# The potential and limitations of big data in development economics: The use of cell phone data for the targeting and impact evaluation of a cash transfer program in Haiti

Oscar Barriga-Cabanillas[1], Joshua Blumenstock[2], Travis J. Lybbert[1], and Daniel Putman[3]

[1] *University of California, Davis*
[2] *University of California, Berkeley*
[3] *Innovations for Poverty Action*

August 17, 2021

Please click here for the latest version of this paper.

**Abstract**

Call Detail Records (CDRs) can reveal surprising details about mobile phone users' wealth when subject to machine learning techniques. The predictive power of these metadata has opened new frontiers of empirical analysis, particularly in developing countries. Can they identify potential beneficiaries of means-tested programs and thereby enhance targeting? Can they detect *changes* in well-being well enough to use in impact evaluation? We address these questions using an emergency cash transfer program in Haiti. A conventional regression discontinuity-based impact evaluation using survey data shows positive impacts of the transfer on household food security and dietary diversity and provides the benchmark for testing a CDR-based evaluation alternative for the same cash transfer. We extract features from the CDR data and use them in machine learning models to replicate the targeting of the transfers and then substitute for the survey-based outcomes before re-estimating impact evaluation results. Predicted outcomes are too noisy to differentiate between targeted beneficiaries or to detect changes in food security, and the CDRs therefore fail as an alternative basis for either targeting or evaluation. In a postmortem, we identify likely causes of suspects in this prediction failure and discuss implications and limitations for predicting welfare outcomes using big data in poor countries.

*JEL Classification*:
*Keywords*:

# 1 Introduction

Understanding whether a given product, program, or intervention improves livelihoods is as important as it is challenging. While established impact evaluation techniques can provide credible evidence of impact, they can also be expensive because they typically require active survey-based data collection. As individuals – rich and poor – generate data through their use of information and communication technologies, new opportunities emerge to leverage these passive records to understand behavior, preferences, and well-being. Blumenstock, Cadamuro and On (2015) first demonstrated such an opportunity. Using metadata associated with mobile phone usage, the study predicted wealth levels with surprising accuracy and triggered a cottage industry of similar machine learning-based approaches to extracting meaningful signals from these Call Detail Records (CDRs). This empirical success raised a host of intriguing possibilities. Two such possibilities, here phrased as questions, motivate our analysis in this paper: Can CDRs identify potential beneficiaries of means-tested programs and thereby enhance the cost-effectiveness of targeting? Can CDRs also capture *changes* in wealth or well-being and thereby enable near real-time evaluation of the impact of new programs, products or policies?

In 2016, the World Food Program (WFP) responded to a third consecutive year of drought in Haiti with an emergency unconditional cash transfer program to protect household food security. These transfers, which targeted the poorest households in drought-hit areas, consisted of three consecutive monthly disbursements via mobile money. The value of these transfers was significant, each representing 130% of monthly average per capita income.[1] We collaborated with WFP to evaluate the impact of the cash transfers on household food security, dietary diversity and consumption. Additionally, we partnered with the mobile network operator that facilitated these mobile money-based transfers to obtain access to cellphone transactions data of the people that participated in the targeting and evaluation phase of the program. In this paper, we provide a conventional survey-based impact evaluation of the program and then use these results as a benchmark for an alternative CDR-based impact evaluation of the same program.

As in the benchmark evaluation, we leverage the targeting threshold used by the WFP to implement a regression discontinuity (RD) design that estimates the impacts of the unconditional cash transfer program using survey data. We find the program increased household food consumption and food expenditure in 0.35 and 0.31 standard deviations, respectively. For the CDR-based alternative, we use feature engineering and a structured combinatorial method to generate several hundred CDR metrics from the volume, intensity, timing, direction, and location of communication, as well as the household's position in the cellphone network.[2] With this rich set of features in hand, we deploy machine learning algorithms to predict the targeting status of beneficiaries, as well as household food security outcomes, which we then use to replicate the RD impact evaluation.

At the outset, this cash transfer program seemed to provide an ideal setting in which to test the viability of a CDR-based targeting and evaluation. On the targeting side, WFP uses a standard and simple scorecard to quickly rank would-be beneficiaries in terms of wealth, and CDRs have shown promise in predicting precisely such outcomes. On the evaluation side, the monthly transfers were large relative to household income and were sent for three consecutive months, so it is reasonable to expect the impacts to readily detectable. The use of a scorecard threshold for targeting beneficiaries enables RD identification of these impacts for both survey- and CDR-based measures of well-being.

[1] Based on the 2012 ECVMAS households survey.

[2] For our main results we use Bandicoot, an open-source Python toolbox. See section 4.1 for details.

Finally, the transfers were distributed via mobile money in collaboration with Haiti's largest mobile network operator, which enjoys a dominant 80% market share in the country. This facilitates our ability to link scorecard responses and cash transfers to specific mobile network users.

Despite these advantages, we were not able to replicate the survey-based RD results using these metatdata or the targeting. Our failure to replicate these impact evaluation results runs deep as all of our the predictions of the household-level food security outcomes using our CDR features and machine learning algorithms present low levels of accuracy. That is, in the horse race we set out to run between conventional survey-based and novel CDR-based impact evaluation, the CDR horse stumbled out of the gate, which ended the race before it really began. There is, however, much to learned from this failed prediction attempt about the limits of CDR-based analyses such as the targeting and impact evaluation applications that motivate this work.

In a postmortem assessment, we discuss why CDR predictions failed in this seemingly-ideal setting. First, given high levels of poverty and vulnerability, cellphone ownership and, in some regions, access to the cellphone network is low. Only 34% of the participants in the scorecard survey reported owning a cellphone, with cellphone ownership concentrated among less vulnerable individuals. Solely relying on cellphone data limits the population we can observe (target) and can introduce bias as higher vulnerability levels correlate with lower levels of cellphone ownership. Furthermore, the practice of sharing devices, more common in low income households, creates additional challenges as it dilutes any potential signal about wealth levels in the data. Second, the targeted geographic areas in this program were pre-selected based on drought-induced and general vulnerability. Within these pre-selected communes, the WFP administered the scorecard to identify specific beneficiary households. While this two-tiered approach can improve targeting, it restricts the statistical variation we can observe in household wealth and outcomes and limits our ability to predict outcomes using CDRs. Third, the exogenous eligibility cut-off is ideal for RD impact evaluation, but its primary identifying assumption is that those on either side of the cut-off are statistically indistinguishable, limiting the variation in the CDR usage patterns on both sides of the eligibility cut-off. Finally, the primary outcomes of interest in this study are flow variables (e.g., food consumption) rather than stock variables, and the promising evidence of CDR predictions has so far concentrated on the prediction of stock variables such as assets and wealth.

We empirically explore these postmortem considerations from a variety of perspectives. We compare our data to nationally-representative data from Haiti to understand the implications of tight (effective) targeting. The statistical variation in the primary outcomes for our WFP sample is indeed significantly lower than for the general Haitian population. The scorecard sample, unsurprisingly, is more vulnerable with food insecurity levels double the national average and the total food expenditure distribution below the 70th percentile of the overall distribution. These factors limit the variation that machine learning models can use to recreate both the targeting and the outcome variables. With additional tests, we find that CDR predictions improve as we increase the variability in outcomes. To explore our ability with greater variation in the data to predict flow variables such as food consumption, we obtained informed consent from a subsample of the participants in the nationally representative survey and predicted their wealth levels and food expenditures. We show that while we can predict wealth levels relatively well in this broader sample, the predictions are much less reliable when it comes to flow outcomes like consumption.

In the next section, we provide a broader introduction to the use of passive data in development economics and then describe additional details about the WFP cash transfer program. In the methods section, we present the RD design and the various data sources we use in our analysis.

Section 5 provides the survey-based impact estimates that serve as a benchmark for the alternative CDR-based impact estimation. In section 6, we present our CDR-based predictions and discuss the associated implications for CDR-based targeting and impact evaluation. We conclude with a detailed postmortem discussion and broader reflections.

## 2 Passive Data in Development Economics

Reliable and up-to-date data is a key factor in the effective implementation of public policies. The absence of official data is more acute in poor and developing countries, forcing governments to implement public programs with limited information (Blumenstock, 2016). Collecting these data is expensive in both monetary and administrative terms, and in many situations, it cannot be produced with the necessary speed to attend to extraordinary demands such as relief programs following a natural disasters.

The last few years have seen a large increase in the amount of data produced daily by private digital transactions. This digital footprint contains information on billions of individuals, including those living in poverty. One of these novel sources of information is the transaction data that mobile phone subscribers create every time they make or receive a call. Users do not create this information to contribute to any policy or research objective; instead, the data are passively created as part of the regular network operation. While these datasets may be a step removed from the on-the-ground outcomes policy-makers care about, advances in feature engineering and machine learning allow us to use this information to circumvent data limitations.[3]

With more than 95% of the global population with mobile-phone coverage, CDR data create a unique opportunity to address a major challenge policy-makers and researchers face in contexts where reliable quantitative data are scarce (Blumenstock, Cadamuro and On, 2015; Blumenstock, 2016). The first applications of these methods studied how regional socioeconomic conditions create signals that can be detected on aggregated the CDR data trail, in particular regional poverty levels (Hernandez et al., 2017) and food security indicators (Decuyper et al., 2014). In areas where official statistics are not recent, CDR data and machine learning methods provide an tool for updating indicators in-between household survey rounds.

Building on the previous results, several papers demonstrate that mobile phone data can be use to estimate outcomes at the individual level. The logic behind this approach is that phone usage captures many behaviors that have some intuitive link with socioeconomic indicators, allowing researchers to differentiate the most vulnerable households. As described by Björkegren and Grissen (2015), a phone account is a financial account that captures part of a person's expenditure, with most of the calling behavior being an indicator of how a person manages his expenses. For example, individuals with different income streams are likely to have call patterns indirectly linked with socioeconomic status. These include when a person uses his phone, geographic mobility, and the diversity of the calling network. Blumenstock, Cadamuro and On (2015) use CDRs and data from a nationally representative survey in Rwanda to demonstrate how an individuals' socioeconomic status can be inferred from CDR transactions, while Blumenstock (2018) demonstrates similar techniques can also be used to accurately predict wealth levels in an Afghan sample. In both

[3]There are plenty of remote sensing applications that rely on data other than cellphone records. For example, Jean et al. (2016) uses satellite imagery, nightlights to infer poverty in Nigeria, Tanzania, Uganda, Malawi, and Rwanda, and Goldblatt et al. (2018) use satellite imagery and remote sensing data to characterize land cover and urbanization in India, Mexico, and the United States.

applications, the authors find their behavioral features can explain about 46% of the total variation of a wealth composite index.[4]

The success in predicting individual-level outcomes potentially opens the door to new applications to improve public policy.[5] The first application is in complementing/replacing conventional methods used to target social programs. The importance of targeting to make anti-poverty programs more cost-effective has been widely studied (Alatas et al., 2012; Coady, Grosh and Hoddinott, 2004; Brown, Ravallion and Van de Walle, 2016). Common targeting protocols rely on administrative and survey-based measures of assets or consumption. These information is not available in many developing, and if it is available, usually has reliability problems. Moreover, for most practical applications to small and medium scale programs collecting this information adds a large administrative cost. The passive nature of CDR data may provide an additional tool to improve targeting efforts with shorter deployment times. Aiken et al. (2020) studies the extent that mobile phone data can be used to identify ultra-poor households in the context of an anti-poverty program in Afghanistan. In their study, a community wealth ranking and a deprivation survey provide the ground-truth data that classifies a household as ultra-poor.[6] An advantage of this study is that a household survey was collected independently of the ultra-poor classification survey, providing a much richer set of indicators to compared the CDR-based method accuracy. Using six months of cellphone data, the authors find CDR-based methods have an accuracy of 70%, very similar to standard survey-based consumption (73%) and asset-based measures (70%). Combining the information from the CDR data and the household survey into a single classification problem provides the best results, with an accuracy of 79%. However, as discussed in the paper, using several data sources might not a possibility in most real-life applications.

The second potential application is the use of mobile phone transaction records to enable new approaches to impact evaluation and program monitoring. An initial application of CDR data to complement impact evaluation studies used the data as part of the identification strategy. One example is (Olivieri et al., 2020) who uses cellphone records to identify the geographical distribution of Venezuelan migrants in Ecuador to understand their impact on labor market outcomes. Other papers have used the mobility patterns captured by the usage of mobile phones to understand how local-level policies affected refugees' mobility in Turkey (Beine et al., 2019), population movement in the wake of natural disasters (Wilson et al., 2016), and the spread of infection disease (Wesolowski et al., 2012; Milusheva, 2020). We go one step further by trying to use cellphone data to estimate changes in welfare over time in the context of an impact evaluation study.

# 3 The EMOP program

In 2016, in response to a third consecutive year of drought in Haiti, the WFP conducted an Emergency Food Assistance Program (EMOP), with the objective to improve food security of households in the affected areas. During the program's lifespan, it provided 46,163 households

[4]Applications of these methods already exist for commercial applications, for example, providing credit scores in settings where credit bureau are not present (Björkegren and Grissen, 2015).

[5]Blumenstock, Cadamuro and On (2015) provides a complete list of potential applications of CDR data for social research.

[6]The ranking used the following questions 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. The household does not own any productive assets.

with three monthly transfers of 3050 HTG (around 50 USD) delivered via mobile money; a large sum with each payment roughly equivalent to the minimum wage or 130% of monthly average per capita income.[7]. The rapid-response nature of the program guided key decisions on the targeting and implementation protocols, including the use of a simple scorecard mechanism to measure vulnerability and relying on the existing mobile money infrastructure to deliver funds.

The EMOP used a three-step targeting process to select beneficiaries. During the first stage the program's geographical targeting was determined. The WFP, together with the Haitian government and other organizations, conducted a nationwide emergency food security assessment to determine the areas with the largest concentration of food-insecure households.[8] To avoid duplication of aid initiatives, the WFP focused on areas where no other actors were intervening at the time. This led to the prioritization of 21 rural communes across the country.[9]

In the second stage, the WFP engaged with local stakeholders to construct a list of the families living in the area, build trust in the program's implementation, and tailor the scorecard questionnaire to local conditions. The result of this process was the sampling frame used to conduct the scorecard survey, and slight modifications to the questionnaires, tailoring them to the specific farm animals present in each region.[10]

The selection of beneficiaries took place during the third stage of the program's targeting process. The WFP administered a short survey, that official documents call scorecard survey, and used the information to calculate the value of a vulnerability index for every household interviewed. The objective of the vulnerability index is to measure how susceptible a household is to suffer from food insecurity and hunger. A higher number corresponds to higher levels of vulnerability, and therefore, greater need of the program's assistance.[11] The overall score of a household is an integer number based on responses to questions (most of which were simple "yes" or "no" questions) about land, animal ownership, and the presence of vulnerable people in the household, with possible scores ranging from zero to eight.[12] The scorecard exercise is a standard practice for the WFP operations worldwide, especially when rapid delivery of funds is necessary to provide aid.

The scorecard vulnerability index follows a proxy means test approach towards identifying the most vulnerable households in the absence of consumption based poverty indicators (Grosh and Baker, 1995). As explained by Gazeaud (2020), a complete proxy means test requires a two step process. First, an in-depth survey is administered to a sample of households to collect data on consumption as well as some easily observable and verifiable correlates of consumption. These data are used to estimate a regression of log consumption per capita on correlates of consumption.

[7]Based on the 2012 ECVMAS households survey.

[8]The WFP carried out the data collection together with the Haitian National Coordination for Food Security and in collaboration with its partners (FAO, FEWS NET, OCHA and others) between November and December 2015. The assessment measured the impact of the drought on the number of moderately and severely food insecure households. The analysis also entailed assessing the functioning of markets and their response capacity in case of providing aid in the form of cash. See https://docs.wfp.org/api/documents/808de753fc264d318ec818204f5c71bc/download/?$_g a$ = 2.133089166.1686038113.1625782808 − 592851595.1621481902

[9]These communes are located in prioritize the Nord-Est, Artibonite, Nord, Centre, Ouest, Nippes, Grande-Anse, Sud and Sud-Est Departments. Full details on the program's implementation are available at: https://docs.wfp.org/api/documents/a2e0fd284cf2431289120c36f83ffe28/download/

[10]Depending on the region, questions about animal ownership included different combinations of cows, horses, sheep, goats, and pigs.

[11]The scorecard exercise is part of the Household Economic Analysis (HEA) used by the WFP. The HEA arose from a collaboration in the early 1990s as a tool to improve the FAO ability to predict short-term changes in a population's access to food, see (Holzmann et al., 2008).

[12]This includes children, pregnant and lactating women.

Second, a short survey is administered to all potential beneficiary households to collect information on the same correlates of consumption, and compute a new score using short survey information. Unfortunately, the EMOP implementation lacks the necessary data to implement the first stage of the identification and weighting of the correlates of the indicators, relying instead on previous WFP experience in to choose the variables used in the construction of the index.

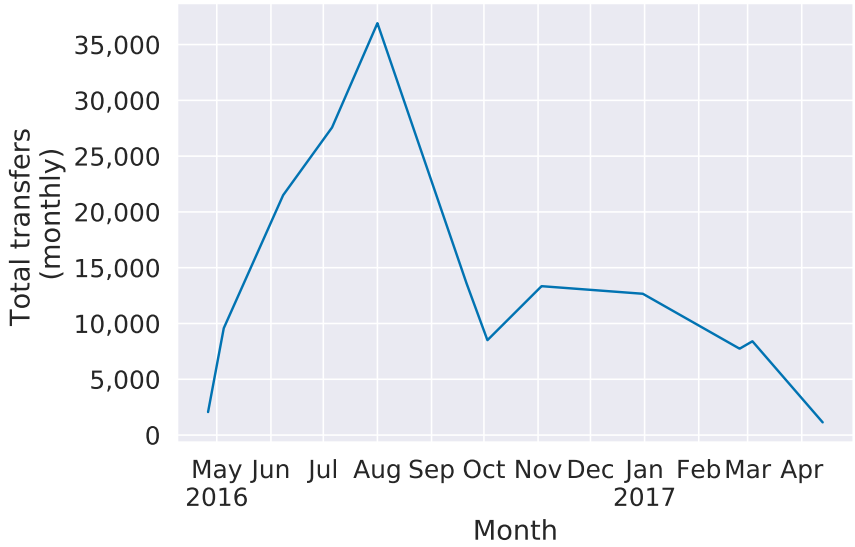
The implementation of the program used the vulnerability score to target aid. At the commune level, the cut-off to be eligible for the program depended on the total funds assigned to the region.[13] Since the eligibility cut-offs were selected after scores were assigned taking into account the total funds assigned to each area, households had not opportunities to try to manipulate their vulnerability score and precise treatment status. This constitutes the basis for the evaluation of the program using a Regression Discontinuity design, see section 5. Under the selection criteria, there are several instances where a person with a specific score did not receive aid given the commune where he lived but would have been eligible with the same score in another area. Most communes required a score between three to five to grant eligibility (81% of the covered population), with some communes requiring a vulnerability score as high as 6 (17% of the covered population). We provide more details on how this affects our replication of the eligibility criteria using CDR data and machine learning methods in section 6.

The use of mobile money to distribute funds makes every transaction linked to a cellphone number. If beneficiary household did not have a SIM card to receive transfer payments, WFP and Digicel provided one.[14] Records show a total of 63,201 phone numbers receiving a transfer during the period. As Figure 1 shows, transfers occurred at different times over the period, reaching their peak in August 2016; however, within geographical locations (e.g., commune), transfer patterns tend to be similar between households.

[13]The total funds assigned to each communed where the results of the total population in an area, with communes with higher population levels receiving more funds independently of the distribution of the vulnerability scores in the commune.

[14]We have access to all the mobile money transaction records, where the EMOP transfers are just one of the many transactions.

Figure 1: Count of WFP Transfers: 2016-2017




Note: Author's calculations using mobile money transaction logs.

# 4 Methods

In this section, we discuss the validity of the RD design and explain how we combine the CDR data with the formal evaluation of the program. Figure 2 provides a diagrammatic representation of our approach towards using CDRs combined with the formal evaluation of the EMOP program. The first part consists of the traditional approach to impact evaluation. The program's implementation relied on identifying households with the highest levels of vulnerability using the scorecard exercise described above. The use of a vulnerability score to determine eligibility allows us to evaluate the program's impact using an RD design. We measure the outcomes using an additional in-person survey administered seven months after the start of the program. We use both surveys as the ground-true data to feed the prediction models. The second part of Figure 2 shows the steps we follow to combine the CDR data with the information from the surveys. We start by processing the CDR data to create hundreds of behavioral features for each cellphone number and match them to individuals participating in the surveys. We face several challenges as a large percentage of participants lacked a cellphone and, in several cases, several household heads reported the same number as their own. After connecting the survey information with the behavioral features, we test different algorithms to try to replicate the program assignment determined by the scorecard vulnerability score and the program's main outcomes as captured during the in-person survey. We use cross-validation to limit the possibility of overfitting these models and choose the one with the best out-of-sample performance.

The rest of this section is organized as follows. Section 4.1 explains the feature engineering process we implement to extract useful information from the raw cellphone data. Section 4.2 describes the different surveys available for the impact evaluation of the program and as ground-truth data

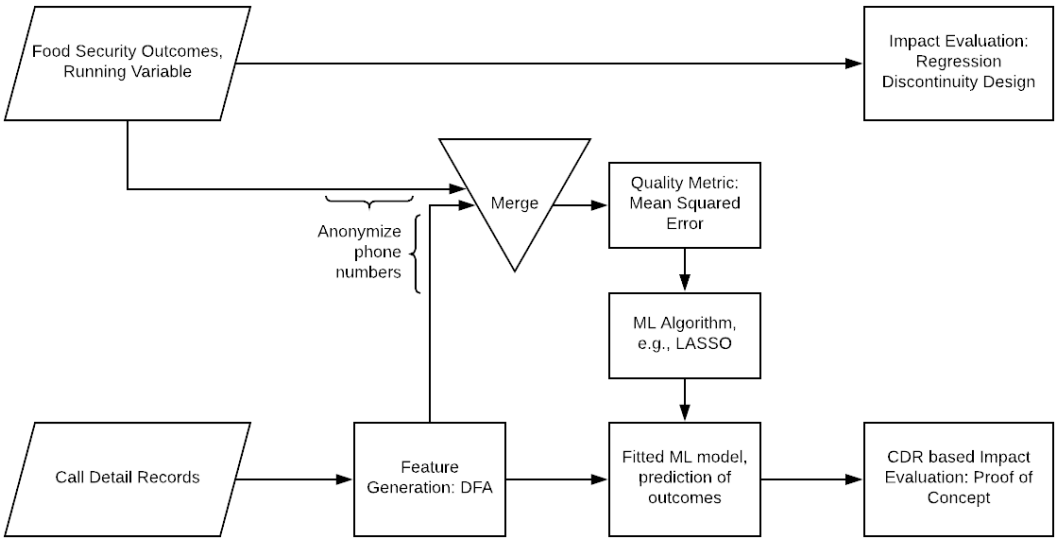



Figure 2: Our Approach to Running a Call Detail Record Based Evaluation

for the prediction. Additionally, this section explains the process to integrate individual survey responses with the corresponding the CDR data. Finally, section 4.3 discusses the identification strategy using an RD design.

### 4.1 Feature engineering

Our objective with the cellphone transaction data is two-fold. First, we want to predict the eligibility status from the scorecard survey. Second, we want to predict the program's main outcome variables and used the prediction to replicate the results from the impact evaluation of the program. In its raw format, CDR data provides a detailed account of each cellphone number activity. We want to extract information about each user's behavioral patterns that correlate with his socioeconomic characteristics. For this, we use feature engineering on the cellphone transaction data to compute behavioral indicators that capture aggregate aspects of each individual's mobile phone usage. In essence, for each phone number, we construct a vector of values that represent usage patterns and link them back to the survey responses of individual $i$. We generate the features using Bandicoot, an open-source toolbox for CDR analysis. The program creates metrics similar to other feature engineering methods used in the literature, such as Blumenstock, Cadamuro and On (2015). Its main advantage is that it provides a ready-to-use and computationally optimized method to extract features from cellphone metadata.[15]

Indicators include information about an individual's overall behavior (average call duration and percent initiated conversations, percent of nocturnal interactions, inter-event time between two phones calls), spatial patterns based on cell tower locations (the number of unique antennas visited and the radius of gyration), social network (the entropy of their contacts and the balance of interactions per contact), and recharge patterns (including the average amount recharged and

[15]For a full description of the method, see De Montjoye, Rocher and Pentland (2016).

the time between recharges) (De Montjoye, Rocher and Pentland, 2016). Each feature is calculated as a week-level mean, standard deviations, median, min, max, kurtosis, and skewness, as well as additional statistics disaggregated over weekdays, weekends, and working and night hours. Figure 1A explains the process to calculate individual features.

Since the features represent week-level statistics, we must decide how many days of cellphone transaction data to include in the feature generating process. Two forces are at play. First, a wider time window provides a more diverse set of transactions from which to extract information. On the other hand, we must consider how much the outcome we want to predict varies during this time window. In principle, outcomes with little variation over time benefit from using long series of cellphone data as the different features can capture more information. Applications that predict individual-level wealth levels, a variable that during normal times presents little variation over time, extract features from one year to six months of CDR data (Blumenstock, Cadamuro and On, 2015; Aiken et al., 2020). There is evidence on how changes in the size of the ground-true data affect the prediction capacity of CDR-based models and how a models' accuracy decays over time (Blumenstock, Cadamuro and On, 2015; Lazer et al., 2014). However, to the best of our knowledge, there is no evidence on how the length of the time series of CDR data affects predictive capacity, and how that relates to the variability over time of the predicted outcomes. Considering that we are interested in food security outcomes from individuals who live in regions affected by a natural disaster, we expect that the observed levels at the time of the survey were affected by both the shock and coping strategies. We have no prior about the optimal number of months of pre-survey CDR data to create the features. Instead, we opt to compute features using three different time windows preceding the date a person was surveyed: Fifteen days, one month, and six months. To calculate the time window, we omit the week preceding the survey to avoid changes in calling patterns in expectation to be interviewed.

For each individual, we extract a total of 2,148 behavioral features in each of the three time windows. We implement two initial filters on the features that make each time window includes a slightly different set of variables. First, we drop any feature with more than 50% of missing values. Second, we eliminate those with a variance below 0.02. Finally, to avoid models that contain highly correlated features that do not provide additional information, we calculate a correlation matrix for all the features and eliminate all but one for those with a correlation above 0.98.[16] Table 1 shows the number of features available for each time window and survey combination.

### 4.2 Data: Description and integrated sample

Since our main application uses CDR to replicate the program's targeting and impact evaluation, we are constrained by our ability to match the survey responses with the cellphone data. To explain this process, we first describe the available CDR data. Then for each survey, we detail the sample we can match and provide descriptive information to understand how the remaining sample compares with the overall population. Table 1 summarizes the samples of the scorecard and in-person surveys, and shows the specific samples we are able to match to the cellphone data.

[16] This approach is similar to Aiken et al. (2020)

Table 1: Data sources

| | Survey | |
|---|---|---|
| | **Scorecard** | **In-person** |
| **Sampling** | | |
| Unit of analysis | Household | Household |
| Data collection period | Apr. - Sept 2016 | Dec. 2016 |
| Location | 21 communes | 2 communes |
| Observations | 58,881 | 1,137 |
| Beneficiaries | 46,163 (78.4%) | 697 (61.3%) |
| **Cellphone related data (% of total sample)** | | |
| Owns a phone | 20,190 (34.3%) | 1,076 (94.6%) |
| Phones matches | 13,780 (68.3%) | 872 (81.0%) |
| **Numbers that match (% of matched numbers)** | | |
| Beneficiaries | 10,491 (76.1%) | 506 (58.0%) |
| Non-beneficiaries | 3,289 (23.9%) | 366 (42%) |
| **Sample with features** | | |
| Fifteen-days | 12,739 | 797 |
| Thirty-days | 13,142 | 825 |
| Six-months | 13,682 | 855 |
| **Features available** | | |
| Fifteen-days | 1,355 | 1,265 |
| Thirty-days | 1,457 | 1,394 |
| Six-months | 1,315 | 1,303 |

Note: Features available after filtering those with more that 50% missing values, a variance of less that 0.02, and correlations above 0.98

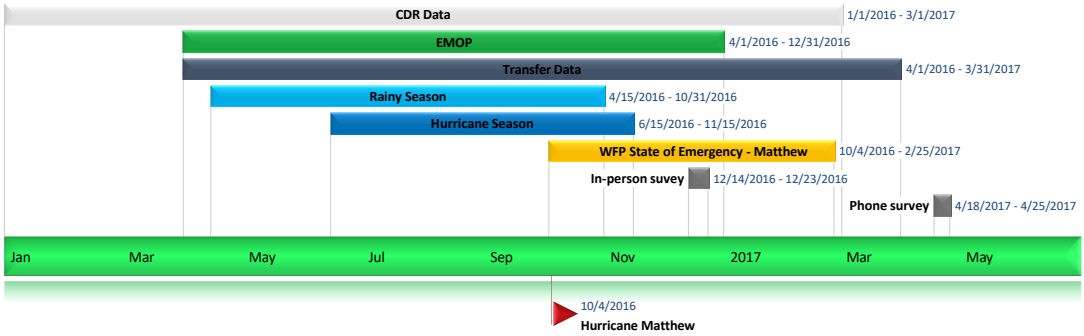
### Call Detail Records

Digicel is the largest cellphone company in Haiti. Besides providing cellphone services, the company operates the largest mobile money platform in the country. In 2016, it partnered with the WFP to deliver funds to beneficiaries using its mobile money platform. As part of the partnership with the WFP, Digicel provided cellphone records containing all the transactions made, in the network, between August 2015 and May 2017. All the information is anonymized, and we cannot see the content of calls or text messages. The CDR data covers all the evaluation cycle, including several months before each survey took place. Figure 3 shows how all the data sources overlap. For each type of transaction, we can observe:

- Calls: Anonymized numbers for the caller and the receiver, time of the call, duration of the call, and cellphone towers connecting the call.
- Text message: Anonymized numbers for the sender and the recipient, time of the text message, and the closest tower to the party that sent the text message.

During the period, a total of 29,907,850 transactions took place, broken down into 12,523,717 calls, 17,384,133 text messages. From this transaction, we extract behavioral features for three different time windows.

Figure 3: Program's timeline: Data availability and implementation




## Scorecard

This survey provides the basis for eligibility for the program. The WFP conducted the scorecard exercises between April and mid-September 2016.[17] It constitutes the final step of the WFP three-step targeting strategy. The survey took place in regions the WFP had previously identified as the hardest hit by the 2016 drought, with only households living in the pre-screened candidate communes participating in the survey. Eligibility depends on a commune-specific score. Due to the rapid response nature of the survey, the scorecard survey is not as thorough as an LSMS survey and lacks a full module on assets and consumption. [18]

A total of 58,881 households participated in the scorecard survey. Using the vulnerability score and the program's financial constraints, the WFP determined 46,163 (78.4%) eligible for the program. Considering that cellphone ownership in Haiti was below 50%, it is not surprising that in a vulnerable population, such as ours, only 20,190 (34.3%) of the participants reported having a cellphone number.[19] Since our application relies on matching individual survey responses to cellphone records, we are limited to participants with a cellphone number that matches our records. We are able to recover the CDR data for a total of 13,780 (68.3%) of the respondents with a phone, a number that is close to the market share of Digicel. As Table 2 shows, the share of beneficiaries in the sample of cellphone owners is marginally lower than in the overall sample (76.1%), as a consequence of cellphone ownership being positively correlated with wealth and negatively correlated with the vulnerability score. Additionally, we find a statistically significant

[17] Data collection was planned from April until August. The schedule was extended to account for initial delays and the inclusion of additional areas.

[18] Similar sampling methodology and questionnaires are standard for the WFP operations worldwide, especially when rapid delivery of funds is necessary to provide aid. The WFP website provides more details about how interventions are implemented and current programs: https://www.wfp.org/emergency-programming

[19] A minority of participants in the survey (174) reported a phone number that was also reported by another household. We omit their answers from the study.

difference in the vulnerability score within the group of cellphone owners. In particular, numbers that match our records have the lowest average vulnerability levels; see last two columns of Table 2. We are not aware of the cause of this apparent self-selection into different carriers. However, this highlights the intrinsic limitations of CDRs as a targeting tool.

We extract behavioral features for the 13,780 individuals with matched cellphone records. As Table 1 shows, the number of subscribers with features increases with the length of the time window, and it is always below the total number of numbers in the network. In the presence of sparse cellphone usage, short time windows are less likely to capture any activity.[20] For the six-month window, the only reason for a number not presenting activity is that the respondent activated the number in the week before participating in the survey, and therefore is not included in the feature extraction process.[21]

Table 2: Vulnerability score by cellphone ownership status

| | **Cellphone Ownership** | | | **Sig. diff.** | |
|---|---|---|---|---|---|
| | (i) No | (ii) Other operator | (iii) In network | (i) -(iii) | (ii)-(iii) |
| **Panel (a): Scorecards survey** | | | | | |
| Vulnerability Score | 0.74 | 0.41 | 0.32 | *** | *** |
| | (1.55) | (1.28) | (1.27) | | |
| Beneficiary (%) | 79.19 | 78.52 | 76.13 | | |
| Households | 38,691 | 6,410 | 13,780 | | |
| **Panel (b): In-person survey** | | | | | |
| Vulnerability Score | 0.54 | 0.17 | -0.11 | *** | *** |
| | (1.18) | (1.25) | (1.28) | | |
| Beneficiary (%) | 85.25 | 68.14 | 58.03 | | |
| Households | 61 | 204 | 872 | | |

Note: A higher vulnerability score makes a household more likely to participate in the program. We only have access to the transaction data for the number in the participating network.

### In-Person Survey

To measure the impact of the program, we collected an in-person survey in December of 2016. The survey collects information for a total of 1,137 people in two regions in the south of Haiti: Aquin (801) and Fonds-Verrettes (336).[22] The sampling of this survey targeted people around the eligibility threshold as it is the main tool to evaluate the program's impact using a RD strategy. The original scorecard survey in those two communes included 16,637 people. In contrast to the scorecard survey, this survey includes detailed questions on food consumption and expenditures, occupation, livestock holdings, and family composition. The share of beneficiaries in the sample is 61.3%, which is 17 percentage points lower than in the general population. The rate of cellphone ownership stands at 94.6%; a number that represents a large increment from the 44% captured by

[20] Sparse activity (low demand for communication) correlates with low-income levels that limit cellphone usage.
[21] Ownership of a number is lost if no recharge takes place during three consecutive months.
[22] Aquin is on the southwestern peninsula, and Fonds-Verrettes, which is inland, southeast of Port Au Prince.

the scorecard survey in the same communes.[23] We are able to recover the CDR data for a total of 872 of the respondents (81.0% of the respondents with a phone). In this sample, the share of beneficiaries is 3 percentage points lower than in the general population (58.0%). We also observe that respondents with a cellphone have lower vulnerability levels, see Panel (b) of Table 2. As expected, the number of individuals with features increases with the time window, see Table 1.

**Additional Data: Nationally representative survey**

As mentioned before, the emergency nature of the intervention implies that the survey data is concentrated in areas affected by the 2016 drought. We use a nationally representative survey of 4,267 households interviewed between May and October 2018 (FinScope, 2018). These data set provides a sample with greater socio-economic variation, which allows to put the participating population in context, and test the extend that a narrow sampling limits our capacity to predict the main outcomes of the program. [24] The questionnaire is close to an LSMS survey but has several modules focused on access to financial services. Using the questions available, we harmonized food consumption and food security variables to reflect as close as possible the WFP surveys.

A key advantage of using this survey is that for a subset of participants we are able to link their responses with their individual CDR data. For this, we conducted a follow-up call where participants provided 'informed consent' to access their cellphone records. Since participation in the Finscope survey was in person, we only contacted those who provided a phone number at the time of the survey. Between January and March 2021 we made four attempts to contact each number. Out of 2,870 participants with a phone number, we can use the CDR data for 1,132 cellphone lines.The Finscope survey has information for a 4,267 people, but only 2,870 (67%) provided a phone number had a phone, of which only 1,960 were part of the Digicel network. The IRB approved that we could link survey answers with CDR data in the case where the original survey participant was contacted and provided informed consent (519 cellphone numbers), and when a line had been disconnected for more than six months, making impossible to re connect with the original survey respondent (613 cellphone numbers). A total of 167 people refused to participate, and 661 active numbers did not answer to any of our contact attempts.

Table 3 shows differences in household composition, location, wealth, and food security by phone ownership status (see columns 2 and 3). As expected, people who own a phone tend to be younger and are more likely to live in urban areas. Phone ownership is correlated with higher levels of wealth and lower levels of food insecurity.[25] When we look for differences across the population with a phone an interesting dynamic appear. First, active lines at the time of our follow-up survey, but whose data we lack permission to use in this study, present highest levels of wealth (columns 5 and 6). These lines include people who were contacted and explicitly declined to participate and numbers that did not answer in any of the three contact attempts. Second, data suggests there is a correlation between joining a particular cellphone network and wealth, with cellphone owners who belong to a network other than Digicel (column 3) presenting lower wealth levels and being more

[23]The rate of phone ownership is close to 95% for both beneficiaries and non-beneficiaries. This larger than average rate of cellphone ownership is caused by the sampling strategy and not the result of the WPF providing SIM cards to eligible households.

[24]The sample can be disaggregated at the urban/rural level and seven regions, including the Port-au-Prince metropolitan area. Given the limited data sources available in Haiti, this is the survey closest to the time of the program. The latest LSMS survey in Haiti took place in 2011.

[25]In this case, the PCA to compute the wealth index omitted cellphone ownership from the calculation.

likely to live in rural areas. There is an apparent self-selection by location and wealth levels into a person's decision to contribute his data and choosing a particular network. This raises several questions about who we are able to predict socioeconomic indicators using cellphone data, specially when most studies only gather data from one cellphone provider.

Table 3: Descriptive stats Finscope sample

| | Owns phone | | | Owns phone: Number in Network- | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | | (3) | (4) | (5) | (6) | | |
| | No | Yes | Sig. diff. Yes-No | (i) No | Yes | | | Sig. diff. | |
| | | | | | (ii) Matched | (iii) No consent | (iv) No answer | (i)-(ii,iii,iv) | (ii)-(iii,iv) |
| Age HH head | 49.62 | 44.64 | *** | 43.34 | 44.4 | 47.07 | 46.19 | *** | *** |
| | (16.24) | (14.61) | | (14.07) | (14.48) | (12.93) | (15.73) | | |
| HH size | 3.34 | 3.88 | *** | 3.82 | 3.93 | 3.66 | 3.96 | | |
| | (1.86) | (1.99) | | (2.08) | (1.92) | (1.99) | (1.97) | | |
| Urban | 0.62 | 0.86 | *** | 0.83 | 0.89 | 0.93 | 0.85 | *** | |
| | (0.49) | (0.34) | | (0.38) | (0.31) | (0.25) | (0.35) | | |
| Food insecure | 0.54 | 0.49 | *** | 0.5 | 0.49 | 0.41 | 0.49 | | |
| | (0.5) | (0.5) | | (0.5) | (0.5) | (0.49) | (0.5) | | |
| Wealth index | -0.42 | 0.2 | *** | 0.05 | 0.19 | 0.64 | 0.32 | *** | *** |
| | (0.7) | (0.89) | | (0.84) | (0.73) | (0.92) | (1.11) | | |
| Obs. | 1,398 | 2,869 | | 909 | 1,132 | 167 | 661 | | |

Note: Author's calculations using Finscope 2018 and CDR data. Phone ownership is determined at the household head level. Columns 3 to 6 include only those with a cellphone. Column 3 includes cellphone owners with a line in a different network. Column 4 includes numbers that agreed to participate as well as those that were deactivated in the six months prior to our follow-up survey; we can only match the survey answers of these numbers with their CDR transaction data.

### 4.3 Regression discontinuity

The WFP conducted its standard scorecard survey to calculate a vulnerability score for each household in the areas affected by drought. The running score provides us with the basis to evaluate the impact of the intervention using a RD evaluation. The WFP's scorecard surveys were administered before the cutoffs were determined, elimination the possibility that households could precisely manipulate their eligibility status. In this way, we achieve randomization around the cut-off to assess the impact between beneficiaries and non-beneficiaries using a traditional regression discontinuity design. Due to the running variable being discrete, it is relatively difficult to run the regular diagnostics to detect precise manipulation of scorecard responses. This type of targeting is one of the bluntest used by WFP – used when information needs to be quickly gathered so fund or food assistance can be distributed. While it is manipulable in that those interviewed can misrepresent their asset holdings, household demographics, and occupations, doing so precisely would be extremely difficult without either the cut-off score and question set (which differed from area to area) or the locally committed budget of the program, the per household expenditure, and a prior about the responses of all others taking the same survey. Figure 4 shows the score distributions for each commune.

Our Regression Discontinuity design takes the form:

$$Y_i = f(S_i) + \rho_1 T_i + \varepsilon_i \tag{1}$$

where $Y_i$ is the outcome variable $f(S_i)$ is some function of the running variable (in this case, $S_i$ is the normalized targeting score), and $T_i$ is an indicator variable which tracks whether $S_i$ is above or below the threshold which determines treatment status. Conditional on local randomization at the boundary between beneficiary and non-beneficiary status, $\hat{\rho_1}$ estimates the Local Average Treatment Effect (LATE) of those at the margin between beneficiary and non-beneficiary status (Imbens and Lemieux, 2008; Lee and Lemieux, 2010).[26]

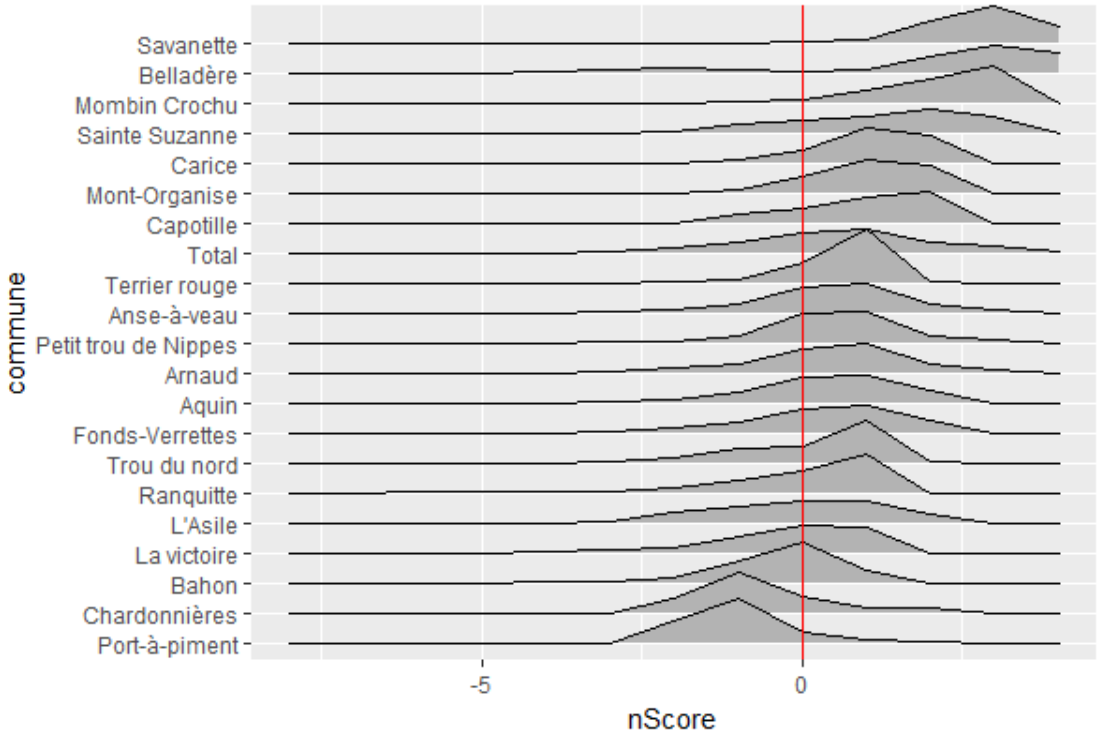



Figure 4: Distributions of Normalized Scores by Commune

Note: Author's calculations using scorecard survey. Cut-off (at zero) is marked with red. The heterogeneity of normalized scores by location around the cut-off does not suggest a consistent pattern of bunching near the cut-off.

We are limited in the forms $f$ can take due to the discrete nature of the running variable. We commit to a linear functional form as our preferred specification.[27]

[26] In this case, we estimate the LATE for a household with a score of zero. A possible robustness check would instead estimate the LATE for a household with a score of -0.5 or 1. Given the linear specification described below, we expect this choice should not matter much.

[27] Despite the advantages of splines and local linear regressions, we suggest that these usual choices for $f$ will not be worth exploring. In fact, even third order polynomials may be subject to issues at the boundary point, due to a limited number of values of the running variable.

$$Y_i = \mu_1 + \delta_1 S_i + \gamma_1 S_i T_i + \rho_1 T_i + \varepsilon_{1i} \tag{2}$$

The main assumption for the RD design is that those who are just beneficiaries should have similar observable characteristics to the those who just miss the cut-off for beneficiary status. A common approach to falsifying regression discontinuity designs is to use a density test for manipulation of the running variable. These allow us to test for discontinuities in the distribution, which would suggest that beneficiaries have bunched, for example, directly above the cut-off (Mccrary, 2008; Cattaneo et al., 2017). Boilerplate falsification tests for manipulation of the running variable tend to over-reject in the case of a discrete running variable. These tests tend to depend on the assumption that as sample size increases; the sample size adjacent to the cut-off will also increase. With coarse variables, however, the area local to the cut-off does not "fill in." Thus the density estimation tends to yield precise but misspecified density estimates on either side of the discontinuity. Moreover, the issues caused by the violation of this assumption become worse as sample sizes increase and the running variable coarsens. Hence, these tests are of little use in a falsification exercise and in our context will tend to reject even when manipulation is not present (Frandsen, 2016).

We use a test proposed by Frandsen (2016) which adjusts the McCrary test for use with a discrete running variable. The intuition here is the same as in other similar tests – changes in the rate that probability mass is accumulated at a cut-off will indicate bunching in or out of treatment. We test for continuity of baseline covariates around the treatment threshold. Baseline characteristics were collected only for treated households, making them of little use. Therefore, we utilize plausibly invariant characteristics of households from the in-person survey. Balance tests on the selected variables can be found in Table 4.[28] We are not able to reject the null hypothesis that there is no discontinuity in the distribution at the cutpoint. This result gives us confidence that households could not precisely target their locations in terms of normalized score.[29]

[28]Specific animal assets contribute to the treatment status. These depend on cut-offs about exactly how many animals of a given type a household owns. Hence, we include in the balance test a continuous animal stocks aggregated into tropical livestock units. Poultry is not used to determine eligibility so it is also included in balance test.

[29]This test is controlled by a parameter $k \geq 0$ which scales the bound the second order finite difference of the probability mass function. A larger value of $k$ allows for more deviation from linearity in point local to the cut-off, while $k = 0$ requires the distribution to be linear at the cut-off. Too low of a $k$ will tend to result in overrejection and too low will be underpowered. For arbitrarily small values of $k > 0$ (e.g., $k = 1e - 64$), We fail to reject the null hypothesis ($p = 1.000$)

Table 4: Balance Table

| Variable | Normalized Score $< 0$ | $\geq 0$ | t-stat on diff | RD estimate | t-stat on diff |
|---|---|---|---|---|---|
| Respondent | | | | | |
| Age | 39.050 | 41.444 | 2.674 | 0.195 | 0.11 |
| Gender | 0.695 | 0.706 | 0.421 | 0.015 | 0.26 |
| Household | | | | | |
| Size | 5.862 | 6.100 | 1.446 | 0.452 | 1.38 |
| Children | 1.013 | 0.941 | -1.156 | -0.148 | -1.14 |
| Pregnant or nursing | 0.335 | 0.322 | -0.3907 | -0.068 | -1.03 |
| Monoparental | 0.283 | 0.364 | 2.839 | 0.078 | 1.34 |
| Livestock | | | | | |
| TLU | 0.478 | 0.357 | -1.213 | -0.051 | -0.23 |
| Poultry | 2.357 | 2.449 | 0.385 | 0.199 | 0.39 |

Note: Outcomes variables are part of the in-person questionnaire collected seven months after the cash transfer implementation. The vulnerability score comes from the scorecard survey collected before the program started and used as targeting instrument. First two columns are means by beneficiary status. Third column is the t-statistic on difference in means. Fourth column is the coefficient on beneficiary status from our preferred RD specification. Fifth column is the t-stat related to this coefficient. Gender is coded 1 if female, 0 if male. Children are age 5 or younger. Poultry is number of birds, TLU is in cattle equivalent units.

# 5 Conventional Survey-Based Impact Evaluation as Benchmark

Based on the program's goals of improving food security outcomes, the evaluation considers five main outcomes. A full description of how each outcome variable is constructed can be found in Appendix 9.1.

1. Food expenditure: Food purchased by the household (in HTG).
2. Food consumption: Food purchased plus home production and food received through informal or NGO assistance (in HTG).
3. Food Consumption Score (FCS): The index measures the nutritional content of food eaten in the past week. Food is weighted by the nutritional content.
4. Dietary Diversity Score (DDS): Number of food categories consumed in the past week. Categories with no nutritional value have a weight of zero.
5. Coping Strategy Index (CSI): This index assesses a household's food security status. It is possible to see a positive impact in FCS or DDS which is the product of coping strategies i.e.,

households consuming food across more days, but restricting meal sizes or number of meals.[30]

Results of the RD estimations are presented in Table 5. We find an increase in the expenditure and consumption of food (measured in HTG) as well as the nutritional intake of beneficiary households in the week prior to the survey. In particular, the program increase food expenditure in 224.5 HTG, and food consumption (including donations and food consumption) in 282.1 HTG. However, we do not just see more spending by beneficiary households; we also see greater nutritional intake over the week before the survey as measured by the FCS. In particular, we see a 5.9 unit increase in the FCS of a base FCS of 40.6 units, significant at a 95% confidence level. To get a sense of magnitude in terms of food consumption, we can think about this in terms of food categories over the course of the week. This should be equal to an additional day and a half of proteins or dairy, two days of pulses, three days of cereals, or six days of fruits and/or vegetables. This increase in food consumption and nutritional intake is coupled with an increase in the diet diversity. We see an approximately 0.50 unit increase in the DDS, significant at the 99% confidence level. This magnitude corresponds to every other household consuming an additional category of food during the week, and a 5.8pp decrease in the share of consumption going to cereals, see Table 5.

We do not find evidence that the transfer program reduces the usage of coping mechanisms. For the CSI, beneficiary households have a 0.80 unit reduction in CSI (not significant at any standard level). Similarly, we see little evidence of impact on disaggregated coping strategies. We suspect our measurement of the impact on coping mechanisms is limited by the occurrence of hurricane Matthew. The acute stress may have wiped out access to certain strategies for coping with risk, while ensuring that accessible strategies are used almost universally. Figure 2A presents a graphical validation of the RD results at the score cut-off.

Table 5: RD Results: Effect of Cash Transfer on Indices of Food Security

| | (1) FCS | (2) DDS | (3) CSI | (4) Food Expenditure | (5) Food Consumption | (6) Share Cereal |
|---|---|---|---|---|---|---|
| Beneficiary | 5.880* | 0.499** | -0.796 | 224.5* | 282.1** | -0.0582* |
| | (2.494) | (0.164) | (1.483) | (91.78) | (96.36) | (0.0269) |
| Normalized Score | -1.844 | -0.155 | 0.938 | -48.36 | -66.84 | 0.0393** |
| | (1.432) | (0.0973) | (0.840) | (47.69) | (48.41) | (0.0146) |
| Beneficiary X | 1.411 | 0.0549 | -1.153 | 62.42 | 54.99 | -0.0418* |
| Normalized Score | (1.753) | (0.117) | (1.040) | (62.72) | (65.93) | (0.0182) |
| Observations | 1132 | 1132 | 1132 | 1132 | 1132 | 1130 |
| $R^2$ | 0.206 | 0.192 | 0.100 | 0.224 | 0.208 | 0.161 |

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Outcomes variables are part of the in-person questionnaire collected seven months after the cash transfer implementation.

[30] Additionally, only some of the improvement in household welfare is captured by the FCS or DDS because cash transfers might serve as a substitute for coping strategies otherwise used by the household (Maxwell and Caldwell, 2008).

# 6 CDR-based Targeting and Evaluation

## 6.1 Predicting eligibility status

An emergency response intervention requires a fast mechanism to identify potential beneficiaries. In a context with limited resources, the necessity for speed leads to the implementation of rapid surveys that identify the most vulnerable with a restricted set of questions. Our objective is to replicate the program's assignment of eligibility status using CDR data. We assume the scorecard exercise represents a "gold-standard" against which we evaluate the capacity of CDR-based methodology to replicate the identification of beneficiaries. The vulnerability score acts a proxy means test for the consumption level of participant households, providing a cost-effective way to target aid (Del Ninno and Mills, 2015).[31]. As the program lacks any other targeting mechanisms, it is not possible to assess any miss-classification of potential beneficiaries either by inclusion or exclusion[32] For our purposes, the scorecard exercise provides the ground-truth data for our analysis, and we assume it perfectly identifies different vulnerability levels, with households with higher scores being objectively more vulnerable.

There are two main concerns about the usage of the scorecard survey as a targeting tool. First, this methodology assumes the underlying regressions are error-free or measured with random error. The program's implementation of the scorecard survey did not validate the scores against food consumption –the proxied outcome–and relied instead on previous WFP experience to choose the variables included in the scorecard survey. This makes it impossible to quantify how effective the scorecard survey is in distinguishing vulnerability levels across households. We recognize that the validation of the targeting tool is a relevant concern to guarantee the proper implementation of social programs. However, assessing the targeting tools is outside the scope of this paper since our objective is to explore how CDR-based methods, can complement existing strategies. Second, the WFP used commune-specific eligibility thresholds that were determined by the availability of resources for each region, allowing for cases where two households with the same vulnerability score have different eligibility status. Considering that the differences in the cut-offs are small and households could not precisely target their eligibility status, we assume this to be an additional form of random miss-classification error that introduces noise into the CDR-based models.[33]

To evaluate the extent that cellphone usage patterns differ by eligibility status, we perform a two-sided t-test comparing the means across each sample. We follow (Khaefi et al., 2019) and classify features into five groups that reflect similar information content. Table 4A provides more details of the classification:

1. Basic phone usage, e.g., call, text, interactions
2. Active user behavior, e.g., call duration, percent initiated conversations, response delay
3. Spatial behavior, e.g., frequent antennas, number of antennas, the radius of gyration

[31]There is ample evidence that similar methods are more efficient and cost-effective than a universal allocation (Houssou et al., 2019).

[32]Systematic reviews of the methodology reveal that it tends to yield relatively low inclusion errors but high exclusion errors (Brown, Ravallion and Van de Walle, 2018).

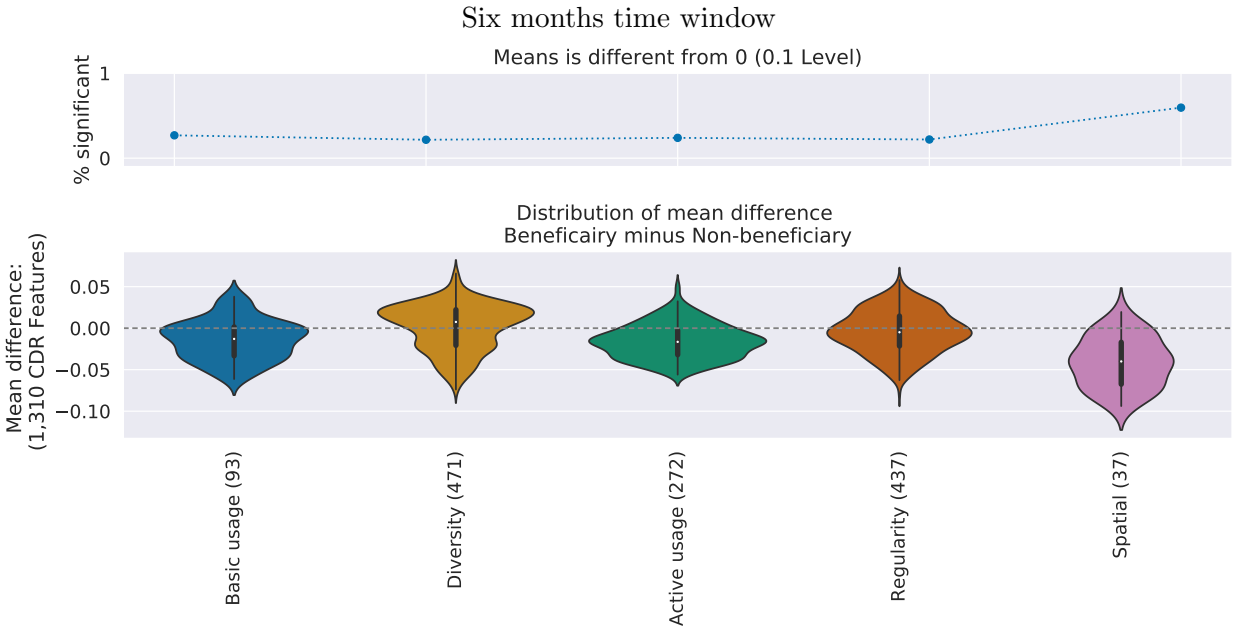
[33]To reduce the miss-classification error from the community-specific thresholds, we implement additional tests where we only use the information from the households at the tails of the vulnerability index distribution. Since these households have the highest (lowest) levels of vulnerability, they would have been eligible (ineligible) independently of the commune they live. We do not see significant differences in our results.

4. Regularity, e.g., inter-event time, percent nocturnal, the entropy of contacts

5. Diversity, e.g., number of contacts, the balance of contacts, interactions per contact

First, we want to check if the eligible population presents different patterns of behavior that the CDR can capture. For this, we compute the average difference between the two groups for each individual feature. In order to provide a similar scale we normalized all the features. Figure 5 shows the distribution of the mean difference for each group, as well as the percentage of features that present statistically detectable differences for the six-month time window. A negative value indicates that non-beneficiaries have a higher average for a given feature. It is not easy to characterize a general pattern since the size of the coefficients and the percentage that are significant varies depending on the time-window. Overall, results suggest beneficiary households are active fewer days, with fewer contacts and lower mobility; however, they interact more frequently and are more likely to start the calls.[34] We find that using a six-month window provides the highest number of statistically significant features (458), with this number going down with the length of the time window (316 for the one-month window, 172 in the two-week window). Figure 3A shows the results for the one-month and two-week time windows. Additionally, Figure 4A shows only the statistically significant coefficients.

Figure 5: Mean difference CDR features by eligibility status

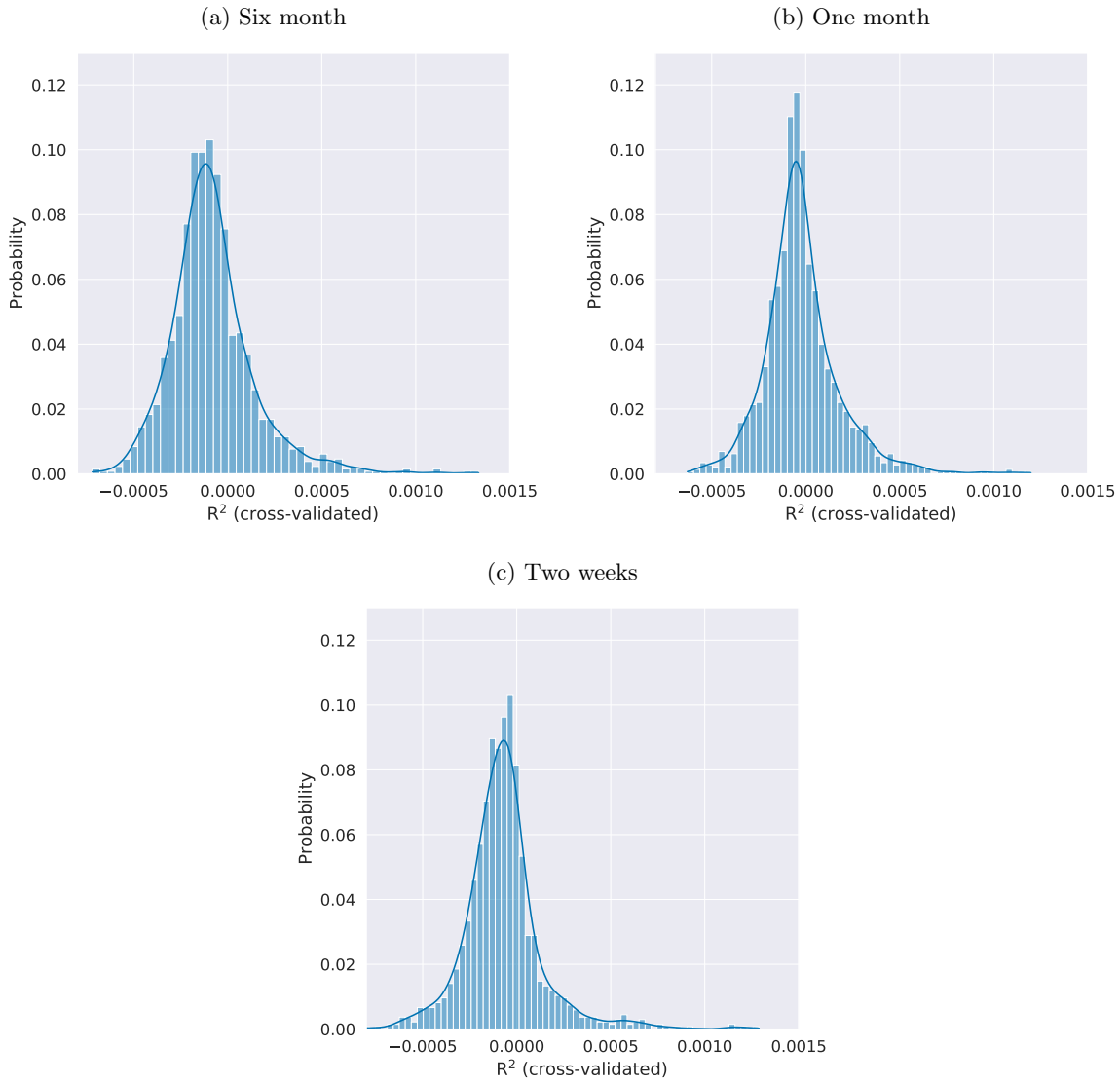



Note: Author's calculations using scorecard survey sample with records matching CDRs. Violin plot shows for each feature the mean difference between the average of beneficiaries and non-beneficiaries. All variables normalized. A negative value indicates that for a feature non-beneficiaries have a higher average. Number of features on each group in parentheses.

[34]Location is captured at the tower level. Mobility is detected only when a person uses his phone and switches the towers connecting the transactions. Therefore, usage and mobility are are deeply interconnected.

The previous results show the program's participants present different network usage patterns. Despite this, once we account for the individual features out-of-sample predictive capacity using cross-validation, no feature reaches an R square above 0.015. Figure 6 shows the distribution of R squares for each time window.

Figure 6: Beneficiary status: individual features predictive power




Note: The distribution of $R^2$ values from separate regressions of the beneficiary status on each available feature, showing average accuracy on the test set after 5-fold cross validation

We evaluate the model performance of the different machine learning methods using 10-fold cross-validation sets. Considering that a large percentage of the sample is eligible for the program, we stratified each fold to preserve class balance. For each fold, we train different machine learning

models that include Random Forest Classifier, XGBoost, and Elastic Nets. Technical Appendix 9.2 provides details about each model the model's hyperparameter tuning. To account for class imbalance, we rank individuals using their predicted probability to be in the beneficiary class and use a cut-off such that the model identifies the correct proportion of beneficiaries. To capture the trade-off between inclusion and exclusion errors for varying values of this threshold, we consider the area under the curve (AUC) score to measure targeting quality.

Results are not encouraging. As Panel A of Table 6 shows, predicting the beneficiary status using the sample in the scorecard survey produces an AUC no greater than 0.52, with little differences independently of the time window used to calculate the features. In practical terms, these results imply that if we select two random phone user from the network, with only one of them being eligible, our model will rank the eligible person as more likely to be a potential beneficiary with a probability of 52%. This means our models are not better than randomly assigning eligibility. Results are no different if we try to predict beneficiary status on the in-person sample, in which case our best performing model reaches an AUC of 0.52.

Table 6: Predicting binary outcomes

| | **Random Forest** | | **XGBoost** | | **Elastic Nets** | |
|---|---|---|---|---|---|---|
| | Full sample | Restricted | Full sample | Restricted | Full sample | Restricted |
| **Panel A: Scorecard survey** | | | | | | |
| Six months | 0.51 (0.01) | 0.52 (0.02) | 0.50 (0.01) | 0.51 (0.02) | 0.51 (0.01) | 0.51 (0.01) |
| One Month | 0.49 (0.01) | 0.49 (0.02) | 0.50 (0.00) | 0.52 (0.02) | 0.50 (0.00) | 0.51 (0.01) |
| Two weeks | 0.49 (0.01) | 0.48 (0.02) | 0.50 (0.01) | 0.51 (0.02) | 0.50 (0.01) | 0.51 (0.02) |
| **Panel B: In-person survey** | | | | | | |
| Six months | 0.50 (0.04) | 0.50 (0.05) | 0.50 (0.03) | 0.49 (0.06) | 0.52 (0.02) | 0.51 (0.05) |
| One Month | 0.50 (0.03) | 0.54 (0.06) | 0.52 (0.04) | 0.49 (0.04) | 0.51 (0.02) | 0.53 (0.04) |
| Two weeks | 0.52 (0.03) | 0.52 (0.05) | 0.51 (0.04) | 0.50 (0.05) | 0.51 (0.04) | 0.52 (0.04) |

Note: Accuracy for predicting beneficiary status and food expenditure. Binary measures are evaluated using the AUC score. Results are averages over 10-fold cross validation, with standard deviations in parentheses. Restricted sample contains a subset of the survey participants at the extremes of the vulnerability score.

Table 7 shows the exclusion errors (a beneficiary classified as non-beneficiary), inclusion errors (a non-beneficiary classified as beneficiary), and different classification metrics like accuracy, precision and recall for the best performing model. We see that low AUC translate into large errors of inclusion and exclusion. For example, in the six-month time window, out of a total of 10,189 beneficiaries identified by the scorecard survey, only 75% were incorrectly classified (false negatives); at the same time, 21% of the non-beneficiaries were deemed eligible by the model (false positive). Even if accuracy seems high, we see that the large number of false positives and false negatives produce a relative low precision and recall.[35] Depending on how the WFP weights inclusion and

[35] Accuracy denotes the percentage of correct predictions, in a dataset where the samples in one class are highly skewed (like ours), a high accuracy not necessarily reflects a well performing model since it might not be classifying

exclusion errors, the metric guiding the classification performance can rely on these additional evaluation metrics. A similar situation happens when we predict beneficiary status for the participants in the in-person survey.[36]

Table 7: Classification metrics for beneficiary status.

| Survey | True Pos. | True Neg. | False Pos. | False Neg. | Accuracy | Precision | Recall |
|---|---|---|---|---|---|---|---|
| Scorecard | 2,452 | 2,450 | 7,572 | 708 | 0.372 | 0.245 | 0.776 |
| In-person | 209 | 208 | 275 | 131 | 0.507 | 0.432 | 0.615 |

Note: Results using the best performing model for each sample.

Two elements of the WFP targeting process can affect out capacity to predict a person's beneficiary status. First, the WFP multi-step targeting process only interviews people living in the regions hardest hit by the 2016 drought. This makes that the sample only includes very vulnerable households, limiting the variation in the levels of vulnerability we can use to train our models. We do not have information for people outside of the targeted areas to increase the available information. Second, since the eligibility cut-off changes depending on the fund available for each region there are instances where two people with the same vulnerability score present different beneficiary status. We assume this constitutes an additional source of random noise in the eligibility criteria that should not invalidate out approach.

With the information at our disposal, we try to solve these two problems by restricting the sample to respondents at the extreme tails of the vulnerability score. In this sample the amount of variation in the vulnerability levels is maximized, while at the same time the overlapping eligibility cut-offs are eliminated. To make the results comparable, we reweight the new sample to represent the same number of observations and share of eligible households. The sample we can use for this greatly reduced as in both the scorecard and in-person surveys a large share of respondents are concentrated around the eligibility threshold.[37] It is important to clarify that this exercise does not rule out that lack of variability in the outcome variable drives our low predictive capability. It can still be the case that, even after subsampling in the tails of the vulnerability distribution, the resulting sample has no detectable differences in their true underlying vulnerability. As Table 6 shows, the changes in the AUC are marginal at best; with still large levels of false positives and negatives, see Table 8.[38]

the class with few observations correctly but still display a high accuracy level. Formally, accuracy is described by $\frac{TP}{TP+TN+FP+FN}$. Precision is defined as the proportion of beneficiaries that are correctly classified divided by the total number of predicted beneficiaries; a higher precision implies the classification returns more correctly classified outcomes than incorrect one. Formally, it is described by $\frac{TP}{TP+FP}$. Recall shows the proportion of beneficiaries that are correctly classified divided by the total number of individuals beneficiaries; a high recall implies the classifications identifies a higher proportion of the actual beneficiaries.Formally, it is described by $\frac{TP}{TP+FN}$.

[36]Figure 5A shows the ROC for each model

[37]In the scorecard sample, out of the 12 levels of vulnerability, we keep we keep scores from [-8,-3] below and [2,4] above the eligibility cut-off, representing 15% of the original sample. The in-person survey only contains people with vulnerability scores between -3 and 2. We keep scores between [-3,-2] and [1,2], representing 49% of the original sample.

[38]Figure 6A shows the corresponding ROC.

Table 8: Classification metrics for beneficiary status: Restricted sample

| Survey | True Pos. | True Neg. | False Pos. | False Neg. | Accuracy | Precision | Recall |
|---|---|---|---|---|---|---|---|
| Scorecard | 215 | 212 | 1,587 | 20 | 0.210 | 0.119 | 0.915 |
| In-person | 109 | 107 | 68 | 126 | 0.527 | 0.616 | 0.464 |

Note: Results using the best performing model. Sample restricted to the tails of the vulnerability levels.

## 6.2 Predicting food consumption and expenditure

We follow a similar approach to predict the main two outcomes from the program: Food consumption and food expenditure. These outcomes are only present in the in-person survey and, in particular, to participants who had a valid cellphone at the time of the survey.[39] We evaluate model performance using the cross-validated correlation (r) between true and predicted outcome. As Table 9 shows, in all cases our models perform poorly, with correlation close to zero. As a point of reference, CDR-based prediction of a wealth index in Rwanda show performances around 0.68 (r) (Blumenstock, Cadamuro and On, 2015).

Table 9: Predicting continuous food outcomes

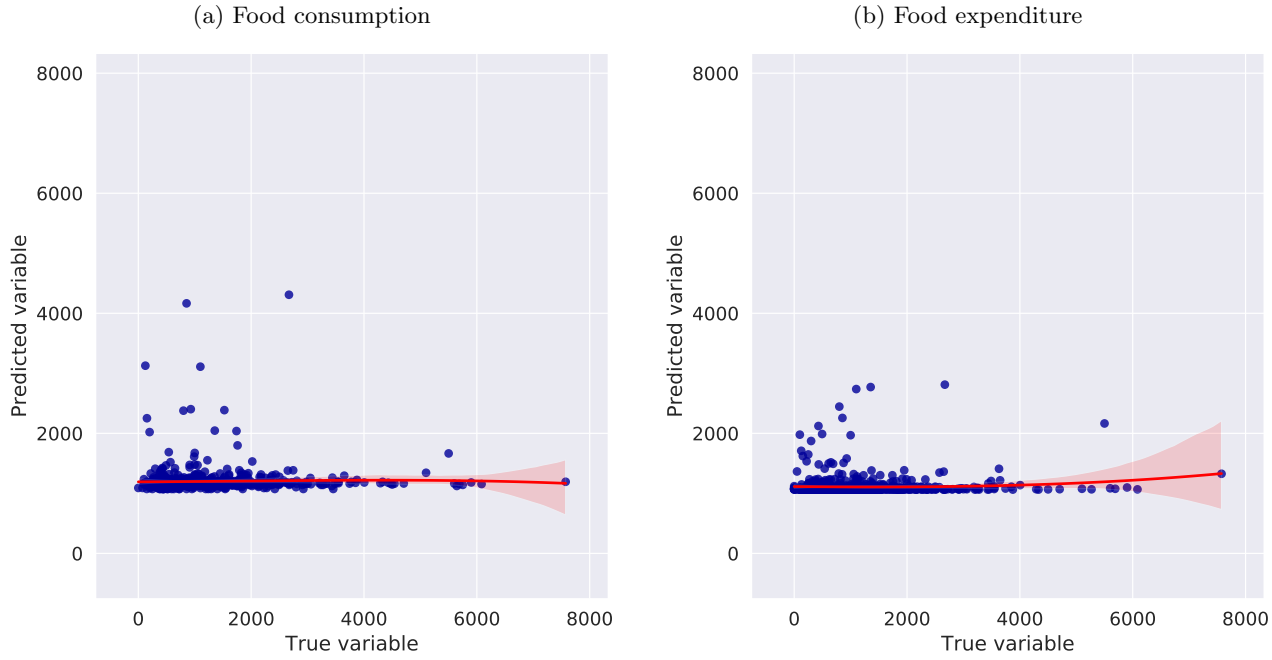
| | **Random Forest** Full sample | **XGBoost** Full sample | **Elastic Nets** Full sample |
|---|---|---|---|
| **Panel A: Food expenditure (r)** | | | |
| Six months | 0.08 (0.10) | 0.08 (0.14) | 0.00 (0.00) |
| One Month | 0.05 (0.14) | 0.06 (0.09) | 0.00 (0.00) |
| Two weeks | 0.07 (0.10) | 0.06 (0.12) | 0.00 (0.00) |
| **Panel B: Food consumption (r)** | | | |
| Six months | 0.05 (0.09) | 0.07 (0.13) | 0.00 (0.00) |
| One Month | -0.01 (0.08) | 0.00 (0.09) | 0.00 (0.00) |
| Two weeks | 0.09 (0.07) | 0.12 (0.06) | 0.00 (0.00) |

Note: Data form in-person survey. Models fit measures using correlation coefficient between true and predicted values. Results are averages over 10-fold cross validation, with standard deviations in parentheses. Restricted sample contains a subset of the survey participants at the extremes of the vulnerability score.

[39] Valid means a cellphone that was part of Digicel's network and that had activity covered during the time window used to extract behavioral features. Section 4.2 provides details on this sample.

Using the best performing model, in Figure 7 we plot the true and predicted values for the food consumption and expenditure outcomes. As we can see, the low predictive power of our models makes that there are no observable differences in the predicted values for households with high and low levels of consumption. This constrains any attempt to used the results to replicate the program's impact, as identified by RD results.

Figure 7: Predicting food consumption with phone data




Note: Relation between actual weekly consumption (expenditure) on food (as reported in the in-person survey) and predicted consumption (expenditure) as inferred from mobile phone data for survey participants with a valid phone. We use the best performing model using features extracted for the six-month time window.

### 6.3 Replicating RD results

The most compelling application of mobile phone data to predict socioeconomic indicators is to enable new approaches to impact evaluation and program monitoring (Blumenstock, Cadamuro and On, 2015). However, complementing well-understood impact evaluation and monitoring methods presents several challenges. First, even if a large body of work indicates cellphone data present fingerprints that are unique to certain socioeconomic indicators, at the individual level, prediction tends to work best for outcomes with little variation over time, limiting its usage in the context of impact evaluation and monitoring. Second, mapping digital data to welfare outcomes is both population and time-period specific, with evidence suggesting that predictive capacity deteriorates quickly (Lazer et al., 2014; Blumenstock, Cadamuro and On, 2015). Third, the quality of the prediction affects the value of the outcomes of interest reducing the power of any test run using the predicted outcomes.

Our previous results show that in a setting where we use data from a real-world impact, cellphone data does not replicate the program's main outcome. Despite these unimpressive results, we use the predicted food consumption and food expenditure to test if we can detect similar average differences between beneficiaries and non-beneficiaries as the impact evaluation results by estimating equation 2. For this, we use our best-performing model on the six-month time window features. Considering the power analysis at the end of this section, we provide all the impact effects in standardized units. Columns (1) and (2) of Table 10 shows effect of the program for the whole sample, including households both with and without a cellphone. We can see the program increase the food consumption and food expenditure among participants. In the next two columns, we replicate the same regression but only including households with a cellphone we were able to match to the CDR data. As we have discussed throughout the document, households with a cellphone present better baseline indicators. Using this sample, the program's impact on food consumption goes down while disappearing for food expenditure. Finally, using the predicted outcomes the best performing model, columns (5) and (6) detect no impact for the program. This is not surprising given that out low predictive capacity eliminates most variation across the food expenditure (consumption) space.

Table 10: Replicate RD results using predicted outcomes

| | | | Cellphone in network | | | |
|---|---|---|---|---|---|---|
| | Full sample | | Observed | | Predicted | |
| | Food Consumption | Food Expenditure | Food Consumption | Food Expenditure | Food Consumption | Food Expenditure |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Beneficiary | 0.354*** | 0.318** | 0.254* | 0.210 | 0.058 | 0.058 |
| | (0.128) | (0.128) | (0.149) | (0.148) | (0.149) | (0.149) |
| Normalized Score | -0.088 | -0.067 | -0.042 | -0.015 | -0.094 | -0.094 |
| | (0.076) | (0.076) | (0.090) | (0.090) | (0.090) | (0.090) |
| Beneficiary X Normalized Score | 0.039 | 0.038 | 0.048 | 0.046 | 0.181* | 0.181* |
| | (0.092) | (0.092) | (0.109) | (0.109) | (0.109) | (0.109) |
| Observations | 1,137 | 1,137 | 823 | 823 | 823 | 823 |
| $R^2$ | 0.011 | 0.011 | 0.010 | 0.011 | 0.004 | 0.004 |

*p<0.1; **p<0.05; ***p<0.01

Note: Full sample includes all the participants in the in-person survey. Numbers in cellphone network restricts sample to participants with a valid phone number. Result in standard deviations.

We want to understand how the minimum detectable impact changes as the precision of the machine learning models deteriorates. When we include all the survey participants, we are able to detect a change of 0.17 standard deviation over the mean of the control group.[40] Only including household with a valid phone number increases the minimum detectable effect to 0.21. Under these circumstances, it is not surprising we are less likely to find significant impacts in the restricted sample, as it is the case in the results in columns (3) and (4). The sample of individuals with a valid phone number represents the ground-truth data we use to measure the predictive capacity

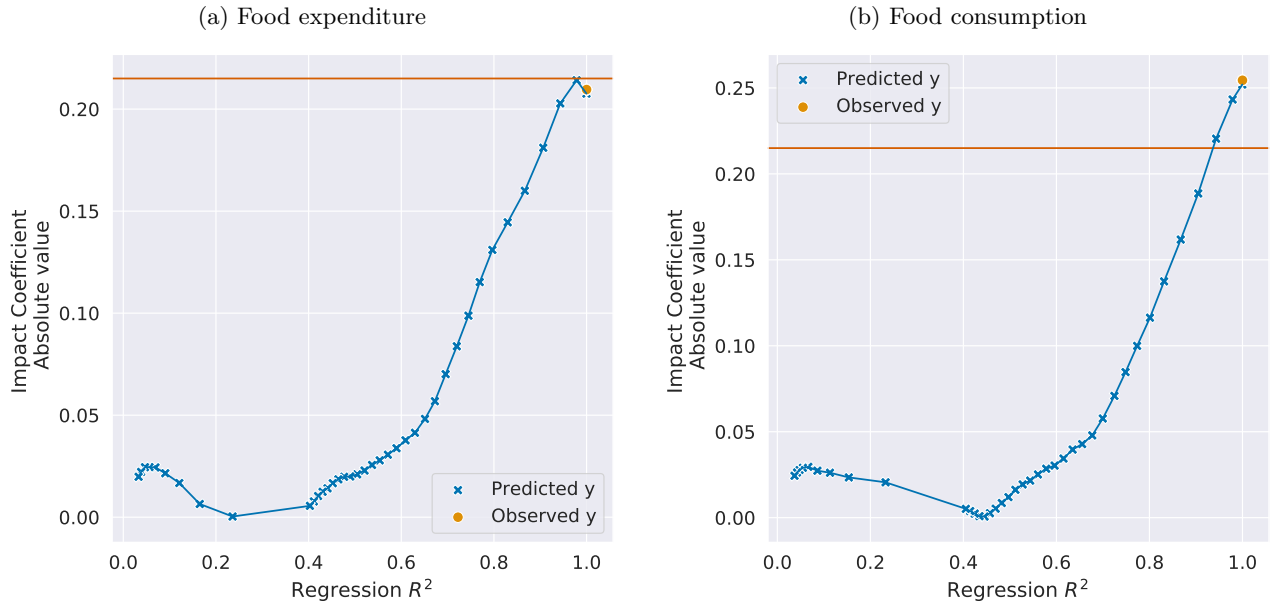
[40] Assuming an alpha level of 0.05, and a power of .8

of any machine learning exercise. Using different combinations of feature extraction and machine learning algorithms will create an unique combination of predicted values. Each combination will produce its own $R^2$, with the predicted values of the different algorithms converging as the $R^2$ approaches one.

Figure 8 shows how prediction quality affects our capacity to detect statistically significant impacts. To overcome that our predictive models have a very low $R^2$, we use a Lasso model with different regularization levels to predict the different food consumption measures. A penalty of 0 will produce a model that perfectly (over)fits the observed data and provides the same regression coefficients as the original sample. In contrast, increasing the penalization reduces the number of features entering the model and reduces the $R^2$. Using the the predicted values, we proceed to estimate regression 2, showing how they change as the quality of the predictions goes down. The horizontal line provides the minimum detectable effect in our sample. Therefore, we cannot identify as significant regression coefficient below this level.

Two main themes appear. First, at least in this case, the estimated impact using predicted outcomes rapidly declines with reduction of the $R^2$. Of course, the minimum detectable impact can be reduced by predicting values on a large sample. This is often the case in social research, where the number of observations in the ground-truth data is small compared with the total population. However, this does not address the apparent underestimation of the true impacts. Second, for $R^2$ values below 0.4 we see that the predicted data signals a decline in the food expenditure and consumption levels of the programs' beneficiaries. Considering that high-performing models tend to have an $R^2$ around 0.4, this results raises questions about how to properly benchmark machine learning predictions in the context of an RD design.

Figure 8: Changes on minimum detectable impact for different levels of $R^2$




Note: The horizontal line shows the minimum detectable impact for a sample equivalent to the in-person survey respondents with a valid phone number. The figure plots the impact coefficients from regression 2 using the predicted outcomes from a Lasso regression with different penalty level ($\alpha$). Orange dot represents the effect observed in the actual data.

# 7 Postmortem: What went wrong?

Why were we unable to replicate the success of others in using CDRs to predict household-level well-being? There are several insights that emerge from exploring this question. In this section, we discuss these insights and associated implications for predicting welfare outcomes using big data in poor countries. We use data from a nationally representative sample (Finescope, 2018), which includes many individual responses that can be matched to our CDR data, as the empirical basis for conducting this postmortem.[41]

We identify three reasons our prediction ability is limited in this specific context. First, the nature of cell phone ownership and usage among the rural poor in Haiti can undermine the signal value of CDRs for this population. Second, predicting flow variables like food expenditure and consumption is likely more challenging than predicting stock variables like wealth and asset indices. Finally, several features of this cash transfer program conspire to reduce statistical variation of outcome variables, which complicates prediction. We empirically explore each of these in turn.

[41] Section 4.2 provides details on the protocol to obtained informed consent for this sample.

### 7.1 Cellphone Penetration among Rural Poor Limits Value of CDRs

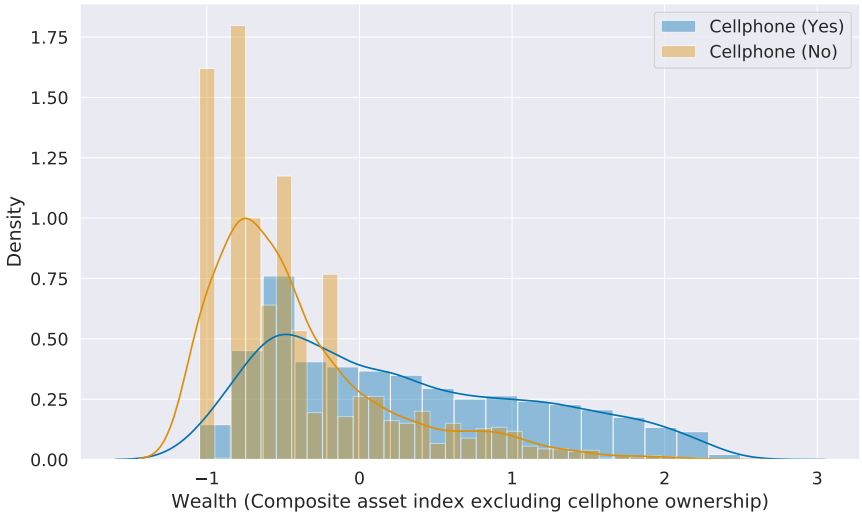
A natural limitation of using cellphones for social research is that having a phone in itself indicates socioeconomic status. Despite the impressive growth in the number of users in developing countries during the last decade, at the time of this cash transfer program less than 60% of Haitian households possessed a cellphone number (with large disparities across age groups and locations). The sample that participated in the WFP scorecard process lives in the poorest rural areas of Haiti where cellphone ownership is particularly low: Only 34% of scorecard respondents reported having a cellphone number. This lack of cellphone penetration into these households raises two primary challenges.

First, households without a cellphone are systematically poorer than those with cellphones. In the WFP data we use in this paper, we see that households with cellphones are less vulnerable on average (see Table 2).[42] This pattern unsurprisingly holds with nationally representative data: As well as shown in Figure 9, households without a cellphone are concentrated in the bottom of the wealth distribution.[43] While the observation that cellphone ownership is correlated with wealth is not surprising, it has important implications for the use of CDRs among the poor populations that are the intended beneficiaries of development or assistance programs.It is also worth noting that relying on self-reported cellphone ownership as the basis for eligibility for an assistance program like the WFP cash transfer may introduce under-reporting incentives since cellphones are easy to hide as assets. Much like some face incentives to strategically manipulate their cellphone usage to gain access to resources allocated using CDR-based algorithms (e.g., nano-loans) (Björkegren, Blumenstock and Knight, 2020), those whose access to benefits hinge on the kind of scorecard targeting used in this program face very well-documented incentives to under-report assets.

[42] Figure 7A shows the difference in the distribution of the vulnerability scores for households with and without a phone.

[43] Wealth is represented by a composite wealth index calculated using principal component analysis with cellphone ownership omitted from the calculation.

Figure 9: Comparison wealth index for Households with and without a phone




Note: Author's calculations using Finscope 2018. Wealth index is the first principal component of household assets. We exclude cellphone ownership form the calculation.

Second, relatively low cellphone penetration rates among the rural poor make the resulting CDR data systematically less plentiful and less useful as a source of statistical signal. Naturally, where cellphones are sparse so too will be CDR data, which obviously undermines CDR-based prediction. Even if a meaningful prediction could be coaxed from these sparse CDRs, using them for targeting and evaluation would be complicated by the fact that CDRs are most likely to be missing for the poorest households. Low penetration rates also change the way cellphones are used in these rural settings as they are much more likely to be shared within households and even among households. Such sharing introduces considerable noise into resulting CDRs and renders these data less informative as the basis for prediction.[44]

## 7.2 Flow Variables Harder to Predict than Stock Variables

The main objective of the WFP intervention was to improve food security by enhancing households' ability to purchase food. We therefore used total consumption and expenditure on food as leading impact indicators for the program evaluation. Our use of CDR-based methods for predicting these outcomes was motivated in part by encouraging evidence from the literature, which suggests that CDRs can be used to predict not just wealth but also food security indicators at the census-tract level (Blumenstock, Cadamuro and On, 2015; Hernandez et al., 2017; Decuyper et al., 2014). The sharp contrast between these prior prediction successes and this analysis suggests that estimating

[44]We find evidence of cellphone sharing in all the surveys available. In the case of the scorecard survey, 205 out of 13,780, and in the in-person survey 56 out of 872 households reported the same phone number as their own. Considering that we can only identify shared numbers when two or more households report the same line, these percentage of shared numbers represents a lower bound on the real level of number sharing.

individual consumption levels is far more difficult. As a final reason for this failure to predict, we explore important differences between predicting consumption flow variables and stock variables such as wealth.

We lack the necessary information to replicate a wealth index based on the WFP survey data. Instead, we use the nationally-representative.[45]. This survey of diverse individuals includes measures of wealth, which span the entire wealth distribution. Results in Table 11 show that, at least in the case of Haiti, CDR-based behavioral features perform a far better job at predicting wealth than food and total expenditure. We see that across our models the predicted correlation between true and predicted wealth is above 0.4, in line with previous findings in the literature, see (Blumenstock, Cadamuro and On, 2015). In the case of food expenditure our model clearly underperforms. Even when presented with a sample that covers the whole distribution of expenditures in Haiti, we are unable to properly predict either food and total expenditure with any degree of confidence. Although not definitive, this evidence suggests that flow variables are indeed harder to predict than stock variables.

Table 11: Predicting wealth and consumption using a nationally representative sample

| | **Random Forest** | **XGBoost** | **Elastic Nets** |
|---|---|---|---|
| **Panel A: Wealth composite index (r)** | | | |
| Six months | 0.45 (0.03) | 0.45 (0.02) | 0.49 (0.03) |
| **Panel B: Food expenditure(r)** | | | |
| Six months | 0.09 (0.09) | 0.14 (0.04) | 0.16 (0.04) |
| **Panel C: All expenditure (r)** | | | |
| Six months | -0.02 (0.03) | 0.02 (0.06) | 0.00 (0.00) |

Note: Models fit measures using correlation coefficient between true and predicted values. Results are averages over 10-fold cross validation, with standard deviations in parentheses.

For reference, figure 3A compares predicted and actual outcomes for the wealth asset index, food expenditure, and total expenditure. Here, we can see the degree that of variance in the outcome our model is able to capture when predicting wealth levels.

### 7.3 Effective Targeting Restricts Variation in Outcome Variables

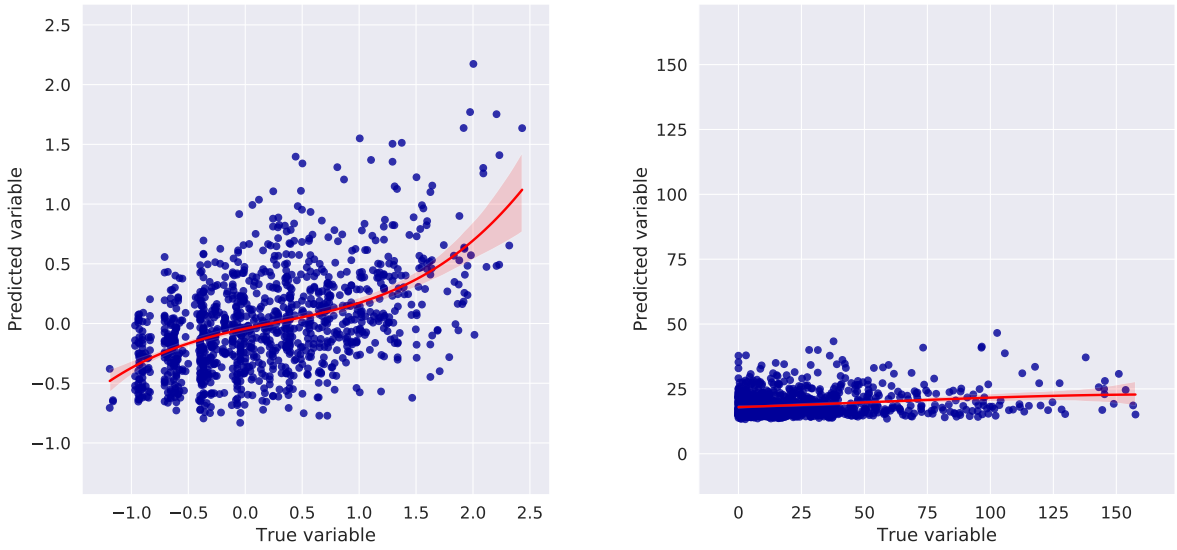
WFP used a standard multi-stage process to identify beneficiaries for this cash transfer program. After pre-selecting rural regions and communities most directly affected by drought in the years prior to the program, the WFP staff constructed lists of vulnerable households in collaboration with local authorities. These potential beneficiaries then participated in a scorecard survey described previously. WFP set a cut-off vulnerability score based on the overall budget for the program and

[45]See Section 4.2 for details on this survey

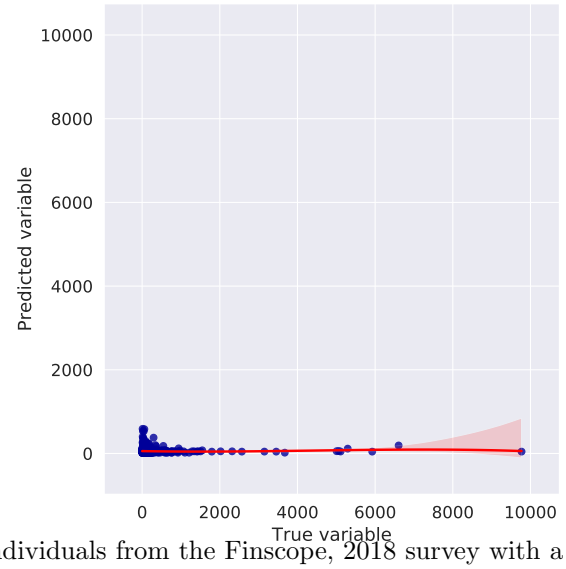
Figure 10: Mean difference CDR features by eligibility status:

(a) Wealth index

(b) Food Expenditure




(c) Total expenditure




Note: Results using individuals from the Finscope, 2018 survey with a valid phone number who agreed to participate in the study.

included all households with vulnerability scores greater than this threshold in the cash transfer program.

While this targeting process can be cost-effective as part of program implementation, it can simultaneously restrict the statistical variation of key outcome variables. Indeed, the more effective the approach targets potential beneficiaries for inclusion in the scorecard survey, the less useful the resulting scorecard data are for training CDR-based algorithms to predict these outcome variables. Compared with nationally-representative data, these pre-screened households are worse off in every aspect. For example, among these households per-capita food expenditure levels are one-fourth of the national average, food insecurity levels reach almost 100%., and food deprivation is higher (see Table 12). As Figure 11 shows, the cumulative distribution of food expenditures for the in-person survey is far below both the urban and rural distributions from the nationally-representative data. This is compelling evidence of the effectiveness of the early stages of the WFP targeting process, which is good news for programmatic operations but implies a very narrow statistical basis for predicting outcomes among this population.

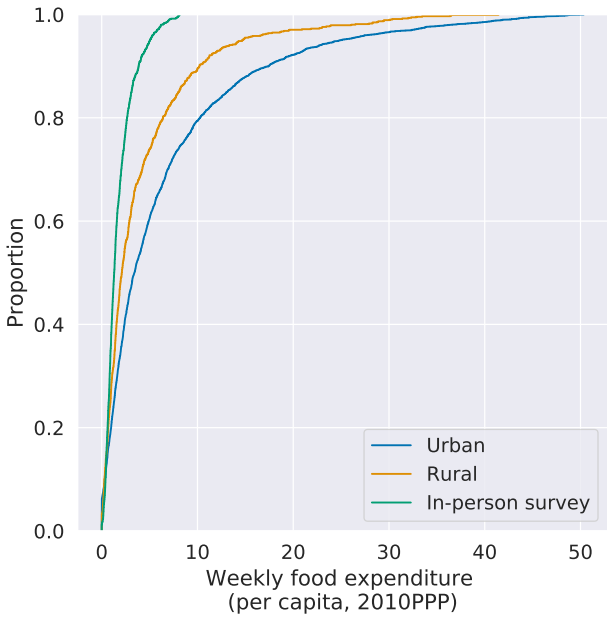
Table 12: Comparing In-person survey with a nationally representative sample

| | In person | National survey | Sig. diff. |
|---|---|---|---|
| Food expenditure per capita | 1.95 | 8.06 | *** |
| | (2.05) | (20.89) | |
| Food insecure | 0.97 | 0.5 | *** |
| | (0.17) | (0.5) | |
| **If food insecure (Last seven days with:)** | | | |
| Smaller meals | 3.12 | 1.74 | *** |
| | (1.91) | (2.18) | |
| Adults ate fewer meals | 2.47 | 1.37 | *** |
| | (2.27) | (2.04) | |
| Fewer meals | 3.01 | 1.77 | *** |
| | (2.0) | (2.23) | |

Note: Monetary values are expressed in 2010 USD PPP

A similar prediction problem emerges with the scorecard cut-off that determines final eligibility status. Pre-screened households for which vulnerability scores are available are much poorer than the average Haitian, as already demonstrated. Thus, the eligibility cut-off differentiates between those who are very vulnerable and those who are just slightly less vulnerable. This is precisely the appeal of the RD design to evaluating impacts of the cash transfer. But as greater similiarity between households on either side of the threshold strengthens the RD case, it simultaneously (and potentially dramatically) restricts the statistical basis for machine learning algorithms to predict outcomes and therefore their performance.

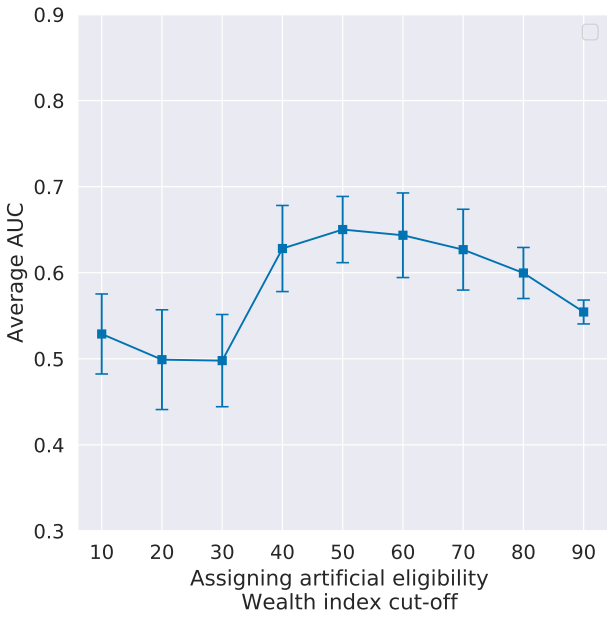
Figure 11: CDF food expenditure




Note: Author's calculations using the in-person survey. Information for the urban and rural population comes from Finscope 2018. All values are deflated to 2010 prices.

We want to provide some evidence about how highly targeting the data collection process affects the model's predictive ability. For this, we use the nationally representative and exploit that we have a higher capacity to predict an asset wealth level index. We implement two exercises. Our first simulation want to verify how for a given eligibility threshold the classification AUC changes. We take the full sample and artificially identify as "eligible" people with a wealth index below several cut-offs. For each cut-off value, we proceed to estimate a classification model using elastic nets and obtain the corresponding AUC. Figure 12 shows how the model performance does not behave linearly with respect to the eligibility cut-off. At least in our case, the signals from the CDR behavioral features appear to be better at identifying who are the wealthy individuals – wealth index above the 50th percentile–, than discerning who are the poorest –cut-offs below the 30th percentile. If we were able to extrapolate these results to the classification of WFP beneficiaries, a sample concentrated on poor individuals, this could further explain our model's low capacity to identify beneficiaries.

Figure 12: Classification AUC: Assigning an artificial eligibility threshold




Note: Author's calculations using Finscope data; only numbers that provided consent to match their responses with their CDR data. Classification uses an elastic nets model. Bars show the standard error across folds. Figure shows how using different eligibility thresholds on the asset wealth index affect the model classification AUC.

Our second simulation looks directly to the impact of restricting the sample to individuals in a narrow band of the wealth distribution. Training the model on a restricted bandwidth on the wealth distribution is problematic as performance declines quickly with sample size.[46] We take the model's prediction of an artificial eligibility status defined by the 50th wealth percentile and study the how missclassification behave as we mode away from the threshold.[47] Figure 13 shows the percentage of individual correctly classified on each wealth decile. Under perfect classification, any individual left (right) of the threshold should be classified as eligible (non-eligible). We see the percentage of individuals correctly classified increases as we move away from the eligibility cut-off; a result that holds both for eligible and non-eligible individuals.

[46] In an additional exercise, we show how the models performance worsen as the number of training instances goes down, see Figure **??**. Results show that the model's performance declines rapidly with the reduction of the sample size.

[47] We use that specific model as it provides the highest performance, see Figure 12.

Figure 13: Classification error around the eligibility threshold.


Note: Author's calculations using Finscope data; only numbers that provided consent to match their responses with their CDR data. Classification uses an elastic nets model. We assume an artificial eligibility threshold at the 50th percentile of the wealth distribution, with individuals to the left (right) of the threshold assumed to be eligible (non-eligible). The figure shows how the percentage of the population on each decile is correctly classified, revealing that those with low (high) wealth levels have a higher probability of being correctly assigned to their simulated eligibility status.Bars show the standard error across folds.

# 8 Conclusions

In recent years, the field of development economics has been invigorated by the prospects, potential and promise of big data and machine learning. Empirical breakthroughs using these novel data sources and methods have pushed research frontiers and deepened our understanding of key questions, many with direct and compelling policy relevance. This paper is specifically motivated by impressive and exciting results that suggest that machine learning techniques can extract surprising insights about mobile phone users from the CDR metadata they generate as they engage with the cellular network. Building on evidence that CDRs can predict wealth levels, we explore the limits of CDR-based analysis in the context of programmatic targeting and impact evaluation as second-generation applications of these methods. Such uses of these data and methods are especially intriguing as they would enable cost-effective and near real-time targeting and monitoring,

evaluation and learning. Testing the viability of these applications is our objective in this paper.

Current targeting methods require that every potential beneficiary participates in a survey to inform his eligibility for the program. Imagine a scenario where CDR-based features can predict a vulnerability score. Once a model has been trained on a passive data source, such as cellphone records, out-of-sample predictions of the targeting score allow for the immediate scoring of large swaps of the population at almost no additional cost. Building in these methods can provide real-time monitoring of socioeconomic indicators, opening the door to complement the impact evaluation of social programs. After training a model to predict a specific outcome, these methods make it possible to observe how an intervention changes the outcome of interest in populations for which the follow-up information was not collected.

The training of CDR-based models will require careful calibration for every context, and there is ample evidence that well-performing models tend to decay quickly over time. We provide evidence of how there are inherent limitations in using these CDR data when a program has been highly targeted. By showing the limitations of these methods under the specific circumstances we study, we hope to inform the design of future applications.

In collaboration with the WFP and the major mobile network operator in Haiti, we first conducted a conventional survey-based impact evaluation of a large, emergency cash transfer program administered and implemented by WFP. Using an RD design based on a threshold used to determine eligibility for receipt of the transfers, we find that the program indeed achieved its goal of improving food security, increasing food expenditure and expanding dietary diversity in the wake of a punishing drought. We then turn to the CDR-based analysis and assess how well these passive metadata and machine learning techniques can replicate the program's targeting and our survey-based impact evaluation. This alternative CDR approach fails in this case on both the targeting and the evaluation front for one simple reason: Even advanced machine learning methods are unable to generate useful predictions of phone users' well-being.

We identify and discuss several explanations for this prediction failure. First, both CDR-based targeting and prediction of food security outcomes apply only to households with mobile phones. Despite rapid growth in cellphone ownership rates, the most vulnerable, and therefore the intended beneficiaries of programs like the WFP cash transfer, have the lowest cellphone ownership rates and often share phones, which undermines the usefulness of CDR data. Second, the WFP program effectively targeted the poorest and most vulnerable households in the rural areas hardest hit by drought, which severely limits the statistical variation in key outcome variables and therefore undermines our ability to predict these outcomes. Third, we find that CDR-based methods perform better in predicting asset-based wealth measures (stock variables) than consumption (flow) variables. Although we cannot directly test it in our setting, this finding may further imply that using CDRs to predict *changes* in wealth over time may be deceptively more difficult than predicting wealth levels in a given cross-section.

The fact that we fail to replicate the survey-based impact evaluation using CDR-based predictions - and the reasons behind this failure - suggest that the excitement generated by CDRs as a data source in development economics should be tempered by some very real limitations. A vision of passive data taking the place of more costly conventional survey data is irresistible in many ways, but it is also likely unrealistic. While information and communication technologies, combined with administrative and big data, may continue to improve the data collection process and the quality of data collected, active data collected through these improved survey techniques will certainly continue to play a central role. This is as true for researchers as for program managers.

More specifically, the analysis in this paper reveals a fundamental empirical tension between algorithmic predictions using machine learning methods, on the one hand, and causal identification, targeting and program evaluation on the other. This tension is multi-faceted, but based on the notion that effective outcome prediction requires statistical variation across a broad support of the underlying outcome distribution. This requirement runs counter to causal identification, which hinges on counterfactuals for comparable individuals or households. This is most clearly on display with RD identification strategies that leverage similarities between treated and untreated units on either side of a threshold and within a narrow bandwidth. The narrower the bandwidth and the more similar the units are, the more compelling the RD estimates, but the more challenging outcome prediction becomes. In the context of targeting, the same pattern emerges: The more effective targeting is at identifying a sub-population with relatively homogeneous needs or characteristics, the less effective outcome prediction will be among this sub-population. Navigating this empirical tension will require a better appreciation for and understanding of its nuances - something to which we hope this paper can contribute.

## References


**Aiken, Emily L, Guadalupe Bedoya, Aidan Coville, and Joshua E Blumenstock.** 2020. "Targeting Humanitarian Aid with Machine Learning and Mobile Phone Data: Evidence from an Anti-Poverty Intervention in Afghanistan." *Mimeo*.

**Alatas, Vivi, Abhijit Banerjee, Rema Hanna, Benjamin A Olken, and Julia Tobias.** 2012. "Targeting the poor: evidence from a field experiment in Indonesia." *American Economic Review*, 102(4): 1206–40.

**Beine, Michel, Luisito Bertinelli, Rana Cömertpay, Anastasia Litina, Jean-François Maystadt, and Benteng Zou.** 2019. "Refugee Mobility: Evidence from Phone Data in Turkey." In *Guide to Mobile Data Analytics in Refugee Scenarios*. 433–449. Springer.

**Björkegren, Daniel, and Darrell Grissen.** 2015. "Behavior Revealed in Mobile Phone Usage Predicts Loan Repayment." 1–10.

**Björkegren, Daniel, Joshua E Blumenstock, and Samsun Knight.** 2020. "Manipulation-proof machine learning." *arXiv preprint arXiv:2004.03865*.

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

**Blumenstock, Joshua E.** 2018. "Estimating economic characteristics with phone data." Vol. 108, 72–76.

**Blumenstock, Joshua, Gabriel Cadamuro, and Robert On.** 2015. "Predicting poverty and wealth from mobile phone metadata." *Science*, 350(6264): 1073–1076.

**Brown, Caitlin, Martin Ravallion, and Dominique Van de Walle.** 2016. *A poor means test? Econometric targeting in Africa.* The World Bank.

**Brown, Caitlin, Martin Ravallion, and Dominique Van de Walle.** 2018. "A poor means test? Econometric targeting in Africa." *Journal of Development Economics*, 134: 109–124.

**Cattaneo, Matias D, Ann Arbor, Michael Jansson, and Ann Arbor.** 2017. "rddensity : Manipulation Testing Based on Density Discontinuity." *The Stata Journal*, 1–24.

**Coady, David, Margaret Grosh, and John Hoddinott.** 2004. "Targeting outcomes redux." *The World Bank Research Observer*, 19(1): 61–85.

**Decuyper, Adeline, Alex Rutherford, Amit Wadhwa, Jean-Martin Bauer, Gautier Krings, Thoralf Gutierrez, Vincent D. Blondel, and Miguel A. Luengo-Oroz.** 2014. "Estimating Food Consumption and Poverty Indices with Mobile Phone Data." 1–13.

**Del Ninno, Carlo, and Bradford Mills.** 2015. *Safety nets in Africa: Effective mechanisms to reach the poor and most vulnerable.* Washington, DC: World Bank; and Agence Française de Développement.

**De Montjoye, Yves-Alexandre, Luc Rocher, and Alex Sandy Pentland.** 2016. "bandicoot: A python toolbox for mobile phone metadata." *The Journal of Machine Learning Research*, 17(1): 6100–6104.

**FinScope.** 2018. "FinScope: Consumer Survey Highlights, Haiti." *https://http://finmark.org.za/wp-content/uploads/2019/04/Haiti_French_17-04-2019.pdf*, Accessed: 2020-09-21.

**Frandsen, Brigham R.** 2016. "Party Bias in Union Representation Elections : Testing for Manipulation in the Regression Discontinuity Design When the Running Variable is Discrete." 1–42.

**Gazeaud, Jules.** 2020. "Proxy Means Testing vulnerability to measurement errors?" *The Journal of Development Studies*, 56(11): 2113–2133.

**Goldblatt, Ran, Michelle F. Stuhlmacher, Beth Tellman, Nicholas Clinton, Gordon Hanson, Matei Georgescu, Chuyuan Wang, Fidel Serrano-Candela, Amit K. Khandelwal, Wan Hwa Cheng, and Robert C. Balling.** 2018. "Using Landsat and nighttime lights for supervised pixel-based image classification of urban land cover." *Remote Sensing of Environment*, 205(November 2017): 253–275.

**Grosh, Margaret E., and Judy L. Baker.** 1995. *Proxy means tests for targeting social programs.* The World Bank.

**Hernandez, Marco, Lingzi Hong, Vanessa Frias-Martinez, and Enrique Frias-Martinez.** 2017. *Estimating poverty using cell phone data: evidence from Guatemala.* The World Bank.

**Holzmann, Penny, et al.** 2008. *Household Economy Approach, the Bk: A Guide for Programme Planners and Policy-Makers.* Save the Children UK.

**Houssou, Nazaire, Collins Asante-Addo, Kwaw S Andam, and Catherine Ragasa.** 2019. "How can African governments reach poor farmers with fertiliser subsidies? Exploring a targeting approach in Ghana." *The Journal of Development Studies*, 55(9): 1983–2007.

**Imbens, Guido W., and Thomas Lemieux.** 2008. "Regression discontinuity designs: A guide to practice." *Journal of Econometrics*, 142(2): 615–635.

**Jean, Neal, Marshall Burke, Michael Xie, W Matthew Davis, David B Lobell, and Stefano Ermon.** 2016. "Machine Learning To Predict Poverty." *Science*, 353(6301): 790–794.

**Jensen, Robert T., and Nolan H. Miller.** 2008. "Giffen behavior and subsistence consumption." *American Economic Review*, 98(4): 1553–1577.

**Khaefi, Muhammad Rizal, Dharani Dhar Burra, Rio Fandi Dianco, Dikara Maitri Pradipta Alkarisya, Muhammad Rheza Muztahid, Annissa Zahara, George Hodge, Rajius Idzalika, et al.** 2019. "Modelling Wealth from Call Detail Records and Survey Data with Machine Learning: Evidence from Papua New Guinea." 2855–2864, IEEE.

**Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani.** 2014. "The Parable of Google Flu: Traps in Big Data Analysis." *Science*, 343(6176): 1203–1205.

**Lee, David S, and Thomas Lemieux.** 2010. "Regression Discontinuity Designs in Economics." *Journal of Economic Literature*, 48(2): 281–355.

**Maxwell, Daniel G., and Richard Caldwell.** 2008. "The Coping Strategies Index Field Methods Manual." January.

**Mccrary, Justin.** 2008. "Manipulation of the running variable in the regression discontinuity design : A density test." *Journal of Econometrics*, 142: 698–714.

**Milusheva, Sveta.** 2020. *Using Mobile Phone Data to Reduce Spread of Disease.* The World Bank.

**Olivieri, Sergio, Francesc Ortega, Ana Rivadeneira, and Eliana Carranza.** 2020. "Shoring up Economic Refugees: Venezuelan Migrants in the Ecuadorian Labor Market."

**Wesolowski, Amy, Nathan Eagle, Andrew J Tatem, David L Smith, Abdisalan M Noor, Robert W Snow, and Caroline O Buckee.** 2012. "Quantifying the impact of human mobility on malaria." *Science*, 338(6104): 267–270.

**Wilson, Robin, Elisabeth Zu Erbach-Schoenberg, Maximilian Albert, Daniel Power, Simon Tudge, Miguel Gonzalez, Sam Guthrie, Heather Chamberlain, Christopher Brooks, Christopher Hughes, Lenka Pitonakova, Caroline Buckee, Xin Lu, Erik Wetter, Andrew Tatem, and Linus Bengtsson.** 2016. "Rapid and near real-time assessments of population displacement using mobile phone data following disasters: The 2015 Nepal earthquake." *PLoS Currents*, 8(Disasters).

**World Food Programme (FAO).** 2008. "Food Consumption Score - Construction of the FCS: Interagency Workshop Report WFP - FAO Measures of Food Consumption - Harmonizing Methodologies Rome, 9 - 10 April 2008." April.

# 9 Technical Appendix

## 9.1 Outcomes specification

### Food Consumption and food expenditure

The survey asks respondents for seven day recall of spending across several categories of food, including total spending on fruits and vegetables, meat products (fish, offal, meat), eggs, dairy, legumes and tubers, cereals, sugars, oils and fats. Respondents are asked for spending over the past week first and then are asked for any other main sources of food consumed. If respondents report home production or donation as one of these sources, they are prompted to estimate the amount consumed but not purchased. From these data, we construct expenditure over the past week, as well as a measure of consumption, which includes food obtained by gifts, assistance, or home production. Finally, we construct the share of expenditure going to staples – in this case cereals – by dividing the past week's expenditure on these by total expenditure across all categories to proxy for proximity to subsistence (Jensen and Miller, 2008).

### Food Consumption Score (FCS)

Let $w_j$ be the weight for food category $j$ and $days_j$ be the number of days food category $j$ was consumed (see Table 1A for weights). Suppressing the household subscript, $FCS$ is defined as

$$FCS = \sum_j w_j \times days_j. \tag{3}$$

When food subcategories are present (e.g., green v. orange vegetables), the number of days each food category is consumed is measured by summing food subcategories, capped at seven days. Let $days_{jk}$ be the number of days food subcategory $k$ in category $j$ was consumed (World Food Programme , FAO).[48]

$$days_j = \min \left\{ \sum_k days_{jk} \right\} \tag{4}$$

### Dietary Diversity Score (DDS)

The second index, the Diet Diversity Score (DDS) is usually presented alongside the FCS to give a better sense of extensive changes in the FCS. In particular, the DDS measures the number of food categories consumed in the past week. Categories are weighted zero in the computation if they are not nutritionally valuable – in this case sugars and condiments have weights of zero.

[48]This method clearly gives a best case of the FCS where subcategories of food are substitutes day to day as opposed to complements day to day. That is, in this view, households consume fish or beef or chicken, but will not consume both on a given day unless every day already features a protein product. This should not matter for measurement of impact as long as the degree of substitutability of goods does not differ between beneficiaries and non-beneficiaries or generally, beneficiaries do not have some kind of taste for variety within the day. Given that these individuals are drawn from very similar populations, this would seem unlikely.

**Coping Strategies Index (CSI)**

Additionally, to further qualify increases in spending and our measures of nutritional quality, we look at the WFP defined Coping Strategies Index for Food, which is built from questions about "survival strategies."

To contextualize the food security indices and consumption outcomes above, we measure strategies used to cope when food insecurity arises. These outcomes encompass a set of decisions that affect future household welfare. Seven of fourteen of these outcomes look explicitly at food related coping strategies. These strategy questipns occur as a seven day recall of how many days the household ate smaller, fewer, or less preferred meals, borrowed food from friends or relatives, or restricted adult's food intake so that children could eat. From five of these outcomes we construct a Coping Strategies Index (CSI). For CSI specific strategies see Table 2A and for all additional strategies see Table 3A. Much like the FCS, the CSI is computed as a weighted sum of days over the past week that a certain coping strategy was undertaken.

$$CSI = \sum_j q_j \times days_j \tag{5}$$

Looking at various coping strategies as outcomes, the coefficient on treatment can be meaningful in various ways. We expect negative coefficients on treatment status as evidence of the impact of beneficiary status. In this case, a negative coefficient would suggest that the cash transfers would reduce the degree to which households needed to use coping strategies as part of an income effect. Income from beneficiary transfers would serve as a substitute for these strategies, in such a context. Because of the costly nature of strategies used to cope, such an impact could potentially still mark an improvement in welfare, even if substitution away from coping strategies reduced the overall impact of program on measured food security. To this end, coping strategies are chosen to be dynamically harmful to the household.

However, a null coefficient might arise in two scenarios. If household's budgets are strained and/or access to coping strategies is limited, households would operate at their "coping frontier." In this case, we would expect to see a negative impact on coping strategies only if transfer payments were sufficiently large. This case is consistent with a story where the measured increases in food security are not due to a simultaneous coping decision made at the household level for beneficiary households e.g., reduction in meal sizes or number of meals per day. Alternatively, if the transfer program was not effective in increasing food consumption, we would similarly see a null effect on coping mechanisms. However, in this case, we would not expect to see impacts on food consumption or nutritional intake. Finally, positive impacts on the use of coping strategies would be highly counterintuitive. This would suggest that households are made worse by receiving transfers, or perhaps that some other omitted variable is discontinuous at the boundary, i.e., that households that receive transfers are qualitatively different than those who do.

Table 1A: WFP Food categories, subcategories and weights for FCS and DDS Indices

| | | Weights | |
|---|---|---|---|
| Category | Subcategories | FCS | DDS |
| Cereals, Roots and Tubers | | 2 | 1 |
| Vegetables | Orange, Green Leafy, Other | 1 | 1 |
| Fruits | Orange, Other | 1 | 1 |
| Protein | Meat, Offal, Fish, Eggs | 4 | 1 |
| Pulses, Nuts, Seeds, and Legumes | | 3 | 1 |
| Dairy | | 4 | 1 |
| Oil and Fats | | 0.5 | 1 |
| Sugars | | 0.5 | 0 |
| Condiments | | 0 | 0 |

Table 2A: Disaggregated strategies (weekly recall) and weights for CSI

| Coping Strategy: How many days out of the last seven days did your household adopt the following strategies due to lack of food or money? | CSI weight |
|---|---|
| Eat less preferred or less expensive foods | 1 |
| Reduced number of meals per day | 1 |
| Reduce meal size | 1 |
| Restrict adult food consumption to feed young | 3 |
| Borrow food or rely on help from friends or relatives | 2 |

Table 3A: Other Disaggregated strategies (weekly recall)

| | |
|---|---|
| Days in the last week: | |
| Borrow food or rely on help from friends or relatives | Integer (0-7) |
| Didn't Eat | Integer (0-7) |
| Went to bed hungry | Integer (0-7) |
| Worried the household would not have enough food to eat | Integer (0-7) |
| In the past seven days have you: | |
| Used grain stock meant for the agricultural season, for food | Indicator |
| Withdrawn children from school | Indicator |
| Reduced health expenses | Indicator |
| Taken on debt to buy food or bought food on credit | Indicator |
| Had a household member migrate | Indicator |
| Sold livestock (cow, chicken, sheep, goat, etc.) | Indicator |
| Sold other productive assets (mill, agricultural land, etc.) | Indicator |

### 9.2 Hyperparameter tuning

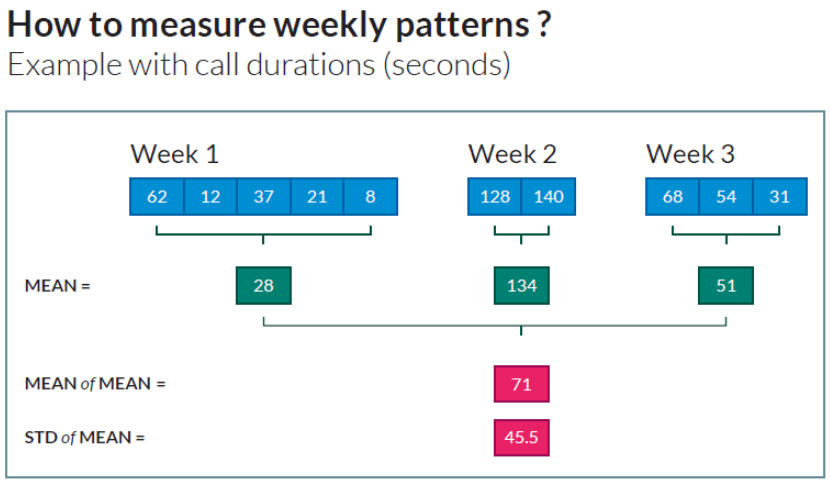
In the case of the random forest classifier (an ensemble of 100 decision trees) to predict the probability of beneficiary status on the training set and evaluate its out-of-sample performance on the test set. The maximum depth of the random forest is selected from $\{2, 4, 8, 16, 32\}$ via 3-fold cross-validation on the training set.

We also try to predict food expenditure using elastic net regression and three flexible two-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 from the $\{0, 0.2, 0.4, 0.6, 0.8, 1\}$. The maximum tree depth is selected for each tree-based model via 3-fold cross-validation from the set $\{2, 4, 8, 16, 32\}$. As in the previous exercise, we implement a 10-fold cross validation to account for over-fitting. Since the outcome is continuous, we use the $R^2$ coefficient to evaluate the predictions.

# 10 Additional figures and tables

## 10.1 Bandicoot feature engineering

Figure 1A: Bandicoot feature extraction process




Note: Taken from De Montjoye, Rocher and Pentland (2016). The figure explains how information for each transaction type is calculated at the week level, and the computed as a single aggregate indicator.

Table 4A: Classification of features into similar information groups

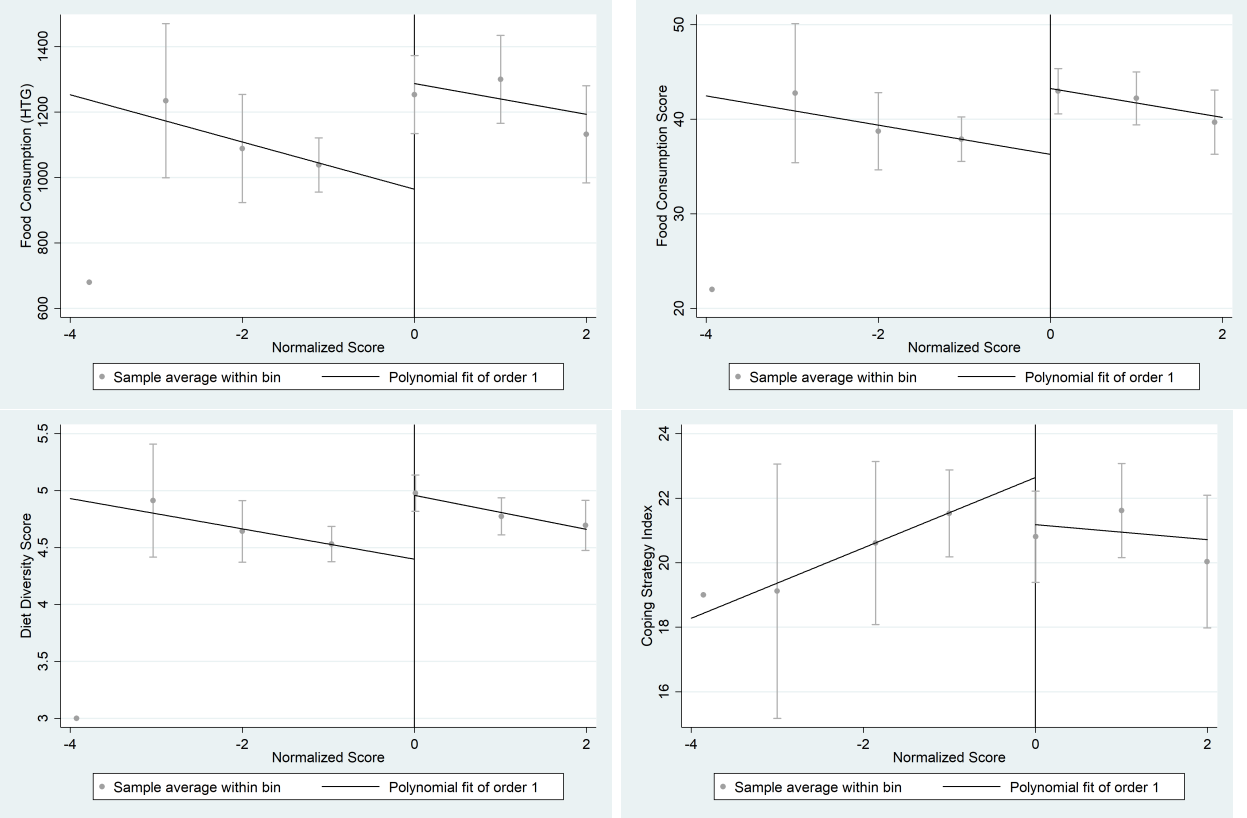
| Category | Name | Description |
|---|---|---|
| B | Number of records | Number of actual records produced by user's mobile phone activity. |
| B | Active days | Number of days during which the user was active. |
| B | Number of interactions | Number of interactions by the user (incoming and outgoing). |
| B | Percent records missing location | Percentage of records different with home location |
| A | Call duration | Duration of the user's calls. |
| A | Percent initiated conversations | Percentage of conversations that have been initiated by the user. |
| A | Percent initiated interactions | Percentage of calls initiated by the user. |
| A | Response delay | Response delay of the user within a conversation (in seconds) |
| A | Response rate | Response rate of the user (between 0 and 1). |
| S | Percent at home | Percentage of interactions the user had while he was at home. |
| S | Radius of gyration | Returns the radius of gyration, the equivalent distance of the mass from the center of gravity, for all visited places. |
| S | Frequent antennas | Number of location that account for 80% of the locations where the user was. |
| S | Churn rate | Computes frequency spent at every tower each week, and returns the distribution of the cosine similarity between two consecutive weeks. |
| S | Number of antennas | Number of unique places visited. |
| S | User locations | LLG districts where the user resides, calculated using the tower geolocation (latlong). |
| R | Percent nocturnal | Percentage of interactions the user had at night. |
| R | Entropy of contacts | Entropy of the user's contacts. |
| R | Entropy of antennas | Entropy of visited antennas. |
| R | Interevent time | Interevent time between two records of the user. |
| R | Percent pareto interactions | Percentage of user's contacts that account for 80% of its interactions. |
| R | Percent pareto durations | Percentage of user's contacts that account for 80% of its total time spend on the phone. |
| D | Number of contacts | Number of contacts the user interacted with. |
| D | Balance of contacts | Balance of interactions per contact. |
| D | Interactions per contact | Number of interactions a user had with each of its contacts. |

Note: Taken from Khaefi et al. (2019). The table shows the classification of features into different categories that reflect similar information content. B: Basic phone usage, A: Active users, S: Spatial behaviors, R: Regularity, D: Diversity.

# 11 Additional Results

## 11.1 Additional Impact Evaluation Results

Figure 2A: Effect of Cash Transfer on Food Consumption, Diet Diversity and Coping Strategies

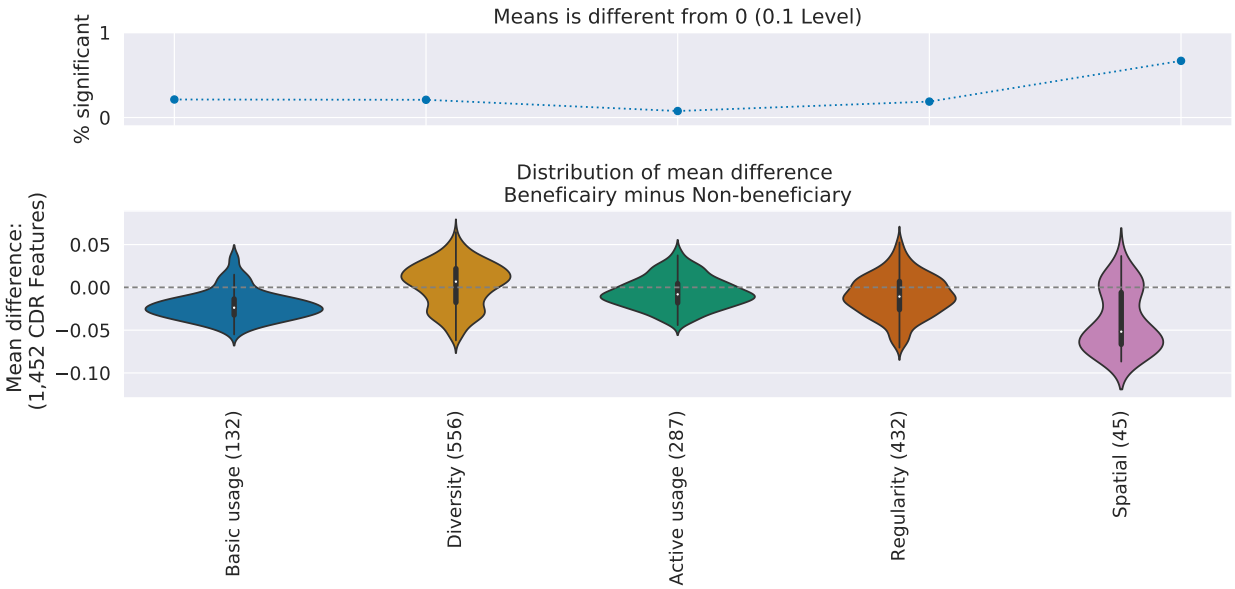



Note: Author's calculation using the in-person survey to measure outcomes, and the scorecard survey to create the vulnerability score. Figures show the RD impact at the bin level.

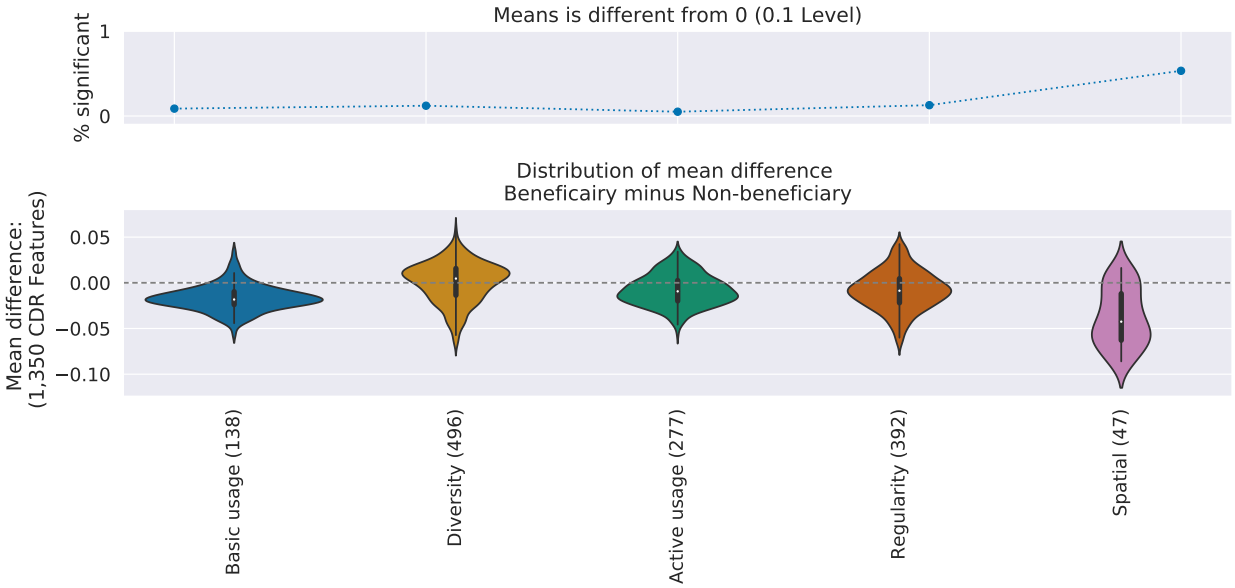
## 11.2 Additional Results Machine Learning

Figure 3A: Mean difference CDR features by eligibility status:

(a) One month



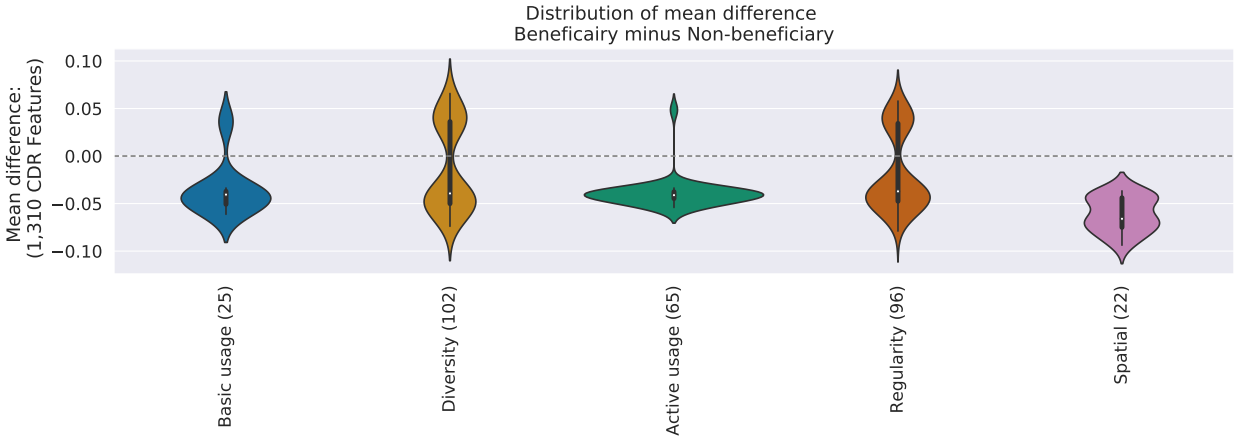

(b) Two weeks



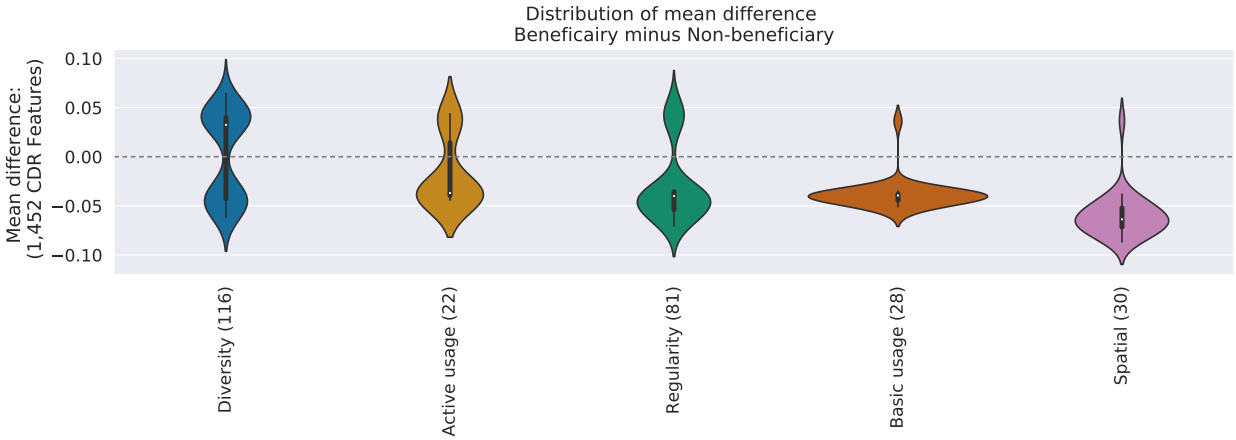

Note: Author's calculations using scorecard survey sample with records matching CDRs. Violin plot shows for each feature the mean difference between the average of beneficiaries and non-beneficiaries. A negative value indicates that for a feature non-beneficiaries have a higher average.

Figure 4A: Mean difference CDR features by eligibility status: Significant features only.

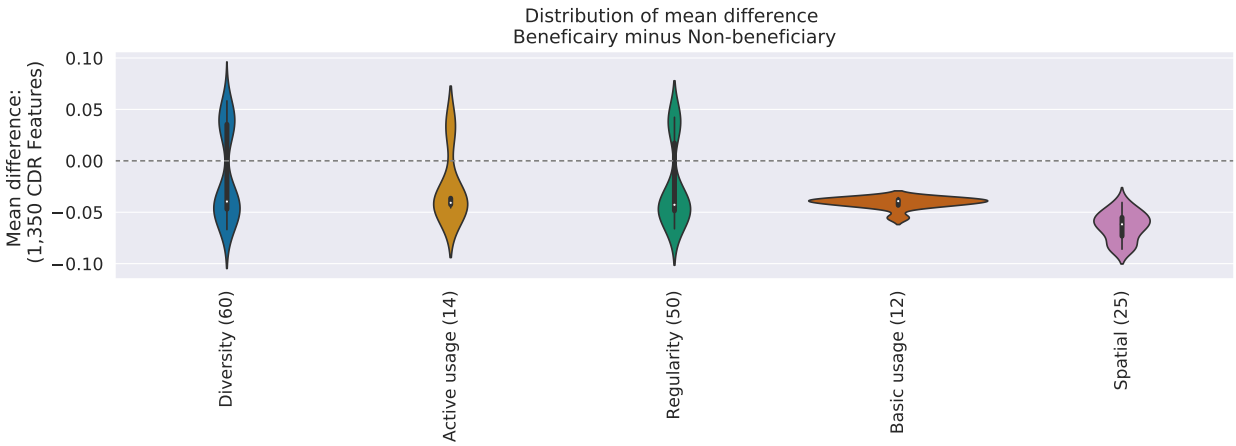
(a) Six months




(b) One month



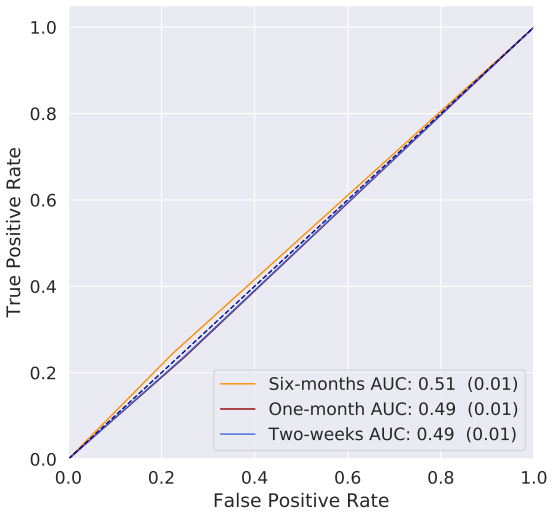

(c) Two weeks



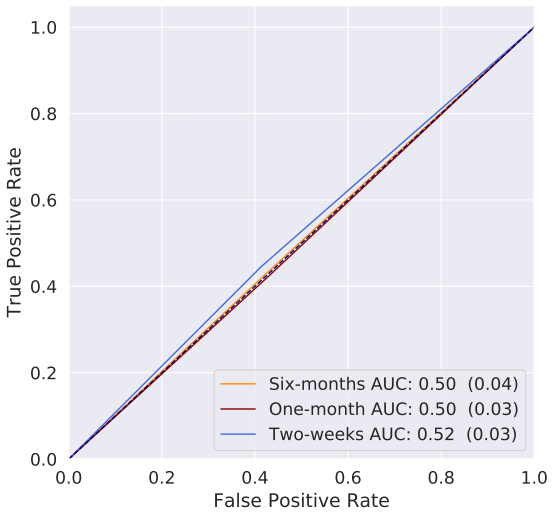

Note: Author's calculations using scorecard survey sample with records matching CDRs. Violin plot shows for each feature the mean difference between the average of beneficiaries and non-beneficiaries. All variables normalized. A negative value indicates that for a feature non-beneficiaries have a higher average. Only features with statistically significant difference shown.

Figure 5A: ROC Curves for CDR-based targeting of beneficiaries

(a) Scorecard survey



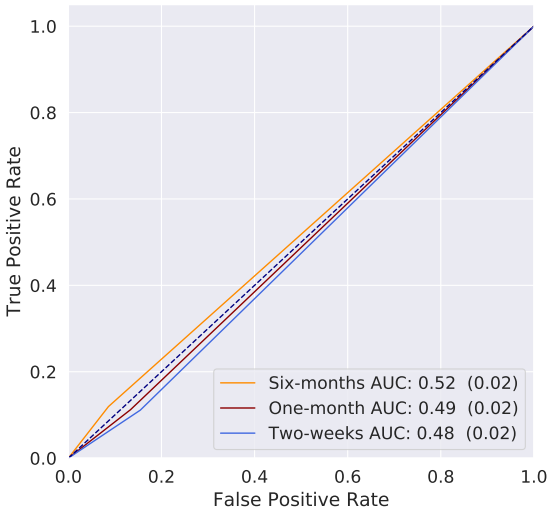

(b) In-person survey



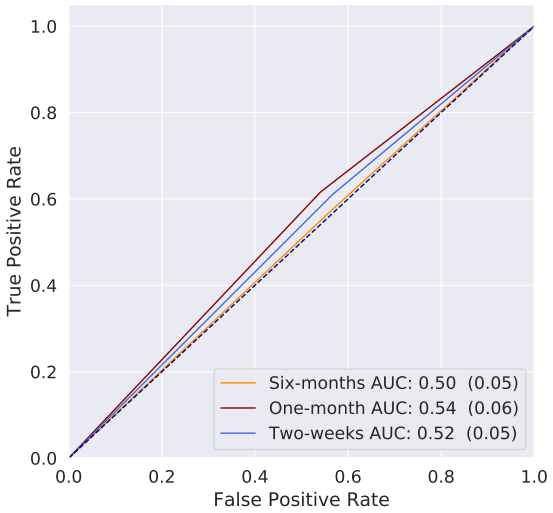

Note: ROC curves for classifying beneficiary status using CDR-data. Features extracted for the six-month time window for survey participants with a valid phone.

Figure 6A: ROC Curves for CDR-based targeting of beneficiaries: Restricted sample

(a) Scorecard survey

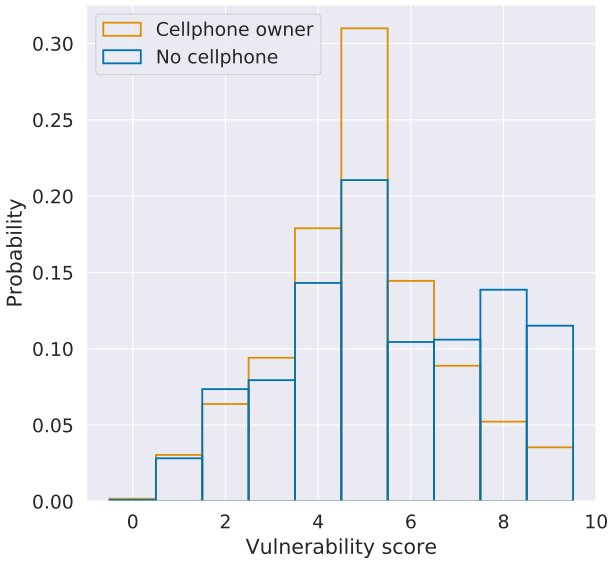



(b) In-person survey




Note: ROC curves for classifying beneficiary status using CDR-data for survey participants at the tails of the distribution of the vulnerability scores. Features extracted for the six-month time window for survey participants with a valid phone.

Figure 7A: Distribution of the vulnerability score by cellphone ownership




Note: Author's calculations using scorecard survey. A higher vulnerability score make a household more likely to be eligible for the program.