Targeting Humanitarian Aid with Machine Learning and Mobile Phone Data: Evidence from an Anti-Poverty Intervention in Afghanistan

Emily L. Aiken[*] Guadalupe Bedoya[†] Aidan Coville[†]
Joshua E. Blumenstock[*]

March 24, 2020

`Preliminary version. Please do not cite or distribute without permission.`

Abstract

Recent papers demonstrate that non-traditional data, from mobile phones and other digital sensors, can be used to roughly estimate the wealth of individual subscribers. This paper asks a question more directly relevant to development policy: Can non-traditional data be used to more efficiently target development aid? By combining rich survey data from a "big push" anti-poverty program in Afghanistan with detailed mobile phone logs from program beneficiaries, we study the extent to which machine learning methods can accurately differentiate ultra-poor households eligible for program benefits from other poor households deemed ineligible. We show that supervised learning methods leveraging mobile phone data can identify ultra-poor households as accurately as standard survey-based measures of poverty, including consumption and wealth; and that combining survey-based measures with mobile phone data produces classifications more accurate than those based on a single data source. We discuss the implications and limitations of these methods for targeting extreme poverty in marginalized populations.

[*] School of Information, University of California, Berkeley
[†] Development Impact Evaluation Group, World Bank

1 Introduction

Program targeting—the task of determining who is eligible and who is ineligible for humanitarian aid—is a major source of inefficiency in anti-poverty program administration (Coady et al., 2004). Typically, programs make targeting decisions based on administrative data such as tax records or survey-based measures of assets or consumption. But the quality of these data is in rapid decline (Meyer et al., 2015), and in many developing countries, reliable data for targeting do not exist and would be prohibitively expensive to collect (Jerven, 2013). However, over the past several years, a handful of studies have shown that non-traditional "digital trace" data—behavioral indicators recorded in everyday interactions with technology—are predictive of wealth in developing contexts (e.g. Blumenstock et al., 2015; Sheehan et al., 2019). There is optimism in the machine learning and development communities that these data could provide a quick and low-cost alternative to standard field-based targeting methods (e.g. Blumenstock, 2016; De-Arteaga et al., 2018).

This paper evaluates the extent to which digital trace data can be used for program targeting. Specifically, we match mobile phone transaction logs (call detail records, or CDR) to household survey data from a World Bank-led impact evaluation of the Afghanistan government's Targeting the Ultra-Poor (TUP) program (Bedoya et al., 2019). In the TUP program, ultra-poor households are targeted for "big push" interventions based on the combination of a community wealth ranking and a measure of multiple deprivation. We evaluate the accuracy of machine learning methods leveraging CDR data in comparison to two standard targeting methods—asset-based wealth and consumption expenditure—for differentiating the ultra-poor households deemed eligible for the TUP intervention from ineligible non-ultra-poor households.

Our analysis produces three main results. First, by comparing errors of inclusion and exclusion, we find that a CDR-based method is comparable in accuracy to standard survey-based measures of welfare for identifying the phone-owning ultra-poor in this sample. We emphasize, however, that the CDR-based method is inherently limited to households owning mobile phones and its utility is compromised by incomplete mobile phone penetration among TUP program beneficiaries and in the developing world more generally. Second, we find that methods combining CDR data with measures of assets and consumption are more accurate than methods using any one of the data sources for identifying the ultra-poor. Third, we find that, in contrast to their success for classifying the ultra-poor, machine learning methods leveraging CDR data have little ability to estimate asset- or consumption-based measures of welfare in this sample, suggesting that CDR-based wealth prediction methods are not an "off-the-shelf" tool that can be applied without adjustment to any population or prediction task.

1.1 Related Work

1.1.1 Poverty Targeting

Most targeted anti-poverty programs in the developed world use means tests, restricting program benefits to those below a certain income or consumption threshold. In the developing world, however, means tests are frequently impractical, particularly in areas where most employment is in the informal sector or records of income and expenditures are limited. Most poverty targeting schemes in the developing world therefore rely on proxy measures of wealth. While some programs are targeted based on general geographic or demographic data, many use asset ownership as a proxy for income or consumption, either via a proxy means test (PMT) (Grosh & Baker, 1995) or an asset-based wealth index constructed with principle component analysis (Filmer & Pritchett, 2001). An increasingly popular alternative to asset-based proxies for wealth is community-based targeting, in which community members or community leaders select beneficiaries. However, there is a growing consensus in the literature that both asset-based and community-based wealth measures are limited by low-quality data, and in a subset of cases targeting based on these measures is found to be regressive or no better than universal allocation of benefits (e.g. Coady et al., 2004; Karlan & Thuysbaert, 2019; Brown et al., 2018).

Hanna & Olken (2018) present two empirical cases in which targeting based on imperfect data is superior to universal allocation under a budget constraint. By comparing the cost-effectiveness of PMT-targeted cash transfer schemes and universal allocation using data from two nationwide anti-poverty programs in Indonesia and Peru, the study finds that narrowly-targeted programs are significantly more cost-effective than widely-targeted programs or universal allocation, even when targeting is based on imperfect PMT proxies for consumption (PMT $r^2 = .56 - .66$). The study notes, however, that adding noise to the asset data for the PMT leads to drops in targeting accuracy and projected poverty outcomes.

Wider reviews suggest that low-quality data for PMT and other targeting schemes lead to poor targeting outcomes for a wide range of programs. Coady et al. (2004) provide a review of 122 targeted anti-poverty programs in 48 countries in the years 1985-2003, finding that while most programs transfer more to the poor than to the wealthy, a staggering 25% of programs are regressive, transferring less to the poor than would universal allocation. Many of the regressive programs use coarser forms of targeting than the household-level targeting mechanisms we focus on here (such as geographic targeting, demographic targeting, or self-selection), but several of them involve means-tests, proxy-means tests, or community-based methods. Brown et al. (2018) present similar conclusions, reviewing nine PMT-targeted anti-poverty programs in Africa and finding that the PMT yields only modest gains in poverty impacts over geographic targeting or universal allocation.

Leaving behind the question of targeting over universal allocation, Alatas et al. (2012) explore in depth the comparison between proxy-means tests and community-based targeting using data from a field experiment in Indonesia. The study finds that the PMT and community targeting methods perform similarly in terms of targeting accuracy (70% accuracy for the PMT, 67% accuracy for community-based targeting). Importantly for our work, this study and related papers conclude that communities apply a concept of poverty outside of income, consumption, or assets in their poverty assessments, and that community-based targeting yields higher rates of community satisfaction than a PMT (Alatas et al., 2012; Banerjee et al., 2007; Karlan & Thuysbaert, 2019).

Regardless of their accuracy (or inaccuracy), the current methods available for poverty targeting in the developing world are time and resource intensive, and moreover may be infeasible in fragile or conflict-affected areas. We turn, therefore, to the literature on using passively collected digital trace data to estimate wealth, and later evaluate whether this data is useful for identifying the poor.

1.1.2 Wealth Estimation from Mobile Phone Metadata

Research to date on wealth prediction from digital trace data has focused on mobile phone metadata as cell phones have become increasingly ubiquitous worldwide, projected to reach a global penetration rate of 73% in 2020 (GSMA, 2017). Recent work has shown that machine learning methods leveraging call detail records (CDR) can produce useful estimates of wealth and well-being at a fine spatial granularity. This body of work focuses largely on poverty, typically quantified by an asset-based wealth index (e.g. Blumenstock et al., 2015; Hernandez et al., 2017), but related papers explore a wider set of well-being measures, including literacy (Schmid et al., 2017), food security (Decuyper1 et al., 2014), and infrastructure (Blumenstock et al., 2017).

While most of this work addresses spatially granular poverty mapping, to our knowledge two papers to date cover the individual-level wealth prediction task that is more relevant to poverty targeting. Blumenstock et al. (2015) show that CDR data are predictive of an individual-level asset-based wealth index in Rwanda. More specifically, the study matches ground-truth survey data to call details records covering two years of mobile phone activity for 856 geographically stratified individuals, extracts a suite of thousands of behavioral indicators from the CDR covering different aspects of mobile phone use, and applies a supervised learning algorithm to generate wealth predictions from behavioral indicators. Model accuracy is evaluated with cross-validation to ensure that the wealth prediction model generalizes out-of-sample (cross-validated $r = 0.68$). Blumenstock (2018) performs the same experiment for 1,234 male heads of households in the Kabul and Parwan districts of Afghanistan, yielding similar predictive accuracy.

1.2 Our Contribution

Given the success of wealth prediction from mobile phone data and the need for cheaper and more accurate targeting mechanisms in the developing world, it makes sense to experiment with poverty targeting based on CDR data. Here we build on the existing literature on poverty targeting and wealth prediction from digital trace data to evaluate whether CDR data can be used in practice to target the poor for anti-poverty programs. By matching CDR to household survey data from 537 households in 80 of the poorest villages in the Balkh province of Afghanistan, we assess whether CDR-based methods can replicate household-level targeting from a recent anti-poverty program. We find that CDR-based targeting is as accurate as targeting based on assets or consumption for the subset of households that own a mobile phone in this sample, and that targeting based on a combination of assets, consumption, and CDR data is more accurate than targeting based on any one data source. We find, however, that CDR are less predictive of standard measures of poverty (asset-based wealth index and consumption) than in previous work, and hypothesize that sample homogeneity may contribute to low predictive accuracy for standard wealth measures.

2 Methods

2.1 Ground Truth Data

Our ground-truth survey data was collected as part of the Targeting the Ultra-Poor (TUP) program implemented by the government of Afghanistan with support from the World Bank. The TUP program included a randomized controlled trial evaluating the impact of a "big push" program for lifting the ultra-poor out of poverty with multi-faceted interventions (Bedoya et al., 2019). The TUP dataset contains wealth and welfare measures for 2,899 households in 80 of the poorest villages in Afghanistan's Balkh province, surveyed once at baseline (November 2015 - April 2016), and once at endline in 2018, a year after the conclusion of interventions for treated households. Baseline and endline data were collected in two in-person interviews, one with the primary woman of each household (lasting two hours), and one with the primary man (lasting around 45 minutes).

Our main analysis is limited to baseline survey data for the 537 households in the TUP sample that possess at least one cell phone that placed a call between November and April 2016 appearing in our CDR dataset (see section 2.2 for details of the CDR data). Data from the TUP survey include several outcome measures that we use in our analysis: 1) an asset-based wealth index, 2) household consumption, 3) the designation of households as above or below Afghanistan's poverty line (based on consumption), and 4) the designation of households as ultra-poor or non-ultra-poor (which ultimately determined households' eligibility for the TUP intervention).

Asset-based Wealth Index The asset-based wealth index is calculated as the first principal component of variation in household asset ownership for the following items: Radios/CDs/cassettes, televisions, dish antennas, VCRs/DVD players, refrigerators, generators, mattresses, mobile phones, non-mobile phones, clothes irons, bed frames, pieces of jewelry, mosquito nets, mosquito-repellent candles, fans, and cameras. The wealth index principle component analysis (PCA) is calculated over the entire dataset of 2,899 households. Figure 1 shows the distribution of each of the individual components and of the composite wealth index for the 537 households in our sample. We find that the wealth index explains 26.2% of the variation in asset ownership in our sample. To remove outliers we winsorize the wealth index with a 99% limit before applying supervised learning methods.

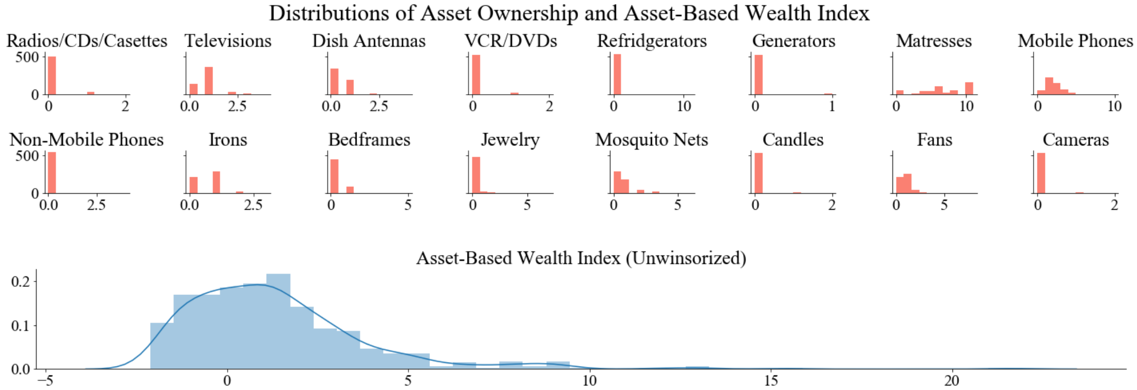

Figure 1: Distribution of asset-based wealth index and individual asset ownership across 537 households.

Consumption The consumption module of the TUP survey gathers information on the value of food expenditures for the week prior to the interview and non-food expenditures for the year prior to the interview. Non-food expenditures are "personal and household items, education and medical expenses, household repairs, social expenses (e.g. weddings, funerals and other ceremonies), and temptation goods" (Bedoya et al., 2019). Based on this data, we construct as an outcome measure the logarithm of per-capita household monthly consumption. Figure 2 shows the distribution of per-capita consumption and log-transformed per-capita consumption in our 537-household sample.

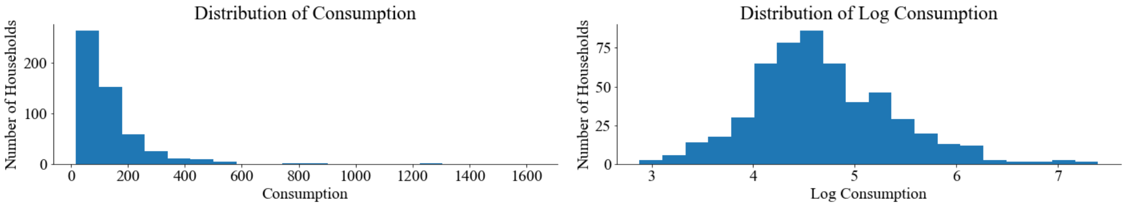

Figure 2: Distribution of consumption across 537 households pre- and post- log transformation.

Poverty Line The Afghanistan national poverty line for 2016-2017 is AFN 2,064 (equivalent to USD 112 PPP) per person per month. In our dataset, determination of poverty line status is based on a poverty-line consistent measure of consumption, which includes food consumption, consumer durables, housing costs, and a subset of the non-food costs included in the standard consumption calculation. 53% of our 537-household sample is below the poverty line and 47% is above the poverty line.

Ultra-Poor Designation Ultra-poor designation determined eligibility for the TUP program, and was based on a community wealth ranking and a follow-up in-person verification evaluating whether households met a set of qualifying criteria. Community wealth rankings were conducted separately in each village, coordinated by a local NGO and village leaders. Community wealth rankings divided households into four categories: well-off, better-off, poor, and ultra-poor, with 44% of households falling into the ultra-poor category. Community wealth rankings were followed by in-person verification of ultra-poor status by government representatives, based on a measure of multiple deprivation. To be designated as ultra-poor and intervention-eligible, a household had to meet at least three of the following criteria:

1. Household is financially dependent on women's domestic work or begging.
2. Household owns less than 800 square meters of land or is living in a cave.
3. Targeted woman is younger than 50 years of age.
4. There are no active adult men income earners.
5. Children of school age are working for pay.
6. Household does not own any productive assets.

Ultimately 11% of the households classified as ultra-poor in the community wealth ranking met the verification standards. In our 537-household sample, 146 households (27%) are designated as ultra-poor (UP), while the remaining 391 are non-ultra-poor (NUP).

Comparison of Outcome Measures While all the wealth measures in our dataset are correlated with one another, none are completely overlapping. Even the poverty line indicator, which is a consumption-based measure, differs slightly from the consumption outcome since it is based on a somewhat different set of costs. Table 1 displays the correlations among wealth measures, while Figure 3 visualizes the overlap among the different outcome measures. It is particularly important to note the characteristics of the ultra-poor: while the ultra-poor population makes up 27% of the overall sample, less than half of the ultra-poor fall into the bottom 27% of the sample by wealth index or consumption.

	Wealth	Log Consumption	Below PL	Ultra-Poor
Wealth Index	1.00			
Log Consumption	0.34	1.00		
Below PL	-0.23	-0.68	1.00	
Ultra-Poor	-0.31	-0.29	0.25	1.00

Table 1: Correlations between outcome measures.

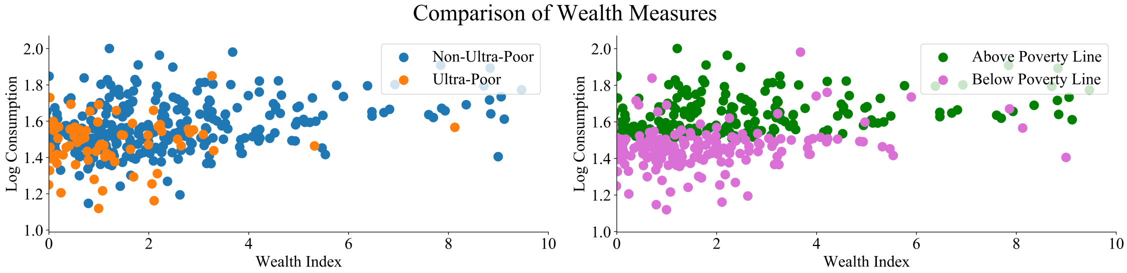

Figure 3: Comparison of wealth measures: wealth score vs. consumption by ultra-poor and poverty-line status.

Selection into Sample It is important to note that the 2,899 households in the TUP survey are not nationally representative of Afghanistan as a whole, and the 537 households that we analyze in this paper are not representative of the overall sample from the TUP survey. Table 2 compares key outcomes between the households that are included and excluded from our 537-household subsample. We also use summary statistics from the Afghanistan Living Conditions Survey of 2016-2017 (ALCS, 2017) for general comparisons of our sample to nationally representative data. There is evidence of systematic selection into our subsample: the 537-household sample we analyze is similar in average poverty to Afghanistan as a whole (53% of households are below poverty line in our sample, 55% of individuals in the nationally representative data), but significantly richer on average than households surveyed in the TUP study (62% of households in the TUP study are below the poverty line). Households in our subsample must own at least one mobile phone, so it is not surprising that they tend to be richer than the average household surveyed in the TUP project.

Panel A: Ultra-poor households

	In our sample	**Outside our sample**	**Difference**
Asset-based Wealth Index	0.10 (1.50)	-0.88 (1.10)	0.98***
Consumption	97.39 (108.33)	78.19 (69.82)	19.21**
Below Poverty Line	0.73 (0.44)	0.81 (0.39)	-0.08*
Own Mobile Phone	1.00 (0.00)	0.67 (0.47)	0.33***
N	146	1073	1219

Panel B: Non-ultra-poor households

	In our sample	**Outside our sample**	**Difference**
Asset-based Wealth Index	1.83 (2.75)	0.13 (1.94)	1.70***
Consumption	164.69 (174.38)	139.93 (135.56)	24.76**
Below Poverty Line	0.45 (0.50)	0.50 (0.50)	-0.05
Own Mobile Phone	1.00 (0.00)	0.87 (0.34)	0.13***
N	391	1289	1680

Table 2: Differences between households included and excluded from our 537-household sub-sample. Standard deviations are shown in parentheses; significance of difference in means between the samples is determined with a two-sided t-test (* indicates p < .05, ** indicates p < .01, *** indicates p < .001).

Moreover, every household in our sample owns at least one mobile phone, but the group of households surveyed in the TUP project and the ALCS data better reflect realistic patterns of phone ownership in Afghanistan. In the full TUP sample, 80% of households own at least one phone, and only 72% of ultra-poor households own at least one phone. In contrast, only 43.3% of individuals in the ALCS survey report owning a mobile phone (though at least part of this difference can be attributed to the inconsistency of household vs. individual-level data).

2.2 Mobile Phone Metadata

We obtain informed consent from survey respondents in the TUP project to merge their survey responses with their CDR. We match the labeled survey responses to CDR obtained from one of the largest Afghan cell providers for the months of November 2015 to April 2016. For households with multiple phones (N=84), we use data from only the phone belonging to the household head. For households without a recorded household head and multiple phones (N=17), we use data from one of the households' associated phones at random.[1] CDR data contain the following transactions:

- **Call:** Phone numbers for the caller and receiver, time of the call, duration of the call, and cell tower through which the call was placed
- **Text message:** Phone numbers for the caller and recipient, time of the text message
- **Recharge:** Time of the recharge, amount of the recharge

In total, for the 537 households in our sample, 629,543 transactions took place in the months of November 2015 to April 2016, broken down into 310,883 calls, 305,756 text messages, and 12,904 recharges.

From these CDR, we compute a set of behavioral indicators capturing aggregate aspects of each individual's mobile phone use using bandicoot, an open-source toolbox for CDR analysis (de Montjoye et al., 2016). These indicators include features relating to an individual's overall behavior (for example, average call duration and percent initiated conversations), their network of contacts (for example, the entropy of their contacts and the balance of interactions per contact), their spatial patterns based on cell tower locations (for example, the number of unique antennas visited and the radius of gyration), and their recharge patterns (including the average amount recharged and the time between recharges). Each indicator is computed separately for weekdays, weekends, daytime, and nighttime. In total, 869 behavioral features are computed for each individual; after dropping features which are missing for over 50% of households and those that have variance below 0.01, 623 features remain. We standardize each feature to zero mean and unit variance.

[1] We recognize that there is an alternative matching between household-level survey data and individual-level phone data which focuses on predicting individual-level poverty from mobile phone metadata (the sample size, in this case, would be 641 individuals, where each individual's CDR record is matched to the ground-truth poverty outcomes for the household they are associated with). In results not shown, we evaluate our CDR-based methods for this prediction task and find that they are slightly more accurate than for predicting household-level outcomes. We choose, however, to focus on household-level poverty prediction here as it is consistent with the unit of targeting from the TUP study.

2.3 Analysis Methods

Identifying the Ultra-Poor Our main analysis evaluates whether machine learning methods leveraging CDR data can identify the ultra-poor in the TUP targeting scheme, and compares their performance to standard asset and consumption-based targeting methods. We evaluate the performance of each method over 10-fold cross validation, stratified to preserve class balance across folds. For each fold, we train a random forest classifier (an ensemble of 100 decision trees) to predict the probability of ultra-poor status from CDR data on the training set, and evaluate its out-of-sample performance on the test set. In each fold, the maximum depth of the random forest is selected from {2, 4, 8, 16, 32} via 3-fold cross validation on the training set. For consistency, we also evaluate the accuracy of asset- and consumption-based targeting methods using 10-fold cross validation.

Each targeting method is evaluated based on classification accuracy, errors of exclusion (ultra-poor households misclassified as non-ultra-poor) and errors of inclusion (non-ultra-poor households misclassified as ultra-poor). To evaluate accuracy metrics for the CDR-based method, we pool out-of-sample predictions across the ten cross-validation folds, so that every household in our dataset is associated with a CDR-based predicted probability of ultra-poor status produced out-of-sample. To account for class imbalance, we evaluate model accuracy by selecting a cut-off threshold for ultra-poor qualification (a maximum wealth index, maximum consumption, and minimum predicted ultra-poor probability based on CDR data) such that each method identifies the correct proportion of ultra-poor households; this cut-off also balances inclusion and exclusion errors. To capture the trade-off between inclusion and exclusion errors for varying values of this threshold, we also evaluate the receiver operating characteristic (ROC) curves for each method and consider the area under the curve (AUC) score as a measure of targeting quality. ROC curves and AUC scores are evaluated for the mean score over the ten cross-validation folds.

To evaluate the accuracy of CDR-based targeting methods with incomlete mobile phone penetration, we also experiment with augmenting our sample with synthetic households to reflect the distribution of phone ownership from the TUP sample. Specifically, we add 122 synthetic households to the sample, none of which own mobile phones, and of which 76 are ultra-poor and 46 are non-ultra-poor. We distribute the synthetic households evenly over cross-validation folds, stratified by true ultra-poor status. We then compare ROC curves for classifying the ultra-poor with the CDR-based method in the original sample to those for classifying the ultra-poor in the augmented sample when (1) all non-phone-owning households are identified as ultra-poor and (2) all non-phone-owning households are identified as non-ultra-poor.

Finally, for the phone-owning sample, we evaluate a combined method that classifies the ultra-poor based on a combination of assets, consumption, and information from CDR. Specifically, we train an unregularized logistic regression to classify the ultra-poor from reported assets and consumption as well as their predicted probability of being ultra-poor using our CDR-based method. Like the other classification methods, we evaluate the combined method based on ROC curves over ten-fold cross validation, as well as pooled accuracy and errors of inclusion and exclusion across folds. For comparison, we similarly evaluate logistic regressions that combine only two of the available data sources (assets plus consumption, assets plus CDR, and consumption plus CDR).

CDR Prediction Across Wealth Measures For direct comparison to previous work on wealth estimation from mobile phone data, we are also interested in estimation of each of the wealth metrics available in our dataset (the asset-based wealth index, consumption, and above/below poverty line indicator, in addition to the ultra-poor indicator) from CDR data. We begin with an in-sample analysis, exploring patterns in the CDR data that may relate to wealth estimation. We compare the distributions of CDR features in ultra-poor to non-ultra-poor households and above vs. below poverty line households, and explore correlations between behavioral CDR features and the asset-based wealth index and consumption.

In our out-of-sample analysis, we evaluate a suite of supervised learning methods, including elastic net regression and three flexible tree-based machine learning methods: a decision tree, a random forest, and XGBoost. For elastic net, the L1 penalty is chosen via 3-fold cross validation from the set $\{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 0, 1\}$ and the mixing parameter is chosen from the set $\{0, 0.2, 0.4, 0.6, 0.8, 1\}$. For each tree-based model, the maximum tree depth is selected via 3-fold cross validation from the set $\{2, 4, 8, 16, 32\}$. Since CDR features are highly collinear, to evaluate issues of overfitting we also implement a linear regression (or logistic regression for classification tasks) using the first principal component of variation in the CDR features as a predictor. As in the previous analysis, we evaluate the accuracy of each machine learning method over 10-fold cross validation. We evaluate methods for predicting binary outcomes (the indicators for being above or below the poverty line and ultra-poor) based on the distribution of AUC scores across 10-fold cross validation. We evaluate methods for predicting continuous outcomes (the asset-based wealth index and consumption) based on the distribution of Pearson's correlation coefficient (r) between true and predicted wealth across 10-fold cross validation.

3 Results

3.1 Identifying the Ultra-Poor

We first compare the accuracy of our CDR-based targeting methods to methods based on assets and consumption for identifying the ultra-poor households from the TUP targeting scheme. Due to class imbalance (27% of the dataset are truly ultra-poor) we evaluate errors of inclusion and exclusion by choosing a threshold where the correct number of households are identified as ultra-poor. We find that the CDR-based method (errors of exclusion and inclusion of 54%) is close in accuracy to methods relying on assets (errors of exclusion and inclusion of 51%) or consumption (errors of exclusion and inclusion of 56%).

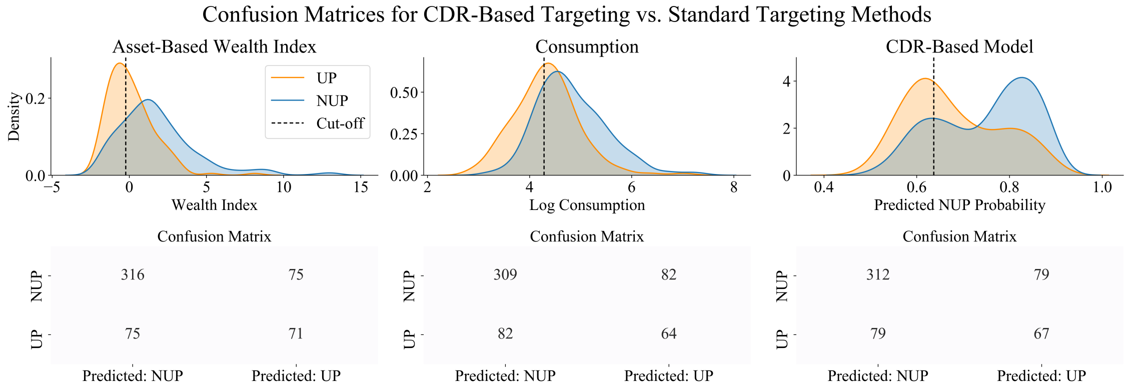

Figure 4: Comparing the predictive accuracy of assets, consumption, and CDR-based methods for identifying the ultra-poor in our 537-household sample. To adjust for class balance, thresholds for classification (shown in dashed black vertical lines) are selected such that the correct number of households are identified as ultra-poor.

To evaluate the trade-off between inclusion errors and exclusion errors resulting from selecting alternative cut-off thresholds, Figure 5 shows the ROC curve associated with each classification method. Area under the curve (AUC) scores are comparable among methods, with assets (cross-validated AUC = .73) slightly superior to consumption and the CDR-based method (cross-validated AUC = .70).

Although the CDR-based method is comparable in accuracy (and presumably much cheaper) than the asset and consumption-based targeting methods, it is limited to households owning a phone. As noted earlier, only 80% of households in the TUP survey own a phone, and only 72% of the ultra-poor own a phone. To evaluate a CDR-based method for identifying the ultra-poor under conditions of incomplete phone ownership, we generate a sample with phone ownership reflective of the demographics of phone ownership in the overall TUP sample by adding synthetic households in proportion to the size of the non-phone-owning ultra-poor and non-ultra-poor. In Figure 5, we consider the ROC curves of methods which use CDR-based targeting for those with phones and classify the remaining households as either all ultra-poor or all non-ultra-poor. We find that classifying all non-phone-owning households as ultra-poor maintains a similar standard of accuracy to our original benchmark (cross-validated AUC = .73), but note that this method may be unrealistic in practice. Classifying all non-phone-owning households as non-ultra-poor yields significantly worse targeting outcomes (cross-validated AUC = .51).

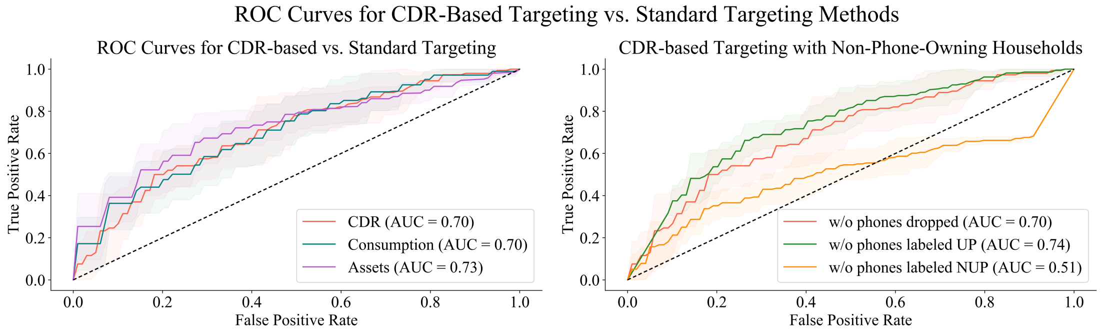

Figure 5: Left: Comparing the ROC curves for classifying the ultra-poor based on of assets, consumption, and CDR-based methods. Right: ROC curves for identifying the ultra-poor with CDR-based methods when households without phones are included in the sample.

Comparison of Errors Across Methods To test for systematic misclassification of a certain type of household, Table 3 compares how the ultra-poor misclassified as non-ultra-poor (errors of exclusion, or false negatives) compare to the correctly classified ultra-poor (true positives), and how the non-ultra-poor misclassified as ultra-poor (errors of inclusion, or false positives) compare to the correctly classified non-ultra-poor (true negatives). We find that ultra-poor households misclassified by CDR data are not significantly different from correctly-classified ultra-poor households, but misclassified non-ultra-poor households tend to have a significantly lower asset-based wealth index than correctly classified non-ultra-poor households.

Panel A: How are the ultra-poor classified?

	Assets			**Consumption**			**CDR**		
	True Pos.	False Neg.	Dif.	True Pos.	False Neg.	Dif.	True Pos.	False Neg.	Dif.
Consumption	4.25	4.41	0.15	3.79	4.75	0.96***	4.25	4.40	0.15
	(0.69)	(0.58)		(0.34)	(0.47)		(0.59)	(0.66)	
Wealth Index	-1.03	1.18	2.21***	-0.24	0.38	0.62*	-0.14	0.32	0.46
	(0.49)	(1.34)		(1.19)	(1.66)		(1.16)	(1.71)	
# Mobile Phones	0.89	1.63	0.74***	1.00	1.48	0.48**	1.13	1.38	0.25
	(0.68)	(1.12)		(0.71)	(1.14)		(0.69)	(1.19)	
N	71	75	146	64	82	146	67	79	146

Panel B: How are the non-ultra-poor classified?

	Assets			**Consumption**			**CDR**		
	True Neg.	False Pos.	Dif.	True Neg.	False Pos.	Dif.	True Neg.	False Pos.	Dif.
Consumption	4.86	4.65	-0.21*	5.04	4.01	-1.03***	4.84	4.77	-0.07
	(0.68)	(0.66)		(0.59)	(0.25)		(0.70)	(0.63)	
Wealth Index	2.49	-1.08	-3.57***	2.03	0.97	-1.05**	2.00	1.07	-0.93**
	(2.42)	(0.50)		(2.73)	(1.80)		(2.71)	(1.96)	
# Mobile Phones	2.10	0.96	-1.14***	1.99	1.46	-0.53**	1.97	1.53	-0.44*
	(1.43)	(0.76)		(1.47)	(0.99)		(1.45)	(1.11)	
N	316	75	391	309	82	391	312	79	391

Table 3: Above: Differences between ultra-poor households correctly classified as such and those misclassified as non-ultra-poor (errors of exclusion). Below: Differences between non-ultra-poor households correctly classified as such and those misclassified as ultra-poor (errors of inclusion). Standard deviations are shown in parentheses; significance of difference in means between the samples is determined with a two-sided t-test (* indicates p < .05, ** indicates p < .01, *** indicates p < .001).

To evaluate the consistency of misclassifications across methods, Table 4 displays the overlap in errors of exclusion and inclusion between methods. Note that overlap rates should be interpreted relative to the expected overlap in errors for random classifiers with the same cut-off threshold for ultra-poor classification. More specifically, based on our selection of thresholds such that 27% of the sample is identified as ultra-poor, our three classifiers misidentify 19-21% of the non-ultra-poor and 51-56% of the ultra-poor. If these classifiers were random, we would expect approximately 20% overlap in errors of inclusion and 55% overlap in errors of inclusion. Our results therefore suggest that the three classifiers misidentify the same households at a rate only slightly above random.

Panel A: Errors of Exclusion

	Assets	**Consumption**	**CDR**
Assets	100%	57%	52%
Consumption	63%	100%	66%
CDR	55%	63%	100%

Panel B: Errors of Inclusion

	Assets	**Consumption**	**CDR**
Assets	100%	26%	27%
consumption	28%	100%	20%
CDR	28%	20%	100%

Table 4: Above: Overlap between ultra-poor households that are misclassified as non-ultra-poor (errors of exclusion) for each targeting method. Below: Overlap between non-ultra-poor households that are misclassified as ultra-poor (errors of inclusion).

Combining Data Sources Since the asset-based, consumption-based, and CDR-based classifications of the ultra-poor are not perfectly overlapping, a natural extension of our cross-method comparison is to identify the ultra-poor based on a combination of an asset-based wealth index, consumption, and CDR data. As shown in Table 5, we find that a logistic regression using the wealth index, consumption, and predicted ultra-poor probability from CDR data is more accurate (cross-validated AUC = .79) than methods using any one data source (cross-validated AUC = .70-.73) or any two of the data sources (cross-validated AUC = .76-.77) for identifying the ultra-poor in our 537-household sample.

	AUC	**Accuracy**	**False Pos.**	**False Neg.**
Assets	.73 (.06)	72%	75	75
Consumption	.70 (.07)	69%	82	82
CDR	.70 (.06)	71%	79	79
Assets + Consumption	.76 (.02)	74%	70	70
Assets + CDR	.77 (.06)	75%	67	67
Consumption + CDR	.76 (.06)	73%	73	73
Assets + Consumption + CDR	.79 (.04)	76%	65	65

Table 5: Accuracy of classifying the ultra-poor in our 537-household sample based on an asset-based wealth index, consumption, and CDR data, and combinations of the three data sources. AUC scores are evaluated for the mean over 10-fold cross validation, with standard deviations in parentheses. Accuracy, false positives (errors of inclusion), and false negatives (errors of exclusion) are evaluated over all folds together.

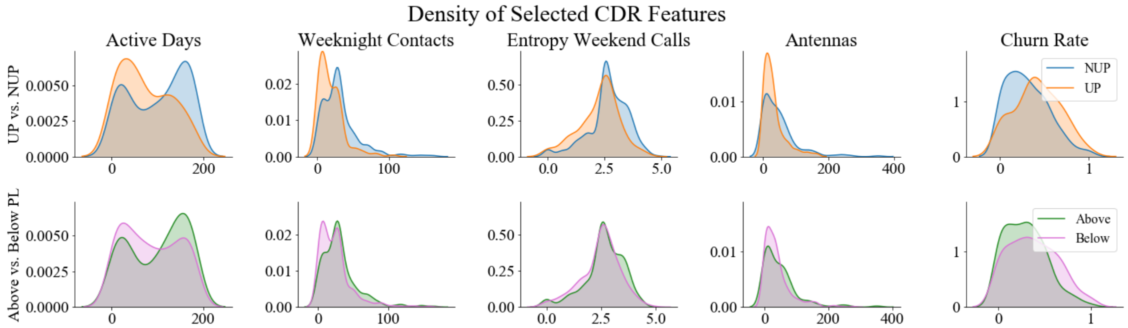

Figure 6: Comparing distributions of selected CDR features for ultra-poor vs. non-ultra-poor households (above) and above vs. below poverty line households (below). From left to right: number of active days of phone use, number of unique weeknight contacts (call or text), entropy of weekend contacts (call or text), number of unique antennas, and the churn rate (level of dissimilarity in spatial patterns between weeks, based on antenna locations).

3.2 Comparison of Predictive Accuracy Across Wealth Measures

For comparison with past work on poverty estimation from CDR data, we also evaluate the ability of CDR-based methods to predict the asset-based wealth index, consumption, and an indicator for being above or below the poverty line in our sample. We begin with an in-sample analysis, comparing the relationship between behavioral features from CDR data and ground-truth survey-based measures of welfare across our 537-household sample. We then evaluate the accuracy of machine learning models leveraging CDR data for out-of-sample prediction of the four wealth measures based on CDR data.

In-Sample Analysis We perform a two-sided t-test to compare the mean of CDR features in the ultra-poor and non-ultra-poor samples, and find that 237 of the 623 CDR features have significantly different means in the two populations on a 0.05 level. With the same methodology, we find that 158 CDR features have significantly different means for populations above and below the poverty line. Kernel density estimates comparing the distributions of five of these behavioral features are included in Figure 6.

We also find that there is significant correlation between many of the behavioral CDR features and the continuous measures of wealth (wealth index and consumption);

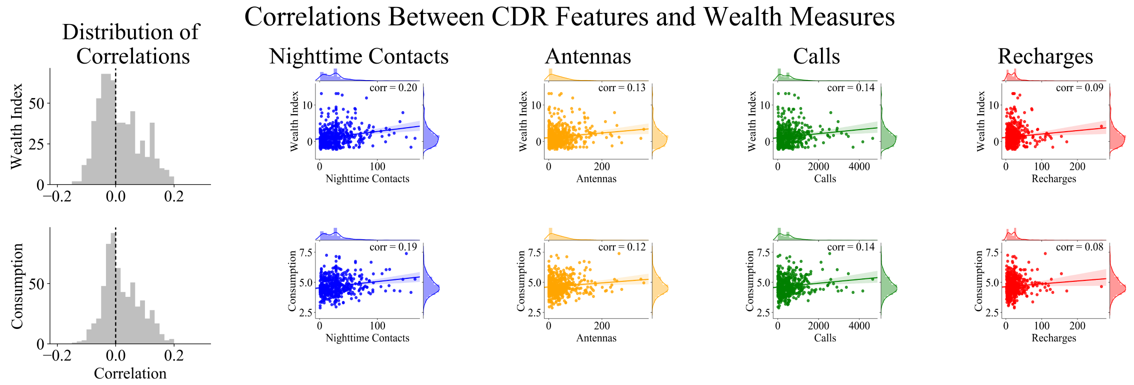

Figure 7: Summary of correlations between CDR features and continuous wealth measures (wealth index above and consumption below). Left: Distribution of correlations between behavioral features extracted from CDR data and ground truth wealth measures. Right: Scatterplots showing the relationship between selected CDR features and ground truth wealth measures, with linear trend lines. From left to right: number of unique nighttime contacts (call or text), number of unique antennas, total calls, and total recharges.

the distribution of correlations is shown in Figure 7. However, we find that many of the CDR features most correlated with wealth are highly collinear. Moreover, many of the most useful CDR features are structurally related to one another: among the top 10 features most correlated with the wealth index, for example, nine are related to number of contacts. Surprisingly, in comparison to previous work (Steele et al., 2017), we find weak correlations between recharge behavior and wealth (for example, the correlation between number of recharges and the wealth index is 0.09; there is no correlation between the average amount recharged and wealth index). Figure 7 shows the relationship between several intuitively interpretable CDR features and the wealth index and consumption in our sample.

Out-of-Sample Prediction While the differences in CDR feature distributions between ultra-poor and non-ultra-poor households and the correlations between CDR features and the wealth index suggest that there is promise for predicting wealth from behavioral CDR features in this dataset, it is possible that the in-sample model is overfit and has no out-of-sample predictive power. Out-of-sample prediction, in this case through cross-validation, is a better measure of the model's ability to predict an arbitrary individual's wealth from their mobile phone history.

As shown in Table 6, and in contrast to previous work, we find little predictive power for an asset-based wealth index (cross-validated r = .03-.17), consumption (cross-validated r = .01-.13), or the poverty line indicator (cross-validated AUC = .47-.58). As above, we find significantly more predictive power for the ultra-poor indicator (cross-validated AUC = .62-.70).

	Elastic Net	**Decision Tree**	**Random Forest**	**XGBoost**	**OLS on 1st PCA Component**
Panel A: Continuous wealth measures (r)					
Wealth Index	**0.17 (0.13)**	0.03 (0.09)	0.11 (0.14)	0.16 (0.14)	0.14 (0.13)
Consumption	0.09 (0.15)	0.01 (0.11)	0.09 (0.14)	0.03 (0.15)	**0.13 (0.15)**
Panel B: Binary wealth measures (AUC)					
Below Poverty Line	0.52 (0.07)	0.56 (0.05)	0.57 (0.08)	0.47 (0.09)	**0.58 (0.08)**
Ultra-Poor	0.64 (0.07)	0.62 (0.06)	**0.70 (0.06)**	0.67 (0.06)	0.65 (0.07)

Table 6: Accuracy for predicting four measures of wealth from CDR data. Continuous wealth measures (asset-based wealth index and consumption) are evaluated based on correlation (r) between true and predicted wealth, binary wealth measures (indicator for being below the poverty line and indicator for being ultra-poor) are evaluated for AUC score. Results presented are averages over 10-fold cross validation, with standard deviations in parentheses. The most accurate machine learning method for predicting each measure is bolded.

4 Discussion

Our key finding is that in a sample of 537 phone-owning households in a set of poor villages in one province of Afghanistan, machine learning methods leveraging behavioral indicators computed from CDR data are as accurate as standard asset- and consumption-based methods for identifying ultra-poor households. Further, we find that methods combining information on assets and consumption with CDR perform better than any of the methods using a single data source. These results extend past work on wealth estimation from mobile phone data to suggest that CDR and other digital trace data could be used in practice to target anti-poverty programs or other development interventions with some accuracy. Moreover, recent reviews of standard field-based poverty targeting schemes find that targeting is limited by low-quality ground truth data on poverty across programs and regions (Brown et al., 2018; Coady et al., 2004). CDR-based methods like the one presented here could provide a lower-cost complement to standard targeting methods without sacrificing accuracy, and would be particularly useful in unstable or conflict-affected regions where field-based targeting is near-impossible.

We emphasize, however, that CDR-based targeting applies only to households that own a mobile phone, so CDR-based methods are inherently limited by mobile phone penetration rates. As mobile phone penetration rates continue to rise in the developing world (GSMA, 2017), CDR-based methods will become increasingly relevant, but an understanding of the distribution of mobile phone ownership among the poor is essential to practical deployment of the methods introduced here. Given this limitation and our promising results on combining CDR data with standard survey measures for increased classification accuracy, CDR-based methods may be best deployed in conjunction with standard targeting methods so that survey-based data on poverty is complemented by digital trace data.

As a secondary finding, we note that in contrast to the relative success of CDR-based methods for identifying the ultra-poor in this sample, our CDR-based methods have little predictive power for estimating an asset-based wealth index or consumption among the TUP households. We hypothesize that there may be aspects of the ultra-poor classification that make it easier to estimate than other wealth measures. Recall that ultra-poor classification in our sample is based on a community wealth ranking followed by verification of qualifying criteria that cover several aspects of poverty and deprivation. It is possible that ultra-poor status is simply a less noisy than the other ground-truth wealth measures: it is based on multiple components, which helps reduce measurement error, and qualifying criteria were validated in-person by ministry staff, which adds an element of objectivity. An alternative hypothesis is that the ultra-poor are a segment of the population that is easier to identify through behavioral traces in mobile phone data. Alatas et al. (2012) suggest that communities apply a concept of wealth beyond income, consumption, or assets in their wealth rankings; it could be that this concept is better reflected in mobile phone traces than the standard quantified welfare measures. Further, the additional qualifying criteria for ultra-poor status cover heterogeneous aspects of welfare relating to income, housing, assets, and education. It is possible that this holistic measure of wealth is better reflected in the digital footprint of the ultra-poor than specific asset or consumption-based indicators.

We also present several hypotheses for why the prior success of CDR-based wealth index prediction methods (Blumenstock, 2018; Blumenstock et al., 2015) does not replicate to the sample we analyze. First, machine learning methods perform better with large amounts of training data, and the sample size of this dataset (537 households) is significantly smaller than the samples used in previous CDR work, which include datasets from Rwanda with 856 individuals (Blumenstock et al., 2015) and Afghanistan with 1,234 households (Blumenstock, 2018). Second, the components used to construct wealth indices vary between wealth prediction projects, and it is possible that the wealth index used here is less useful (or less predictable) than those used in past work. Although early papers indicate that asset-based wealth indices are robust to changes in the underlying set of components (Filmer & Pritchett, 2001; Wagstaff & Watanabe, 2003), subsequent work has found that household wealth categorization can be sensitive to the assets selected for inclusion (Houweling et al., 2003; Michelson et al., 2013). Third, it may be that structural differences between the type of sample we analyze here and those used in previous work can account for differences in predictive power. Most significantly, the sample of households here is drawn from poor villages in a single rural province of Afghanistan, whereas previous samples have included households across the socioeconomic spectrum from several or all provinces in a country. We hypothesize that much of the variation picked up by CDR-based models in previous work captures stark rich-poor inequalities and urban-rural divides; it may be much more challenging to train models that pick up on variation in a relatively homogeneous sample of poor households using CDR data. This hypothesis is consistent with previous work showing that distinguishing between the very wealthy and the very poor is easier than discerning finer variations in welfare, even with standard survey-based measures of wealth (Karlan & Thuysbaert, 2019).

Our results suggest that there is potential for using CDR-based methods to target vulnerable members of society for economic aid or interventions, significantly reducing program targeting overhead and costs. Our results also provide evidence that CDR-based methods could complement and enhance existing survey-based methods for better targeting accuracy. However, we emphasize that, as demonstrated by their low predictive power for wealth and consumption in our sample, CDR-based methods are not one-size-fits-all and cannot be blindly adapted to new contexts without careful tuning and validation. There is much room for future work demonstrating the use of CDR-based wealth prediction and targeting in new contexts to build the literature on when and how these methods can be effectively and appropriately leveraged for policymaking.

References

Alatas, V., Banerjee, A., Hanna, R., Olken, B., & Tobias, J. (2012). Targeting the poor: Evidence from a field experiment in Indonesia. *American Economic Review*, *102*(4), 1206-1240.

ALCS. (2017). *Afghanistan living conditions survey 2016-17.*

Banerjee, A., Duflo, E., Chattopadhyay, R., & Shapiro, J. (2007). Targeting efficiency: How well can we identify the poor? *Institute for Financial Management and Research Centre for Micro Finance, Working Paper Series No. 21*.

Bedoya, G., Coville, A., Haushofer, J., Isaqzadeh, M., & Shapiro, J. (2019). No household left behind: Afghanistan targeting the ultra poor impact evaluation. *World Bank Policy Research Working Paper*, *8877*.

Blumenstock, J. (2016). Fighting poverty with data. *Science*, *353*, 753-754.

Blumenstock, J. (2018). Estimating economic characteristics with phone data. *American Economic Review: Papers and Proceedings*, *108*, 72-76.

Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone data. *Science*, *350*, 1073-1076.

Blumenstock, J., Maldeniya, D., & Lokanathan, S. (2017). Understanding the impact of urban infrastructure: New insights from population-scale data. *Proceedings of the 9th IEEE/ACM International Conference on Information and Communication Technologies and Development, ICTD '17).*

Brown, C., Ravallion, M., & van de Walle, D. (2018). A poor means test? econometric targeting in Africa. *Journal of Development Economics*, *134*, 109-124.

Coady, D., Grosh, M., & Hoddinott, J. (2004). Targeting outcomes redux. *The World Bank Research Observer*, *19*(1).

De-Arteaga, M., Herlands, W., Neill, D., & Dubrawski, A. (2018). Machine learning for the developing world. *ACM Transactions on Management Information Systems (TMIS)*, *9*(2).

Decuyper1 et al., A. (2014). Estimating food consumption and poverty indices with mobile phone data. , *arXiv preprint arXiv:1412.2595*.

de Montjoye, Y., Rocher, L., & Pentland, A. (2016). bandicoot: a python toolbox for mobile phone metadata. *Journal of Machine Learning Research*, *17*, 1-5.

Filmer, D., & Pritchett, L. (2001). Wealth effects without expenditure data—or tears: An application to educational enrollments in states of India. *Demography*, *39*, 115-132.

Grosh, M., & Baker, J. (1995). Proxy means tests for targeting social programs: Simulations and speculation. *Living Standards Measurement Study Working Paper No. 118*.

GSMA. (2017). *Mobile economy.* https://www.gsma.com/mobileeconomy/global/2017/.

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

Hernandez, M., Hong, L., Frias-Martinez, V., & Frias-Martinez, E. (2017). Estimating poverty using cell phone data: evidence from Guatemala. , *World Bank Policy Research Working Paper Series No. 7969*.

Houweling, T., Kunst, A., & Mackenbach, J. (2003). Measuring health inequality among children in developing countries: does the choice of the indicator of economic status matter? *International Journal for Equity in Health*, *2*(8).

Jerven, P. (2013). *Poor numbers.* Cornell University Press.

Karlan, D., & Thuysbaert, B. (2019). Targeting ultra-poor households in Honduras and Peru. *The World Bank Economic Review*, *33*(1), 63-94.

Meyer, B., Mok, W., & Sullivan, J. (2015). Household surveys in crisis. *Journal of Economic Perspectives*, *20*(4), 199-256.

Michelson, H., Muniz, M., & DeRosa, K. (2013). Measuring socio-economic status in the Millennium Villages: The role of asset index choice. *The Journal of Development Studies*, *40*(7), 917-935.

Schmid, T., Bruckschen, F., Salvati, N., & Zbiranksi, T. (2017). Constructing sociodemographic indicators for national statistical institutes by using mobile phone data: estimating literacy rates in Senegal. *Journal of the Royal Statistical Society*, *180*, 1163-1190.

Sheehan, E., Meng, C., Tan, M., Uzkent, B., Jean, N., Lobell, D., … Ermon, S. (2019). Predicting economic development using geolocated wikipedia articles. *Proceedings of the 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Steele, J., Sundsøy, P., Pezzulo, C., Alegana, V., Bird, T., Blumenstock, J., … Bengtsson, L. (2017). Mapping poverty using mobile phone and satellite data. *Journal of the Royal Society Interface*, *14*.

Wagstaff, A., & Watanabe, N. (2003). What difference does the choice of SES make in health inequality measurement? *Health Economics*, *12*, 885-890.