Micro-level structural poverty estimates for southern and eastern Africa

Elizabeth Tennant[a,1], **Yating Ru**[b,c], **Peizan Sheng**[d], **David S. Matteson**[e], **and Christopher B. Barrett**[f,g]

This manuscript was compiled on May 23, 2024

For many countries in the Global South traditional poverty estimates are available only infrequently and at coarse spatial resolutions, if at all. This limits decision-makers' and analysts' ability to target humanitarian and development interventions and makes it difficult to study relationships between poverty and other natural and human phenomena at finer spatial scales. Advances in Earth observation and machine learning-based methods have proved capable of generating more granular estimates of relative asset wealth indices. They have been less successful in predicting the consumption-based poverty measures most commonly used by decision-makers, those tied to national and international poverty lines. For a study area including four countries in southern and eastern Africa, we pilot a two-step approach that combines Earth observation, accessible machine learning methods, and asset-based structural poverty measurement to address this gap. This structural poverty approach to machine learning-based poverty estimation preserves the interpretability and policy-relevance of consumption-based poverty measures, while allowing us to explain over 70% of cluster-level variation in a pooled model and over 50% even when predicting out-of-country.

Keywords: assets — expenditures— machine learning — poverty maps — small area estimation

Accurate estimates of the number of people deprived of a minimum acceptable standard of living are available infrequently and only at the first- or second-level administrative unit, if at all, for many places in the Global South. These aggregate estimates can mask pockets of extreme poverty and quickly become outdated. This limits policymakers' ability to recognize and respond to the most urgent human needs, to study the processes that cause and perpetuate poverty, and to evaluate the effectiveness of interventions. The scarcity of poverty estimates in low-resource settings persists because high quality household surveys of income and consumption expenditures are difficult and expensive to administer, and therefore under-supplied. This gap is particularly stark in many African countries (1).

Recent research seeks to address this gap through modeling efforts that leverage advances in machine learning (ML) and Earth observation (EO)(1–10). Scientific progress in this space has focused on improving the out-of-sample predictive accuracy of asset-based poverty (or wealth) measures through advances in algorithms or in the feature sets used to explain outcomes. For these advances to translate into greater uptake and impact, however, the measures predicted must also be policy relevant. The maps of asset wealth indices prevalent in this literature do not readily translate to the consumption-based poverty measures more often used by policy makers, such as the share of people living below national or international poverty lines.

Our goal is to improve the relevance of the dependent variable, or object, for ML poverty mapping without compromising our ability to predict it. We do this using a two-stage modeling approach that first calibrates a model of 'structural' poverty (11). Structural poverty is defined as the expectation that a household will on average have a (non-)poor level of consumption expenditure given their durable characteristics, such as productive assets. We then train EO-based models on aggregates of the fitted structural poverty estimates from the first-stage, allowing us to predict into un-surveyed areas.

Data fusion for micro-level poverty estimation. Multiple data fusion methods have been developed to address gaps in the availability of survey-based poverty estimates. For decades, researchers and practitioners have used and refined techniques that leverage census data on the covariates of poverty to produce more precise and unbiased small area estimates (SAEs) (12–15). First, a model of household or area-level characteristics on poverty is estimated using sample-based survey data that includes consumption expenditures (or income). The resulting parameterized model is then used to predict poverty at more granular scales from the same

Significance Statement

The first UN Sustainable Development Goals target is for all people to clear the expenditures-based extreme global poverty line ($2.15/person-day in 2017 prices) by 2030. Yet survey estimates of when and where people fall below this poverty line are often unavailable. One strategy to fill that gap is to train machine-learning models to estimate poverty from Earth observations data. But most models train on asset data, generating maps of relative wealth that do not map to poverty lines. We pilot a two-step modeling procedure that harnesses the accuracy gains of prevailing methods, but then maps those predictions to more policy relevant poverty measures. This allows us to compute stable and forward-looking estimates of where people live in poverty.

Author affiliations: [a]Department of Economics, Cornell University, Ithaca, NY 14850; [b]Asian Development Bank, Mandaluyong, Metro Manila, Philippines; [c]Department of City and Regional Planning, Cornell University, Ithaca, NY 14850; [d]Harris School of Public Policy, University of Chicago, Chicago, IL 60637; [e]Department of Statistics and Data Science, Cornell University, Ithaca, NY 14850; [f]Charles H. Dyson School of Applied Economics and Management, Cornell University, Ithaca, NY 14850; [g]Cornell Jeb E. Brooks School of Public Policy, Cornell University, Ithaca, NY

All authors designed research; E.T. & Y.R. & P.S. curated data; E.T. & P.S. analyzed data; E.T. prepared visualizations; E.T. wrote the paper; all authors reviewed/edited the paper.

The authors have no competing interests to declare.

[1]To whom correspondence should be addressed. E-mail: ejt58cornell.edu

household or area-level characteristics available for the entire population in the census data. These SAEs offer insight into spatial patterns of poverty, but are published infrequently and often with long lag times.[*] Further, such poverty maps are not designed for inter-country comparability and cannot be easily customized because of the proprietary nature of the underlying data.

Newer machine learning (ML) based methods that harness Earth observation (EO) and other geospatial 'Big Data' have proved capable of generating more granular estimates of relative deprivation within as well as across low- and middle-income countries (1–3, 6, 8).[†] Instead of using census data, researchers derive area-level characteristics from cell phone records, satellite imagery, and various EO-based data products – including publicly available sources (1–3, 8). These data are used in concert with machine learning (ML) methods that are well suited to handle large feature sets and model non-linear relationships.

Amongst efforts to leverage geospatial data to interpolate and extrapolate into unsurveyed places, individual country studies in the SAE tradition have frequently retained flow-based monetary measures as their object or predictand. In contrast, multi-country studies have favored indices of asset holdings to proxy spatial patterns in poverty, with several advantages. Survey data collection of asset stocks is easier, cheaper, and less prone to substantial measurement error than of flow measures of well-being like expenditures or income. As a result, high-quality asset data are more often available to train ML models. Productive assets are the stocks that generate income flows that enable consumption expenditure. Thus the connection between asset-based wealth indicators and income- or expenditure-based poverty measures follows intuitively. Indeed, the literature on asset-based, structural poverty demonstrates that, especially in poor places subject to multiple market failures that impede consumption expenditure smoothing, productive asset holdings reflect expected, permanent income (11, 19–21). Household assets and their correlates may also be more easily observed from EO. Satellite imagery can detect the size and quality of buildings, vehicles, and infrastructure but may overlook many short-term drivers of community-level consumption expenditures, such as disease outbreaks, labor market conditions, or price shocks. For these (and other) reasons, ML models trained on assets are more prevalent and consistently outperform models of monetary poverty and other well-being measures (2, 3, 22).

The result is that the poverty mapping literature has primarily produced maps of relative asset wealth. Meanwhile, practitioners predominantly use monetary poverty measures based on flows of income or consumption expenditures that can be anchored to interpretable normative thresholds, such as national and international poverty lines representing a minimum acceptable standard of living as defined by governments and multi-lateral institutions. For example, under the first Sustainable Development Goal to "end poverty in all its forms everywhere", the first target is to bring all people above the $2.15 (2017 purchasing power parity, PPP) per person per day extreme global poverty line by 2030 (23).[‡] Progress toward such a goal cannot be tracked using an asset wealth index, which has no direct conversion to monetary poverty. While asset wealth indices may seem an intuitive proxy for consumption-based poverty, this assumed correlation is not always empirically well supported (24).

Unlike unit-less asset indices, estimates of monetary poverty can be compared over place and time using PPP conversions. Monetary measures are also flexible. For example, consumption expenditures data can be used to estimate the Foster-Greer-Thorbecke (FGT) class of distribution-sensitive poverty measures, including the 'poverty gap' and 'poverty gap squared' (25). The FGT measures take into account how far below the poverty line people's incomes or consumption expenditures fall, and satisfy a range of desirable axiomatic properties (25, 26). The advantages of consumption-based poverty measures are balanced by the expense (and therefore scarcity) of high quality training data and the stochastic nature of consumption. Snapshots of monetary poverty may be dominated by transitory shocks or seasonality in income or expenditure patterns. This may obscure the chronic or structural deprivations of first-order humanitarian concern (11, 27–29).

As ML poverty mapping gains traction, it is a timely moment to consider these trade-offs in policy relevance, comparability, and accuracy that follow the choice of an asset- vs. consumption-based poverty map, as well as how they might be mitigated. One way forward, which we set forth in this paper, is to leverage *both* asset and consumption data to train the ML models that predict poverty from geospatial features. We propose a set of structural poverty measures – based on the expectation of consumption expenditures given household asset holdings (11) – as predictands for micro-level multi-country poverty estimation. This paper describes the conceptual advantages of these measures, and develops a two-stage approach to ML structural poverty mapping. We also evaluate this approach empirically using data from 13 Living Standards Measurement Studies (LSMS) household surveys conducted in Ethiopia, Malawi, Tanzania, and Uganda between 2008-2020, spatially and temporally matched to geospatial data on population density, building footprints, remoteness, night lights, elevation, slope, rainfall, temperature, and the Normalized Difference Vegetation Index.

Results

We propose a set of structural poverty measures with desirable properties as the object (or predictands) of ML poverty mapping. These structural poverty measures are stable and forward-looking because they are anchored to the stock of productive assets. They are also expressed in familiar flow-based units tied to a normatively meaningful standard of living (e.g., the share of people living below a poverty line). They can be compared across countries and over time. These attributes respond to the needs of humanitarian and development programming, which require an understanding of both absolute and relative levels of deprivation, and must be responsive to poverty now and into the future.

[*] This is in part because contemporaneous censuses and household surveys are scarce. Methods for SAE with disjoint census and consumption surveys have been developed (16, 17). See (18) for a discussion of variations of the SAE approach suited to different data availability scenarios.

[†] Country-level work that uses geospatial data to build more directly on SAE methods is also gaining traction (7, 9, 10).

[‡] When the 2030 Agenda for Sustainable Development was released in 2015 this goal referenced the $1.25 per person per day (2005 PPP) extreme poverty line. This was later updated to the $1.90 (2011 PPP) and most recently to $2.15 (2017 PPP). These updates are implemented primarily to adjust for inflation. The empirical portion of this paper employs the $1.90 (2011 PPP $) poverty line, which was in effect at the start of this research.

To construct these structural poverty measures in the training data, we begin by introducing and modeling a more durable, asset-based analogue to flow-based monetary measures: structural consumption (see Methodology). Structural consumption is the expectation of consumption expenditures for a given portfolio of household assets. We use the household-level structural consumption estimates from these models to construct the FGT poverty headcount ($P_s^{\alpha=0}$), poverty gap (P_s^1), and poverty gap squared (P_s^2) aggregates for each survey cluster sampling unit, where subscript s denotes *structural* and superscript $\alpha = 0, 1, 2$ is the FGT poverty aversion parameter. These cluster-level structural poverty measures become the training data for the EO-based models. Because structural poverty is a latent variable, the performance of our EO models is validated against our estimates of structural poverty. These estimates in turn rely on assumptions about the strength and stability of the relationship between productive assets and consumption expenditures, and the stochastic nature of shocks to consumption.

We evaluate the strength of these assumptions for our study area prior to proceeding to train EO-based models from our structural poverty estimates. Our empirical assessment confirms the premise that productive assets are strong predictors of consumption expenditures, but with some limitations. In particular, we find that asset-expenditures relationships vary even across our study countries, which are geographically proximate and share many social and economic characteristics. Differences in the distributions of our structural estimates versus realized consumption, which should be similar in expectation, emerge and model fit declines when we predict structural consumption out-of-country.

In addition to its conceptual advantages, we hypothesize that structural poverty can be more accurately proxied than realized consumption expenditures using ML models and EO data. This is supported by the comparative success in predicting assets over consumption in the literature (1, 2, 22). Our empirical results corroborate this expectation. We find that models of structural poverty consistently outperform models of comparable realized poverty measures, by multiple performance metrics and by a substantial margin. A multi-country ML model predicts approximately 72% (50% out-of-country) of cluster-level variation in the structural poverty headcount, compared to 57% (12% out-of-country) for a comparable realized poverty measure.[§]

Our results allow us to expound current limitations of EO-based poverty mapping, in particular weaker model performance at the bottom of the wealth distribution and the risks of bias when predicting spatially out-of-sample. These problems persist, but do not appear to be exacerbated by the structural poverty estimation approach.

Estimating structural poverty from productive assets. As we consider candidate structural consumption models, traditional performance metrics have the potential to mislead. Perfect or near-perfect correlation between structural predictions and realized consumption expenditures would signal over-fitting. The advantage of a structural measure is that it filters out the 'noise' of classical measurement error and stochastic shocks from which a household may have already recovered by the time data are published and an agency had time to assimilate and act upon the poverty estimates. We instead look for a balance of fit and stability, as well as evidence of unbiasedness as we compare regressions of consumption on assets, including parametric first- (OLS-1) and second-order (OLS-2) polynomials and a random forest (RF-)regression. We also consider an RF-classification model for (non-) poor status, suitable only for poverty headcount (P^0) estimation. Separate sets of models are estimated for each individual country and for the pooled (all-country) data set, and compared to a test set of held-out EAs in the country or else the held-out country.

Our structural consumption models confirm our prior from the literature that productive assets are strongly predictive of consumption expenditures, using both linear and non-linear models. As depicted in Figure 1, structural predictions and realized consumption estimates agree on the (non-)poor

[§]This is based on standard cross-validation and a global extreme poverty line of $1.90 per person per day, 2011 PPP.

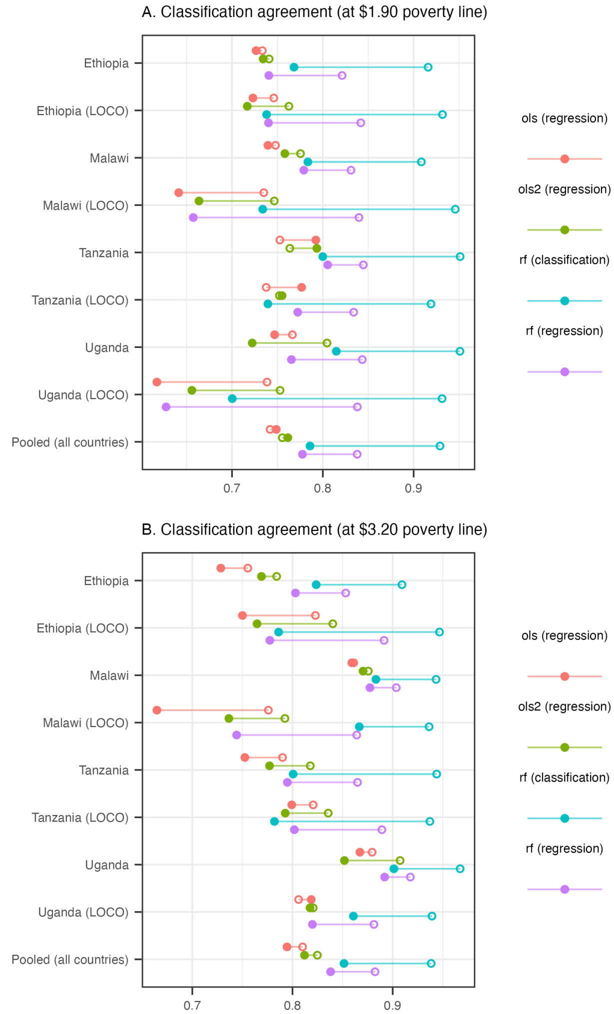


Fig. 1. Comparison of realized versus structural (non-)poor classification. The circle indicates the share of households with the same classification in the training (open circle) and test (filled circle) data. The horizontal line is the difference between agreement in the training and the test set. Country models are based on training and test data from the same country, while leave-one-country-out (LOCO) models are trained on the pooled dataset excluding the test country.

classification of 62-82% of households at the global extreme poverty line (z = \$1.90 a day in 2011 PPP), and 66-90% for the global poverty line (z = \$3.20 a day in 2011 PPP).[¶] The RF models for structural poverty have overall stronger agreement with realized poverty estimates in the test set than the OLS models (see SI Figure S10 for additional measures of fit). However, while agreement with the test set is highest in the RF models they are also less stable, with greater differences in these statistics from training to test set.

In theory, our structural poverty models should produce unbiased estimators of realized consumption. Thus, while we anticipate lower variance in the structural versus realized consumption distributions, we expect similar means at higher levels of aggregation (e.g., at the country level). In Figure 2, we compare the realized and predicted distributions of consumption expenditures by country and by model. For the single country models (left column), we observe no distinguishable difference in means for Malawi, Tanzania and Uganda. In Ethiopia, predictions may be biased slightly upwards (the mean of the RF predictions is \$3.66, versus \$3.48 for realized consumption). For the pooled models (middle column), we still see no difference in means for Malawi and Tanzania, but the Uganda RF model now predicts mean consumption at \$2.04 compared to the realized \$1.87 and in Ethiopia the difference is more pronounced compared to the single country model. As we move to leave-one-country-out (LOCO) validation (right column) we detect differences in most models; the largest of these for Malawi (\$3.53-\$3.68 versus the realized \$2.70) and Tanzania (\$3.02-\$3.13 versus the realized \$3.98).

Our approach assumes a stable relationship, across space and over time, between productive assets and consumption expenditure. Our empirical assessment suggests that this is a strong assumption, more likely to hold when models are trained on same-country data versus the data of neighboring countries. This may reflect substantive differences across our study countries: for example, the returns to land or livestock depend on the asset quality as well as local agro-ecology, labor and agricultural markets, the quality of institutions and social safety nets, and other factors. Observed differences may also reflect inconsistencies in measurement: how assets and consumption are surveyed and aggregated by different national statistical agencies.

Importantly, the quality and productivity of assets may also vary systematically with poverty. Poor households may have lower quality assets, or may live in places where the productivity of those assets is lower due to lack of access to markets, production technologies, institutions, or physical infrastructure. If so, our models might over-predict consumption for the poorest households and under-predict consumption of the wealthiest. Empirically, we cannot distinguish this heterogeneity in the asset-consumption relationship from differences that arise due to stochastic variation in consumption. For example, if the lowest and highest realizations of consumption expenditures arise due to classical measurement error or stochastic shocks, rather than true structural poverty, we would expect to see a reduction in the variance of the distribution. This reduction in variance is observed (see SI Figure S11). The households with the

[¶] For brevity, we refer to these simply as the \$1.90 and \$3.20 poverty lines or the global extreme and global poverty lines henceforth. As previously noted, these thresholds have more recently been updated to 2017 PPP values of \$2.15 and \$3.65, respectively, which are substantively similar.

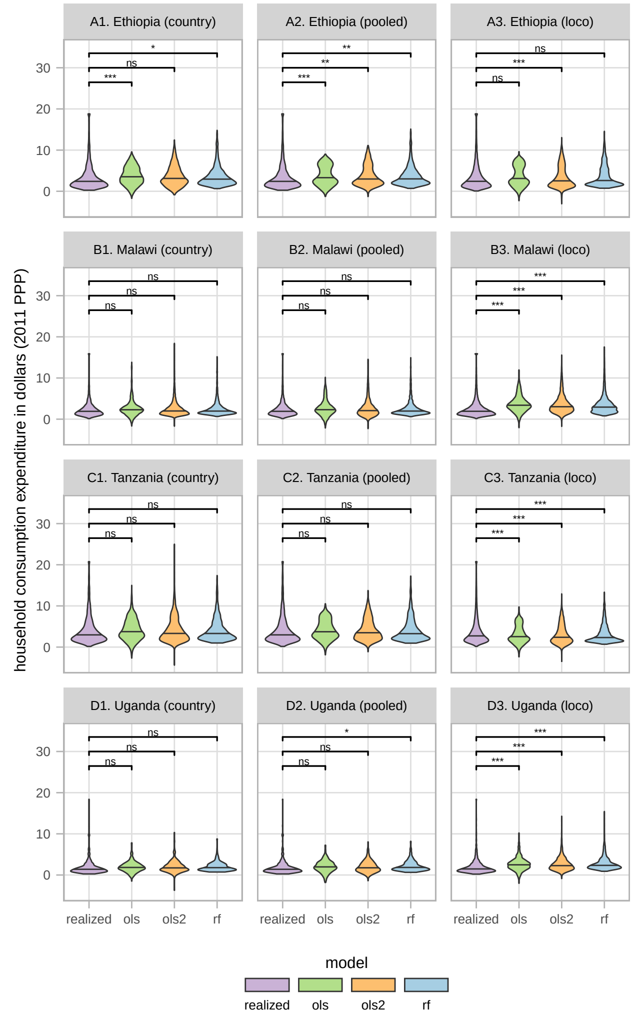


Fig. 2. Distributions and comparison of means for realized consumption and structural estimates. The mirrored density plots represent the distribution of realized consumption expenditures as well as structural estimates from OLS and RF models. Horizontal lines represent the mean of each distribution. The brackets at the top of each plot indicate whether a t-test for the difference in means is statistically significant (statistical significance is indicated by *not significant (ns)*, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

lowest realized consumption expenditures in our data are predicted to be slightly better off in terms of structural consumption. The reverse is also true: the households with the highest realized consumption have relatively lower structural consumption. One way forward, particularly in settings where we have a sense of the magnitude of the undesirable component of this difference, would be to adjust predicted structural poverty estimates ex-post. However, here we are unable to parse the desirable reduction in transient shocks and noisy data offered by structural poverty estimation from undesirable risk of bias due to model errors that correlate with poverty. We thus proceed using un-adjusted estimates from RF structural consumption models to build the training data set for the EO models. However, we urge that this be kept in mind when interpreting final estimates, particularly for the distribution-sensitive P^1 and P^2 measures that will be more affected by such biases.

Predicting structural poverty from Earth Observation. EO models trained on structural poverty (P_s^α) demonstrate consistently superior predictive performance over models trained on realized poverty (P_r^α), with higher out-of-sample r^2 values, lower Root Mean Squared Error (RMSE), and higher Spearman's rank correlation coefficients (ρ). RF models consistently outperform comparable OLS specifications, but the main result is qualitatively similar for the linear models. Results from the RF models are summarized in Table 1 and results by geography and for the benchmark OLS models are reported in Figures S12-S17 and SI Table S1. To ensure that the superior performance of the structural models is not simply a product of the noisier consumption data (of concern particularly for the r^2 metric), we also compare the EO-based model trained on realized consumption against the test set of structural poverty estimates. Our main result is robust to this alternative validation: the structural EO model consistently outperforms the realized EO model when both are evaluated against structural poverty. In other words, the EO-based model trained directly on consumption expenditures does not appear to be indirectly learning about structural poverty.

Our models consistently perform best when trained on data that spatially overlaps with the test set.[‖] For example, using standard (vs. LOCO) cross-validation, our pooled model for the poverty headcount P_s^0 at $z = \$1.90$ has an average r^2 of 0.72 (vs. 0.50), RMSE of 0.17 (vs. 0.23), and ρ of 0.84 (vs. 0.73). To visualize this, Figure 3 plots the first and second-stage out-of-sample structural predictions for all three spatial approaches to cross-validation (see Materials and Methods: Data Splitting), as well as realized consumption expenditures (left-most panel) for comparison.

This result is consistent with the literature as well as expectations; what poverty 'looks like' from a satellite view varies somewhat across even neighboring countries as the natural, social, and economic systems differ across contexts. Measurement error may also be correlated by country, survey, and even spatially within surveys due to the enumerators or the way that people answer questions about consumption. In sum, we may have both true differences and differential ability to detect these relationships across settings. Accordingly, performance across our country-specific models is heterogeneous, with Uganda standing out for its weak performance across evaluation metrics and specifications. For the P_s^0 at $z = \$1.90$ predictand with standard cross-validation, Uganda (vs. other countries) has an average r^2 value of 0.48 (vs. 0.52-0.82), RMSE of 0.23 (vs. 0.16-0.17), and ρ of 0.62 (vs. 0.68-0.86). It does especially poorly in spatial cross-fold validation, with one test fold predicted so poorly that the r^2 is negative (see SI Figures S12-S17 for additional country-specific results). There may be several reasons for this, including the aforementioned issues of data quality or a fundamentally weaker correlation between our EO features and structural poverty in Uganda. However, we suspect that it at least in part reflects the small sample size for Uganda: with only 245 clusters, we may simply not have enough data to train a reliable model. We have substantially more data for the remaining study countries, with 1047 (Ethiopia), 1691 (Malawi), and 1642 (Tanzania) unique clusters.

[‖] Specifically, containing clusters from the same country and/or region; there is no test/train overlap of the clusters.

Table 1. Summary out-of-sample performance for EO-based random forest models

A. FGT Poverty Measures, extreme global poverty line ($z = \$1.90$)

	Validation	Average R-squared P_r	Average R-squared P_s	Average RMSE P_r	Average RMSE P_s	Average Spearman's ρ P_r	Average Spearman's ρ P_s
P^0	country cv	0.542	0.648	0.209	0.169	0.717	0.736
	country sp-cv	0.272	0.396	0.242	0.190	0.531	0.617
	pooled cv	0.565	0.716	0.210	0.174	0.736	0.837
	pooled loco	0.116	0.502	0.289	0.233	0.518	0.732
P^1	country cv	0.463	0.600	0.109	0.058	0.713	0.751
	country sp-cv	0.130	0.289	0.121	0.067	0.550	0.631
	pooled cv	0.520	0.661	0.106	0.054	0.746	0.858
	pooled loco	0.093	0.248	0.169	0.072	0.509	0.732
P^2	country cv	0.404	0.508	0.073	0.026	0.698	0.731
	country sp-cv	0.061	0.060	0.074	0.030	0.543	0.591
	pooled cv	0.459	0.592	0.069	0.024	0.730	0.850
	pooled loco	0.031	0.033	0.118	0.030	0.490	0.729

B. FGT Poverty Measures, global poverty line ($z = \$3.20$)

	Validation	Average R-squared P_r	Average R-squared P_s	Average RMSE P_r	Average RMSE P_s	Average Spearman's ρ P_r	Average Spearman's ρ P_s
P^0	country cv	0.655	0.762	0.189	0.166	0.693	0.719
	country sp-cv	0.383	0.617	0.193	0.175	0.557	0.602
	pooled cv	0.660	0.769	0.178	0.165	0.704	0.777
	pooled loco	0.256	0.569	0.251	0.214	0.528	0.696
P^1	country cv	0.607	0.759	0.119	0.081	0.761	0.814
	country sp-cv	0.317	0.591	0.145	0.092	0.594	0.650
	pooled cv	0.644	0.800	0.125	0.085	0.783	0.872
	pooled loco	0.135	0.577	0.191	0.114	0.534	0.763
P^2	country cv	0.523	0.695	0.097	0.053	0.746	0.813
	country sp-cv	0.241	0.470	0.113	0.059	0.557	0.633
	pooled cv	0.581	0.765	0.098	0.053	0.776	0.876
	pooled loco	0.131	0.474	0.158	0.071	0.526	0.764

C. Asset Wealth Index

	Validation	R-squared	RMSE	Spearman's ρ
$\bar{A}$	country cv	0.707	7.446	0.780
	country sp-cv	0.508	8.517	0.650
	pooled cv	0.771	7.461	0.816
	pooled loco	0.512	9.665	0.668

Notes: Diagnostic statistics are averaged over folds and geographies. The corresponding disaggregated performance statistics are plotted in SI Figures S12-S17.

For comparison and to better situate our findings in the literature, we also predict asset wealth ($\bar{A}$) using a comparable model and EO feature set. Results are reported in Part C of Table 1. However, while structural and realized estimates are for P^α measures, the asset index is instead aggregated to the EA level using a simple mean. This limits comparability: to which poverty line and to which P^α do we compare? We cannot compare RMSE across the dependent variable types and the r^2 may also be sensitive to differences in the distribution, variance, and quality of different data sources. Still, it is encouraging that for the structural poverty headcount models (P_s^0 at $z = \$1.90$ and $\$3.20$), the r^2 and ρ for $\bar{A}$ are in the same general range: neither dependent

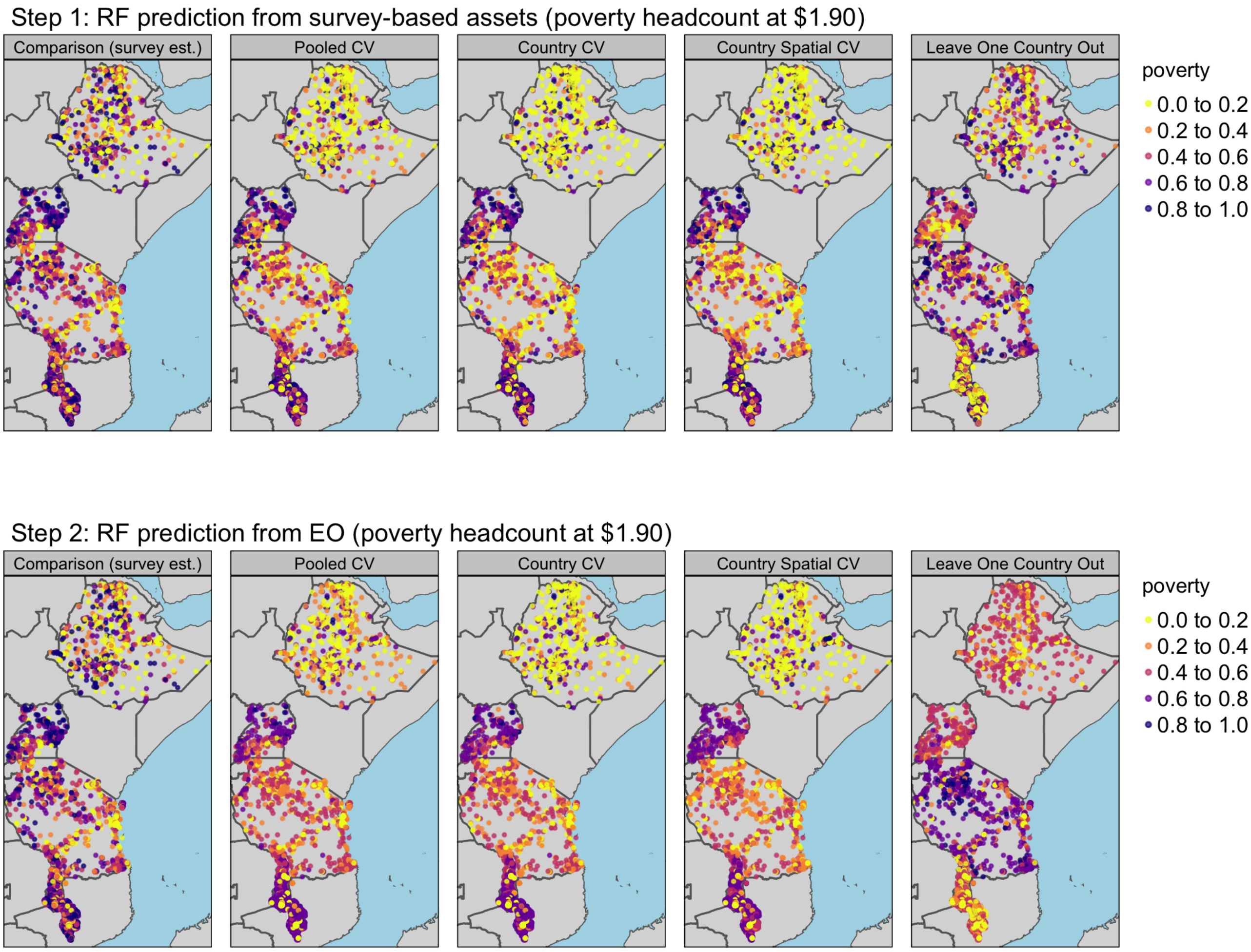


Fig. 3. Maps of Poverty Headcount (P_0) at the $\$1.90$ extreme poverty line. For comparison, the leftmost panel for each row are EA poverty rates estimated directly from realized consumption in the survey data. The remaining panels on the top row are predictions from the asset-consumption models into the test sets (combining model results from cross-validation). The corresponding maps in the bottom row are predicted from EO data trained on the structural poverty estimates.

variable demonstrates a clear and consistent performance advantage vis-a-vis the other for all models.

Our goal is to improve the relevance of the dependent variable for ML poverty mapping without compromising our ability to predict it. To usefully compare and contrast across predictands and samples requires a model and feature set with good predictive performance, which we achieve using an RF model and a suite of EO-derived variables. When predicting the structural poverty measures we achieve an average r^2 value (the most commonly reported metric in this literature) of 0.72 for the pooled cross-validation and 0.50 for the LOCO validation (and a slightly higher r^2 for the comparable asset index models). In comparison, a previous effort using satellite imagery and deep learning to predict consumption and assets using a LOCO approach for a very similar study area acheived r^2 for consumption (and assets) of 0.36 (0.46) for Malawi, 0.39 (0.63) for Nigeria, 0.52 (0.54) for Tanzania, and 0.44 (0.62) for Uganda (2). Another study that trained ML models on asset wealth data from 23 African countries achieved an average r^2 of 0.70 for held out country-years (1). An asset-wealth model trained on data from 56 low- and middle-income countries (LMICs) achieved an average r^2 of 0.70 using basic cross-validation and 0.59 using LOCO cross-validation (6). Using an approach that combines inference from interpretable features and satellite imagery from 25 countries in Africa, another recent study achieved an average r^2 of 0.85 for country-level CV and 0.88 for LOCO prediction of an asset wealth index (8).

In sum, it appears that our models and feature set offer solid performance despite the comparative simplicity and accessibility of our data and methods. We suspect that the small size of the clusters in the LSMS data (from 6-16 households) is also a limiting factor for model performance (7). While it is useful to situate our performance within the literature and r^2 is an intuitive metric, we caution that differences in the data, study areas, and approaches to validation complicate comparison of these values across studies. The r^2 may also not be the most important metric. For example, the relative ordering of clusters (as captured by metrics like the rank correlation coefficient) might matter more for targeting the distribution of a limited aid budget. Further, all three of the performance metrics considered thus far (r^2, RMSE, and ρ) are agnostic to heterogeneity in predictive performance. Previous studies have shown that performance at the low end of the wealth distribution tends to be the weakest; a promising average r^2 statistic may reflect the ability to distinguish "wealthier clusters from poorer clusters rather than in separating the poor from the near poor" (1). To investigate this important issue, we consider heterogeneity as well as how model performance changes when predicting the more distributionally sensitive $P^{1\&2}$ measures.

We find that our models systematically under-predict the poverty headcount and over-predict asset wealth for the poorest clusters. Consider again our benchmark pooled EO-models for P_r^0 and P_s^0 at $z = \$1.90$ and the model predicting average asset wealth $\bar{A}$. All three of these models have good overall predictive performance and appear to be relatively unbiased estimators, with predicted means similar to the reference test sets. Yet the realized, structural, and asset models all predict that the poorest clusters are better off, in an absolute sense, than in the reference data. For the bottom quintile, the EO-model of realized consumption predicts (vs. the 'ground-truth') a poverty rate of 66% (vs. 89%), the structural model predicts a poverty rate of 67% (vs. 84%), and the asset model predicts a wealth index of 16.6 (vs. 10.7).[**]

Our findings across measures also suggest that these models' ability to predict the magnitude of the gap below an extreme poverty line is weak. The proportion of the variation that we can predict out-of-sample (r^2) declines as we move from the poverty headcount (P^0) to the poverty gap (P^1) and poverty gap squared (P^2) measures, especially for the \$1.90 poverty line and for the spatially out-of-sample predictions (spatial CV and LOCO). This is not unique to the structural poverty estimates; r^2 also declines for the realized poverty measures.

In contrast, the rank correlation coefficients are relatively stable across the P^{0-2} predictands. This may arise because in our data relative poverty rankings appear to be relatively stable across the P^{0-2} measures: in the realized survey data the rank correlation coefficient (ρ) between $P_r^{0\&1}$ is 0.95 and between $P_r^{0\&2}$ is 0.91. Rank correlations are similarly high for the structural estimates. Arguably, these rank correlation coefficient estimates are the most salient for policymakers or program managers in geographic targeting of the distribution of scarce resources.

Discussion

We argue that structural poverty holds promise as a policy relevant and predictable object for machine learning poverty mapping. It is expressed in the same units as national and global policy objectives such as "[e]radicating extreme poverty for all people everywhere" under the first UN Sustainable Development Goal (23). Structural poverty is stable and forward-looking by construction; it is less sensitive to the classical measurement error and stochastic shocks that may quickly render maps based on realized consumption outdated. This interpretability and durability makes structural poverty estimates well-suited to inform development agendas that require medium- and long-term planning. They also have potential for measuring the geography of progress, but further research is needed to understand the dynamics of structural poverty mapping.

Of course, for applications such as the targeting of humanitarian aid in response to shocks policy makers need to understand *both* patterns of chronic deprivation and short-term impacts. We do not argue for structural poverty mapping to the exclusion of other efforts, especially important recent progress on mapping and forecasting shocks to consumption, food insecurity, or undernutrition (30, 31). Maps of asset wealth indices can be predicted quasi-globally and are useful complements to consumption-based poverty maps, even if they are imperfect substitutes. Human flourishing and deprivations are multidimensional and contextual, and pursuing a rich landscape of data products can eventually help us to understand the geographic intersections and discontinuities across measures.

In addition to its conceptual advantages, for a sample of four countries in southern and eastern Africa, we find that structural poverty is more easily predicted than realized poverty from an EO-based feature set. These differences are substantial. In our benchmark pooled multi-country model for the \$1.90 poverty headcount, the structural poverty measure has a higher r^2 (0.716 vs. 0.565), a lower RMSE (0.174 vs. 0.210), and a higher rank correlation coefficient (0.837 vs. 0.736) compared to models predicting realized poverty. In some specifications, asset indices may maintain a slight predictive advantage over structural poverty, but at the cost of interpretability and relevance to anti-poverty policy.[††]

The predictive accuracy of our models fall within the range of recently published multi-country poverty mapping efforts, but short of recent work that combines interpretable features and image-based deep learning (8). Our approach prioritizes accessibility: we use open-source data and models that can be run on a personal computer.[‡‡] Combining structural poverty and deep- and transfer-learning could be a productive avenue for future research.

Our results suggest that bias in our structural poverty estimates is likely modest in the context of interpolation: for example, when we are predicting poverty using models trained on survey data from adjacent communities in the same country. But the likelihood of bias increases when we extrapolate, for example, into another country that is not represented in the training data. This has potential implications for the coverage of structural poverty estimation, as well as for other methods of FGT poverty estimation. The type of data we utilize to train the structural consumption models is of limited availability, while asset indices are available for more countries.

In theory, we could leverage our trained asset-consumption models to predict structural poverty in settings where *only* asset data are available. But our results give us pause about undertaking such extensions, given the substantial issues of bias that emerge even predicting into a neighboring country

[**] The values for the asset wealth index range from approximately 1 to 80.

[††] The benchmark pooled multi-country model for average asset wealth has an r^2 of 0.771. However, comparative performance of asset vs. structural models varies across models, and measures such as r^2 and RMSE are sensitive to the scale and distributions of the respective data.

[‡‡] With the possible exception of the grid searches used to tune the household structural consumption models. It is therefore worth noting that the performance of these models is not highly sensitive to the choice of hyperparameters. Models still perform well with simpler (including software default) approaches to model tuning. For those interested in further reducing computation time, we note that first-stage models using second-order polynomial OLS regression achieve good performance in the first-stage structural poverty estimation.

with many shared attributes. Going beyond the southern and eastern Africa context it will be necessary to adapt and re-calibrate the structural poverty model. For example, incorporating savings and liabilities may be important in countries where these are more common.

All three sets of models – for realized poverty, structural poverty, and asset wealth – underestimate poverty in the poorest places. That this occurs across all three predictands suggests a fundamental limitation of current methods to predict extreme poverty from EO features. The same constraint appears to affect performance at the low end of the wealth distribution for imagery-based deep learning approaches (1, 2). This may arise because local labor markets, social safety nets, health, and other factors that are difficult to capture from satellite imagery or other geospatial features play a disproportionate role in the well-being of the poorest households. It likely also reflects noise or bias in the training data.

Even before we layer in data fusion and machine learning, survey-based consumption and poverty measurement is a topic of lively debate. Household consumption estimates are known to suffer from measurement errors, and those errors may inversely correlate with consumption or other markers of household welfare such as literacy and asset holdings (32). This has been shown to bias and decrease the accuracy of proxy means testing (33), and could similarly affect our first-stage structural poverty estimates. Long-standing questions around how best to adjust poverty measures for local consumption patterns and economies of scale within households have yet to be resolved (34, 35), and may affect geographic comparisons particularly between urban and rural populations, or across settings with different livelihoods or cultural norms regarding household structure. The small size of the household clusters in the LSMS data introduces random sampling error that will negatively impact our model performance (7). Sample bias is also a concern, particularly if there are fundamental differences between places that are and are not surveyed (36).

In time, improvements in algorithms or the availability of EO and other geospatial data products may improve our ability to detect the features of extreme poverty. But for now, high quality household surveys and survey-based research are needed to accurately understand the depth of deprivation amongst the poorest households and communities. Such data and analyses are similarly critical to the progression of ML micro-level poverty mapping in the future (4, 5).

Materials and Methods

Methodology. A household is defined as structurally (non-)poor if in expectation their portfolio of assets is associated with a (non-)poor consumption expenditure level (11). Here, we are interested in the continuous analogue to this binary concept of structural poverty, which we will refer to as structural consumption and denote by $\{C^A\}$ for household i in period t:

$$C^A_{it} = E[C_{it}|A_{it}] = f(A_{it}), \quad [1]$$

where $\{A_{it}\}$ is a vector of household productive assets. Unlike with a binary (poor and non-poor) classifier, a continuous $\{C^A\}$ measure allows us to assess the depth of a household's structural deprivations and to later construct aggregate poverty measures that capture the magnitude of any such shortfall.

If we assume that the differences between a household's structural consumption and realized consumption are stochastic, due to random shocks and/or classical measurement error, we can estimate a regression model that relates household assets and consumption expenditures to identify the function f:

$$C_{it} = C^A_{it} + \epsilon_{it} = f(A_{it}) + \epsilon_{it}, \quad [2]$$

where $\{\epsilon_{it}\}$ are the idiosyncratic errors. The best function f for prediction is unknown, so we test the comparative performance of parametric (first- and second-order polynomial regressions) as well as non-parametric RF models. We also evaluate a RF classification model, which predicts households' (non-)poor status. In our preliminary analysis we consider the bias-variance (or approximation v. overfit) trade-off between parametric and non-parametric models. We then use the most promising (RF-regression) model to construct estimates of structural poverty at the household level:

$$\widehat{C^A_{it}} = \widehat{f}(A_{it}). \quad [3]$$

Next, we construct the FGT (25) poverty measures $\{P^\alpha_s\}$, $\alpha = 0, 1, 2$, for the survey cluster s of interest. Specifically, we calculate the estimated share of individuals[§§] with consumption expenditures that fall below a national or international poverty line, or poverty headcount $\{P^0_s\}$; the average shortfall, or poverty gap $\{P^1_s\}$; and the average squared poverty gap $\{P^2_s\}$:

$$P^\alpha_s = \frac{1}{n_s} \sum_{\{j:\widehat{C}^A_j < z\}} \left(\frac{z - \widehat{C}^A_j}{z} \right)^\alpha \quad [4]$$

where $\{n_s\}$ is the total number of individuals in the sample and $\{j : \widehat{C}^A_j < z\}$ denotes the sampled households j at location s estimated to be below the poverty line z, and α is the FGT 'poverty aversion' parameter. These are the poverty measures we wish to map; these estimates serve as the training data in the subsequent step.

Next we consider how to project these estimates of structural poverty into areas (that we simulate are) not covered by household surveys. We train OLS and RF regression models to predict structural poverty $\{P^\alpha_s\}$ using open-source EO and other geospatial data $\{Z_s\}$ as:

$$P^\alpha_s = g(Z_s) + \varepsilon_s. \quad [5]$$

All RF models are fit to minimize RMSE. The household structural consumption models use a shared set of hyperparameters based on results of grid search (see SI Figures S1-S9), with the number of $trees = 1000$, the minimum size of terminal nodes $min\_n = 30$, and the maximum number of variables sampled as candidates at each split at $mtry = 8$. For the cluster level models, the number of $trees = 200$ and the model hyperparameters are instead tuned individually using a 10 (min_n) by 10 ($mtry$) grid search. Model tuning is further described in SI Section .

We evaluate model performance using the coefficient of determination, or r^2, the Root Mean Squared Error (RMSE), and a rank correlation coefficient, Spearman's ρ. The r^2 is the most commonly cited performance measure in the poverty mapping literature, and offers an intuitive measure of the degree of variation in the dependent variable that is explained by the model. However, it is sensitive to features of the data (e.g., variance and measurement error) used for validation (37). RMSE may be a more reliable indicator of performance, except (as in our case) if we wish to compare across different types of dependent variables. Finally, rank correlation coefficients may be a particularly useful diagnostic for applications such as the geographic targeting of humanitarian or development aid, when we are most interested in the relative ordering of communities rather than their absolute levels of deprivation.

Data splitting. We use three complementary nested cross-validation approaches that allow us to assess performance in reference to different use cases:

- **k-fold cross-validation** First, we split the data for each country into five folds based on a random draw of the enumeration areas. We also implement a multi-country, or

[§§] Consumption expenditures are estimated at the household level, then weighted based on household size.

pooled, version of the k-fold cross-validation. This approach simulates predictive performance for interpolation in surveyed areas. For example, if we have cluster-sampled household survey data for the country or countries of interest, this approach simulates performance predicting into the un-surveyed clusters.

- **Spatial k-fold cross-validation:** We also implement a spatially stratified variation of the k-fold cross-validation. Here, the test fold is geographically distinct from the training data to avoid overestimating performance due to spatial auto-correlation (6, 38). This would be analogous to a use case where we have survey data for the country of interest, but not for all regions, and therefore need to spatially extrapolate within the same country.
- **Leave-one-country-out cross-validation:** Finally, we test the validity of a pooled model for predicting into a country for which (we simulate that) there are no household survey data available. Here, we leave out each country in the data set in turn, training the model on all other countries' data. Extrapolation into un-surveyed countries requires stronger assumptions, but also has the advantage of more training data.

Household survey data. Our approach requires data on consumption expenditures (or income) and productive assets, geo-referenced at the micro-level. These are obtained from from 13 LSMS surveys: for Ethiopia (2011-2012, 2013-2014, 2015-2016 & 2018-2019), Malawi (2010-2011, 2016-2017 & 2019-2020), Tanzania (2008-2009, 2010-2011, 2012-2013, 2014-2015 & 2019-2020) and Uganda (2011-2012). We use the published consumption expenditure aggregates from the respective datasets, which have been constructed by aggregating across several categories of consumption and then adjusting for regional cost-of-living differences. According to survey documentation, these are broadly consistent in their construction across countries and surveys. We convert all values to 2011 purchasing-power-parity US dollars. Our asset index, which is not pre-constructed in the LSMS data, is calculated following the data reduction techniques used to consolidate and harmonize asset data across Demographic and Health Surveys (39–42). We implement a broad definition of productive assets: the stocks that generate the income that enables consumption expenditure. This includes human capital, land, livestock, capital equipment and buildings, and water and sanitation. Details of the procedure and specific assets are described in the SI.

Geo-spatial features. Our geo-spatial predictors consist of interpretable features, known to correlate with poverty and/or wealth, derived from publicly available data sources. Because our data are geo-referenced at the cluster level with some random displacement to preserve anonymity, we extract survey-year averages of our geo-spatial variables for a 2km buffer radius in urban areas and a 5km buffer radius in rural areas (unless otherwise noted). Feature values are contemporaneous to the survey data unless otherwise noted; we include lags for some variables based on the expected temporal relationship. Large datasets were pre-processed in Google Earth Engine, with dataset construction in R.

Several of our features relating to geography and demography are time-invariant or slow-moving. We include building footprints, obtained from the Open Buildings Project Version 2 (43), which are a reliable indicator of human settlement and socioeconomic conditions on the ground (7, 10, 44). Average slopes and elevations are computed via Google Earth Engine based on data from the NASA Shuttle Radar Topography Mission (SRTM) to capture geophysical constraints on economic development (45). We also include travel time to the nearest urban centre, a known correlate of prosperity, from the Malaria Atlas Project (MAP) (46).

We draw on time-series data for features that vary substantively over the study period. Population count and density, which have been shown to be predictive of asset wealth in previous studies (6, 8), are derived from data by WorldPop (47, 48). A three-year average of nighttime lights is included as a proxy for economic activity (49). Given the time span of our dataset (2008-2020), we use a nighttime lights product that harmonizes data from the Defense Meteorological Satellite Program (1992–2013) and the Visible Infrared Imaging Radiometer Suite (2012–2018) (50, 51).

Climatic conditions and episodes of heat and water stress may impact people's well-being through multiple avenues, especially via conditions for agriculture and livestock. We use the Climate Hazards group Infrared Precipitation with Stations (CHIRPS) to construct variables for long-term rainfall patterns, annual rainfall, and rainfall z-scores (52). Binned temperature variables reflecting the hours above 30 degrees Celsius are constructed from the Modern-Era Retrospective analysis for Research and Applications version 2 (MERRA-2) 2-meter air temperature (53). The Normalized Difference Vegetation Index (NDVI) is derived from the NOAA Climate Data Record (CDR) of Advanced Very High Resolution Radiometer (AVHRR) Surface Reflectance (54). The NDVI is an indicator of greenness that has been shown to correlate with poverty in rural, agriculturally dependent settings (3, 38).

ACKNOWLEDGMENTS. This work was supported by the Cornell Atkinson Center for Sustainability. We thank Luc Christiaensen, Rebecca Chaplin-Kramer, Peter Lanjouw, Yanyan Liu, Daniel Maggio, Linden McBride, Joshua D. Merfeld, Aleksandr Michuda, Rachel Neugarten, David Newhouse, Stephan Schmidt, Linda Steinhübel-Rasheed, and Hantao Wu for helpful conversations and feedback on this work. We also thank Cassian D'Cunha and the Cornell Center for Social Sciences for computational resources and support. Any remaining errors are solely ours.

1. C Yeh, et al., Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nat. Commun.* **11**, 2583 (2020).
2. N Jean, et al., Combining satellite imagery and machine learning to predict poverty. *Science* **353**, 790–794 (2016).
3. C Browne, et al., Multivariate random forest prediction of poverty and malnutrition prevalence. *PLOS ONE* **16**, e0255519 (2021).
4. M Burke, A Driscoll, DB Lobell, S Ermon, Using satellite imagery to understand and promote sustainable development. *Science* **371** (2021).
5. L McBride, et al., Predicting poverty and malnutrition for targeting, mapping, monitoring, and early warning. *Appl. Econ. Perspectives Policy* **n/a** (2021).
6. G Chi, H Fang, S Chatterjee, JE Blumenstock, Microestimates of wealth for all low- and middle-income countries. *Proc. Natl. Acad. Sci.* **119** (2022).
7. R Engstrom, J Hersh, D Newhouse, Poverty from Space: Using High Resolution Satellite Imagery for Estimating Economic Well-being. *The World Bank Econ. Rev.* **36**, 382–412 (2022).
8. K Lee, J Braithwaite, High-resolution poverty maps in Sub-Saharan Africa. *World Dev.* **159**, 106028 (2022).
9. D Newhouse, JD Merfeld, AP Ramakrishnan, T Swartz, P Lahiri, Small Area Estimation of Monetary Poverty in Mexico Using Satellite Imagery and Machine Learning. *World Bank Policy Res. Work. Pap.* **10175** (2022).
10. D Newhouse, Small area estimation of poverty and wealth using geospatial data: What have we learned so far? *Calcutta Stat. Assoc. Bull.* p. 00080683231198591 (2023).
11. MR Carter, CB Barrett, The economics of poverty traps and persistent poverty: An asset-based approach. *J. Dev. Stud.* **42**, 178–199 (2006).
12. M Ghosh, JNK Rao, Small Area Estimation: An Appraisal. *Stat. Sci.* **9**, 55–76 (1994).
13. JNK Rao, Some recent advances in model based small area estimation. *Surv. Methodol.* **25**, 175–186 (1999).
14. C Elbers, JO Lanjouw, P Lanjouw, Micro-Level Estimation of Poverty and Inequality. *Econometrica* **71**, 355–364 (2003).
15. L Christiaensen, P Lanjouw, J Luoto, D Stifel, Small area estimation-based prediction methods to track poverty: Validation and applications. *J. Econ. Inequal.* **10**, 267–297 (2012).
16. M Torabi, JNK Rao, On small area estimation under a sub-area level model. *J. Multivar. Analysis* **127**, 36–55 (2014).
17. S Lange, UJ Pape, P Pütz, Small Area Estimation of Poverty Under Structural Change. *Rev. Income Wealth* **68**, S264–S281 (2022).
18. P Corral, I Molina, A Cojocaru, S Segovia, Guidelines to Small Area Estimation for Poverty Mapping, (World Bank, Washington, DC), Technical report (2022).
19. A Sen, *Inequality Reexamined.* (Russell Sage Foundation, New York), (1992).
20. MR Carter, J May, One Kind of Freedom: Poverty Dynamics in Post-apartheid South Africa. *World Dev.* **29**, 1987–2006 (2001).
21. JX Sullivan, L Turner, S Danziger, The Relationship between Income and Material Hardship. *J. Policy Analysis Manag.* **27**, 63–81 (2008).
22. A Head, M Manguin, N Tran, JE Blumenstock, Can Human Development be Measured with Satellite Imagery? in *Proceedings of the Ninth International Conference on Information and Communication Technologies and Development.* (ACM, Lahore Pakistan), pp. 1–11 (2017).
23. GA UN, Transforming Our World: The 2030 Agenda for Sustainable Development, (United Nations), Technical Report A/RES/70/1 (2015).

24. CJ Pu, et al., How poverty is measured impacts who gets classified as impoverished. *Proc. Natl. Acad. Sci.* **121**, e2316730121 (2024).
25. J Foster, Joel Greer, Erik Thorbecke, A Class of Decomposable Poverty Measures. *Econometrica* **52**, 761–766 (1984).
26. J Foster, J Greer, E Thorbecke, The Foster–Greer–Thorbecke (FGT) poverty measures: 25 years later. *J. Econ. Inequal.* **8**, 491–524 (2010).
27. F Naschold, CB Barrett, Do short-term observed income changes overstate structural economic mobility? *Oxf. Bull. Econ. Stat.* **73**, 705–717 (2011).
28. CB Barrett, MR Carter, The Economics of Poverty Traps and Persistent Poverty: Empirical and Policy Implications. *J. Dev. Stud.* **49**, 976–990 (2013).
29. CB Barrett, T Garg, L McBride, Well-Being Dynamics and Poverty Traps. *Annu. Rev. Resour. Econ.* **8**, 303–327 (2016).
30. EC Lentz, H Michelson, K Baylis, Y Zhou, A data-driven approach improves food insecurity crisis prediction. *World Dev.* **122**, 399–409 (2019).
31. G Martini, et al., Machine learning can guide food security efforts when primary data are not available. *Nat. Food* **3**, 716–728 (2022).
32. J Gibson, B Kim, Measurement Error in Recall Surveys and the Relationship between Household Size and Food Demand. *Am. J. Agric. Econ.* **89**, 473–489 (2007).
33. J Gazeaud, Proxy Means Testing Vulnerability to Measurement Errors? *J. Dev. Stud.* **56**, 2113–2133 (2020).
34. A Deaton, *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy.* (The World Bank), (1997).
35. D Jolliffe, SK Tetteh-Baah, Identifying the Poor – Accounting for Household Economies of Scale in Global Poverty Estimates. *IZA Inst. Labor Econ. Discuss. Pap. Ser.* **IZA DP No. 15615** (2022).
36. JD Merfeld, D Newhouse, *Improving Estimates of Mean Welfare and Uncertainty in Developing Countries*, Policy Research Working Paper. (World Bank Group), Wps 10348 edition, (2023).
37. P Corral, H Henderson, S Segovia, *Poverty Mapping in the Age of Machine Learning*, Policy Research Working Papers. (The World Bank), (2023).
38. B Tang, Y Liu, DS Matteson, Predicting poverty with vegetation index. *Appl. Econ. Perspectives Policy* **44**, 930–945 (2022).
39. SO Rutstein, K Johnson, The DHS wealth index, (United States Agency for International Development), Technical Report DHS Comparative Reports No. 6. (2004).
40. SO Rutstein, The DHS wealth index: Approaches for rural and urban areas, (United States Agency for International Development), Technical Report DHS Working Papers No. 60 (2008).
41. TN Croft, AMJ Marshall, CK Allen, Guide to DHS Statistics, (ICF, Rockville, MD, USA), Technical Report DHS-7 (2018).
42. J Smits, R Steendijk, The International Wealth Index (IWI). *Soc. Indic. Res.* **122**, 65–85 (2015).
43. W Sirko, et al., Continental-Scale Building Detection from High Resolution Satellite Imagery (2021).
44. T Masaki, D Newhouse, AR Silwal, A Bedada, R Engstrom, Small area estimation of non-monetary poverty with geospatial data. *Stat. J. IAOS* **38**, 1035–1051 (2022).
45. TG Farr, et al., The Shuttle Radar Topography Mission. *Rev. Geophys.* **45**, RG2004 (2007).
46. DJ Weiss, et al., A global map of travel time to cities to assess inequalities in accessibility in 2015. *Nature* **553**, 333–336 (2018).
47. CT Lloyd, A Sorichetta, AJ Tatem, High resolution global gridded data for use in population studies. *Sci. Data* **4**, 170001 (2017).
48. WorldPop, Global 1km Population (2018).
49. Q Huang, X Yang, B Gao, Y Yang, Y Zhao, Application of DMSP/OLS Nighttime Light Images: A Meta-Analysis and a Systematic Literature Review. *Remote. Sens.* **6**, 6844–6866 (2014).
50. X Li, Y Zhou, M Zhao, X Zhao, A harmonized global nighttime light dataset 1992–2018. *Sci. Data* **7**, 168 (2020).
51. X Li, Y Zhou, M Zhao, X Zhao, Harmonization of DMSP and VIIRS nighttime light data from 1992-2021 at the global scale (2023).
52. C Funk, et al., The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. *Sci. Data* **2**, 150066 (2015).
53. Global Modeling and Assimilation Office (GMAO), MERRA-2 tavg1_2d_slv_Nx: 2d,1-Hourly,Time-Averaged,Single-Level,Assimilation,Single-Level Diagnostics V5.12.4 (2015).
54. E Vermote, NOAA Climate Data Record (CDR) of AVHRR Normalized Difference Vegetation Index (NDVI), Version 5 (2018).

Supporting Information for

Micro-level structural poverty estimates for southern and eastern Africa

Elizabeth Tennant, Yating Ru, Peizan Sheng, David S. Matteson, Christopher B. Barrett

Corresponding Author: Elizabeth Tennant.
E-mail: ejt58@cornell.edu

This PDF file includes:

Supporting text
Figs. S1 to S18
Table S1
SI References

Supporting Information Text

Data Supplement

Household data cleaning and pre-processing. Household surveys suffer from missing data and measurement error, and these data quality issues may vary systematically with the country and specific survey. Uganda has the highest prevalence of missing values; approximately half of the sample households have at least one missing data field, often consumption expenditure, land, livestock, or other agricultural assets. For Ethiopia, Malawi, and Tanzania 4% - 14% of households are missing necessary data fields. We do not detect patterns of imbalance across households with or without missing data. Observations with missing data are dropped. We also use winsorization to replace extreme values, truncating at the 99th percentile. This removes outliers, especially those that appear erroneous (e.g., owning 124 computers or 111 bicycles). Some rare assets are converted to dummy variables, such as: irrigated land, boats, ploughs, tractors, harvesters, and sprayers.

Asset wealth. We construct our asset wealth index from a common set of assets surveyed in the Living Standards and Measurement Surveys (LSMS). These include tropical livestock units, total land area, irrigated land area, number of rooms per person, number of ploughs, radios, TVs, bicycles, motorcycles, and cellphones, as well as access to electricity, improved drinking water, improved toilet facilities, and improved materials for roof, wall, and floor. Several of these variables are recoded following the relevant literature. For example, tropical livestock units convert all livestock to common units based on (an assumed, based on species) live weight of 250kg per TLU (1). Designations of improved or unimproved facilities or materials are based on DHS standards (2). These same re-coding procedures are used to process the asset variables prior to use in the structural consumption modeling.

To construct a single continuous asset index from these variables, we draw on procedures used to calculate the Demographic and Health Surveys Relative Wealth Index (RWI) (3, 4) and the International Wealth Index (IWI) (5). A common wealth score across the full study area is first calculated, as well as separate urban and rural wealth scores. The common scores, which exclude assets that may have divergent relationship with wealth in rural vs. urban areas, are then used to calibrate the separate urban and rural models.

Consumption expenditures. Consumption expenditure aggregates are pre-computed in the LSMS surveys, but typically provided in nominal local currency values. We convert these values to a common currency and equalize their purchasing power over countries and years using purchasing power parity (PPP) adjustments.* We adjust all consumption expenditures to 2011 PPP dollars per capita per day as follows:

$$\text{Real consumption (\$/day/person)} = \frac{\text{Consumption in local currency}}{\text{PPP conversion factor} \times \text{hhsize} \times 365} \qquad [1]$$

Asset to Consumption Modeling

Model tuning. The structural consumption models utilize a shared set of hyper-parameters: an $mtry = 8$, a $min\_n = 30$, and the number of $trees = 1000$. These were selected based on grid search using nested re-sampling of the training data for a single 80%-20% split of each permutation of the study area (individual country, pooled, and LOCO). Re-tuning the household models for every data-split of the cluster-level analysis, as we do in the second stage of the analysis, would be computationally untenable. Based on the generally small differences in model performance across the hyperparameter grids (see Appendix Figures S1-S9) we anticipate that this would not substantially aid model performance. For those interested in replicating this approach but without the computational resources for a similar grid search, we note that performance does not appear to be highly sensitive to the hyperparameters and that solid performance is achieved with software defaults.

*We use the World Bank's PPP conversion factor, private consumption (LCU per international $), available at: https://data.worldbank.org/indicator/PA.NUS.PRVT.PP

Table S1. Summary out-of-sample performance for EO-based OLS-1 models

A. Poverty Measures

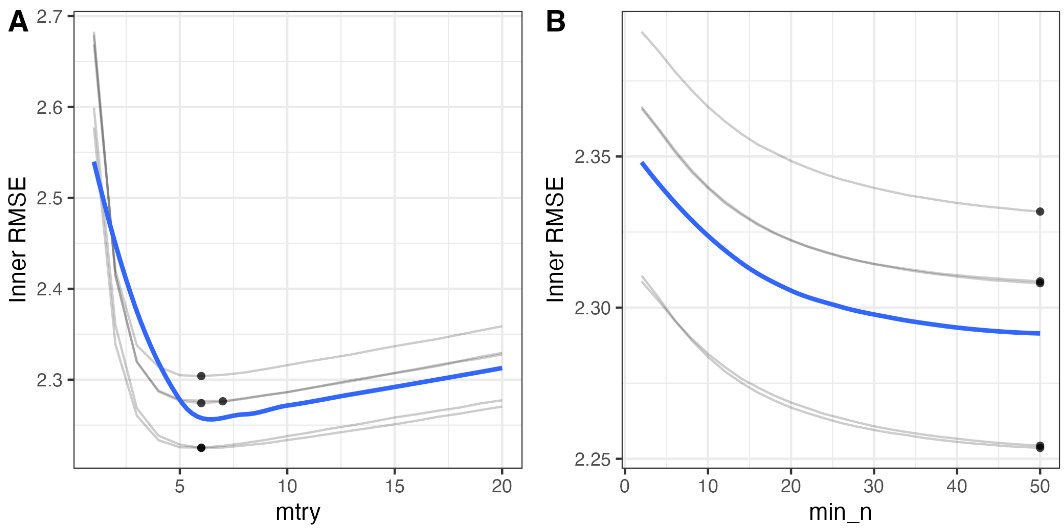
Predictand	Validation	Average R-squared		Average RMSE		Average Spearman's ρ	
		P_r	P_s	P_r	P_s	P_r	P_s
P^0 $z = \$1.90$	country cv	0.453	0.565	0.231	0.198	0.641	0.680
	country spatial cv	0.161	0.358	0.255	0.216	0.457	0.533
	pooled cv	0.486	0.591	0.229	0.212	0.672	0.774
	pooled leave-country-out	0.025	0.309	0.310	0.255	0.518	0.691
P^1 $z = \$1.90$	country cv	0.360	0.387	0.114	0.070	0.657	0.674
	country spatial cv	-0.011	0.216	0.123	0.072	0.418	0.510
	pooled cv	0.390	0.533	0.119	0.065	0.657	0.788
	pooled leave-country-out	-0.014	0.160	0.167	0.081	0.541	0.650
P^2 $z = \$1.90$	country cv	0.270	0.264	0.072	0.031	0.624	0.667
	country spatial cv	-0.055	0.059	0.079	0.034	0.427	0.496
	pooled cv	0.317	0.437	0.077	0.027	0.632	0.772
	pooled leave-country-out	0.007	0.039	0.109	0.032	0.527	0.587
P^0 $z = \$3.20$	country cv	0.603	0.739	0.200	0.181	0.627	0.690
	country spatial cv	0.302	0.548	0.209	0.181	0.490	0.578
	pooled cv	0.607	0.745	0.192	0.171	0.653	0.743
	pooled leave-country-out	0.250	0.510	0.260	0.207	0.524	0.707
P^0 $z = \$3.20$	country cv	0.603	0.741	0.200	0.182	0.627	0.691
	country spatial cv	0.302	0.549	0.209	0.180	0.490	0.582
	pooled cv	0.607	0.747	0.192	0.172	0.653	0.742
	pooled leave-country-out	0.250	0.517	0.260	0.207	0.524	0.710
P^1 $z = \$3.20$	country cv	0.517	0.693	0.131	0.097	0.685	0.750
	country spatial cv	0.152	0.478	0.165	0.107	0.494	0.602
	pooled cv	0.551	0.715	0.140	0.101	0.705	0.815
	pooled leave-country-out	0.012	0.443	0.197	0.129	0.539	0.718
P^2 $z = \$3.20$	country cv	0.441	0.592	0.104	0.066	0.677	0.726
	country spatial cv	0.050	0.415	0.116	0.069	0.442	0.601
	pooled cv	0.461	0.652	0.111	0.065	0.685	0.820
	pooled leave-country-out	0.000	0.344	0.157	0.083	0.562	0.711

B. Asset Wealth Index

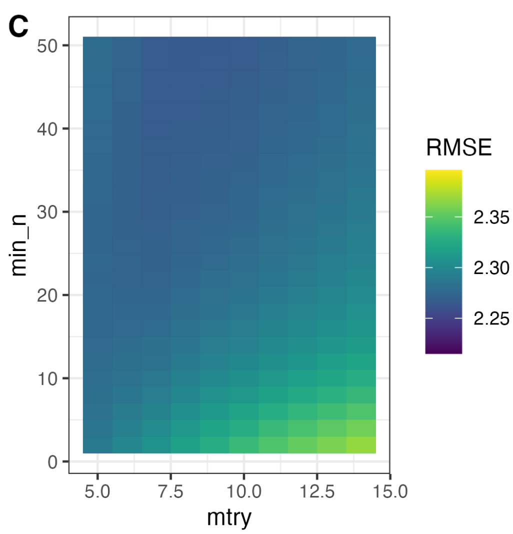
Predictand	Validation	R-squared	RMSE	Spearman's ρ
$\bar{A}$	country cv	0.646	8.240	0.753
	country spatial cv	0.409	8.588	0.588
	pooled spatial cv	0.719	8.249	0.769
	pooled leave-country-out	0.614	9.349	0.709

Notes: Diagnostic statistics are averaged over folds and geographies.

A **B**



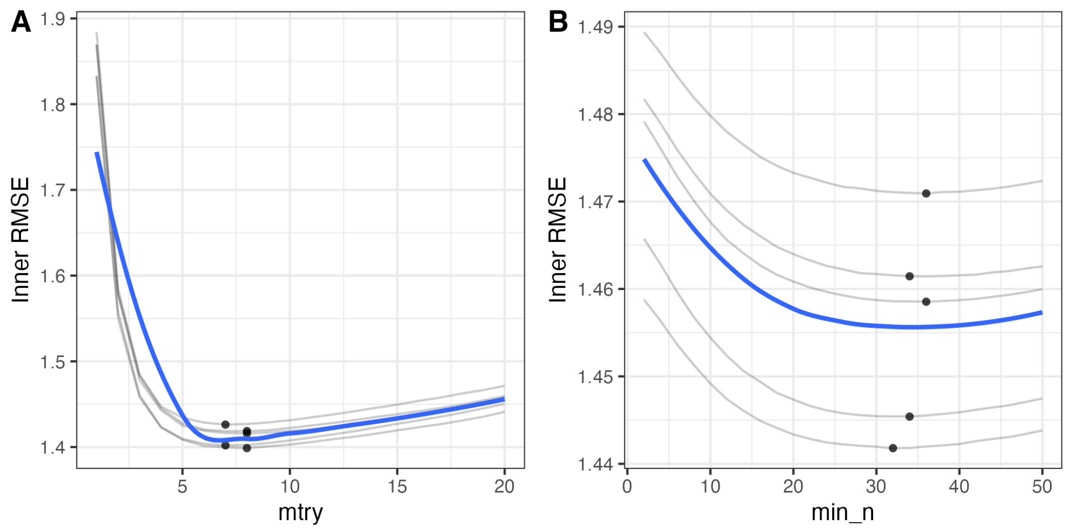
C

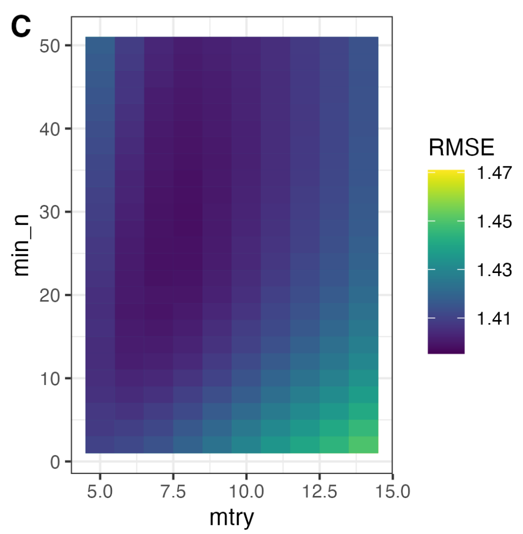


D

mtry	min_n	Inner RMSE (mean)	Outer RMSE
8	50	2.294	2.153
9	50	2.215	2.486
8	50	2.218	2.295
8	50	2.266	2.640
8	48	2.267	2.702

Fig. S1. Results of grid search of hyperparameters for random forest structural poverty estimation for **Ethiopia**.

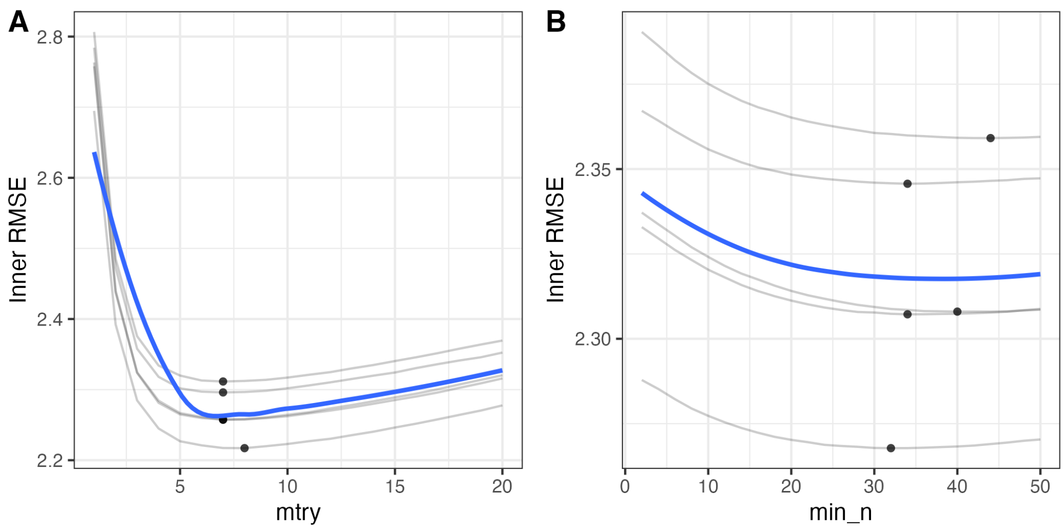


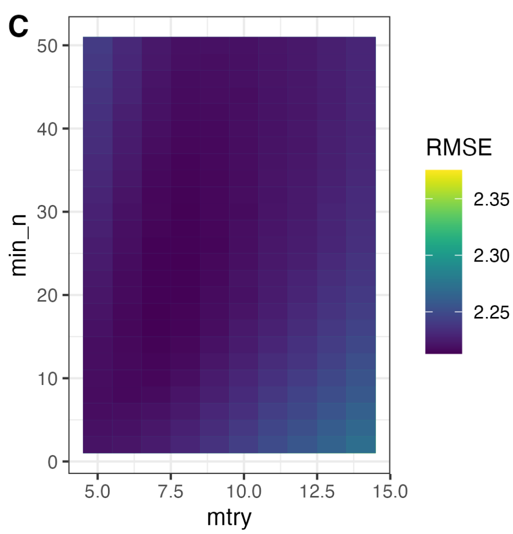


D

mtry	min_n	Inner RMSE (mean)	Outer RMSE
8	28	1.415	1.326
8	26	1.423	1.362
8	28	1.395	1.506
8	30	1.412	1.420
8	28	1.398	1.431

Fig. S2. Results of grid search of hyperparameters for random forest structural poverty estimation for **Malawi**.

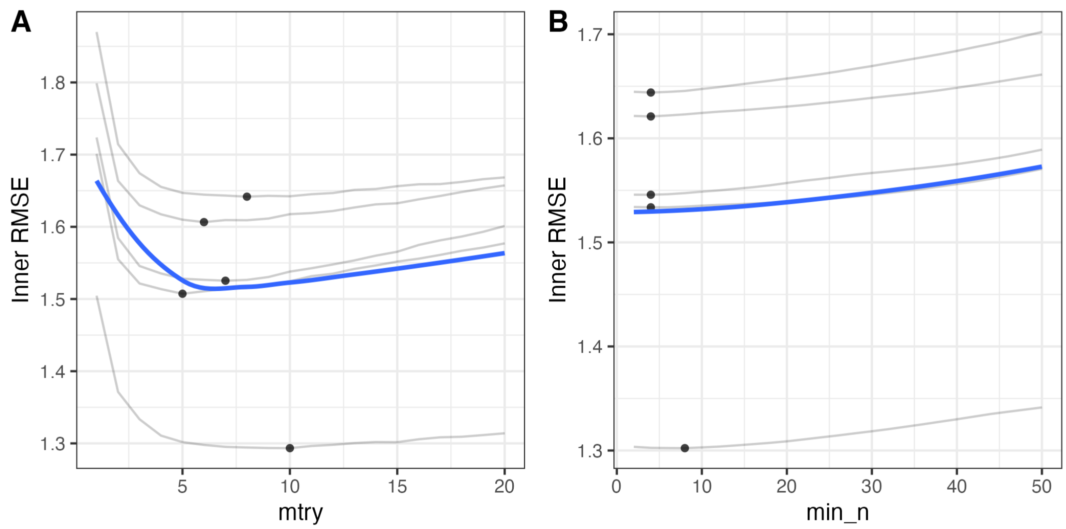


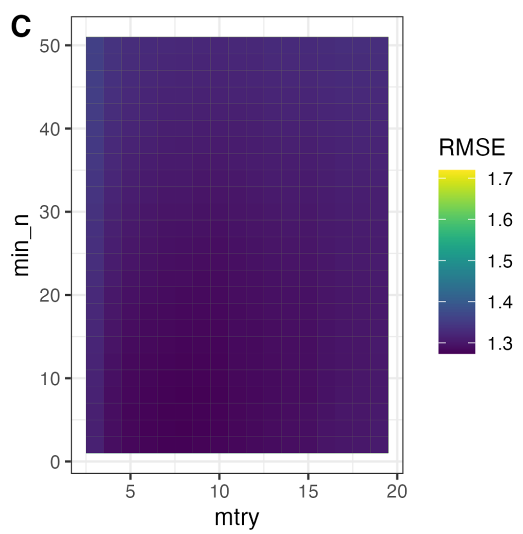


D

mtry	min_n	Inner RMSE (mean)	Outer RMSE
9	38	2.306	2.140
7	20	2.292	2.099
8	32	2.253	2.160
8	30	2.253	2.098
7	18	2.213	2.192

Fig. S3. Results of grid search of hyperparameters for random forest structural poverty estimation for **Tanzania**.

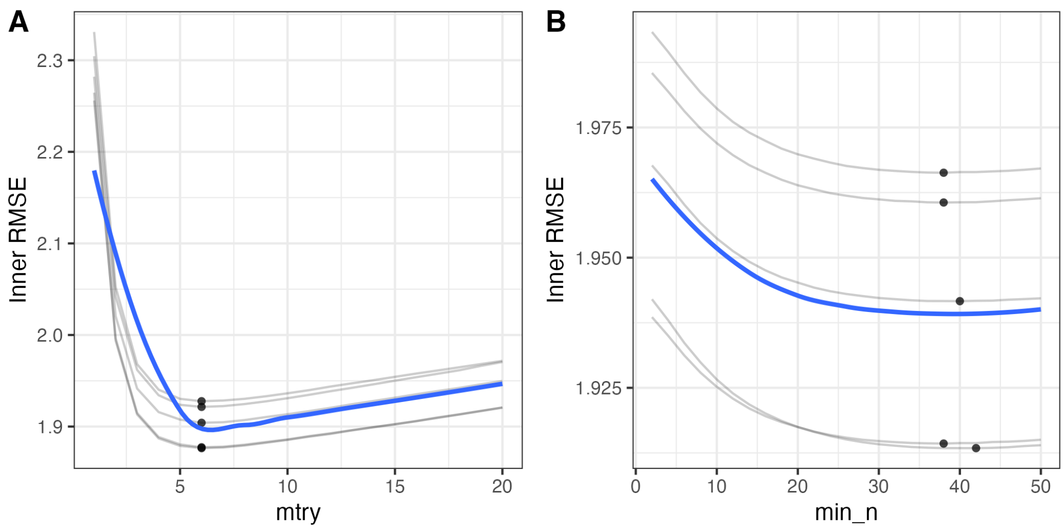


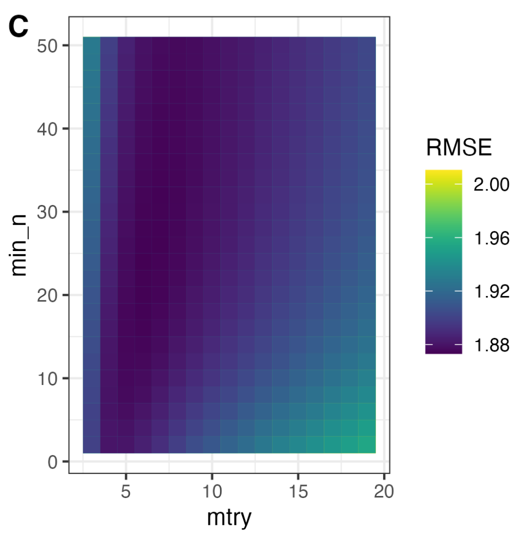


D

mtry	min_n	Inner RMSE (mean)	Outer RMSE
5	2	1.498	1.744
6	4	1.582	1.253
5	2	1.477	1.842
6	4	1.611	1.328
8	4	1.274	1.569

Fig. S4. Results of grid search of hyperparameters for random forest structural poverty estimation for **Uganda**.

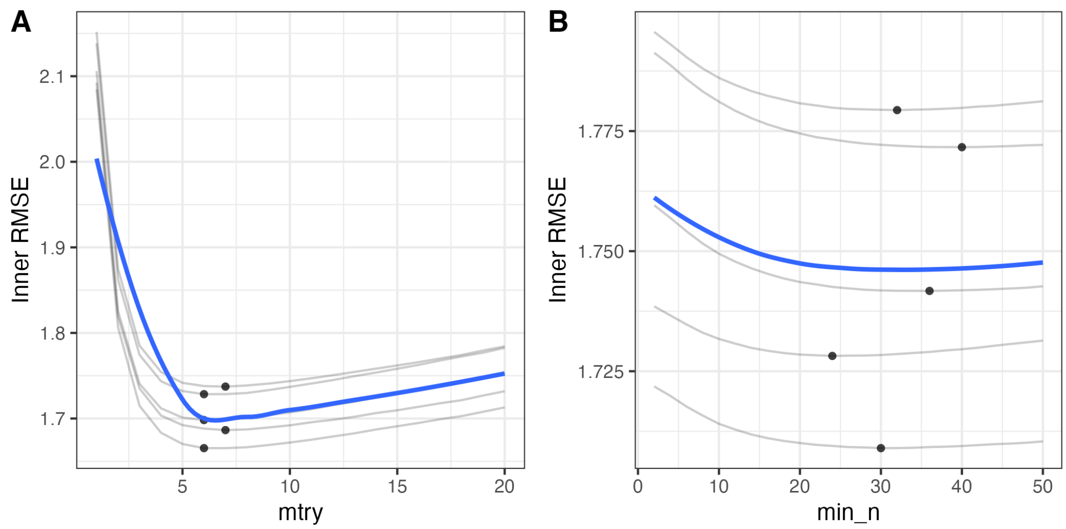


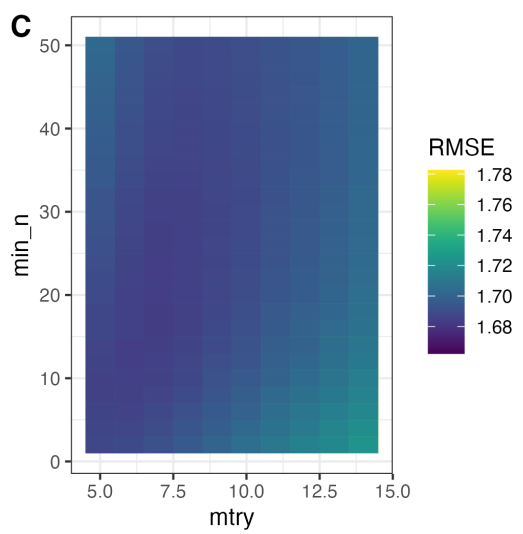


D

mtry	min_n	Inner RMSE (mean)	Outer RMSE
7	28	1.901	1.909
7	32	1.874	2.022
6	20	1.918	1.819
6	24	1.925	1.768
6	24	1.873	2.020

Fig. S5. Results of grid search of hyperparameters for random forest structural poverty estimation for **Pooled model (all countries)**.

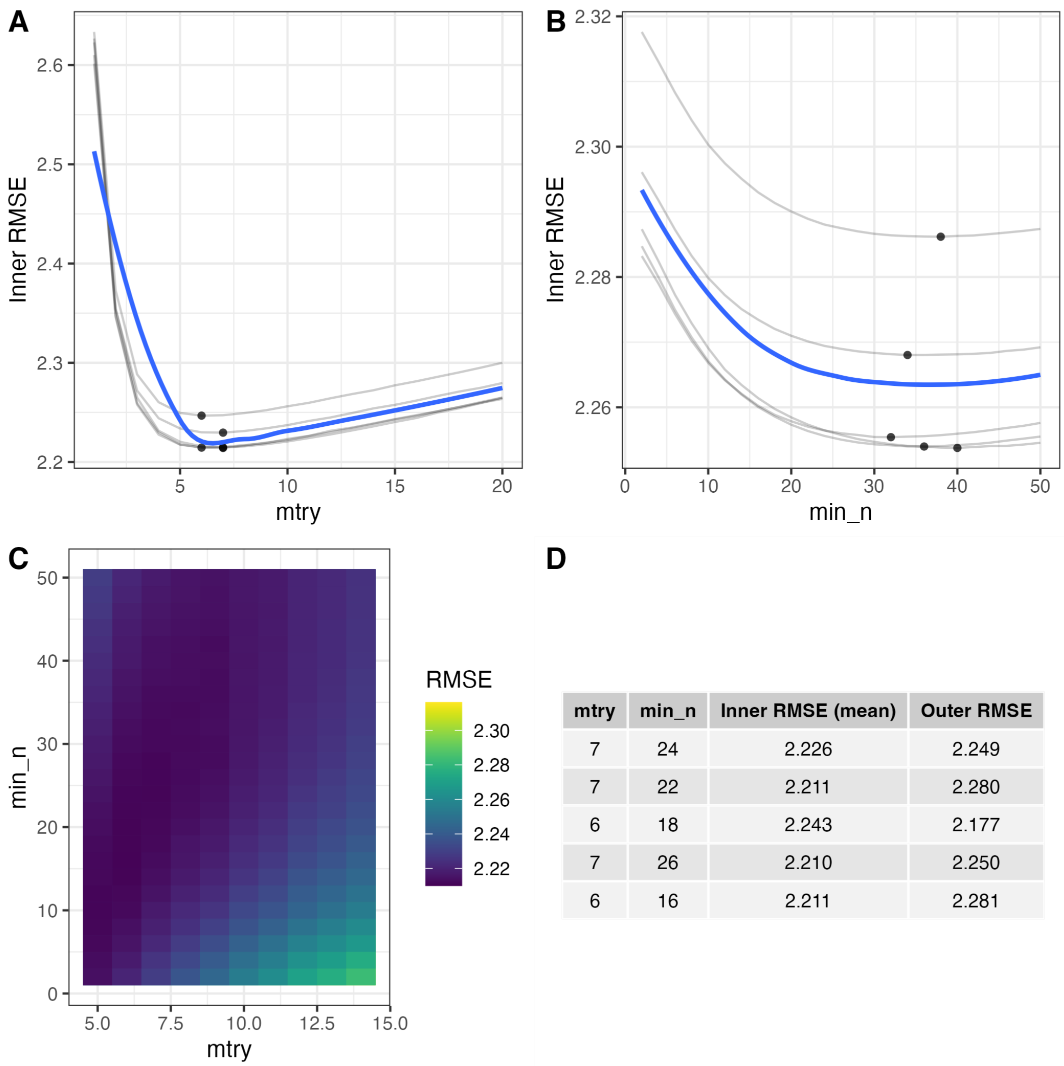




D

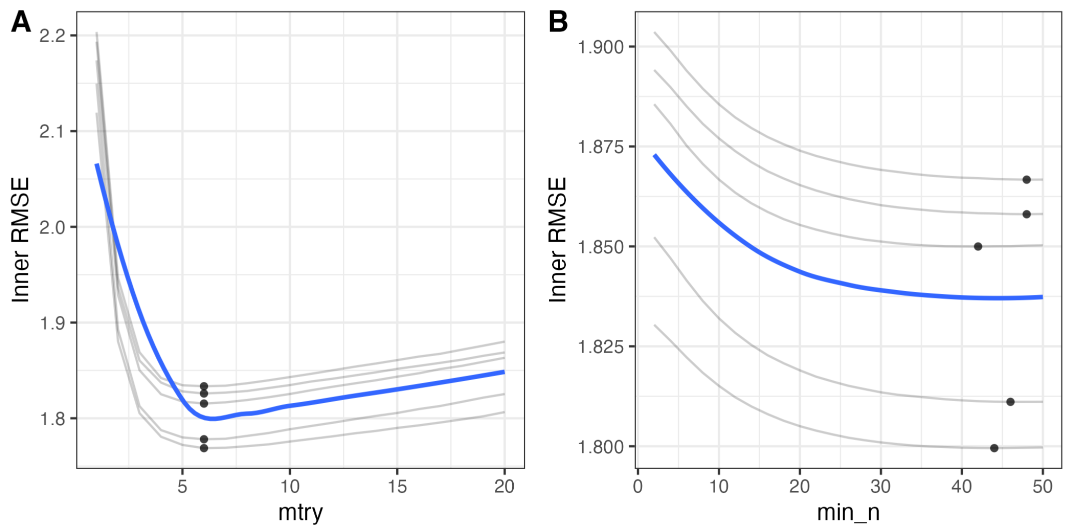
mtry	min_n	Inner RMSE (mean)	Outer RMSE
7	20	1.734	1.588
6	14	1.662	1.854
7	28	1.725	1.643
6	20	1.695	1.735
7	18	1.683	1.771

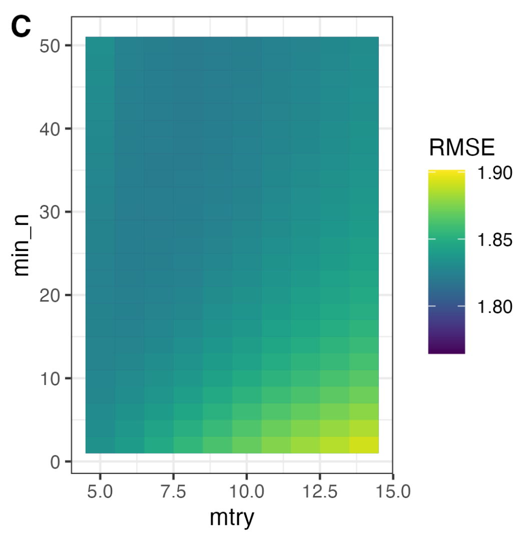
Fig. S6. Results of grid search of hyperparameters for random forest structural poverty estimation for **Leave-one-country-out (Ethiopia)**.



mtry	min_n	Inner RMSE (mean)	Outer RMSE
7	24	2.226	2.249
7	22	2.211	2.280
6	18	2.243	2.177
7	26	2.210	2.250
6	16	2.211	2.281

Fig. S7. Results of grid search of hyperparameters for random forest structural poverty estimation for **Leave-one-country-out (Malawi)**.

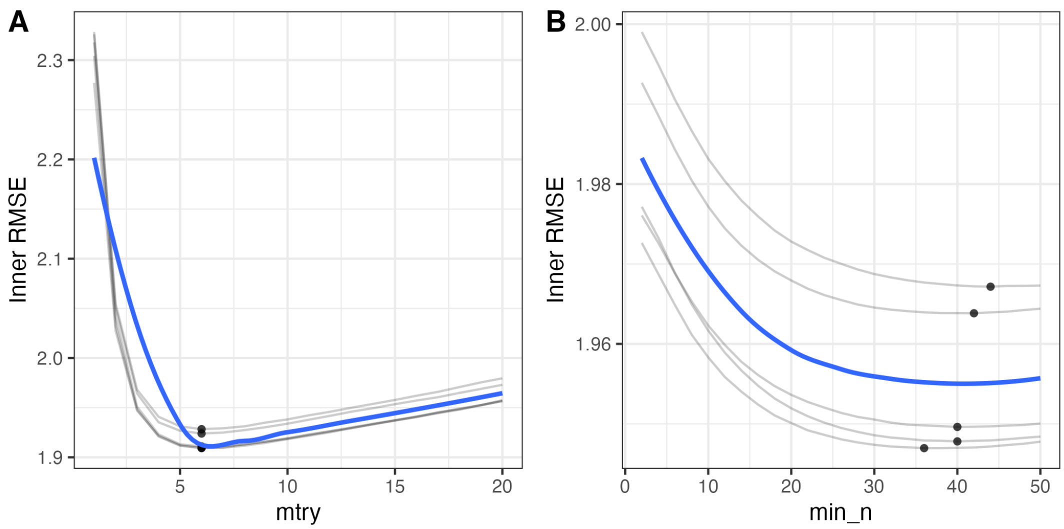


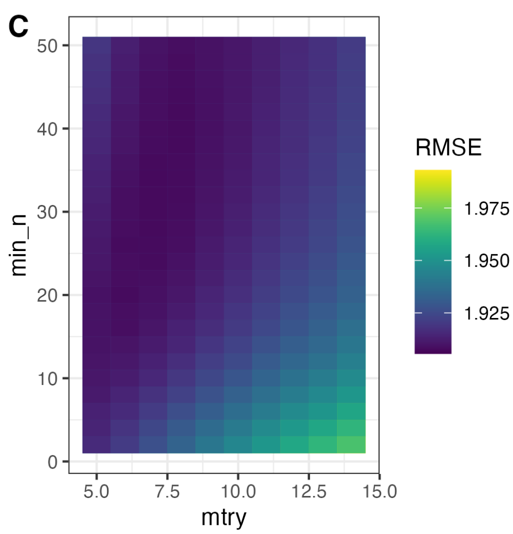


D

mtry	min_n	Inner RMSE (mean)	Outer RMSE
7	38	1.828	1.668
7	34	1.764	1.963
7	40	1.772	1.878
6	30	1.811	1.740
8	42	1.821	1.734

Fig. S8. Results of grid search of hyperparameters for random forest structural poverty estimation for **Leave-one-country-out (Tanzania)**.





D

mtry	min_n	Inner RMSE (mean)	Outer RMSE
6	18	1.906	1.969
7	34	1.906	1.942
8	42	1.925	1.887
6	24	1.921	1.872
7	34	1.908	1.936

Fig. S9. Results of grid search of hyperparameters for random forest structural poverty estimation for **Leave-one-country-out (Uganda)**.

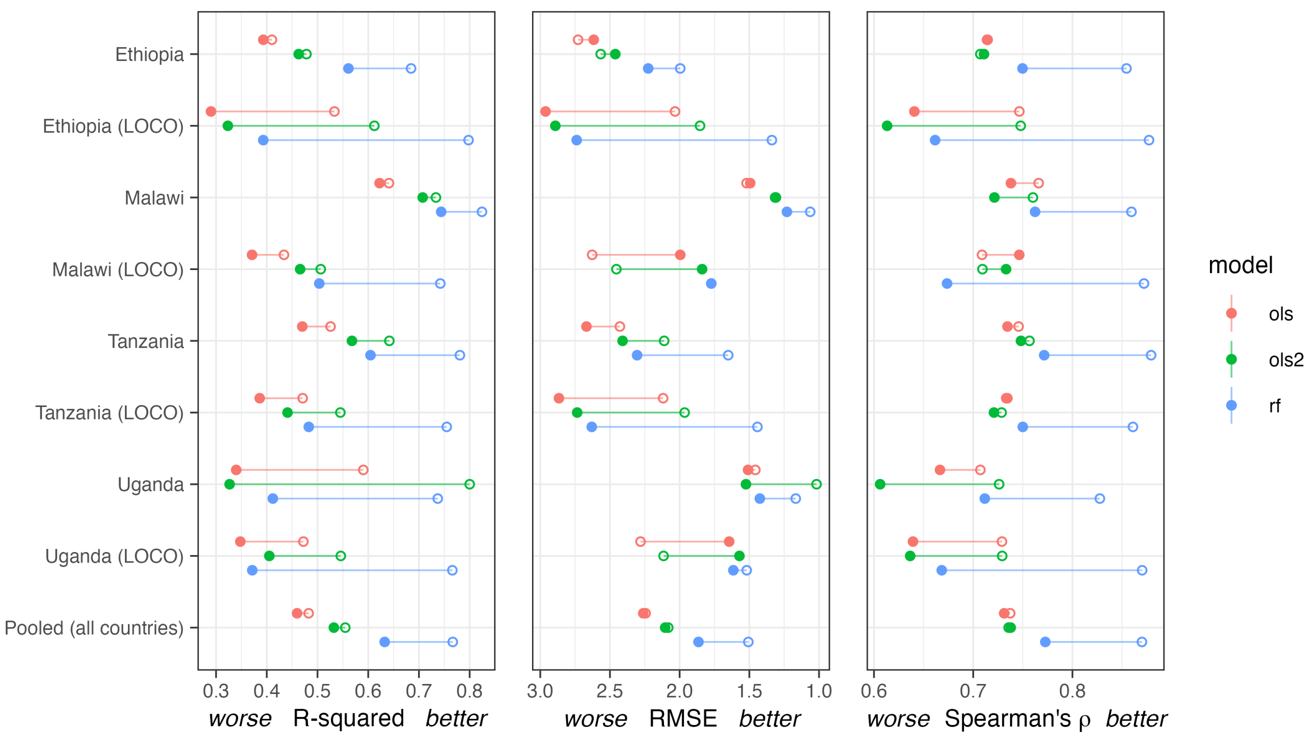


Fig. S10. Comparison of measures of fit for continuous models of structural poverty. The solid circle indicates the fit statistic in the test data, the open circle in the training set, and the line is the difference between these. Wider lines therefore indicate larger differences in fit between the training and test data. Country models are based on training and test data from that country. Leave-one-country-out (LOCO) models are trained on the pooled dataset excluding that country (which serves as the test set).

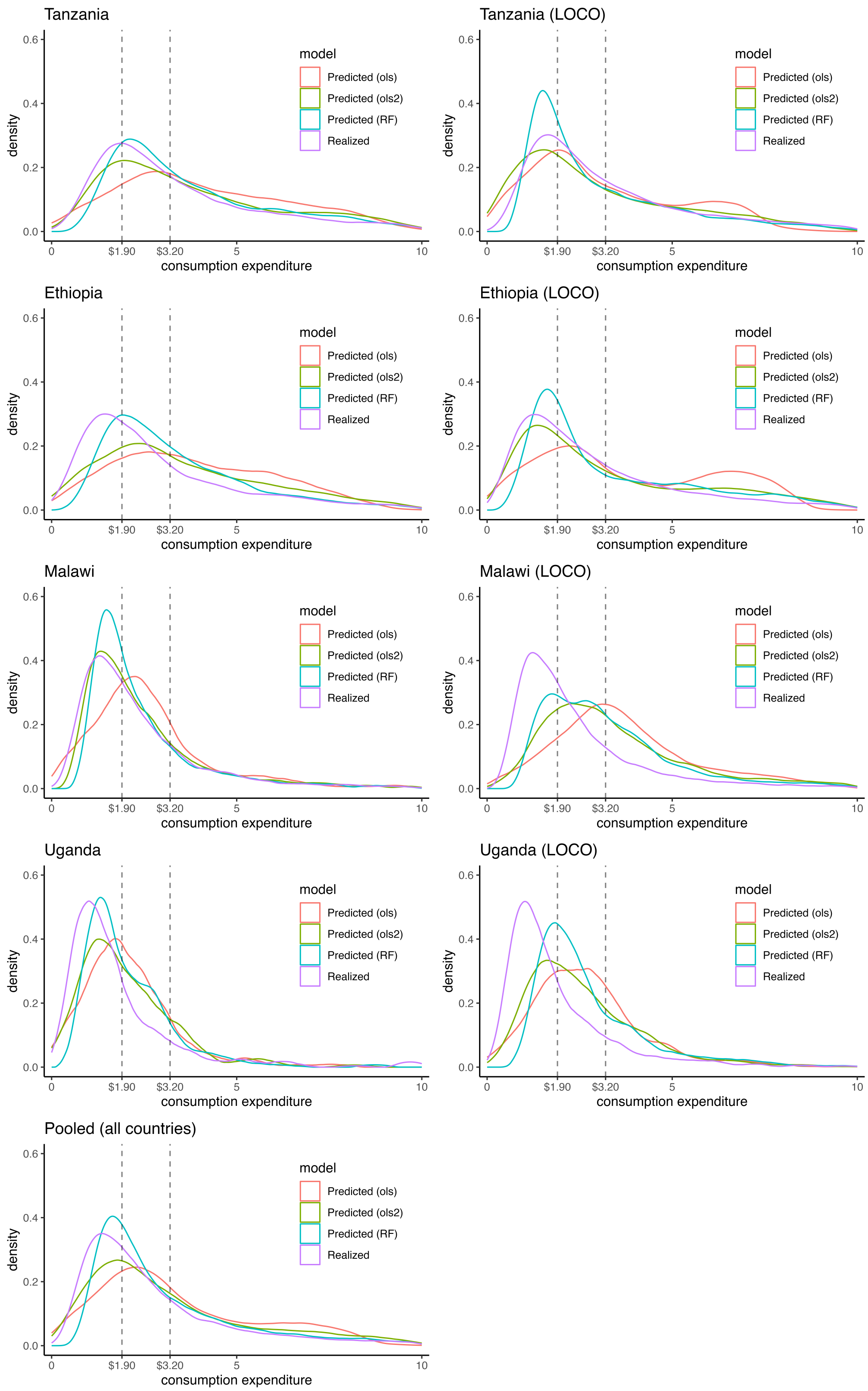


Fig. S11. Kernel density plots of predicted versus realized consumption expenditures.

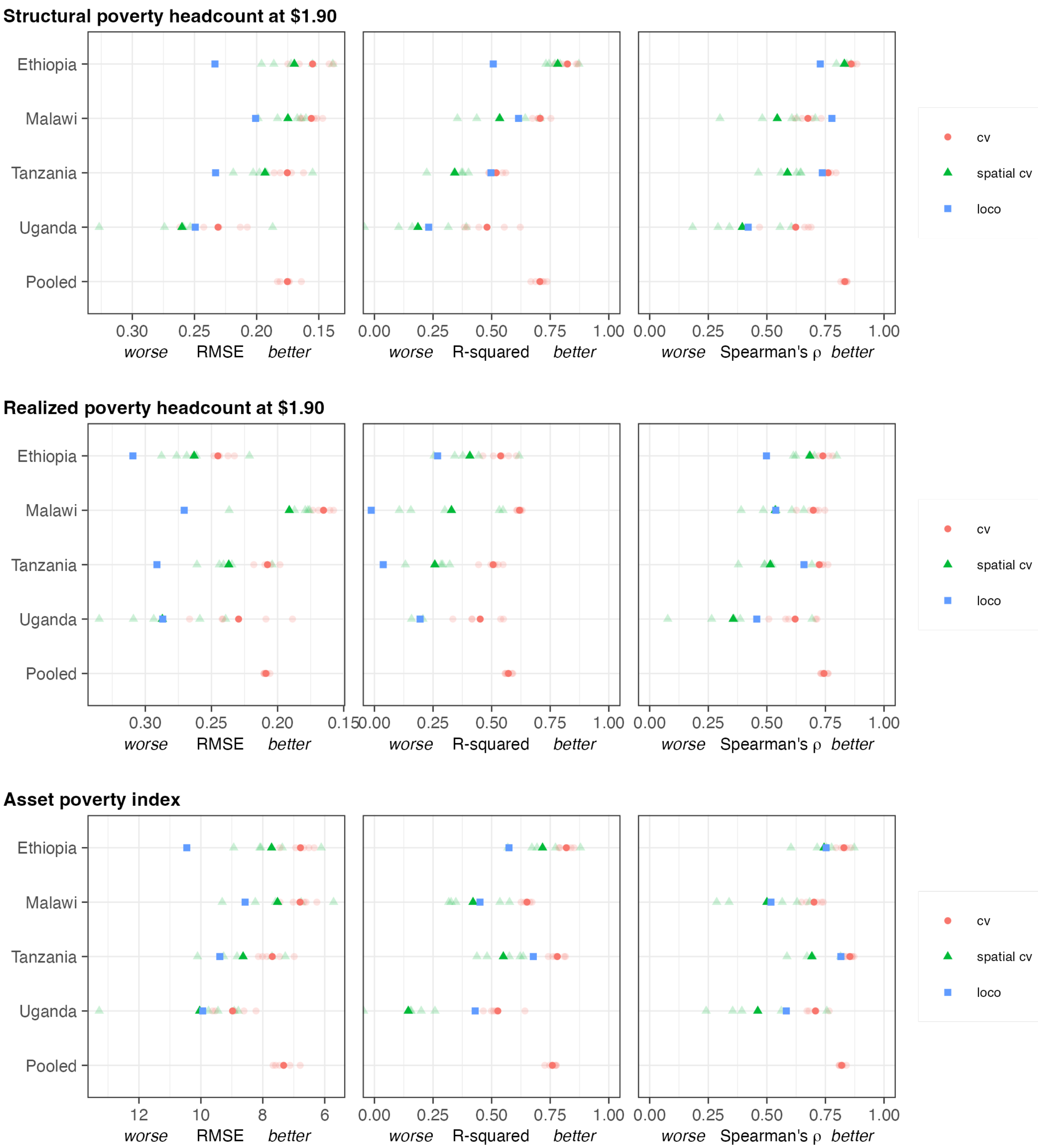
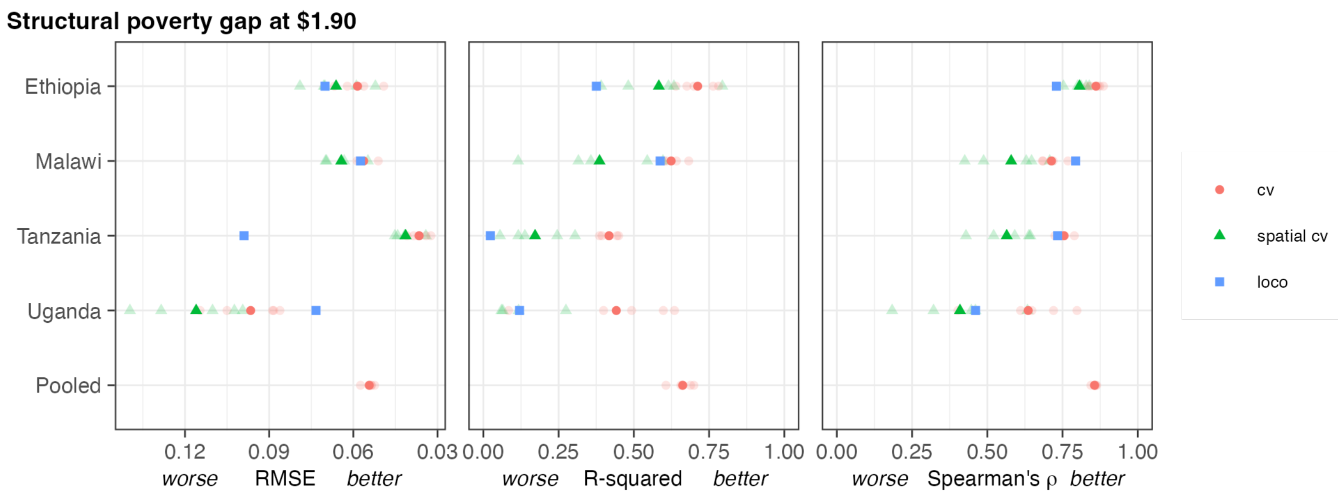


Fig. S12. Performance of EO-Structucal Poverty models in test set, for the: Poverty Headcount (P^0) at a poverty line of $z = \$1.90$. For cross-validated models, the bold symbol indicates mean performance of the shown folds. Country models are based on training and test data from that country. Leave-one-country-out (LOCO) models are trained on the pooled dataset excluding that country, which serves as the test set.



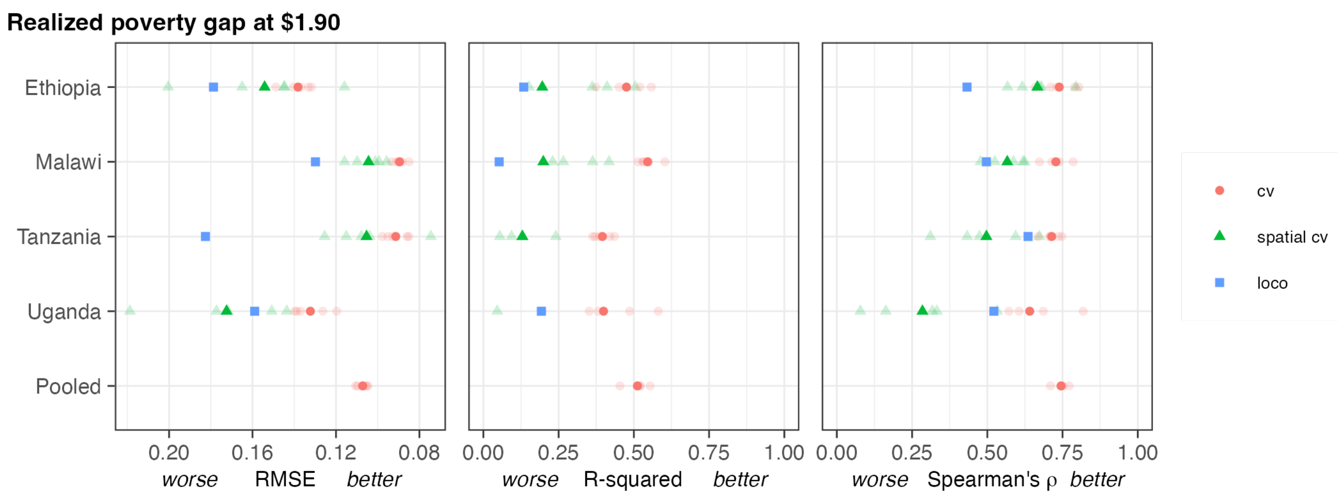


Fig. S13. Performance of EO-Structucal Poverty models in test set, for the: Poverty Gap (P^1) at a poverty line of $z = \$1.90$. For cross-validated models, the bold symbol indicates mean performance of the shown folds. Country models are based on training and test data from that country. Leave-one-country-out (LOCO) models are trained on the pooled dataset excluding that country, which serves as the test set.

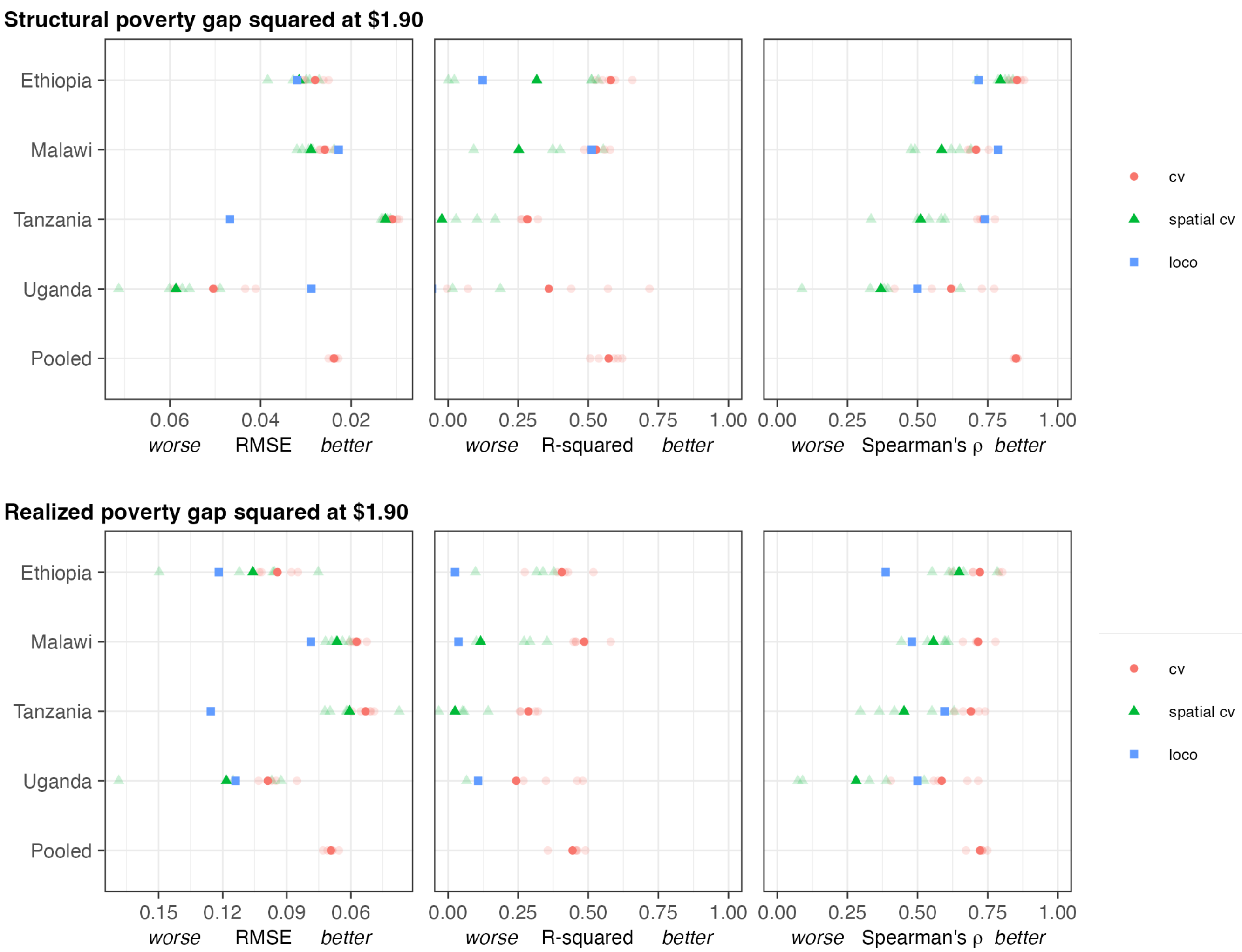


Fig. S14. Performance of EO-Structucal Poverty models in test set, for the: Poverty Gap Squared (P^2) at a poverty line of $z = \$1.90$. For cross-validated models, the bold symbol indicates mean performance of the shown folds. Country models are based on training and test data from that country. Leave-one-country-out (LOCO) models are trained on the pooled dataset excluding that country, which serves as the test set.

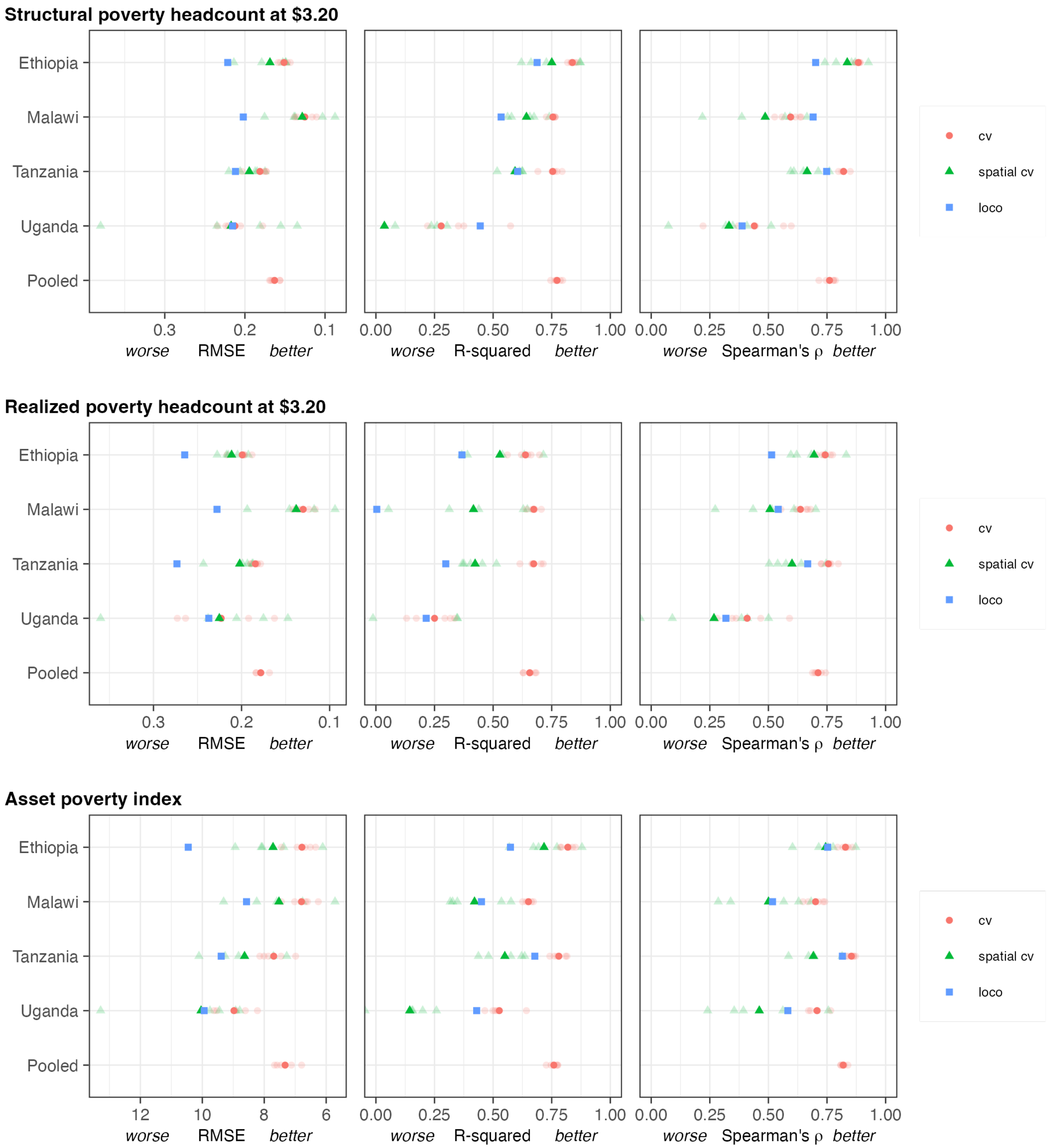


Fig. S15. Performance of EO-Structucal Poverty models in test set, for the: Poverty Headcount (P^0) at a poverty line of $z = \$3.20$. For cross-validated models, the bold symbol indicates mean performance of the shown folds. Country models are based on training and test data from that country. Leave-one-country-out (LOCO) models are trained on the pooled dataset excluding that country, which serves as the test set.

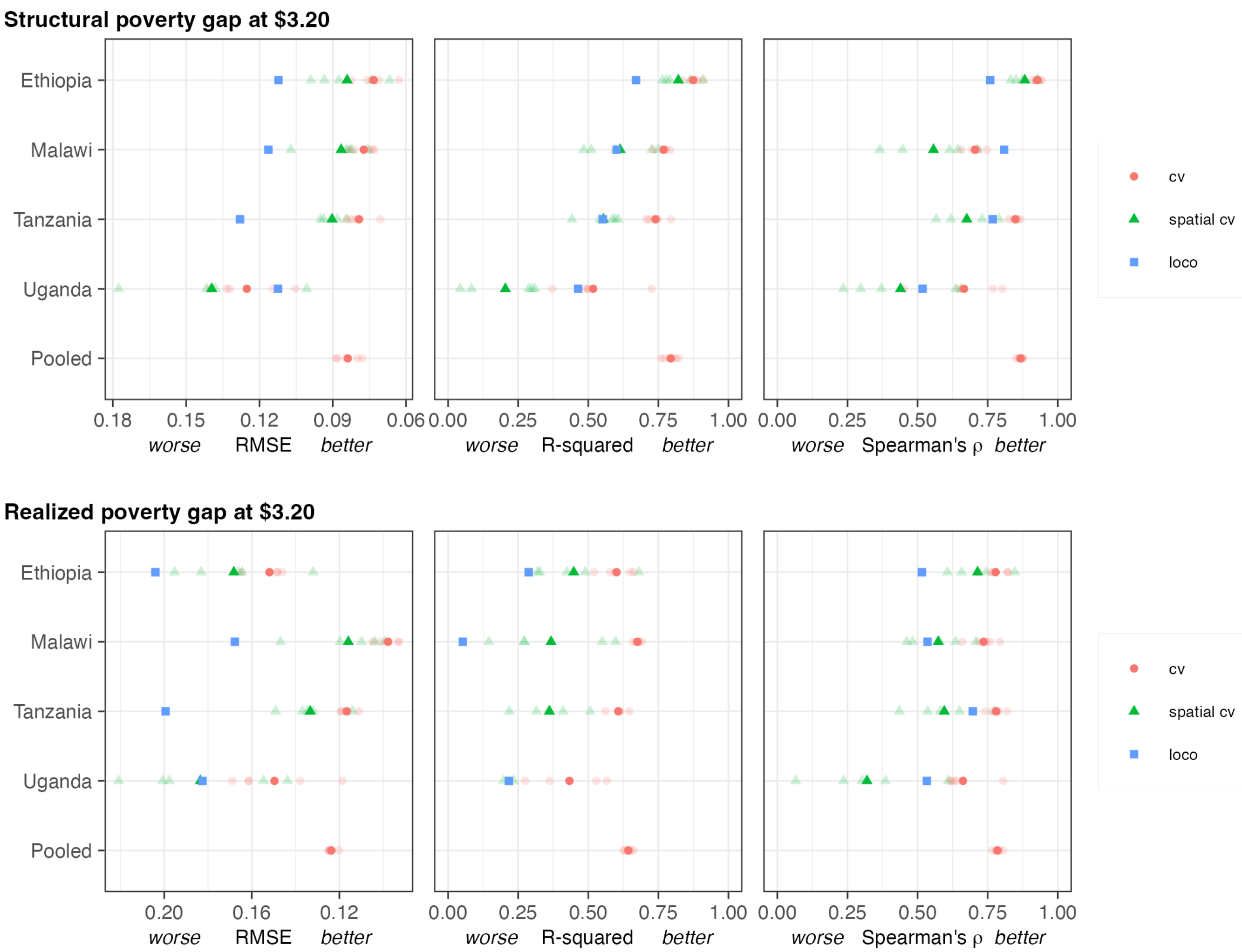


Fig. S16. Performance of EO-Structucal Poverty models in test set, for the: Poverty Gap (P^1) at a poverty line of $z = \$3.20$. For cross-validated models, the bold symbol indicates mean performance of the shown folds. Country models are based on training and test data from that country. Leave-one-country-out (LOCO) models are trained on the pooled dataset excluding that country, which serves as the test set.

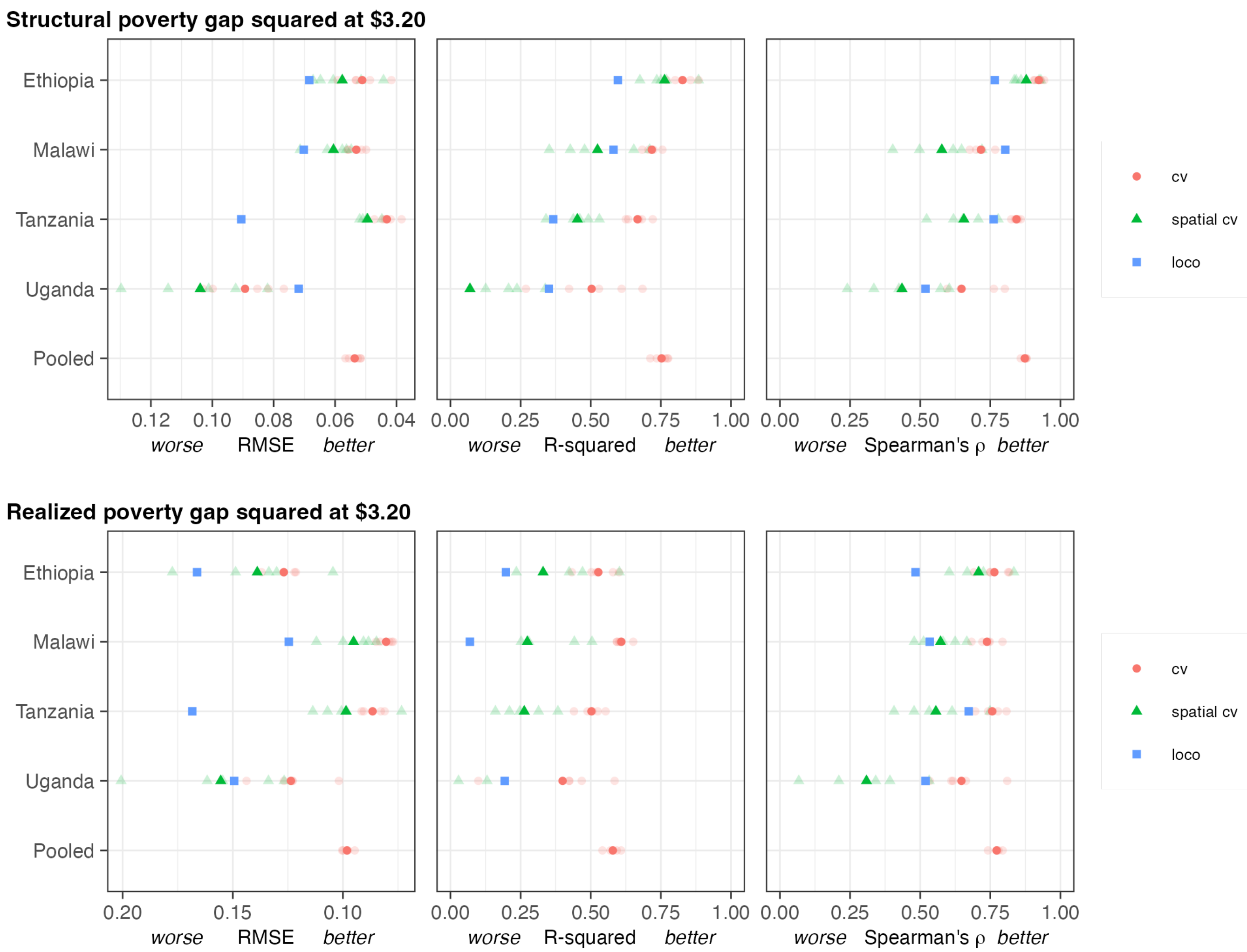


Fig. S17. Performance of EO-Structucal Poverty models in test set, for the: Poverty Gap Squared (P^2) at a poverty line of $z = \$3.20$. For cross-validated models, the bold symbol indicates mean performance of the shown folds. Country models are based on training and test data from that country. Leave-one-country-out (LOCO) models are trained on the pooled dataset excluding that country, which serves as the test set.

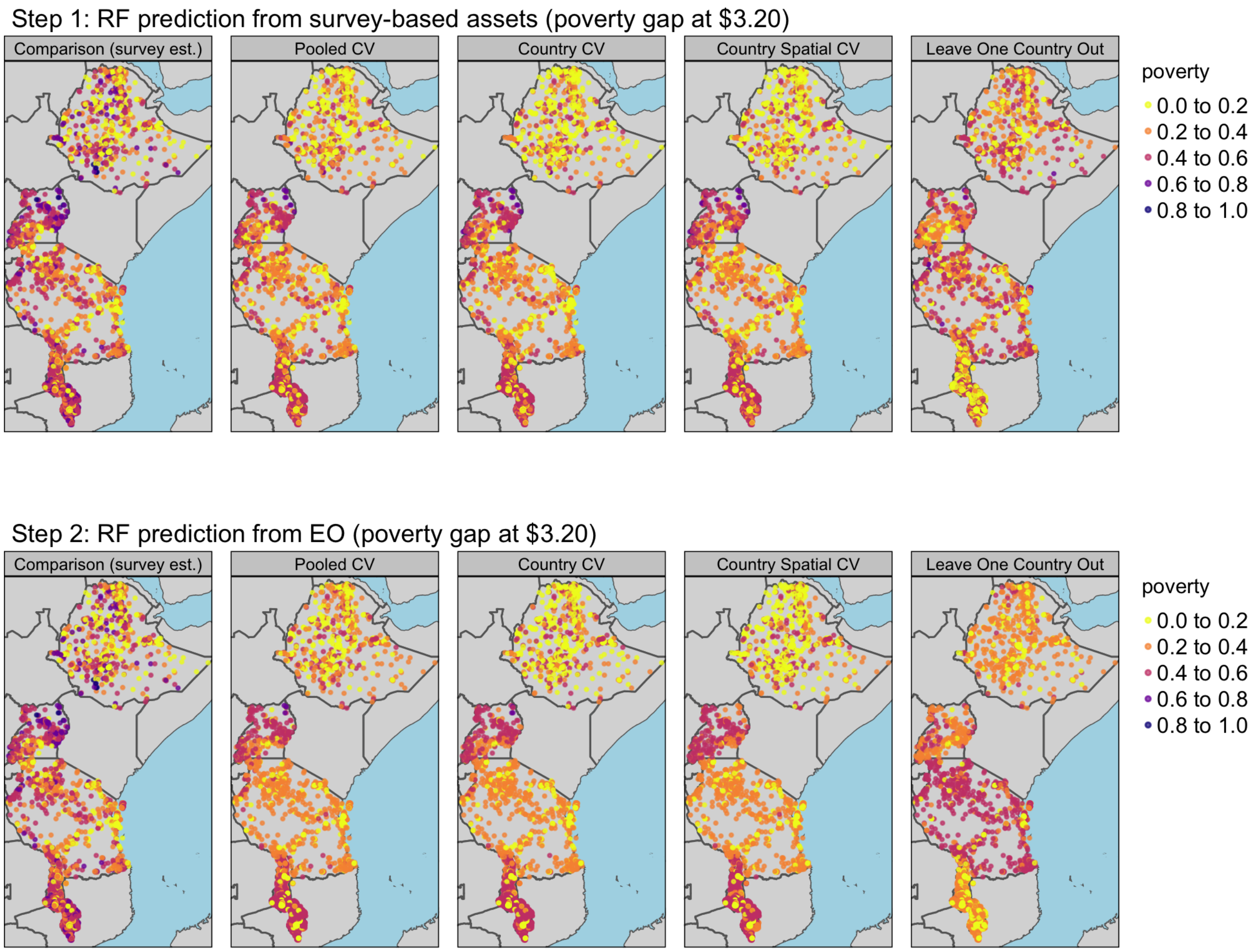


Fig. S18. Maps of Poverty Headcount (P_0) at the $\$3.20$ poverty line. For comparison, the leftmost panel for each row are EA poverty rates estimated directly from realized consumption in the survey data. The remaining panels on the top row are predictions from the asset-consumption models into the test sets (combining model results from cross-validation). The corresponding maps in the bottom row are predicted from EO data trained on the structural poverty estimates.

References

1. P Rothman-Ostrow, W Gilbert, J Rushton, Tropical Livestock Units: Re-evaluating a Methodology. *Front. Vet. Sci.* **7** (2020).
2. TN Croft, AMJ Marshall, CK Allen, Guide to DHS Statistics, (ICF, Rockville, MD, USA), Technical Report DHS-7 (2018).
3. SO Rutstein, K Johnson, The DHS wealth index, (United States Agency for International Development), Technical Report DHS Comparative Reports No. 6. (2004).
4. SO Rutstein, The DHS wealth index: Approaches for rural and urban areas, (United States Agency for International Development), Technical Report DHS Working Papers No. 60 (2008).
5. J Smits, R Steendijk, The International Wealth Index (IWI). *Soc. Indic. Res.* **122**, 65–85 (2015).