Policy Research Working Paper 10257

Integrating Survey and Geospatial Data to Identify the Poor and Vulnerable

Evidence from Malawi

Melany Gualavisi
David Newhouse

WORLD BANK GROUP
Poverty and Equity Global Practice
December 2022

Policy Research Working Paper 10257

Abstract

Generating timely data to identify the poorest villages in developing countries remains a fundamental challenge for existing data systems. This paper investigates the accuracy of four alternative methods for predicting a measure of village economic welfare for approximately 4,500 villages in 10 poor Malawian districts: (1) proxy means test scores calculated from the 2017 social registry, (2) the Meta Relative Wealth Index, (3) predictions derived from a standard household survey and publicly available geospatial indicators, and (4) predictions derived from a two-step approach that first predicts welfare into a hypothetical partial registry of approximately 450 villages, and then predicts welfare into the remaining villages using geospatial indicators. Geospatial indicators include land coverage indicators, weather data, night light data, building patterns, distance to major roads, and population density. Predictions are evaluated against a benchmark village welfare measure, constructed by imputing log per capita consumption from the 2016 integrated household survey into the 2018 household census using gradient boosting. Incorporating the hypothetical partial registry vastly improves the performance of the predictions. When using the partial registry, the rank correlation between the predicted and benchmark welfare measures is 0.75, while those for the other three methods range from –0.02 to 0.2, and similar results are seen when examining the area under the curve. Doubling the size of the partial registry does little to improve predictive performance. The results are robust to using a linear post–Least Absolute Selection and Shrinkage Operator model instead of gradient boosting for prediction. However, predictions using both methods are less accurate when the benchmark welfare measure is derived from a linear post–Least Absolute Selection and Shrinkage Operator model. Overall, the results strongly suggest that collecting partial registries of household-level poverty predictors in low-income contexts can vastly improve the performance of machine learning models that combine survey and satellite imagery for the purpose of village-level targeting.

This paper is a product of the Poverty and Equity Global Practice and the Social Protection and Jobs Global Practice. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://www.worldbank.org/prwp. The authors may be contacted at dnewhouse@worldbank.org.

Integrating Survey and Geospatial Data to Identify the Poor and Vulnerable: Evidence from Malawi[*]

Melany Gualavisi[§]
David Newhouse[†]

Keywords: *poverty, geographic targeting, small area estimation, poverty mapping, satellite data, machine learning*

JEL codes: C51, I32, I38

[*] We are indebted to Lina Cardona, German Caruso, Chipo Msowoya, and Boban Paul for their support in providing data, financing, and support for this project. We thank Ifeanyi Edochie for assistance with obtaining geospatial indicators. We are grateful for Richard Akresh, Sarah Jansen, Nobuo Yoshida, and seminar participants at the World Bank and the University of Illinois for constructive comments.
[†] Department of Economics, University of Illinois Urbana-Champaign
[§] Development Economics Data Group, World Bank, and IZA

1. Introduction

Identifying the poor in developing countries is crucial to inform development policies and programs, particularly those related to social assistance. However, governments and the development community are severely constrained by the high cost of collecting household surveys, censuses, or social registries that are typically used to inform targeting decisions. For example, between 2002 and 2011, 57 countries had conducted zero or one nationally representative household budget survey, preventing them from producing timely poverty estimates, and typically four years pass between nationally representative surveys on consumption or asset wealth in most African countries.[1] Even when household data is collected, data collected by household surveys are typically too small to provide reliable estimates of welfare for small geographic areas such as villages. Estimating poverty at the small area level requires alternative sources of data, traditionally census data, and utilizing this type of auxiliary data for small area estimation can provide resources to the poor more efficiently.[2] Due to this lack of timely and adequate information on measures of well-being indicators in small areas, satellite imagery and other types of non-traditional data have great potential to fill these data gaps and complement traditional household surveys to provide more timely and accurate estimates for local areas.[3]

This paper investigates the benefits of combining traditional data with publicly available remote sensing indicators to predict welfare across approximately 4,500 villages in 10 districts in Malawi. The 10 districts correspond to the ones selected for phase one of the Unified Beneficiary Registry (UBR), which was conducted in 2017, collecting information on living standards to determine the eligibility of households for social programs.

We evaluate four alternative village targeting methods which are compared to a benchmark welfare measure derived from an extract of the 2018 household census: (1) PMT scores calculated in the 2017 UBR administrative data, (2) the Meta Relative Wealth Index,[4] (3) predictions derived from a village-level model estimated using a 2016 household survey and publicly available geospatial indicators, and (4) a two-step procedure that utilizes a hypothetical partial registry of 450 randomly selected villages – 10 percent of the villages in the population -- in addition to the 2016 household survey and publicly available geospatial indicators. This hypothetical partial registry would collect selected proxy welfare indicators such as asset and demographic information from *all* households in the selected villages. The first step entails predicting per capita consumption from the household survey into the partial registry data, which we simulate by sampling from the census extract. The second step uses the partial registry predictions to train a model using publicly available geospatial data to generate estimates for the remaining 90 percent of non-registry villages.

[1] Serajuddin et al. 2015, Yeh et al., 2020.
[2] Van Der Weide et al. (2022).
[3] Burke, 2021, World Bank, 2021.
[4] The Meta relative wealth index is described in Chi et al (2021).

For the purposes of evaluation, a measure of benchmark welfare was constructed by predicting household per capita consumption as a function of household and demographic characteristics in the 2018 census extract using extreme gradient boosting, a popular machine learning algorithm. This model was then used to generate a prediction of per capita consumption for each household in the census extract, which we aggregate to construct a "ground truth" measure of village welfare. We refer to this model as the "census model", because it is used to impute predicted per capita consumption into the census extract. This census model captures over half of the variation in log per capita consumption in the household survey, with an R^2 of 0.537. It is used both to generate "ground truth" using the full census extract and to impute welfare into the simulated partial registry, derived from a subsample of the census extract. To construct a "ground truth" measure of village welfare, the values of predicted household per capita consumption in the 2018 census extract are aggregated to the village level.

For the main set of results, the "ground truth" village welfare measure used is the average predicted per capita consumption of the poorest 50 percent of households in the census, when ranked according to their predicted per capita consumption. We select this as the main measure to generate a clean comparison with the UBR, which only collected information on the poorest 50% of households. Since the PMT scores in the UBR are based on the average predicted welfare of the bottom half of the distribution, evaluating it against a different benchmark could introduce an additional source of noise in the UBR comparison. However, since this is a non-standard measure of village welfare, we also show below that the key results are robust to using a more typical measure, namely the average predicted per capita expenditure across all households in each village.

The 20 percent census extract contains village, traditional authority (TA) and district names but not household or village geocoordinates. Obtaining information on the physical location of census villages is crucial for this exercise. Therefore, we match census village names with UBR administrative data, which contains the names of the administrative areas as well as the geocoordinates of interviewed households. This enabled us to calculate centroids based on the minimum and maximum latitude and longitude of households living in that village in the UBR administrative data as an approximation of the village centroid in the census for about 4,500 villages. These centroids were then matched to a set of grids constructed to cover the country, in order to link remote sensing information to the census.

The remote sensing indicators were obtained from Google Earth Engine, WorldPop, and Meta. For each village, we calculated the average of approximately 40 indicators: landcover indicators (e.g., percentage of vegetation, water, or build-up coverage), global precipitation measurement, soil moisture, nighttime data, and year of the transition from pervious to impervious areas. This was supplemented with gridded maps of building patterns (e.g., number, area, and length of buildings, among others) in 2017, population density indicators, build settlement growth, and distance to major roads taken from Worldpop. Grid-level averages of these satellite-derived features were then linked to the census data using the village centroids obtained from the matched UBR data.

This paper considers three main research questions: (1) How much does a hypothetical partial registry improve predictions of village-level welfare in this context, as compared with the existing UBR and two other feasible alternatives? (2) How much does increasing the size of the partial

registry improve the accuracy of prediction models using geospatial indicators? (3) How do predictions generated using extreme gradient boosting, which are more robust to outliers, compare to those generated using post-LASSO linear models?

The main result is that introducing the simulated partial registry yields predictions of village welfare that are vastly more accurate than the other methods considered. When using a partial registry, the rank correlation, Area Under the Curve (AUC) coefficients, and R^2 are 0.75, 0.89, and 0.57 respectively. In contrast, the rank correlation for the other three methods ranges from -0.02 to 0.2, the AUC scores range from 0.5 to 0.6, and the R^2 of the predictions range from 0.001 to 0.04. These are huge differences in predictive accuracy.

These results are based on a simulated partial registry of approximately 450 villages, about 10 percent of the total number of villages with available data. However, the accuracy of the predictions does not substantially improve when the size of the simulated partial registry increases to 675 or 900 villages, which is 15% or 20% of the census extract. In other words, a partial registry of 450 villages is sufficient to train a high-performing predictive model in this context.

Finally, this exercise provides a useful opportunity to compare the predictive performance of extreme gradient boosting against a post-LASSO model, which has also been used in the literature but imposes a linear functional form on the model. We find that using post-LASSO for the geospatial model – the second step of the partial registry approach -- makes little difference to the accuracy of the predictions. However, using LASSO rather than gradient boosting for the census model used to construct the benchmark village welfare measure greatly reduces the explanatory power of the geospatial model, whether it is estimated using gradient boosting or LASSO. This is because extreme gradient boosting is a tree-based method that is more robust to outliers, and generates a benchmark measure of village welfare that is far easier to predict using geospatial indicators. This suggests that the census data in this case may be susceptible to outliers that introduce noise when using linear prediction models. Overall, the results demonstrate that investing in richer and context-specific training data, such as partial registries, can greatly improve the accuracy of predictions based on geospatial data.

This paper contributes to a growing literature on using satellite imagery to predict welfare. Initially, the most common remote sensing indicator used for this kind of analysis was nighttime lights, which measure the intensity of light in specific areas. Previous studies show strong correlations between night-time lights and GDP.[5] However, the association between night-time lights and other measures of household welfare is weak in most contexts, suggesting the limitation of this indicator for predictions of welfare in small areas.[6] More recent literature has demonstrated that indicators derived from daytime imagery is better suited for predicting welfare.[7] Recent literature has also demonstrated that advances in machine learning and available non-traditional

[5] Henderson et al., 2009, Pinkovskiy and Sala-I-Martin, 2016.
[6] Mellander et al., 2013.
[7] Jean, et al., 2016; Babenko et al., 2017; Engstrom et al., 2015; Engstrom et al., 2017; Head, A., 2017; Yeh, et al.,2020;, Chi et al (2021); Masaki et al., 2022.

data can improve the targeting of social programs.[8] Less attention, however, has been paid to how predictive performance relates to the nature of the training data.[9]

There are a few caveats to consider. The first is that we focus on one context: 10 districts in Malawi. While there can be some common features to other developing countries, additional research would be useful to confirm that the results apply in other contexts. Second, with respect to data availability, imputing per capita consumption into the 20 percent census extract was the best available measure of village-level welfare we could obtain. Nonetheless, because the welfare measure is imputed on the basis of slow-changing characteristics such as household size, head's education, child dependency ratio, and household assets, the benchmark welfare measure is a longer-term measure of welfare that will only partially include transitory welfare shocks. A third caveat is that the sample of villages is constructed by matching village names by hand, which raises the possibility that the sample of villages used in the study is not fully representative. However, as demonstrated in table 3, most of key observed characteristics in the census are similar on average between matched and unmatched villages.

A final methodological issue is that, in villages included in the simulated partial registry, we utilize the same census model and data to construct both the benchmark measure and the partial registry predictions. As a result, the partial registry by construction perfectly predicts benchmark welfare in the 450 villages randomly selected for the registry. To address this issue, we show that replacing the perfectly accurate predictions from the partial registry with the imperfect predictions generated by the geospatial model leads to only a modest fall in predictive performance. This indicates that the vast majority of the improvement from utilizing the partial registry, relative to the other three methods considered, derives from its ability to training a much richer and more accurate predictive geospatial model, rather than the increased predictive accuracy in the villages it covers.

In this context, our findings provide convincing evidence that partial registries that collect a limited set of indicators for all households in a sample of villages, if collected properly, can greatly enhance the accuracy of geospatial predictions of village welfare. Many surveys routinely undertake full listing exercises in sampled enumeration areas, which could be extended to collect information on welfare proxies. Combining household surveys, partial registries, and geospatial data has to our knowledge yet to be implemented. Yet the cost would be relatively modest; a rough estimate is that the marginal cost of interviewing all households in 450 villages would be between $24,300 and $72,900.[10] Moreover, this strategy appears to offer a large improvement over existing feasible methods when targeting social assistance programs to poor villages in contexts where conventional data sources are incomplete or outdated. These partial registries could be integrated into current systems of data collection, to help existing surveys benefit more from the wealth of publicly available geospatial data.

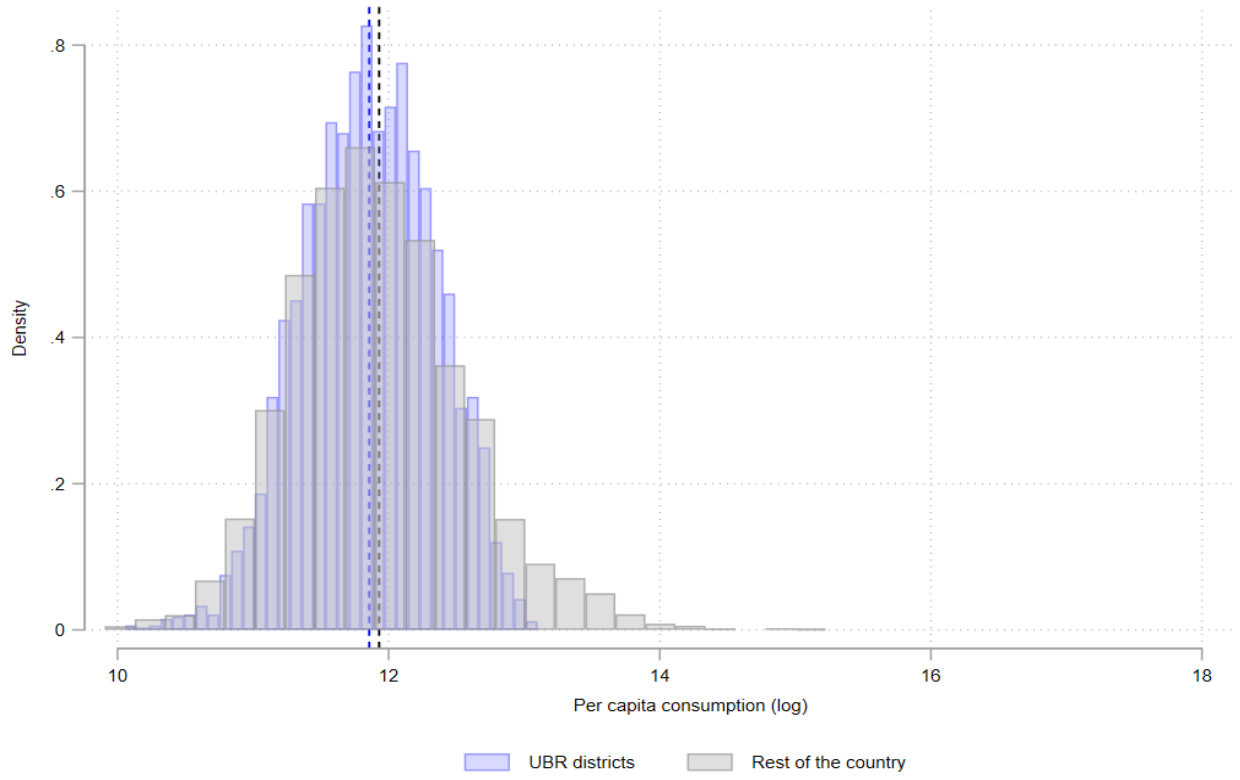
[8] Aiken et al, 2022 and Van der Weide et al. (2022).
[9] Although Engstrom et al (2022) finds that the model performance is very sensitive to the size of the sample training data.
[10] These are based on an estimated marginal cost of $3 to $9 per household to conduct face-to-face surveys in Malawi, and the average of 18 households per village in the census extract.

This paper proceeds as follows: Section 2 describes the data sets used to analyze and construct the benchmark welfare measure. Section 3 presents the statistical methodology used to generate alternative estimators of village level comparison to evaluate against the benchmark. Section 4 describes the main results. Section 5 includes some robustness checks, and finally, Section 6 consists of a discussion and main conclusions.

2. Data

The analysis utilizes data from approximately 4,500 villages in 10 districts in Malawi. The 10 districts correspond to those selected for the first phase of the UBR data collection. These districts are poorer than the rest of the country, according to data from the 2016 integrated household survey. For instance, households in the UBR districts tend to have lower educational attainment, as measured by the share of households where the highest educated male or female completed secondary or tertiary education. Also, households in UBR districts are in rural areas and have lower quality houses in terms of roof, wall, and floor materials. The UBR households are also less likely to have access to piped water or flush toilets and own fewer assets (e.g., cellphone, fridge, computer, cars, radio, television) (see Annex 1 to see the full set of statistics). Finally, as expected, UBR districts have significantly lower per capita consumption (Figure 1).

Figure 1. Distribution of the Log per capita consumption in UBR districts vs. the rest of the country



2.1 Description of the data sets

The primary sources of information are the following: (1) the Unified Beneficiary Registry (UBR), collected in 2017; (2) a 20 percent extract of the 2018 census provided by the National Statistical Office of Malawi; (3) the Integrated Household Survey (HIS) collected in 2016; and (4) publicly available remote sensing indicators.

Unified Beneficiary Registry (UBR)

Malawi's Unified Beneficiary Registry contains information on the households' socio-economic characteristics to determine their eligibility for social programs.[11] For the analysis, we use the data set collected during the first phase of the UBR. These data were collected in 2017 in 10 districts: Lilongwe, Ntchisi, Kasungu, Rumphi, Chiradzulu, Nkhota-Kota, Blantyre, Karonga, Ntcheu, and Dowa. During this phase, half of the households in these districts were registered based on Malawi's average poverty rate. The UBR data set contains apporximately 595,000 households, spread across 14,986 villages in the 10 districts.

The UBR is a crucial data set for this analysis for two reasons. First, it contains the PMT scores for the poorest 50 percent of households, which allows us to evaluate the PMT scores as a targeting mechanism. Secondly, these data include the geocoordinates of sample households, which allows us to merge the satellite data with the census data.

Census data

The analysis uses a 20 percent extract of the 2018 census data for 10 districts provided by the National Statistics Office of Malawi. The study utilizes data from the 4,500 villages that were matched, by name, with the UBR data. The census extract includes 235,600 households in 26,150 villages in the ten UBR districts. The census extract also serves two main purposes. First, it is used, along with the parameters of the census model estimated in the household survey, to generate the benchmark welfare used for evaluation. Second, it provides a randomly selected sample of villages that is used to simulate a partial social registry, as explained in the methodology section below.

Survey data

The analysis uses the fourth Integrated Household Survey (IHS) of 2016, which is made publicly available through the World Bank's Living Standard Measurement Survey (LSMS) program. The survey includes a cross-sectional sample of 12,447 households surveyed in 779 enumeration areas (EAs). Thus, there are roughly 16 sample households per enumeration area. It is considered to be representative at the district level. Because it is an LSMS survey, jittered enumeration area coordinates are also publicly available.[12] The IHS is used in the analysis for two primary purposes. The first is to estimate a model that predicts per capita consumption as a function of household variables common to the survey and the census extract, as a basis for constructing the benchmark measure of welfare in the census. Besides serving as a benchmark for evaluation, this predicted welfare measure is also used to simulate a partial registry in a subsample of villages. The second

[11] See Lindert et al., 2018, for more information on the UBR.

[12] Van der Weide et al (2022) finds that the jittering reduces the correlation between census and geospatial-based estimates for traditional authorities in Malawi by a modest amount.

main purpose of the IHS is to train a model that predicts welfare based on publicly available satellite data, which is one of the candidate prediction methods that is evaluated.

Satellite data

We obtained the satellite data from three different sources: Google Earth Engine, WorldPop, and Facebook. Table 1 contains a summary of the indicators used for the analysis.[13] For indicators for which annual data are available, we collected information for 2017 and 2018 that correspond to the years of the UBR and census, respectively.

This data is collected at the grid level, and then is merged with the village centroids obtained from the UBR.

Table 1. Satellite indicators used in the analysis

Source	Indicators
Google Earth Engine	Land cover type, weather, vegetation, nightlights, year of change to impervious surface. 7 km by 7 km resolution
Worldpop	Population density, build-settlement growth, OSM distance to roads (2016), and building patterns data (2020). Resolution is 0.1 km
Meta	Relative Wealth Index. The resolution is 2.4 km.

Note: Data is for 2017-2018 unless otherwise indicated.

2.2 Matching the census and the UBR data by village

Matching villages between the UBR and the census data is a critical step that enables the linking of census villages with remote sensing indicators. The matching is based on names using an algorithm that matches two text variables and assigns a similarity score. The matching starts at the biggest administrative unit, Traditional Authorities (TA), followed by Group Village Names (GVN), and finally at the village level. This process resulted in 32% (4,727[14]) of the UBR villages being matched with the census villages. Six UBR TAs and 24% of the GVNs are not in the census. Among the non-merged villages (10,181), 38% (3,894) are in non-matched GVNs.

The second panel of Table 2 shows the matching at the district level. The districts with the lowest percentage of matched villages are Rumphi, Nkhotakota, and Lilongwe.

Table 2. Matching results between UBR and Census.

	Merged with census	(%)	Not merged	(%)	Total in UBR	(%)
Traditional authorities	107	95%	6	5%	113	100%
Group village names	1,302	76%	401	24%	1,703	100%
Villages	4,727	32%	10,181	68%	14,908	100%
District						

[13] For more details about the specific names of the indicators, the bands collected, and years see Annex 2.
[14] This number is larger than the villages used in the analysis since we lose some of them for having missing values in some of the features used in the models. For this reason, the analysis focused on approximately 4,500 villages.

Karonga	161	54%	138	46%	299	100%
Rumphi	122	22%	423	78%	545	100%
Kasungu	485	38%	780	62%	1,265	100%
Nkhotakota	119	24%	382	76%	501	100%
Ntchisi	499	51%	473	49%	972	100%
Dowa	481	56%	377	44%	858	100%
Lilongwe Rural	1,203	29%	2,996	71%	4,199	100%
Ntcheu	404	57%	308	43%	712	100%
Chiradzulu	528	77%	161	23%	689	100%
Blantyre Rural	725	74%	249	26%	974	100%
Total*	**4,727**	**43%**	**6,287**	**57%**	**11,014**	**100%**

Notes: (*) the total in the second panel corresponds to the total number of villages in matched GVNs only.

The analysis presented below is based entirely on the final sample of matched villages. Therefore, it is important to check whether there is any systematic selection bias in the sample. Table 3 shows the comparison of census means in the matched and unmatched villages. It shows that in most of the features, the difference in means between matched and unmatched villages is very small and not statistically significant. However, we observe that matched villages have a higher share of households with more educated adults. Also, households in matched villages have a lower child dependency ratio and slightly higher elderly dependency ratio. Finally, a higher share of households in matched villages have better roof quality in their houses.

Table 3. Comparison of census means in matched and unmatched villages

	Mean	**Difference (matched-unmatched)**	**P-value (Difference)**
Highest educated man: primary education	0.27	0.01	0.16
Highest educated man: secondary education	0.11	0.01	0.05
Highest educated man: tertiary education	0.02	0.00	0.22
Highest educated woman: primary education	0.26	0.02	0.09
Highest educated woman: secondary education	0.06	0.01	0.03
Highest educated woman: tertiary education	0.01	0.00	0.30
Share of households with literate house. head	0.73	0.02	0.09
Household size	4.87	(0.16)	0.30
Overcrowding	1.91	(0.10)	0.31
Elderly dependency ratio	0.08	0.01	0.00
Children dependency ratio	0.94	(0.03)	0.09
Firewood for cooking	0.90	(0.00)	0.66
Access to pipe water	0.06	0.00	0.92
Access to flush toilet	0.01	0.00	0.24
Share of HH that own a house	0.91	0.00	0.81
Share of HH with improved walls	0.86	0.04	0.21
Share of HH with improved roof	0.38	0.07	0.04

Share of HH with improved floor	0.18	0.02	0.16
Share of HH with cellphone	0.45	0.01	0.73
Share of HH with fridge	0.02	0.00	0.27
Share of HH with stove	0.02	0.00	0.32
Share of HH with computer	0.02	0.00	0.35
Share of HH with oxcart	0.03	(0.01)	0.24
Share of HH with bicycle	0.33	(0.02)	0.28
Share of HH with motorcycle	0.04	(0.00)	0.77
Share of HH with car	0.01	(0.00)	0.74
Share of HH with radio	0.28	0.01	0.12
Share of HH with television	0.06	0.00	0.50

Notes: the difference in means are calculated using a regression of each variable against the indicator variable equal to 1 if the village was matched and zero otherwise. They are weighted by the number of households in each village and the standard errors are clustered at district level.

3. Methodology

We propose four different targeting methods using census and survey data combined with geospatial data to understand the most effective way to target the poor population: (1) the PMT scores calculated in the UBR data of 2017; (2) the Relative Wealth Index from Meta (3) combining survey data and geospatial indicators to predict average welfare; and (4) a census sample to simulate a *partial registry* data set used to train models using satellite data.

We rely on rank correlations, Area Under the Curve coefficients, and R-squared coefficients to compare the accuracy of each targeting method. In each case, we compare the predictions with a measure of benchmark welfare constructed from the census model.

3.1 Construction of Benchmark Welfare using the Census model

We define a benchmark welfare measure of "ground truth" in order to evaluate different targeting methods. As noted above, the primary welfare measure is the average predicted log per capita consumption of the poorest 50 percent of households in each village. This measure echoes the UBR structure limited to the bottom half of households in each village according to the average Malawi poverty rate. This measure takes advantage of the rich data in the census sample, and it is easier to predict than measured consumption due to reduced measurement error and its inability to capture temporary shocks.[15] To construct the benchmark welfare in the census, we use the IHS 2016 to estimate a model and then impute welfare in the census.

The variables included in the model are selected so that both data sets contain the same information. All the variables are measured at the household level. We include education variables such as the literacy of the household head, the maximum level of education achieved by men and women in the household, dependency ratios, household size and overcrowding,

[15] Due to data availability, we use the per capita consumption as a welfare measure. However, we observe a high rank correlation with other potential measures such as absolute poverty rate (correlation of -0.86) or extreme poverty (correlation -0.796).

house characteristics, and household assets. The dependent variable is household per capita consumption.

W**e** train a machine learning model to predict household per capita consumption using an extreme gradient boosting model with optimal hyperparameters chosen via 5-fold cross-validation. The parameter used in the models corresponds to the average of the selected parameters in each fold. Details on the range of parameters considered are shown in Table 4. Additional technical information regarding these parameters and other aspects of the extreme gradient boosting procedure can be found in Annex 3.

Table 4. Parameters for XGBoost models to estimate benchmark welfare.

Parameter	**Range**
Maximum number of boosting iterations	between 50 and 200
Maximum depth of a tree	2 or 4
Learning rate	0.1 or 0.3
Subsample ratio of the training instance	0.2,0.4, or 0.6
Subsample ratio of columns to construct each tree	0.2,0.5, or 0.7

We estimate four models using different samples in the IHS: (1) The full sample (All districts-all households), (2) All households in the UBR districts (UBR districts-all households), (3) the poorest 50% of the households in all districts (All districts-poorest 50%), and (4) the poorest 50% of households in UBR districts (UBR districts-poorest 50%). Estimating four models allows us to assess the trade-offs along two dimensions: (1) using all districts in the sample rather than restricting to UBR districts, and (2) using all households rather than just the lower half. Annex 4 shows the R-squared for each model and the main explanatory variables in terms of the *gain* measure.[16] Including all households (model 1) leads to higher R-squared, while limiting to the bottom half of UBR districts (model 4) leads to by far the lowest R-squared. All models assign high importance to similar variables, mostly household assets, household size, and the urban or rural location (for the specific gain measures of each variable, see Annex 5).

The model that best matches the UBR sample is model (4) since it uses only UBR districts and the poorest 50% of households; however, the sample size is small, and the R-squared is the lowest. Although Model (1) has the highest R-squared, it also has a wider range of variability in the dependent variable, making R-squared a potentially misleading metric. Models (2) and (3) have similar values of R-squared. We elect to use model (3) as the main census model because it resembles the sample structure of the UBR by using only the poorest 50% of households in each village. However, it also takes advantage of additional data on log per capita consumption by using the full set of sample survey enumeration areas nationwide. Table 5 shows the values for the R-squared and the importance of the top 15 features in the selected model. These features explain around 87% of the model being the "urban/rural" indicator the main contributor. We later present robustness analysis using the other three samples.

[16] Gain reflects the improvement in accuracy brought by a feature to the branches it is on. This means that before adding a new split on a feature X to the branch there were some wrongly classified elements, once the split on this feature is added, there are two new branches, and each of them is more accurate. A higher value of gain indicates that the feature is more important for generating a prediction.

Table 5. Benchmark welfare census model in IHS

	All districts-50% poorest HH
R-squared	53.71
Importance of variables	
Urban (1) or rural (0)	0.24
Child dependency ratio	0.09
Ownership of a cell phone	0.09
House with improved floor	0.09
Household size	0.08
Access to piped water	0.05
Ownership of a television	0.05
Household overcrowding	0.04
Fuel cooking: firewood	0.03
Household size (squared)	0.03
Ownership of a radio	0.02
Access to flush toilet	0.02
Ownership of a car	0.02
HH head literacy	0.01
Highest educated women attained primary	0.01

The selected model is used to predict the benchmark welfare variable in the census extract. This variable is generated at a household level first, then aggregated to a village welfare measure as the average of the bottom half of households in each village.

3.2 Criteria used for evaluating prediction accuracy

We use three main criteria for evaluating the accuracy of the predictions against the benchmark: the Spearman Rank Correlation, the Area Under the Curve (AUC), and the R-squared , which is equal to the share of the variation explained by the prediction. While the latter is a standard measure of prediction accuracy, the first two warrant a brief explanation.

The Spearman rank-order correlation coefficient (r_s) is a statistical measure of the strength and direction of a monotonic relationship between two variables measured on a continuous scale. The rank correlation between two variables, X and Y, is calculated as follows:

$$r_s = \rho_{R(X)R(Y)} = \frac{cov(R(X)R(Y))}{\sigma_{R(X)}\sigma_{R(Y)}}$$

Where $\rho_{R(X)R(Y)}$ denotes the Pearson correlation coefficient applied to variable ranks; $cov(R(X)R(Y))$ is the covariance between two ranked variables, and $\sigma_{R(X)}\sigma_{R(Y)}$ are the standard deviations of the ranked variables.

Finally, the Area under the Curve (AUC) is a measure of the efficacy of a targeting method in identifying the poor population at different targeting thresholds (Wodon, 1997, Olken and Hanna 2018). The curve in question is a receiver operator characteristic (ROC) curve, which indicates the trade-off between true and false positives at different poverty lines. We plot the ROC curve using each percentile of the benchmark village welfare predicted from the census. In other words, for poverty lines defined at each percentile of the benchmark distribution, we plot the true positive rates (TPR) on the Y axis and the false positive rate (FPR) on the X axis. The former is defined as the proportion of poor villages that are correctly predicted to be poor when using predicted village welfare from each candidate prediction method, while the latter is defined as the proportion of non-poor that are incorrectly predicted to be non-poor. Because the true positive rate is on the Y axis, a higher AUC score represents an improvement in true positives for a given level of false positives, or a better targeting method. The 45-degree line, which is what one would expect if villages were ranked randomly, corresponds to an AUC score of 0.5, while a perfectly accurate ranking that correctly identifies poor households under all poverty lines would receive an AUC score of 1.

3.3 Candidate Targeting Methods for Identifying Poor Villages

Proxy Mean Test scores in the UBR

The administrative data from Malawi's UBR provides information on households' characteristics to assess their prospective eligibility for social programs. The data set contains an extensive range of variables such as geographic location, households' assets, food security questions, and economic characteristics.

This information was used by the Malawian government and the World Bank to calculate Proxy-Means Test (PMT) scores to identify poor and vulnerable households. It is used to create a proxy score of weighted variables that are highly correlated with household consumption. The PMT, like our benchmark welfare measure, is a measure of chronic poverty. This is because it uses variables that are less responsive to economic shocks than household consumption, such as assets and household composition (Lindert et al., 2018). Because the PMT variable is available in the data and was used for the UBR, it is useful to evaluate it against welfare predicted into the census extract.

Relative Wealth Index

The Relative Wealth Index predicts the relative standard of living within countries using non-traditional data sources such as satellite imagery, cellular network data, topographic maps, and proprietary connectivity data from Meta. Using supervised machine learning models, the team predicts the relative wealth for grid cells of 2.4 km^2. The estimates of wealth are relatively accurate. Depending on the method used to assess the model's performance, the model explains 56 to 70 percent of the actual variation in household-level wealth in 56 low- and middle-income countries (Chi, et al, 2022). However, the model is trained on a wealth index, which may

perform less well when predicting income or consumption-based poverty measures. For example, the RWI only explains 32 percent of the variation in average per adult equivalent consumption across Cantons in Togo (Aiken et al, 2021). Furthermore, for a microcensus conducted in rural Kenya, the RWI explains 70 percent of the variation in wealth but only 17 percent of the variation in the predicted probability of being poor, defined using household consumption (Chi et al, 2022). Thus, the performance of the RWI varies greatly depending on the context, and particularly depends on whether it is evaluated against a wealth or consumption-based measure of welfare. This study therefore contributes additional information on the performance of the RWI in distinguishing among very poor villages by comparing it to the benchmark measure of village welfare in Malawi.

IHS plus geospatial indicators.

The third alternative method for targeting consists of using the IHS survey data to train a welfare model against satellite indicators. This model can then be used to generate out-of-sample predictions into villages for which matched census data is available, to compare against the benchmark. This method has the advantage of being free, but may suffer from limited training data. We estimate the model only for the poorest 50% of households in each village to resemble the structure of the UBR data set and train it using extreme gradient boosting techniques.

Partial registry

As a final alternative method to target the poor, we consider a hypothetical collection of a partial registry data set from a sample of villages. This exercise would consist of collecting the subset of household welfare proxies used in the census model from all households in a random sample of villages, similar to an expanded sample listing procedure of the type typically carried out for household sample surveys. In practice, for this analysis we simulate a partial registry by drawing a random sample of villages from the census extract. This sample, consisting of all households in the sampled villages, is used as the hypothetical partial registry in the two-step procedure. The first step involves utilizing the parameters from the census model used to predict benchmark welfare (described above) to predict per capita consumption into the simulated partial registry. The predictions are based on 39 independent variables common to both the census and survey. The resulting predictions for households are then aggregated into a measure of village welfare.

The second step of the prediction process entails estimating a second model, the geospatial model. The dependent variable in this model is predicted village welfare from the simulated partial registry, and the independent variables are village geospatial indicators, also estimated using extreme gradient boosting. This "geospatial model" is used to predict out-of-sample welfare predictions for villages not included in the partial registry. The predictions from the two models are combined, by using the census model predictions from the simulated partial registry in the villages for which they are available, and the geospatial model predictions for the remaining villages not included in the registry. In other words, in villages covered by the simulated partial registry, we use predictions from the census model rather than those from the geospatial model. This is because the simulated partial registry provides more accurate predictions than geospatial data, which are in fact exactly equivalent to the benchmark welfare measure by construction. We show below, however, that predictions from this procedure become

only modestly less accurate when using predictions from the geospatial model for all villages. We therefore conclude that, although predictions from the census model predict the benchmark exactly in partial registry villages, this is not a major factor in explaining the improved performance of the partial registry predictions. Instead, the partial registry provides richer data with which to train a more accurate geospatial model.

4. Main Results

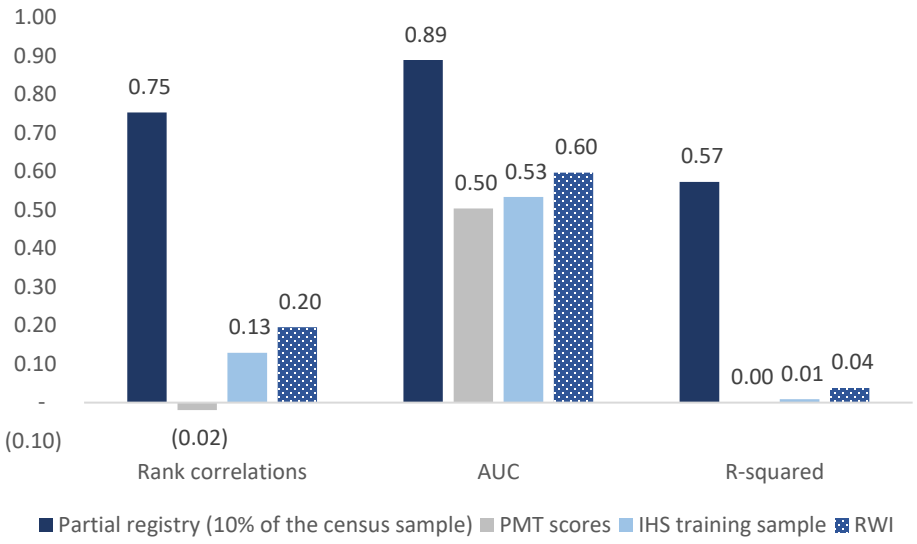
This section shows the results of the four alternative targeting methods. We compare welfare predictions generated using each method to the benchmark predictions using rank correlations, AUC coefficients, and R-squared coefficients.

Table 6 presents the metrics for each method. The partial registry is clearly the most accurate method for targeting the poor villages in the ten UBR districts in Malawi. The second best is using the RWI to rank villages. However, this method performs only moderately well, with an AUC coefficient close to 0.60 and a 0.2 rank correlation. On the other hand, using the PMT scores as a proxy for welfare does not show promising results; the PMT scores show zero correlation with our benchmark welfare and has an AUC coefficient equivalent to guessing poor villages at random. The IHS plus geospatial variable model shows the second-lowest coefficients, only slightly higher than the PMT scores in terms of AUCs. This may be due to the limited sample size. Because the sample size is restricted to the bottom half of households within UBR districts, there are only an average of about 8 households per village available to train the model. One indication of this is that performance improves noticeably when predicting average welfare across all households in the village, as noted below.

Table 6. Rank correlations, AUC, and R^2 of the targeting methods

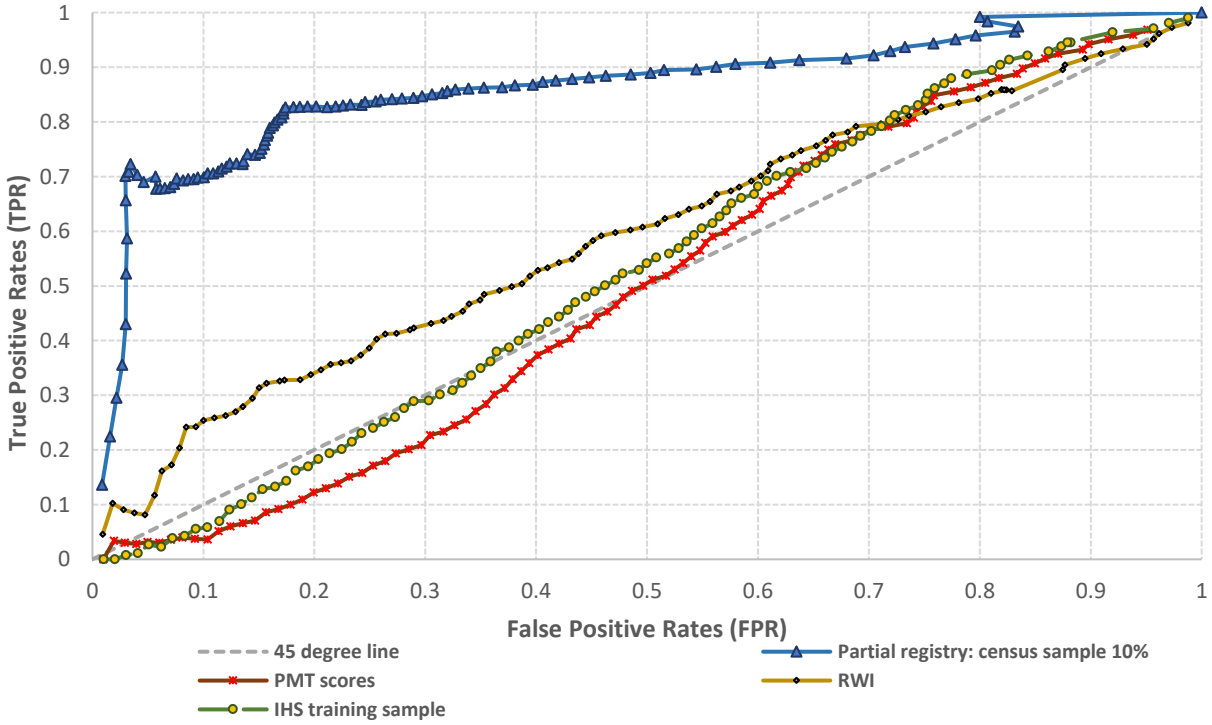
	Rank correlations	**AUC**	**R-squared**
Partial registry (10% of the census sample)	0.75	0.89	0.57
PMT scores	(0.02)	0.50	0.00
IHS training sample	0.13	0.53	0.01
RWI	0.20	0.60	0.04

Figure 2. Rank correlations, AUC, and R^2 of the targeting methods



The ROC curves of the four methods are presented in Figure 3. This is the graphical representation of the AUC results described in Table 6. The partial registry method leads to far superior targeting outcomes, especially at low poverty rates. The RWI is the second-best method but is still far from the partial registry results. The curves for PMT scores and IHS are very close to each other and to the 45-degree line that corresponds to randomly selected villages.

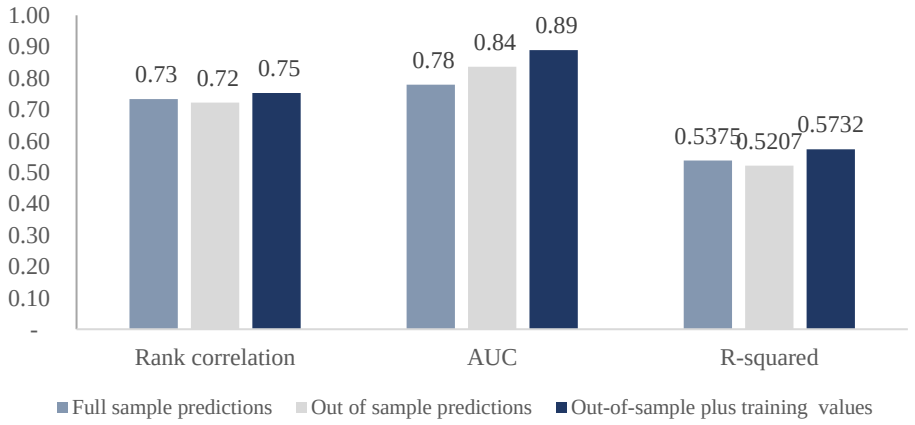
Figure 3. ROC curves of the targeting methods



As noted above, two factors might contribute to the excellent predictive performance of the partial registry method relative to the direct predictions from the household survey. The first is that the geospatial model is trained to a measure of village welfare that is far more precisely estimated than the one used in the sample. The welfare measure is more precisely estimated both because it is based on data from a much large number households from the census extract, and because it is a predicted welfare measure that largely eliminates classical measurement error. More precise training data improves the ability of machine learning to construct a predictive model, by reducing the risk that particular predictive variables will be fit to random noise in the training data, and by improving the accuracy of the cross-validation procedure used to select models.

A second potential reason for the better performance of the partial registry method is that it uses predictions derived from the census model, which is also used as the benchmark measure of welfare. While this is defensible on the grounds that it improves the accuracy of the predictions, using the same values for prediction and evaluation will overstate measured performance. To address this, as a robustness check we consider the accuracy of the geospatial model predictions for all villages, instead of using the partial registry predictions where available as well as out-of-sample predictions. In this case, the rank correlation falls from 0.75 to 0.73 and the AUC coefficient falls from 0.89 to 0.78. However, this still greatly exceeds the performance of the RWI, the second-best method, which has a rank correlation of 0.20 and an AUC of 0.60 (see Figure 4).

Figure 4. Rank correlations, AUC, and R-squared coefficient for the partial registry method.



The RWI, although the second-most accurate method for predicting the benchmark welfare, is far less accurate than the partial registry results, and only explains 4 percent of the variation in our measure of average village predicted welfare. This is probably because the RWI is based on a measure of household wealth instead of consumption or predictive consumption. Wealth may not be as accurate at distinguishing the welfare levels of villages within 10 poor districts. Moreover, the wealth index reflects the full distribution of households, whereas the benchmark welfare measure only pertains to the bottom half.

The RWI, however, performs better than efforts to integrate the household survey with publicly available geospatial data in the absence of the partial registry. This is because of the noise in the household survey data used to train the model. Given that only the bottom half of the household survey data are used to train the model, there are only roughly 8 households per EA with which to generate a measure of consumption. The resulting machine learning model is therefore not particularly accurate.

Finally, the PMT scores from the UBR 2017 are the least accurate proxies for village level welfare, as defined by the benchmark welfare measures. These are not entirely due to outliers. Figure 5 shows the presence of some outliers in the raw scores; however, even after trimming the values, the correlation is low. Some of the relatively poor performance of the UBR PMT targeting might be attributable to problems with data collection, given that it was the initial effort to collect data for the UBR.

Figure 5. PMT scores and benchmark welfare

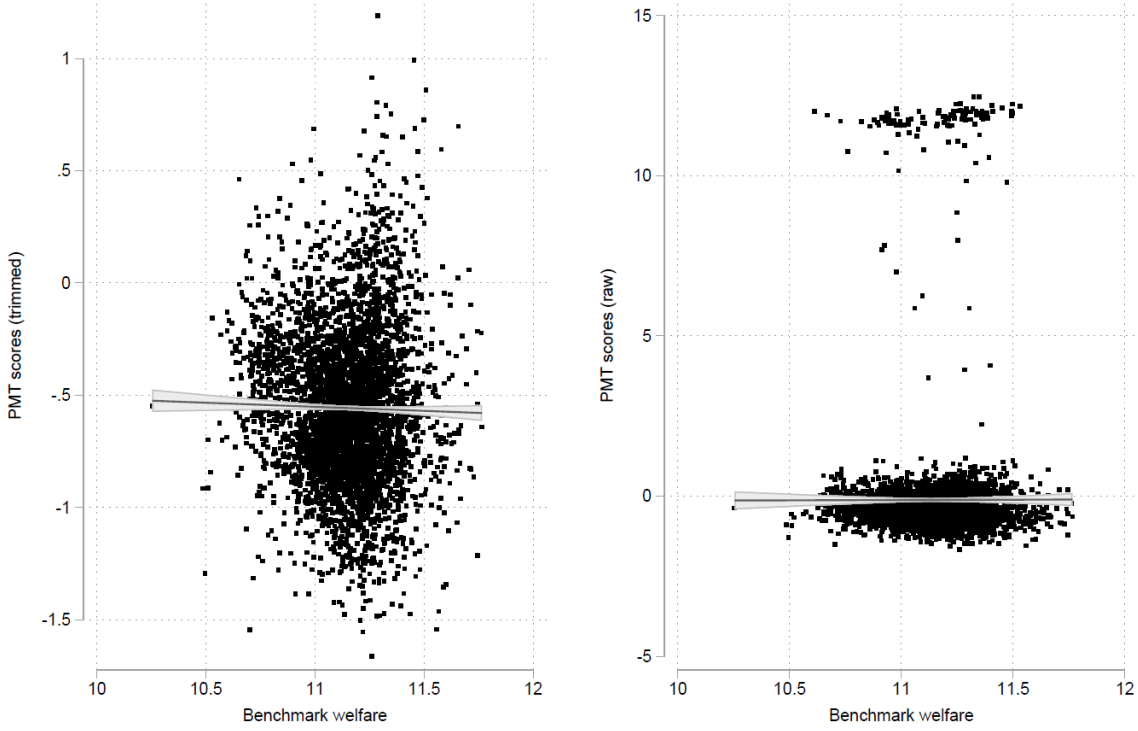
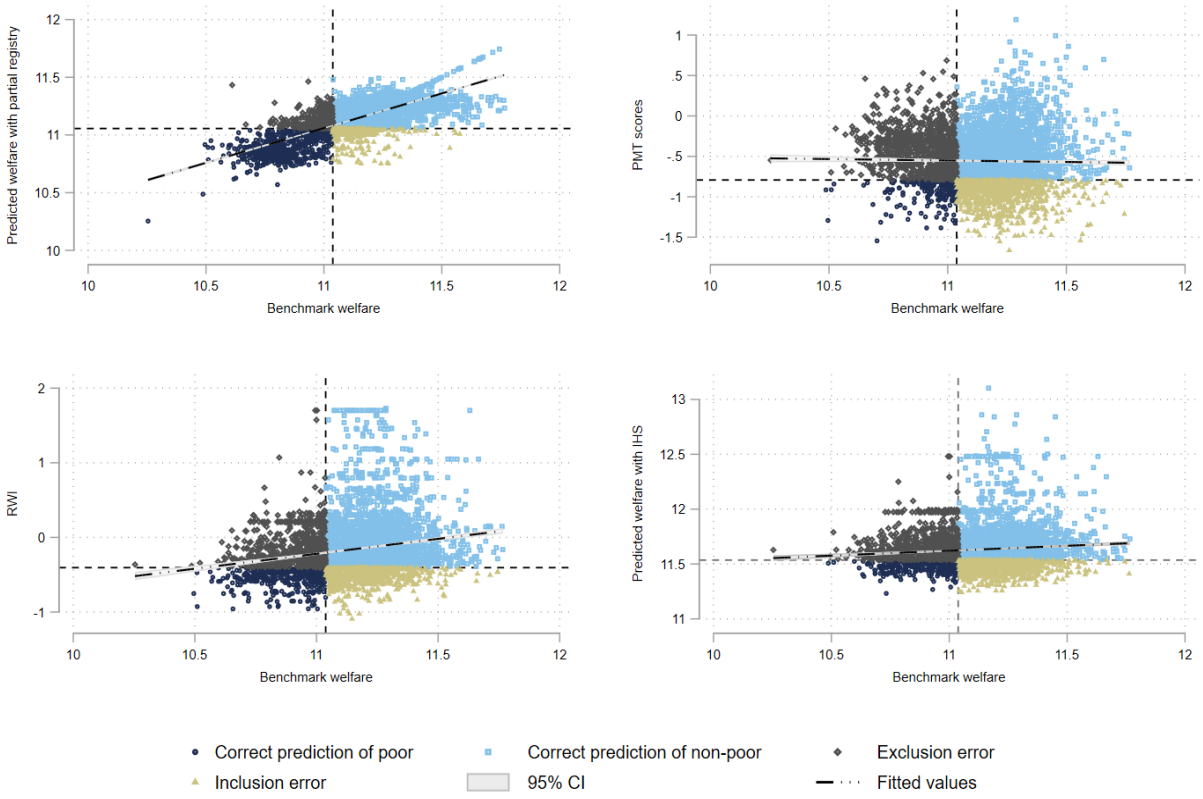


Figure 6 displays another way to present the main results. It plots the predicted welfare measures (on the Y axis) against the actual value of benchmark welfare (on the X axis). Villages are divided into four groups depending on whether their benchmark and predicted welfare fall into the bottom quartile. The bottom left and top right quadrants represent villages that are correctly predicted to be in or out of the bottom quartile. The upper left quadrant shows targeting errors of exclusion and the bottom right shows targeting errors of inclusion. Of the four methods, it is clear that the partial registry approach has by far the lowest prevalence of points in the top left and lower right quadrants, and that these errors are closer to the center. Meanwhile, the PMT has the highest prevalence of errors, especially errors of exclusion. The Meta RWI and the geospatial

household survey model have similar error rates, with the latter slightly less likely to suffer from large exclusion errors.

Figure 6. Predicted vs. Benchmark welfare for the four alternative methods



Notes: Graphs show benchmark village welfare plotted against predicted welfare, for the four predictions methods: The partial registry (top left), PMT (top right), Meta relative wealth index (bottom left) and IHS plus geospatial indicators (bottom right). Each plot is divided into four quadrants with boundaries defined at the 25th percentile of predicted and benchmark welfare. When classifying villages in the bottom quartile as poor, the quadrants represent villages correctly predicted as non-poor (top-right), falsely included as poor (bottom right), correctly predicted as poor (bottom left) and falsely excluded as non-poor (top left)

Finally, Annex 6 presents heat maps of the predicted per capita consumption using each method and the benchmark welfare. In the maps, the PMT scores fail to predict welfare mainly in the central region of Malawi which includes Lilongwe, Dowa, Ntchisi, Nkhotakota, and Kasungu districts. Most methods make accurate predictions in two districts: Rumphi and Chiradzulu.

5. Robustness checks

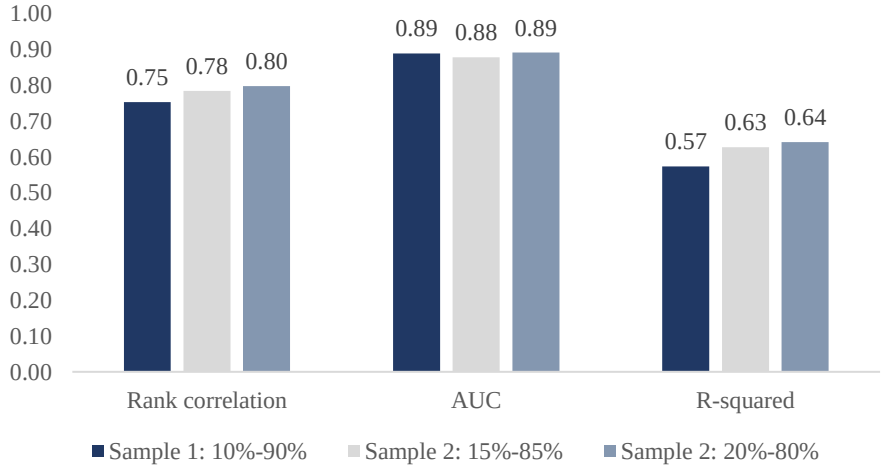
This section presents the results of robustness checks along five dimensions: The size of the partial registry, the nature of the village welfare measure, the geographic composition of the

census model, the estimation method for the geospatial model, and the use of two proprietary geospatial indicators.

A. The Size of the Partial Registry

Given the impressive predictive performance of the partial registry predictions, one might wonder whether increasing its size would further improve performance. Figure 7 compares the performance of the partial registry predictions for partial registries of 444 villages (10 percent), 666 villages (15 percent) and 888 villages (20 percent). Overall, expanding the size of the hypothetical partial registry offers only limited improvements in predictive accuracy, and is not worth the added expense it would entail.

Figure 7. Expanded sample for the partial registry method



B. Village welfare measure

Second, we consider how the results are affected by the choice of welfare measure. This is important because the results until now have used a non-standard welfare measure, namely the average predicted per capita consumption of the bottom half of households in the village. This was based on a conscious decision to match the UBR administrative data, which only contains PMT scores for the bottom 50 percent of households in each village. This section considers how the results change when we consider mean village consumption, taken across all households, as the main welfare measure.

Changing the welfare measure has three main implications for the methodology. First, the census model must now be retrained using all households, not just the bottom half, in the survey. This of course also changes the predicted values of household per capita consumption in the simulated partial registry, which is equal to the benchmark welfare model for villages included in the registry, which entails re-estimating the geospatial model. Finally, we re-estimated the IHS plus geospatial model to train it against average per capita consumption across all households in the survey, rather than just the bottom half.

Table 7 displays the results when using mean village consumption instead of the mean of the bottom half as the village welfare measure. Three main findings are clearly apparent. First, the partial registry approach continues to perform vastly better than the other alternatives when attempting to predict mean village welfare. Second, both the partial registry approach and the RWI suffer moderately when predicting the mean over all households rather than the mean of the poorest half of households, particularly when it comes to rank correlations. This may be because of idiosyncratic positive outliers in the upper half of the household predicted welfare distribution which are more difficult to predict using both geospatial data and predictions trained on asset indices.

Table 7. Metrics of all the methods when using all households in the villages

	All districts-all HH
Rank correlations	
Partial registry (10% of the census sample)	0.61
PMT scores	0.02
IHS training sample	0.19
RWI	0.14
AUC	
Partial registry (10% of the census sample)	0.77
PMT scores	0.50
IHS training sample	0.59
RWI	0.55
R-squared	
Partial registry (10% of the census sample)	0.35
PMT scores	0.00
IHS training sample	0.01
RWI	0.02

Third, the method that combined survey and geospatial predictors without a partial registry (IHS plus geospatial predictors) performs much better when using the full sample of households than when only using the bottom half for each village. The rank correlation increases from 0.13 to 0.19 and the AUC increases from 0.53 to 0.59. This is because on average there are only approximately sixteen households interviewed in each village in the IHS, and average per capita consumption is much more accurately measured when all sample households in each EA are used to train the model rather than only the bottom half. The resulting predictions, when using the full IHS sample, also performs better than the Meta relative wealth index. In this context, when trying to predict the average predicted per capita consumption from a census extract, the fact that the RWI uses additional training data from many countries and proprietary indicators on connectivity does not fully compensate for the fact that it is trained to predict an asset index rather than a consumption-based welfare measure. Therefore, a model that predicts average village per capita consumption directly on the basis of publicly available geospatial characteristics is slightly superior for targeting in this context, though both are far worse than collecting additional partial registry data to train a better geospatial model.

C. The geographic composition of the sample

The benchmark measure of welfare is crucial for evaluating different prediction methods. However, because the census extract is only available for ten districts, it is not immediately clear whether it would be best to use only the survey data from those ten districts, or the full set of survey data to train the census model. The latter takes advantage of a wider set of training data, but the former may better capture the specific relationships between welfare and household characteristics in those poor districts.

Table 8 shows the results when varying the household survey sample used to estimate benchmark welfare. Specifically, we experiment with using only the UBR districts in the household survey data to train the census model, rather than the full sample. While the partial registry method remains the most accurate method by far, it doesn't do nearly as well when the benchmark welfare measure is derived from a model trained on data from only the UBR districts. This is because the sample size used to train the models declines significantly from 6,000 to 2,000 poorest households when limiting the training sample to households in UBR districts, leading to a less informative benchmark welfare model and measure. The partial registry method is particularly sensitive to the weakening of explanatory power in the census model, due to limiting the training data to UBR districts. This is because the predictions from the census model are also used as the dependent variable to train the second stage geospatial model that generates estimates for non-registry villages. Interestingly, however, the predictive performance of the RWI also declines substantially, due to the increase in noise in the benchmark measure of welfare. Nonetheless, the predictive performance of the UBR improves, suggesting that the PMT may have picked up some of the heterogeneity in welfare patterns within the 10 districts.

Table 8. Results when training models on sample data from UBR districts instead of all districts

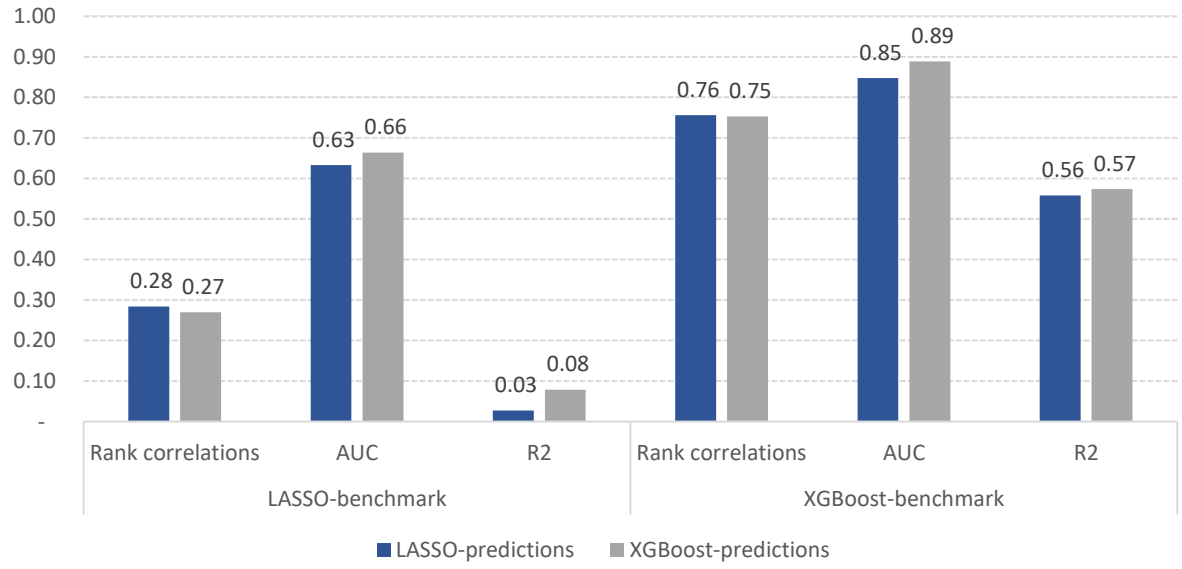
	All districts-poorest 50% HH	**UBR districts-poorest 50% HH**
	Rank correlations	
Partial registry (10% of the census sample)	0.753	0.399
PMT scores	(0.02)	0.15
IHS training sample	0.13	0.04
RWI	0.20	0.11
	AUC	
Partial registry (10% of the census sample)	0.89	0.64
PMT scores	0.50	0.53
IHS training sample	0.53	0.50
RWI	0.60	0.53
	R-squared	
Partial registry (10% of the census sample)	0.57	0.17
PMT scores	0.00	0.01
IHS training sample	0.01	0.00
RWI	0.04	0.01

D. Using LASSO instead of XGboost for the geospatial and census model

This exercise provides a useful opportunity to compare predictions across two different machine learning approaches: Extreme Gradient Boosting, and post-LASSO, which uses LASSO-selected variables in an OLS regression model. The main difference between these two approaches is that the former, as applied here, is a tree-based classification method that uses sequential random forests to predict the dependent variable. Extreme gradient boosting can therefore accommodate highly non-linear relationships. In contrast, post-lasso imposes a linear functional form. An open question in this context is how much this linearity assumption affects the accuracy of the predictions.

The results when using post-LASSO to generate the geospatial predictions based on the partial registry – the second step of the partial registry procedure -- are displayed in Figure 8. These are applied to predictions from the baseline census model, which uses data from all districts but predicts the average welfare of the bottom half of households. Overall, the post-LASSO model performs slightly better in terms of rank correlation, while extreme gradient boosting performs a bit better when looking at AUC and R-squared. Thus, in this context, whether one uses post-LASSO or extreme gradient boosting in the second stage of the partial registry method makes little difference. However, the estimation method for the benchmark welfare seems to affect the results considerably. When the geospatial models are trained using a benchmark welfare that was predicted using LASSO, the metrics for the partial registry methods decrease significantly: rank correlations go from 0.75 to 0.27, AUC from 0.89 to 0.66 and R-squared from 0.57 to 0.08. This could indicate potential measurement errors in the census that affects the performance of LASSO models since they are more sensitive to outliers.

Figure 8. Partial registry predictions using LASSO vs. XGBoost



6. Conclusions

In this paper, we evaluate different alternative methods to identify poor villages in 10 Malawian districts. This is a challenging prediction exercise because villages are highly geographically disaggregated. The results show that a two-step approach utilizing a hypothetical partial registry from 450 villages performs vastly better than the PMT, geospatial prediction based solely on the household survey, or the Meta relative wealth index. The main measure used to identify poor villages is the mean predicted per capita consumption of the bottom half of households in each village, but key results hold when using the mean predicted per capita consumption of all village households as the village welfare measure.

Implementing the partial registry method requires nationally representative survey data, publicly available geospatial indicators, and the collection of a partial registry containing a subset of household characteristics found in the survey data. Several similar household surveys that collect information on welfare proxies have been fielded with the support of the World Bank through the Survey of Well-Being with Instant and Frequent Tracking (SWIFT) program, including in Malawi. Although none have surveyed the full population of households in selected villages, it is quite standard for household surveys to list all surveys in sampled enumeration areas, and we estimate that the cost of collecting approximately 40 variables from all households in approximately 500 villages could be in the ballpark of $24,000 to $73,000. This is a worthwhile investment to greatly boost the accuracy of village welfare measures constructed using geospatial data. Some countries also field periodic community surveys, which could potentially be tweaked to collect partial registries.

This paper also demonstrates the efficacy of applying gradient boosting models in settings with household-level predictors, when sufficient data are available. In particular, using training data from the full sample of households, rather than only the 10 districts of interest, substantially increases predictive performance. Using post-lasso models instead of gradient boosting in the geospatial model, the second step of the partial registry method, only slightly affects the accuracy of the predictions. In contrast, using post-lasso models instead of extreme gradient boosting in the census model, used to impute per capita consumption into the simulated partial registry, greatly reduces the predictive power of the geospatial model. This large reduction in predictive power occurs both when using gradient boosting and post-LASSO for the geospatial model. This suggests that the census data may contain outliers, which introduce more noise into the partial registry predictions when using a linear model of log per capita consumption than when using gradient boosting.

The relatively poor performance of the PMT scores derived from the UBR data is a puzzle. The UBR PMT scores performed a bit better when the benchmark measure of welfare was constructed using data only from the 10 Malawian districts. Even so, the UBR PMT scores do not appear to be consistent with welfare predicted using the census data collected from these districts. Partly this may be due to the UBR data being taken from the initial phase of data collection, although it is also possible that the purpose of the partial registry may have led to measurement error. It would be useful to do these types of evaluations with further rounds of the UBR, even if compared against old census data, to see if later rounds of the UBR produce predictions that are more consistent with the census.

Two limitations of this study are that it only applies to 10 districts in Malawi and is based on a 20 percent extract of the census. Additional work could demonstrate that similar results hold in different contexts and when using a full census. A third limitation is that the hypothetical partial registry is taken from the census, and therefore assumed to match the census exactly. In reality, measurement error in data collection for the partial registry will reduce its performance relative to census-based predictions. Indeed, a partial registry could suffer from some of the same issues in data collection experienced when collecting data for the UBR. A project that pilots the collection of a partial registry for the purpose of training a geospatial model would provide a more realistic test of the partial registry approach and could shed new light on whether such a partial registry would be prone to systematic bias. Finally, future research could leverage household level information on geocoordinates if they can be obtained in census data. This would enable estimating models relating predicted welfare to geospatial indicators at the household level, which may perform better than the village-level models considered in this analysis. Despite these caveats, the results convincingly demonstrate both the limitations of existing methods, and the potential for partial registries to add massive value when using survey and geospatial data to identify the poorest villages in a very low-income setting.

References

Aiken, E., Bellue, S., Karlan, D., Udry, C. R., & Blumenstock, J. (2021). *Machine learning and mobile phone data can improve the targeting of humanitarian assistance* (No. w29070). National Bureau of Economic Research.

Babenko, B., Hersh, J., Newhouse, D., Ramakrishnan, A., & Swartz, T. (2017). Poverty mapping using convolutional neural networks trained on high and medium resolution satellite images, with an application in Mexico. *arXiv preprint arXiv:1711.06323*.

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

Engstrom, R., Sandborn, A., Yu, Q., Burgdorfer, J., Stow, D., Weeks, J., & Graesser, J. (2015, March). Mapping slums using spatial features in Accra, Ghana. In *2015 Joint Urban Remote Sensing Event (JURSE)* (pp. 1-4). IEEE.

Engstrom, R., Hersh, J., & Newhouse, D. (2016). Poverty from space: using high resolution satellite imagery for estimating economic well-being and geographic targeting. *unpublished paper*.

Engstrom, R., Newhouse, D., Haldavanekar, V., Copenhaver, A., & Hersh, J. (2017, March). Evaluating the relationship between spatial and spectral features derived from high spatial resolution satellite data and urban poverty in Colombo, Sri Lanka. In *2017 Joint Urban Remote Sensing Event (JURSE)* (pp. 1-4). IEEE.

Head, A., Manguin, M., Tran, N., & Blumenstock, J. E. (2017, November). Can human development be measured with satellite imagery?. In *Ictd* (pp. 8-1).

Henderson, J. V., Storeygard, A., & Weil, D. N. (2012). Measuring economic growth from outer space. *American economic review*, *102*(2), 994-1028.

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

Limarino, W. (2021) Augmented Proxy Mean Tests. Can Machine Learning Improve Targeting Effectiveness? (Preliminary Draft)

Lindert, K., Andrews, C., Msowoya, C., Paul, B. V., Chirwa, E., & Mittal, A. (2018). Rapid Social Registry Assessment.

Masaki, T., Newhouse, D., Silwal, A. R., Bedada, A., & Engstrom, R. (2022). Small area estimation of non-monetary poverty with geospatial data, *Statistical Journal of the IAOS,* v 37 no. 4

Mellander, C., Lobo, J., Stolarick, K., & Matheson, Z. (2015). Night-time light data: A good proxy measure for economic activity?. *PloS one*, *10*(10), e0139779.

Pinkovskiy, M., & Sala-i-Martin, X. (2016). Lights, camera… income! Illuminating the national accounts-household surveys debate. *The Quarterly Journal of Economics*, *131*(2), 579-631.

Serajuddin, U., Uematsu, H., Wieser, C., Yoshida, N., & Dabalen, A. (2015). Data deprivation: another deprivation to end. *World Bank policy research working paper*, (7252).

Smythe, I., & Blumenstock, J. E. (2021). Geographic micro-targeting of social assistance with high-resolution poverty maps. *In Submission (KDD)*.

Van Der Weide, Roy; Blankespoor, Brian; Elbers, Chris; Lanjouw, Peter. 2022. How Accurate Is a Poverty Map Based on Remote Sensing Data?: An Application to Malawi. Policy Research Working papers;10171. World Bank, Washington, DC. © World Bank. https://openknowledge.worldbank.org/handle/10986/38009 License: CC BY 3.0 IGO."

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

Zar, J. H. (2014). Spearman rank correlation: overview. *Wiley StatsRef: Statistics Reference Online*.

Annexes

Annex 1. Household characteristics in UBR districts vs. the rest of the country

	Total	No UBR	UBR	P-value of the difference
Highest educated male has primary education	0.19	0.18	0.20	0.056
		(0.39)	(0.40)	
Highest educated male has secondary education	0.09	0.10	0.07	0.000
		(0.30)	(0.25)	
Highest educated male has tertiary education	0.03	0.04	0.02	0.000
		(0.19)	(0.12)	
Highest educated female has primary education	0.18	0.17	0.18	0.069
		(0.38)	(0.39)	
Highest educated female has secondary education	0.05	0.05	0.03	0.000
		(0.23)	(0.17)	
Highest educated female has tertiary education	0.02	0.03	0.01	0.000
		(0.16)	(0.10)	
Household head is literate	0.72	0.73	0.71	0.162
		(0.45)	(0.45)	
Household size	4.33	4.33	4.32	0.686
		(2.02)	(1.96)	
Household overcrowding	2.04	2.08	1.96	0.000
		(1.28)	(1.23)	
Urban Household	0.18	0.25	0.05	0.000
		(0.43)	(0.22)	
Elderly dependency ratio	0.07	0.07	0.08	0.000
		(0.20	(0.22)	
Child dependency ratio	0.39	0.39	0.38	0.032
		(0.24	(0.24)	
Fuel cooking: firewood	0.81	0.76	0.91	0.000
		(0.43	(0.29)	
Access to piped water	0.23	0.28	0.12	0.000
		(0.45	(0.33)	
Access to flushtoilet	0.04	0.05	0.01	0.000
		(0.22	(0.12)	
Household owns a house	0.74	0.70	0.81	0.000
		(0.46	(0.39)	
Household has improved walls	0.91	0.94	0.83	0.000
		(0.23	(0.38)	
Household has improved roof	0.50	0.54	0.42	0.000
		(0.50	(0.49)	
Household has improved floor	0.29	0.32	0.21	0.000
		(0.47	(0.41)	

Hosehold has cellphone	0.50	0.51	0.46	0.000
		(0.50	(0.50)	
Hosehold has fridge	0.06	0.08	0.02	0.000
		(0.27	(0.13)	
Hosehold has stove	0.00	0.00	0.00	0.888
		(0.06	(0.06)	
Hosehold has computer	0.03	0.04	0.01	0.000
		(0.19	(0.07)	
Hosehold has oxcart	0.01	0.01	0.02	0.000
		(0.09	(0.14)	
Hosehold has bicycle	0.37	0.36	0.37	0.660
		(0.48	(0.48)	
Hosehold has motorcycle	0.02	0.02	0.02	0.515
		(0.13	(0.13)	
Hosehold has car	0.02	0.03	0.01	0.000
		(0.16	(0.08)	
Hosehold has radio	0.42	0.43	0.39	0.000
		(0.50	(0.49)	
Hosehold has television	0.13	0.16	0.06	0.000
		(0.36	(0.24)	

Source: Integrated Household Survey 2016.
Note: Standard deviation in parenthesis.

Annex 2. Satellite data

Dataset	Bands	Description	Year
Data from Google Earth Engine: the data was collected at grids of 7 x7 km, approximately.			
GPM: Monthly Global Precipitation Measurement (GPM)	mm/hr	Merged satellite-gauge precipitation estimate	Monthly from 2017-2018/annual 2001-2016
Copernicus Global Land Cover Layers	%	Percent vegetation cover for cropland land cover class	Yearly from 2017-2018
	%	Percent vegetation cover for herbaceous vegetation land cover class	
	%	Percent vegetation cover for moss and lichen land cover class	
	%	Percent vegetation cover for shrubland land cover class	
	%	Percent vegetation cover for bare-sparse-vegetation land cover class	
	%	Percent ground cover for built-up land cover class	
	%	Percent ground cover for permanent water land cover class	
	%	Percent ground cover for seasonal water land cover class	
Tsinghua FROM-GLC year of change to impervious surface	[1-34]	Year of the transition from from pervious to impervious. From 34 (year: 1985) to 1 (year: 2018)	Yearly from 2017-2018
MODIS Land Cover Type Yearly Global	Number	Land Cover Type 1: croplands, urban built-up, Cropland/Natural Vegetation Mosaics.	Yearly from 2017-2018
Landsat 7 NDVI Composite	[-1,1]	Normalized Difference Vegetation Index	Monthly from 2017-2018/annual 2001-2016
Landsat 7 NDWI Composite	[-1,1]	Normalized Difference Water Index	Monthly from 2017-2018/annual 2001-2016
NASA-USDA Global Soil Moisture Data	mm	Surface soil moisture	Monthly from 2017-2018
VIIRS Stray Light Corrected Nighttime	nanoWatts/cm2/sr	Average DNB radiance values.	Monthly from 2017-2018
Data from WorldPop repository			
Population density	Estimated population density per grid-cell	30 arc (approximately 1km at the equator).	2017-2018
Distance to OSM major roads	Distance (km) from the cell centre to the nearest feature	3 arc (approximately 100m at the equator)	2016
Global Built Settlement Growth	Built-Settlement Growth Model (BSGM) interpolating for years 2001-2011, 2013 and extrapolating for years 2015-2020.	3 arc (approximately 100m at the equator).	2017-2018
World Pop Open Population Repository/ Gridded maps of building patterns throughout sub-Saharan Africa	9 files that contain data of buildings in Malawi (count, density, area, perimeter, others)	100 m grid cell across the study area	2021

Annex 3. Xgboost models

XGBoost is a gradient boosting algorithm that provides a parallel tree boosting that solves data science problems in a fast and accurate way. It is designed to work with large and complex data sets.

This annex describes the use of Xgboost for regression. The algorithm fits a regression tree to the residuals as gradient boost but uses a unique regression tree. Each tree starts with a single leaf that is called a root, and all the residuals go to the leaf. The algorithm calculates similarity scores and gain to determine how to split the data.

The similarity score for the residuals on each leaf equals

$$Similarity\ Score = \frac{sum\ of\ residuals^2}{number\ of\ residuals + \lambda}$$

Where λ is a regularization parameter intended to reduce the prediction's sensitivity to individual observations and prevent overfitting the training data. If the leaf has several different residuals, the similarity score will be relatively small since they will cancel each other out. In contrast, if the residuals are similar or the leaf has very few residuals, the similarity score will be relatively large.

To quantify how much better the leaves cluster similar residuals than the root, we need to calculate the gain of splitting the residuals into groups. The gain is equal to

$$Gain = sum\ of\ similarity\ scores\ of\ the\ tree\ leaves - similarity\ scores\ of\ the\ root$$

Then the algorithm compares the gain calculated for each split and selects the one with the highest value since that would mean that a particular feature is better at splitting the residuals into clusters of similar values. Then it continues with another split. You can limit the tree depth or the splits to different levels, up to 6 levels is the default.

To determine output values for the leaves, we calculate the following:

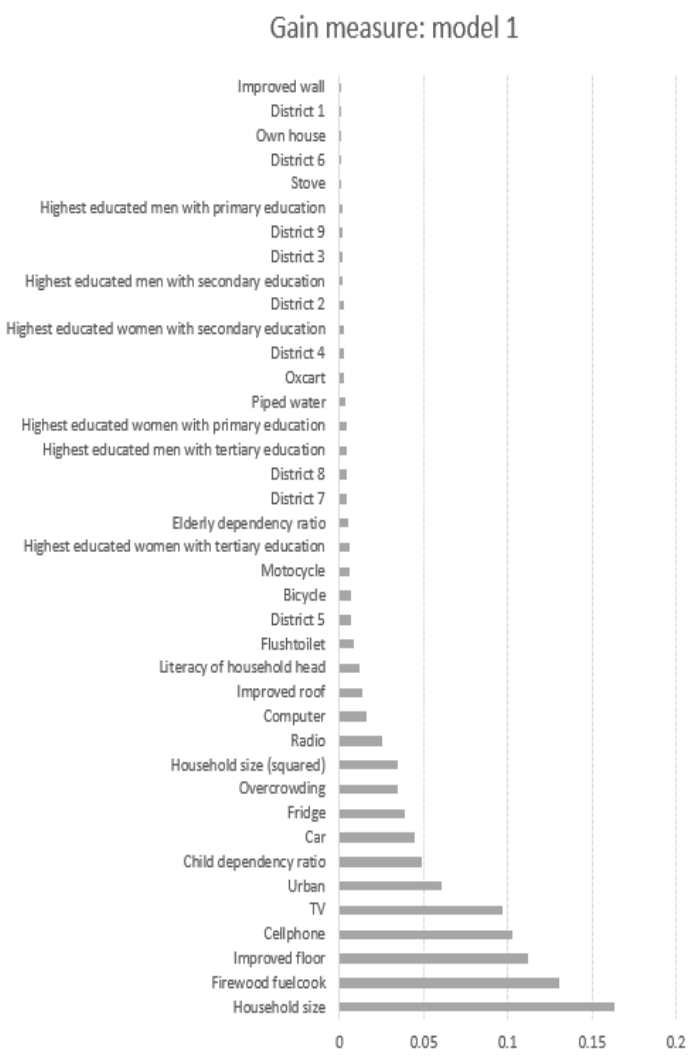
$$Output\ value = \frac{sum\ of\ residuals}{number\ of\ residuals + \lambda}$$

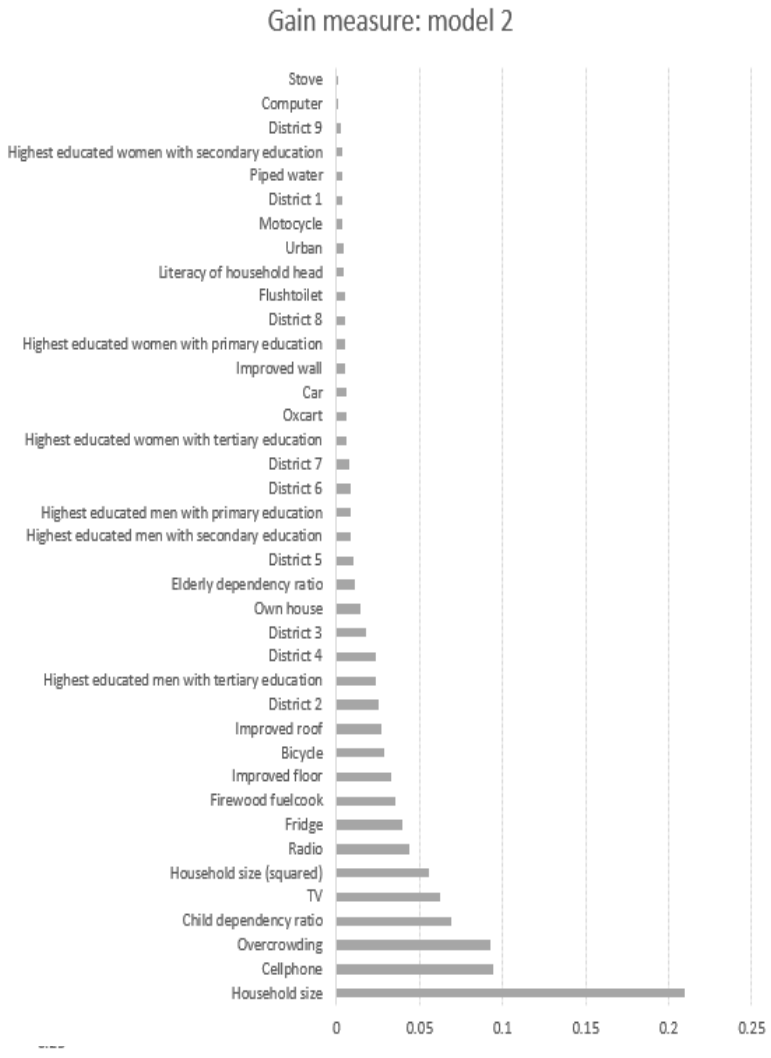
The output value is like the similarity score, except that it does not square the sum of the residuals. After this, the tree can be used for making predictions. Like gradient boost, xgboost makes new predictions starting at the initial prediction and adding the output of the tree scaled by a learning rate ε. The new predictions will have smaller residual values. Then the algorithm builds new trees based on the new residuals until the residuals get very small or it reaches the maximum number of trees.

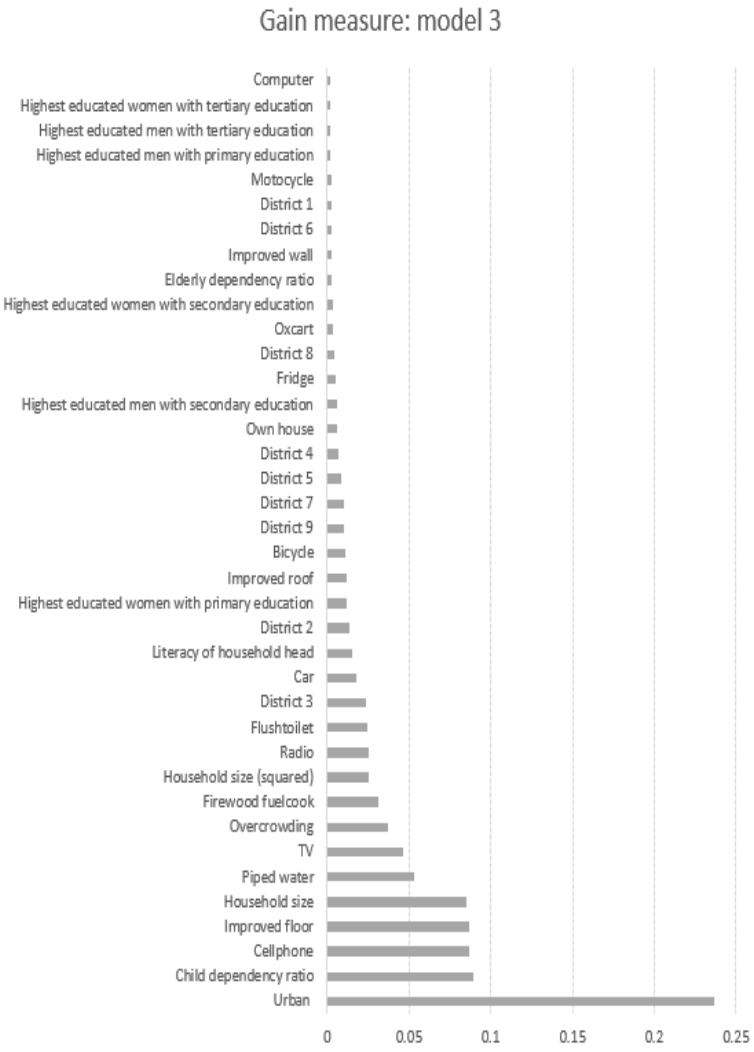
Annex 4. Benchmark welfare models using different training samples

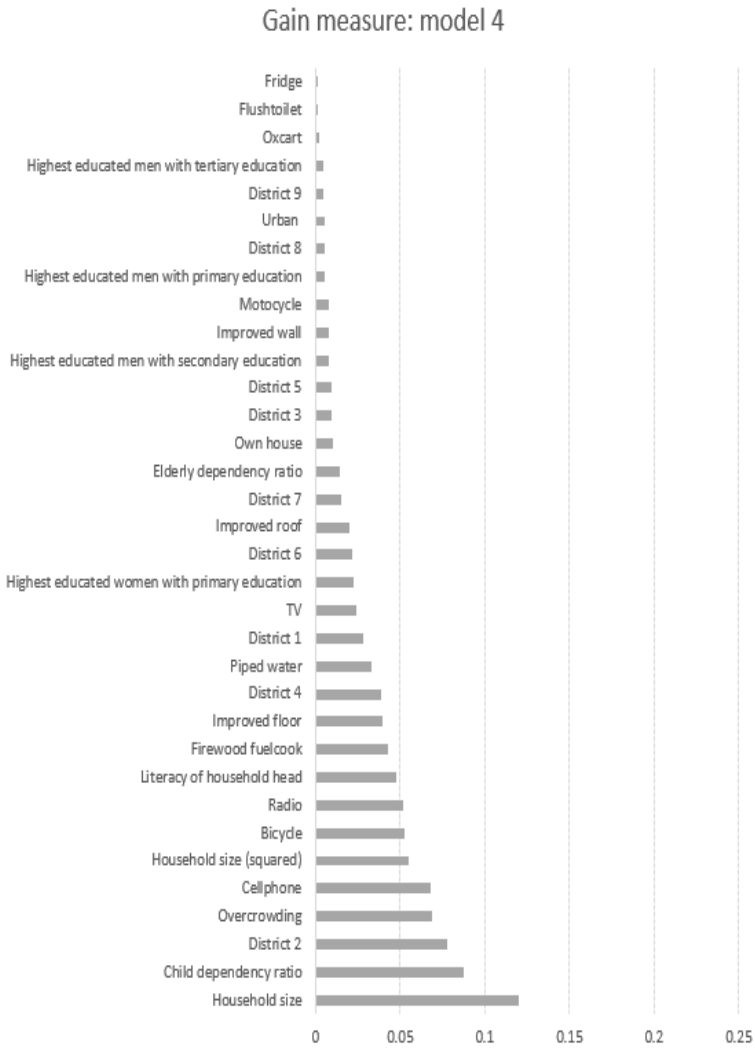
	All districts-All HH	UBR districts-All HH	All districts-50% poorest HH	UBR districts-50% poorest HH
R-squared	65.48	55.41	53.71	25.66
Top 10 variables	Household size, households assets, child dependency, overcrowding, urban/rural	Household size, households assets, child dependency, overcrowding, urban/rural	Household size, households assets, child dependency, overcrowding	Household size, households assets, child dependency, overcrowding, HH head literacy

Annex 5. Importance of variables in benchmark models






Gain measure: model 3
Computer
Highest educated women with tertiary education
Highest educated men with tertiary education
Highest educated men with primary education
Motocycle
District 1
District 6
Improved wall
Elderly dependency ratio
Highest educated women with secondary education
Oxcart
District 8
Fridge
Highest educated men with secondary education
Own house
District 4
District 5
District 7
District 9
Bicycle
Improved roof
Highest educated women with primary education
District 2
Literacy of household head
Car
District 3
Flushtoilet
Radio
Household size (squared)
Firewood fuelcook
Overcrowding
TV
Piped water
Household size
Improved floor
Cellphone
Child dependency ratio
Urban
0
0.05
0.1
0.15
0.2
0.25


Gain measure: model 4
Fridge
Flushtoilet
Oxcart
Highest educated men with tertiary education
District 9
Urban
District 8
Highest educated men with primary education
Motocycle
Improved wall
Highest educated men with secondary education
District 5
District 3
Own house
Elderly dependency ratio
District 7
Improved roof
District 6
Highest educated women with primary education
TV
District 1
Piped water
District 4
Improved floor
Firewood fuelcook
Literacy of household head
Radio
Bicycle
Household size (squared)
Cellphone
Overcrowding
District 2
Child dependency ratio
Household size
0
0.05
0.1
0.15
0.2
0.25

Annex 6. Per capita consumption maps using different prediction methods.