters.

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,

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), f_k \in F, \tag{13}$$

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 $\phi(x_i)$:

$$obj(\phi) = \sum_{i=1}^{n} l\left(y_i, \phi(x_0)\right) + \sum_{k=1}^{K} \omega(f_k), \tag{14}$$

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:

$$obj(t) = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \omega(f_t). \tag{15}$$

This sequentially adds the f_t that provides the largest improvement in performance according to the objective function Equation 14, 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:

$$l\left(y_i, \hat{y}_i + f_t(x_i)\right) = \left(\hat{y}_i^{(t-1)} + f_t(x_t) - y_i\right) \tag{16}$$

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

$$obj(t) = \sum_{i=1}^{n} \left[2\left(\hat{y}_i^{(t-1)} - y_i\right) f_t(x_i) + f_t(x_i)^2\right] + \omega(f_t), \tag{17}$$

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

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

$$\omega(f_i) = \gamma T + \frac{\lambda}{2} \sum_{j=1}^{T} w_j^2, \tag{18}$$

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 (**ref?**)(tab:featuresStats) 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 contain 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 (1) 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:

$$H = \frac{N_{C1}N_{C2}}{N_P^2(\bar{y}_{C1} - \bar{y}_{C2})^2} - \left(\frac{IP}{N_{C1}} + IPN_{C2}\right), \tag{19}$$

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. As such, hyperparameters change across iterations.