UC Berkeley
Recent Work

Title
Targeting Social Protection Programs with Machine Learning and Digital Data

Permalink
https://escholarship.org/uc/item/8wz9q7hv

Author
Aiken, Emily

Publication Date
2024-11-20

eScholarship.org Powered by the California Digital Library
University of California

Targeting Social Protection Programs with Machine Learning and Digital Data

by

Emily Aiken

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Information Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Joshua Blumenstock, Chair
Professor Jennifer Chayes
Professor Hany Farid
Pofessor Solomon Hsiang
Professor Edward Miguel

Spring 2024

Targeting Social Protection Programs with Machine Learning and Digital Data

Copyright 2024
by
Emily Aiken

Abstract

Targeting Social Protection Programs with Machine Learning and Digital Data

by

Emily Aiken

Doctor of Philosophy in Information Science

University of California, Berkeley

Professor Joshua Blumenstock, Chair

Social protection programs are essential to assisting the poor, but governments and humanitarian agencies are rarely resourced to provide aid to all those in need, so accurate *targeting* of benefits is critical. In developed economies, targeting decisions typically rely on administrative income data or broad survey-based social registries. In low-income countries, however, poverty information is rarely reliable, comprehensive, or up-to-date. Novel sources of digital data — from mobile phones and satellites, in particular — are well suited to fill this gap: they are predictive of wealth in low-income contexts and ubiquitously collected. The research studies in this dissertation design and evaluate new methods for targeting aid in low-resource contexts using machine learning, satellite imagery, and mobile phone data, and evaluate these methods in large, real-world interventions. Across social protection programs in Togo, Afghanistan, and Bangladesh, the studies in this dissertation show that targeting methods based on machine learning and digital data sources identify poor households more accurately than methods based on categorical eligibility criteria like geography or occupation, but typically less accurately than traditional survey-based poverty measurement approaches. These results highlight the potential for digital data and machine learning to improve the targeting of humanitarian aid, particularly when traditional poverty data are unavailable or out-of-date and in settings where conflict, environmental conditions, or health concerns render primary data collection infeasible. These studies also provide empirical evidence on the limitations and risks of digital and algorithmic targeting approaches, including privacy, transparency, fairness, and digital exclusion.

Contents

Acknowledgments

I am enormously grateful to my PhD advisor, Professor Joshua Blumenstock, for his academic support, research inspiration, and innumerable rounds of feedback on data analysis and writing. I would like to thank my other dissertation committee members — Professors Jennifer Chayes, Hany Farid, Solomon Hsiang, and Edward Miguel — for helpful feedback on research and career advice.

Each of the chapters in this dissertation would not be possible without the research collaborators that co-authored the papers they are based on, including Anik Ashraf, Guadalupe Bedoya, Suzanne Bellue, Aidan Coville, Raymond Guiteras, Dean Karlan, Mushfiq Mobarak, and Christopher Udry. I am also grateful to the research collaborators I have worked with on papers unrelated to this dissertation, including Susana Costenla-Villadosa, Zoe Kahn, Nitin Kohli, Tim Ohlenburg, Esther Rolf, Satej Soman, Viraj Thakur, and Rachel Warren.

My PhD research would not have been possible without the dedicated work of a number of research assistants and data scientists, including Conner Manuel, Shikhar Mehra, Leo Selker, Adrian Dar Serapio, and Nathaniel Ver Steeg. This research would also not be possible without the hard work of the teams of enumerators from INSEED and ECONS who enumerated the surveys we conducted in Togo and Bangladesh, respectively.

I would like to thank the implementing partners we worked with in Togo and Bangladesh: teams from Togo's Ministry of Digital Transformation (especially Shegun Bakari, Attia Byll, Silété Devo, Kafui Ekouhoho, Morlé Koudeka, Cina Lawson, and Leslie Mills), the Aspire to Innovate program of the government of Bangladesh (especially Abdullah Al Amin, Touhidul Islam, Tanvir Quadar, and Sumaiya Tasnim), and GiveDirectly (especially Han Sheng Chia, Abir Chowdhury, Kristen Lee, Michael Levy, Vera Lumis, Alex Nawar, Daniel Quinn, and Graham Tyler).

Finally, I am grateful for research discussions and feedback from current and former members of the Global Policy Lab at UC Berkeley, including Gabriel Cadamuro, Tamma Carleton, Guanghua Chi, Suraj Nair, Jonathan Proctor, and Simón Ramírez Amaya. Individual chapters in this dissertation have also benefited from feedback from Esther Duflo, Luis Encinas, Tina George, Rema Hanna, Ethan Ligon, Matthew Olckers, and Benjamin Olken.

Chapter 1

Introduction

Each year, roughly three trillion dollars are spent on social protection programs globally [124], making up on average 13% of each country's gross domestic product [94]. The importance of these programs increased dramatically in recent years as a result of the COVID-19 pandemic: In 2020, global extreme poverty increased for the first time in two decades, and most countries expanded their social protection programs, with more than 1.1 billion beneficiaries receiving government-led social assistance payments in 2022 [77].

Determining who should be eligible for program benefits – targeting – is a central challenge in the design of social protection programs [85, 109]. In high-income countries, social protection targeting frequently relies on tax records or other administrative data on income. In low- and middle-income countries (LMICs), where a large fraction of the workforce is informal, programs often require primary data collection. The difficulty and cost of collecting data, and the variable quality of what gets collected, can introduce significant errors in the targeting process [98, 57]. These issues are exacerbated in fragile and conflict-affected countries, where two-thirds of the world's poor are expected to reside by 2030 [53].

To address data gaps in LMICs, targeted aid programs in poor countries typically rely on a proxy measure of poverty, such as asset possessions or housing quality, that is correlated with consumption or income and collected in a universal household census [85], or community-based poverty rankings [11]. However, the construction of such poverty registries via field data collection require resources and infrastructure, and may be infeasible during humanitarian crises such as natural disasters, conflicts, and pandemics – so poverty targeting data is frequently incomplete, outdated, or altogether missing in developing contexts. The COVID-19 pandemic has demonstrated the peril of this data gap: policymakers and humanitarian organizations scrambled to provide much-needed relief, but they possessed little existing data through which to identify vulnerable populations, and could not collect this information in the midst of the pandemic.

In a line of research that has remained separate from poverty targeting until recently, several studies have shown that novel sources of "big" digital trace data – from mobile phones, satellites, and other digital sources – are predictive of wealth in developing

contexts. A set of papers have shown that the wealth of villages and small communities can be inferred at a high spatial resolution from satellite imagery [e.g. 97, 158, 50]; other work has used machine learning to detect patterns of mobile phone use that are predictive of the wealth of individual subscribers [e.g. 36, 38]. A handful of papers have also experimented with using traces from social media [69], Wikipedia [143], and other Internet-based data sources to infer local measures of poverty and well-being. In the absence of accurate and up-to-date household-level poverty data, it is possible that these proxy measures could provide useful information to administrators of targeted social assistance programs.

My research bridges the gap between poverty targeting for social protection programs and research on poverty measurement using digital trace data. My work designs new methods for targeting aid in resource-constrained contexts using machine learning (ML), satellite imagery, mobile phone data, and other sources of digital data, and evaluates these methods in large, real-world interventions. My work aims to evaluate these new methods based primarily on targeting accuracy in comparison to traditional approaches, but also on "softer" policy-relevant criteria, including targeting cost, speed of deployment, privacy, transparency, and perceived fairness and acceptability to beneficiary communities.

1.1 Traditional poverty targeting approaches in low-income contexts

Most targeted social protection programs in the developed world use means tests, restricting program benefits to those below a certain income or consumption threshold. In the developing world, however, means tests are frequently impractical, particularly in areas where most employment is in the informal sector or records of income and expenditures are limited. Most poverty targeting schemes in the developing world therefore rely on proxy measures of wealth.

One simple option is to target on geography, providing benefits to all households in the poorest parts of a country [65]. To facilitate individual or household-level targeting, many social protection programs use asset ownership as a proxy for income or consumption, either via a proxy means test [81] or an asset-based wealth index constructed with principal components analysis [70]. An increasingly popular alternative to asset-based proxies for wealth is community-based targeting, in which community members or community leaders select beneficiaries. However, there is a growing consensus in the literature that both asset-based and community-based wealth measures are limited by low-quality data, and in a subset of cases targeting based on these measures is found to be regressive or no better than universal allocation of benefits [51, 101, 41].

1.1.1 Geographic targeting

In geographic targeting, program benefits are provided to all households located in certain administrative areas of a country (typically the poorest areas). As such, the only data requirements for geographic targeting are an accurate and up-to-date map of poverty levels in each area of the country, and a registry of the general location of each household (which may be collected in voter databases, birth registrations, or similar forms of administrative data). Due to its blunt nature, geographic targeting is generally accepted as less able to prioritize benefits to the poorest households than other approaches [19], but the availability of increasingly granular poverty level may make geographic targeting a more feasible primary targeting approach [19, 65, 146]. Many social protection programs today include a component of geographic targeting prior to individual- or household-level targeting [77].

1.1.2 Survey-based targeting

Survey-based targeting relies on teams of enumerators traveling door-to-door to all households in program-eligible areas (potentially entire countries) to collect household data on poverty for eligibility determination. Ideally income or total consumption expenditure data would be collected in these surveys, but incomes are generally a poor measure of poverty in LMICs, where subsistence agriculture and informal employment are ubiquitous, and consumption data is typically too expensive to collect for a large number of households (requiring a 2-3 hour survey). Instead, survey-based targeting approaches typically rely on simple and readily verifiable information on housing quality, demographics, and asset ownership to proxy consumption expenditures.

In *proxy-means testing* (or PMT), these housing, demographic, and asset variables are collected for all potentially eligible households in a *social registry*, and consumption data are collected for a small and representative sample of households. A machine learning model (traditionally a linear regression) is trained to predict consumption from the poverty scorecard variables, and consumption predictions are produced for all households in the social registry. These predictions are then used to determine program eligibility. Since the introduction of the PMT method [81], PMTs have been deployed in a large number of countries and evaluated in many settings [e.g. 11, 85, 101, 41]. Recent work has also focused on developing more sophisticated machine learning approaches to improve PMT accuracy [116, 126].

An unsupervised alternative to proxy-means testing is using an *asset index* [70]. Asset indices do not rely on a sample survey containing consumption data for supervised learning; instead principal components analysis is used to project variation in asset ownership for roughly 10-30 assets to a unidimensional measure of asset wealth. While not as ubiquitously deployed in social protection programs as PMTs [71], asset indices are notably used for targeting Ecuador's flagship "Selben" poverty scoring algorithm [138].

1.1.3 Community-based targeting

In *community-based targeting* (CBT), members of a neighborhood or village are asked to identify the poorest households in their community to receive social protection benefits. Community-based targeting exercises may be conducted with entire communities [11], or with just selected members of the community [160, 130] or community leaders only [12]. Proponents of community-based targeting argue that it incorporates local knowledge on household conditions, gives communities agency over their own involvement with social protection programs, and may be cheaper than door-to-door survey-based approaches. However, community-based targeting may be subject to elite capture and manipulation [130], and community members may lack complete information on the poverty status of neighboring households [153].

1.2 Evaluating poverty targeting approaches

For the past two decades, development economics and public policy researchers have tried to determine which of these poverty targeting approaches is best. The primary criteria on which to evaluate targeting approaches is targeting accuracy (the extent to which each targeting method selects the poorest households), though a handful of papers have studied other indicators of success, including targeting costs, speed of roll-out, adaptivity to changes in household poverty levels, and acceptability to beneficiary communities.

1.2.1 Accuracy

Targeting accuracy is typically measured based on exclusion errors (what share of poor households were incorrectly excluded from the program?) and inclusion errors (what share program beneficiaries were "too rich"?). Exclusion errors are closely related to the measure of recall in machine learning (what share of poor households received benefits?) and inclusion errors are closely related to the measure of precision (what share of households that received benefits were poor?). In this dissertation I follow recent papers in the targeting literature [41, 140] by focusing on a quota approach to targeting accuracy evaluation: if a program targets 20% of households, I will evaluate the precision and recall of the program for reaching the 20% of households that are truly the poorest (rather than, for example, focusing on coverage of households under local or international poverty lines). In the quota setting, precision and recall are equal by definition [41].

Targeting accuracy is typically evaluated using a sample survey of households in program-eligible areas. For these households, the necessary targeting variables are collected, along with detailed measures of consumption expenditure. "True" poverty is determined based on consumption expenditure, and the targeting accuracy of one or many targeting approaches is simulated using the remaining survey data.

A number of papers have studied the targeting accuracy of geographic targeting, survey-based targeting, and community-based targeting. Of primary interest are direct comparisons between PMT and CBT approaches; in general studies have found that PMTs are slightly more accurate (10-13% lower inclusion and exclusion error rates) than CBTs [11, 130]. However, a number of studies have pointed out that both PMTs and CBTs tend to have high inclusion and exclusion error rates [51, 41, 105, 140], typically on the order of 35-55% inclusion and exclusion errors for a program aiming to target the poorest 20% of households in a country [41]. These high error rates point to the difficulty of targeting poverty using proxy measures in general.

1.2.2 Poverty impacts

A few studies have attempted to compare targeting approaches based on projected poverty impacts of different approaches (with knowledge of the amount of benefits provided to each beneficiary household). Such analyses have been conducted by measuring the extent to which different targeting methods impact the share of households below the poverty line [41, 140] or by using a utility function, such as constant relative risk aversion (CRRA) utility [85]. Several of these papers conclude that differences among similarly-accurate targeting approaches (such as PMT and CBT) are negligible when measured by projected poverty impacts [41, 140].

1.2.3 Costs

Data from social protection program administrators on targeting costs are not widely reported. Geographic targeting is likely to be much cheaper than alternative targeting approaches, as it may rely entirely on existing administrative data and poverty maps constructed with sample surveys. Household targeting strategies requiring primary data collection are likely to be substantially more expensive. Between Alatas et al. (2012) [11], Karlan and Thauysbaert (2019) [101], and Schnitzer and Stoeffler (2021) [140], targeting costs per household are reported for twelve social protection programs in Africa, South America, and Asia using PMTs or CBTs for targeting. In general targeting costs for PMTs (median of $4.00 per household screened) are slightly higher than for CBTs (median of $2.20 per household screened). In general, targeting costs in the programs studied make up roughly 1-6% of total program benefits distributed.

1.2.4 Speed and adaptivity

A concern with all traditional approaches to poverty targeting is that poverty proxies collected in field surveys take a long time to collect, and they are likely to eventually become out-of-date. Most large government-run social protection programs aim to reassess eligibility via primary data collection roughly every 2-3 years, but in reality primary data collection typically occurs once every 5-8 years [96, 26]. As poverty proxies

become out-of-date, the accuracy of targeting on these proxies degrades [41, 90]. Social protection administrators have recently aimed to fill temporal gaps in data collection with government administrative data, on-demand reassessment of eligibility criteria, and digitally-delivered poverty questionnaires, but these innovations in real-time data collection remain nascent [26].

1.2.5 Acceptability to beneficiary communities

Beyond targeting accuracy, the second most studied concern in social protection targeting to date is acceptability to beneficiary communities. Two studies have compared PMTs and CBTs in terms of acceptability and perceived fairness, with contrasting results: Premand and Schnitzer (2021) [130] find that villages in Niger dismiss community-based targeting as less legitimate than survey-based approaches (perhaps due to elite capture), while Alatas et al. (2012) [11] find that neighborhoods in Indonesia perceive CBTs as more accurate and fair than PMTs. One possible explanation for push-back against PMTs is the potential for household "gaming" or manipulation of the scoring criteria by misreporting asset possession and housing characteristics, a phenomenon documented in several PMT-targeted programs [46, 23]. A related line of literature has questioned the premise of selective inclusion in social protection programs altogether, pointing to unintended negative impacts of targeting on social cohesion including stigmatization of beneficiaries [134], jealousy [60], and exclusion of marginalized groups [128]. These papers suggest that some communities would prefer geographic targeting or universal inclusion over targeted benefits – even if the benefits delivered to each household are substantially reduced as a result.

1.3 Measuring poverty with machine learning and digital data

In a new area of research coming mainly from computer science and data science communities, a number of studies in the past decade have documented how poverty can be predicted from digital trace data sources, either spatially or at an individual level. Here we focus on the two main data sources explored in the research literature to date: satellite imagery and mobile phone data.

1.3.1 Satellite imagery

Satellites are recording high-resolution images of the entire Earth's surface on a daily to weekly basis, collecting and storing unprecedented amounts of information about human behavior and planetary conditions. Much of this imagery encodes high resolution information about population density, economic well-being, and environmental conditions that can inform poverty measurement. Jean et al. (2016)

introduced the first transfer learning pipeline to produce high resolution poverty maps from satellite imagery, with relatively accurate results (R^2 = 0.40-0.55 at the village level in Nigeria, Tanzania, Uganda, and Malawi) [97]. Yeh et al. (2020) reproduced the transfer learning pipeline with publicly available satellite imagery [158], and Chi et al. (2022) [50] produced publicly available gridded poverty predictions for all low- and middle-income countries at the 2.4 x 2.4 satellite tile level. Other recent work has focused on improving specific aspects of the satellite-based poverty prediction pipeline, including accessibility [135], explainability [66, 17, 154], and cost of imagery [88].

Recent work has also looked at detecting changes in poverty levels with satellite imagery, though results are mixed. Yeh et al. (2020) [158] find that year-to-year changes in asset-based wealth are difficult to detect over time from satellite images alone (R^2 = 0.15-0.17 in Sub-Saharan Africa), but Huang et al. (2022) [93] show that welfare changes from a very large cash transfer program in Uganda can be detected from high-resolution satellite images of beneficiary households, and Ratledge et al. (2022) [132] show that poverty impacts of electrification in Rwanda can be recovered from sequential years of imagery.

1.3.2 Mobile phone data

Cell phones have become increasingly ubiquitous worldwide, projected to reach a global penetration rate of 73% in 2025 [82]. Recent work has shown that machine learning methods leveraging mobile phone metadata (call detail records, or CDR) can produce useful estimates of wealth and well-being at a fine spatial granularity. This body of work focuses largely on poverty, typically quantified by an asset-based wealth index [36, 38, 145, 89], but related papers explore a wider set of well-being measures, including literacy [139], food security [59], and infrastructure [37].

While most of this work addresses spatially granular poverty mapping, two papers cover individual-level wealth prediction. Blumenstock et al. (2015) show that CDR data are predictive of an individual-level asset-based wealth index in Rwanda [36]. More specifically, the study matches ground-truth survey data to CDR covering two years of phone activity for 856 geographically stratified individuals, extracts a suite of thousands of behavioral indicators from the CDR, and applies a supervised learning algorithm to generate wealth predictions from behavioral indicators. Model accuracy is evaluated with cross-validation to ensure that the wealth prediction model generalizes out-of-sample (cross-validated Pearson's correlation = 0.68). Blumenstock (2018) performs the same experiment for 1,234 male heads of households in the Kabul and Parwan districts of Afghanistan, yielding similar predictive accuracy [38].

1.4 Uses of digital data for targeting social protection programs

Prior to the COVID-19 pandemic of 2020-2022, there was little use of digital data sources in real-world social protection programs: instead programs relied on traditional targeting approaches like PMTs and CBTs. However, the difficulty of primary data collection during the COVID-19 pandemic — and the urgency of social protection benefits distribution during the resulting lockdowns — pushed social protection program administrators to turn to digital data sources for targeting for the first time. The second chapter of this dissertation evaluates one such digitally-targeted COVID-19 aid program in Togo.

Aiken and Ohlenburg (2023) summarize uses of digital data sources for social protection targeting during and after the COVID-19 pandemic [3]. A few examples highlight the variety of novel data sources put to use for social protection targeting in the past three years:

- In Niger, satellite imagery is used to determine areas of drought for famine-responsive cash transfer targeting [42].
- In Costa Rica, satellite imagery is used to identify poor areas with low social registry coverage for outreach campaigns by social workers [3].
- In Togo, satellite-based poverty maps determined eligible areas and poverty predictions from mobile phone records determined individual eligibility [5]. An evaluation of the targeting of Togo's cash transfer program is the second chapter of this dissertation.
- In the Democratic Republic of the Congo, individual estimated home locations based on cell tower usage determined eligibility for an aid program in Kinshasa [122].
- In Colombia, credit scoring data — along with data on mobile phone and financial services usage – were used to inform eligibility for emergency cash transfers [111].
- In South Africa, bank account balances were used as input to a means test to determine eligibility for the social relief of distress grant during the pandemic [80].

1.5 Contributions of this dissertation

This dissertation provides the first rigorous evidence on the suitability of machine learning and digital data sources for targeting social protection programs. In this dissertation, I study real-world social protection programs in Togo, Afghanistan,

and Bangladesh, and assess the accuracy of targeting using poverty metrics inferred with machine learning and digital data (mobile phone data and satellite imagery) in comparison to traditional poverty targeting approaches, including geographic targeting, proxy-means testing, and community-based targeting. Where applicable, I also provide evidence on the fairness of these approaches, and discuss other social and ethical issues relevant to algorithmic targeting, including privacy, transparency, manipulation, social cohesion, and digital exclusion.

The contributions of each chapter in the dissertation are as follows:

- In **Chapter 2** I evaluate the targeting of Togo's COVID-19 cash transfer program, *Novissi*, which used satellite imagery to select program-eligible regions and mobile phone data to target transfers to subscribers estimated to be poor based on phone use patterns. I find that Novissi's phone-based targeting was more accurate than other targeting approaches that were feasible in Togo during the pandemic (such as geographic or occupation-based targeting), but less accurate than traditional poverty targeting methods like proxy-means testing. I also provide empirical evidence on the limitations of Novissi's digital targeting approach in terms of privacy, fairness, and digital exclusion.

- In **Chapter 3** I study an ultra-poverty targeting program in Afghanistan, which used a combination of community-based targeting and verification of categorical ultra-poverty criteria for beneficiary selection. I simulate phone-based targeting, and find that phone-based targeting is as accurate as proxy means testing for recovering the program's original targeting benchmark. I also find that combining phone-based and survey-based information (including proxy-means test components and data on consumption expenditures) improves targeting over using a single data source.

- In **Chapter 4** I compare phone-based targeting to a new alternative for identifying the poorest households: community-based targeting. In the setting of Cox's Bazar, Bangladesh, I show that phone-based targeting is more accurate than community-based targeting for identifying the consumption-poorest households (but, consistent with the results in Chapter 3, less accurate than a proxy-means test). In contrast with the results in Chapter 4, I find little utility in combining targeting methods, and few dimensions of heterogeneity that affect targeting accuracy.

- In **Chapter 5** I return to Togo's *Novissi* program, this time studying whether mobile phone data can be used to track the program's welfare impacts. Leveraging a randomized controlled trial in which some RCT participants received transfers and others did not, I train machine learning models to predict program outcomes (including food security, mental health, and asset-based wealth) from mobile phone data and assess whether differences in phone-predicted outcomes between the treatment and control groups match differences in outcomes collected in a standard survey. I find that mobile phone data is not a useful predictor of any of the outcomes

studied besides asset-based wealth, and as a result mobile phone data does not recover the *Novissi* program's treatment effects.

I conclude the dissertation by discussing five key directions of future research for algorithmic and digitally-driven targeting of social protection programs.

Chapter 2

Targeting aid with machine learning and digital data in Togo

This chapter is based on the paper "Machine learning and phone data can improve targeting of humanitarian aid" [5], written in collaboration with Suzanne Bellue, Dean Karlan, Christopher Udry, and Joshua Blumenstock.

Abstract

The COVID-19 pandemic devastated many low- and middle-income countries, causing widespread food insecurity and a sharp decline in living standards. In response to this crisis, governments and humanitarian organizations worldwide distributed social assistance to more than 1.5 billion people. Targeting is a central challenge in administering these programs: it remains a difficult task to rapidly identify those with the greatest need given available data. Here we show that data from mobile phone networks can improve the targeting of humanitarian assistance. Our approach uses traditional survey data to train machine-learning algorithms to recognize patterns of poverty in mobile phone data; the trained algorithms can then prioritize aid to the poorest mobile subscribers. We evaluate this approach by studying a flagship emergency cash transfer program in Togo, which used these algorithms to disburse millions of US dollars worth of COVID-19 relief aid. Our analysis compares outcomes—including exclusion errors, total social welfare and measures of fairness—under different targeting regimes. Relative to the geographic targeting options considered by the Government of Togo, the machine-learning approach reduces errors of exclusion by 4–21%. Relative to methods requiring a comprehensive social registry (a hypothetical exercise; no such registry exists in Togo), the machine-learning approach increases exclusion errors by 9–35%. These

results highlight the potential for new data sources to complement traditional methods for targeting humanitarian assistance, particularly in crisis settings in which traditional data are missing or out of date.

2.1 Introduction and context

The COVID-19 pandemic led to a sharp decline in living standards across the world, as policies designed to stop the spread of the disease have disrupted ordinary economic activity. Economically vulnerable households in low- and middle-income countries were among the hardest hit, with over 100 million individuals estimated to have transitioned into extreme poverty during the pandemic [107].

To offset the most severe consequences of this sudden income decline, governments and humanitarian organizations around the world mobilized relief efforts. Gentilini et al. (2022) [77] estimate that over 3,300 new social assistance programs were launched during the pandemic, providing over $800 billion dollars in cash transfer payments to over 1.5 billion people (roughly one fifth of the world's population).

The overwhelming majority of COVID-19 response efforts — like the majority of cash transfer programs globally — provided targeted social assistance [85, 109]. However, as described in the introduction, in low and lower-middle income countries (LMICs), where economic activity is often informal and based on home-produced agriculture, governments typically do not observe income for the vast majority of the population [85]. Other potential sources of targeting data are often incomplete or out of date [98, 142]; for example, only half of the poorest countries having completed a census in the past 10 years [158]. In such contexts, data gaps preclude governments from implementing well-targeted social assistance programs [25, 51].

This chapter describes the development, implementation, and evaluation a new approach to targeting social assistance based on machine learning algorithms and non-traditional "big data" from satellites and mobile phone networks. This approach leverages the recent advances in machine learning that show that such data can help accurately estimate the wealth of small geographic regions [97, 66, 148, 129, 50] and individual mobile subscribers [36, 38, 6]. It also builds on a rich economics literature on the design of appropriate mechanisms for targeting social assistance [85, 125, 19, 81, 12, 11, 10, 41, 116]. See Section 1.1 for a discussion of prior work on this topic.

2.1.1 Humanitarian Response to COVID-19 in Togo

Togo is a small country of roughly 8 million in West Africa. Over 50% of the population lives below the international poverty line. Shortly after the first COVID-19 cases were confirmed in Togo in early March 2020, the government imposed economic lockdown orders to prevent the spread of the disease. These lockdowns forced many Togolese to stop working, raising concerns about the potential for rising food insecurity (Figure S1).

On April 8, 2020, the government launched the *Novissi* program, where *Novissi* means "solidarity" in the Ewé language. According to Minister Cina Lawson, Novissi "was built and designed in order to help those people who are the most vulnerable population and the most impacted by the anti-COVID measures."[1] The government of Togo did not have a traditional social registry that could be used to assess program eligibility, however, and had neither the time nor the resources to build such a registry in the middle of the pandemic. The most recent census, which was completed in 2011, did not contain information on household wealth or poverty; more recent national surveys on living standards only contacted a small fraction of all households.

Novissi's first phase

Novissi was initially designed to provide benefits to informal workers in Greater Lomé, the large metropolitan area surrounding the capital city where the lockdown orders were initially focused. The decision to target informal occupations helped prioritize benefits to people who were forced to stop working at the onset of the crisis. However, this approach does not necessarily target benefits to the poorest households in the country (Figure S2). To determine eligibility for Novissi, the government relied upon a national voter registry that was updated in late 2019, in which individuals indicated their home location and occupation. At the time, the voter registry contained 3,633,898 entries, which the electoral commission reports is equivalent to 87% of the total adult population (see Table 2.2 for details).

Receiving Novissi benefits required that individuals register by dialing in to the Novissi USSD platform from a mobile phone. Thus, registration initially required (i) a valid and unique voter ID linked to an eligible occupation from an eligible location; (ii) a valid SIM card, and (iii) access to a mobile phone. A smartphone was not required for registration; the USSD platform was accessible from a basic phone. Since phone sharing is common in Togo, multiple SIM cards could be registered through a single phone (so long as each SIM was then linked to a valid voter ID).[2]

Eligible female beneficiaries were then paid 12,250 FCFA (USD $22.50) per month; men received 10,500 FCFA (USD $20) per month. The payments were disbursed in two bi-weekly installments, for three months, using existing mobile money infrastructure managed by the country's two mobile network operators. The system was designed to be 100% digital, so that registration, eligibility determination, and payment could all be accomplished without face-to-face contact. Novissi was promoted actively through radio advertisements and community leaders, and 4.4 million registration attempts were reported on the day the program launched. In this first phase of Novissi, which focused on Greater Lomé, roughly 510,000 beneficiaries received payments.

[1] https://undp-ric.medium.com/cina-lawson-a-covid-cash-transfer-programme-that-gives-more-money-to-women-in-togo-2386c5dff49

[2] See subsection 2.3.5 for a discussion of the extent to which voter and phone requirements may have led to program exclusions.

Novissi's second phase

Our analysis focuses on a second phase of Novissi, which was initiated after the Novissi program in Greater Lomé had terminated. Specifically, in partnership with the NGO GiveDirectly, the government wished to expand Novissi eligibility to the rural poor. The policy mandate from the government was to (i) prioritize benefits to people living in Togo's 100 poorest cantons (of the 397 cantons nationally), where the number 100 was selected by the government in order to balance the desire to focus on the poorest villages, without focusing excessively on specific regions; and (ii) prioritize the poorest individuals in those 100 cantons.

During the second phase of Novissi, registration and enrollment used several of the same steps described above: individuals were required to have a voter ID registered in one of the 100 poorest cantons, and they had to self-register using a mobile phone with a unique SIM card. However, Novissi's targeting approach changed in its second phase. Our research efforts focused on helping the government expand the targeting of the Novissi program from informal workers in Greater Lomé to poorer individuals in rural regions of the country, and were designed to meet the government's two stated policy objectives: first, to direct benefits to the poorest geographic regions of the country; and second, to prioritize benefits to the poorest mobile subscribers in those regions.[3]

2.1.2 Novissi's targeting approach

We worked with the government of Togo and GiveDirectly to develop the targeting approach for Novissi's second phase, which uses machine learning to analyse non-traditional data from satellites and mobile phone networks, has two distinct steps (Figure 2.1).

In the first step, we obtained public micro-estimates of the relative wealth of every 2.4 km by 2.4 km region in Togo, which were constructed by applying machine-learning algorithms to high-resolution satellite imagery [50]. These estimates provide an indication of the relative wealth of all the households in each small grid cell; we take the population-weighted average of these grid cells to estimate the average wealth of every canton, Togo's smallest administrative unit (see subsection 2.2.2).

In the second step, we estimated the average daily consumption of each mobile phone subscriber by applying machine-learning algorithms to mobile phone metadata provided by Togo's two mobile phone operators. Specifically, we conducted surveys with a large and representative sample of mobile phone subscribers, used the surveys to measure the wealth and/or consumption of each subscriber, and then matched the survey-based estimates to detailed metadata on each subscriber's history of phone use. This sample was used to train supervised machine-learning algorithms that predict wealth and consumption from phone use (see subsection 2.2.5) [36, 38, 6]. This second

[3]Individuals without access to a mobile phone could not receive Novissi payments, which were digitally delivered using mobile money – see subsection 2.3.5 for details.

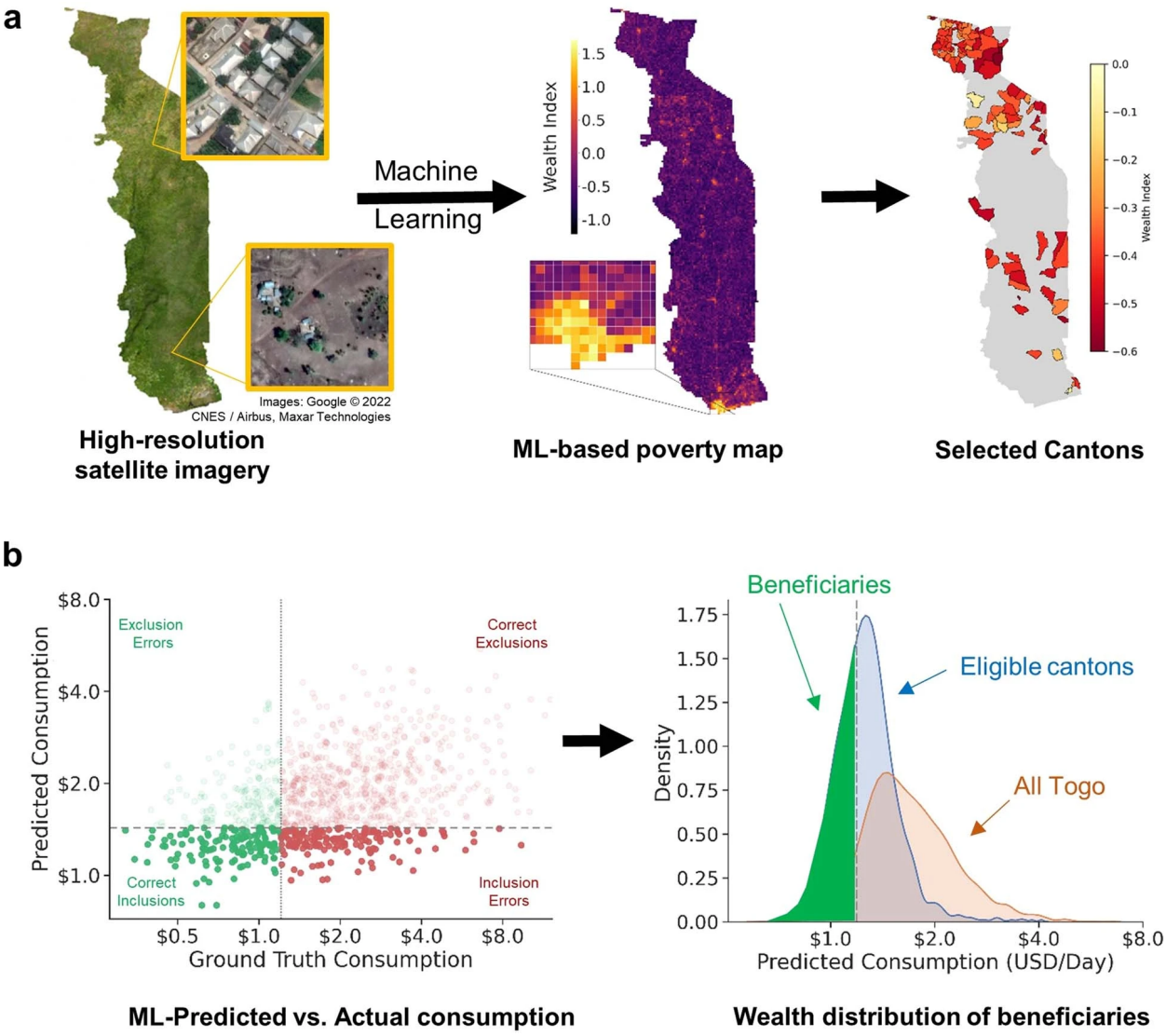

Figure 2.1: (a) Regional targeting. Satellite imagery of Togo is used to construct micro-estimates of poverty (middle) [50], which are overlayed with population data to produce canton-level estimates of wealth. Individuals registered in the 100 poorest cantons (right) are eligible for benefits. (b) Individual targeting. A machine learning algorithm is trained using representative survey data to predict consumption from features of phone use (subsection 2.2.5). The algorithm constructs poverty scores that are correlated with ground-truth measures of consumption (left). Subscribers who register for the program in targeted cantons with estimated consumption less than USD \$1.25/day are eligible for benefits (right). The red distribution shows the predicted wealth distribution of the entire population of Togo; the blue distribution shows the predicted wealth distribution in the 100 poorest cantons; and the green section indicates the predicted wealth distribution of Novissi beneficiaries.

step is similar in spirit to a traditional proxy means test (PMT), with two main differences: we used a high-dimensional vector of mobile phone features instead of a low-dimensional vector of assets to estimate wealth; and we used machine-learning algorithms designed to maximize out-of-sample predictive power instead of the traditional linear regression that maximizes in-sample goodness of fit [116].

2.2 Methods

2.2.1 Survey data

Our core analysis relies heavily on two surveys conducted by Togo's Institut National de la Statistique et des Études Economiques et Démographiques (INSEED). The first survey, which is nationally representative, was conducted in the field in 2018 and 2019 (N = 6,171). The second survey was conducted over the phone in September 2020, and is representative of mobile network subscribers inferred to be living in rural cantons eligible for Novissi aid (N = 8,915). We use these two different survey datasets because neither dataset is sufficient by itself for the analysis we require: the 2020 survey did not collect consumption data, which is important for evaluating certain counterfactuals; the 2018–19 survey is representative only at the prefecture level, and only surveyed a small number of households in the 100 poorest cantons that were eligible for Novissi. (We had planned to conduct a large in-person survey in early 2021 that would provide the single point of focus for this chapter, but were forced to postpone the survey indefinitely owing to a resurgence in COVID-19.)

2018–2019 field survey

Our first survey dataset was obtained from a nationally representative household survey. Specifically, 540 enumeration areas (EAs) were drawn at random from Togo's approximately 6,000 EAs, with weight proportional to the size of the EA in the last national census (conducted in 2011). Twelve households were then drawn at random from each of the selected EAs to be interviewed, for a total of 6,172 households. Surveys, which lasted about 3 hours, were conducted in two waves, with the first wave between October and December 2018 and the second wave between April and June 2019. We removed one observation that is missing consumption expenditure and asset data, leaving 6,171 observations. Interviews took place with the head of household when possible, and alternatively with the most knowledgeable adult present. Answers were recorded by enumerators on tablets using SurveyCTO software.

As part of the survey's recontact protocol, phone numbers were requested from a representative of each household; 4,618 households (75%) of households are matched to a phone number. The data do not include an identifier for which member of the household the phone number belongs to. A total of 4,171 households have phone numbers that

contain at least one transaction in our mobile phone transaction logs in the three months prior to their survey date (90% of households with phone numbers), leading to a matched survey–mobile phone dataset with N = 4,171. Note that this matched dataset is not nationally representative or necessarily representative of mobile phone subscribers, as there is selection in which households and household members provide phone numbers.

2020 phone survey

Our second survey dataset is obtained from a phone survey conducted over two weeks in September 2020. The survey lasted approximately 40 minutes, and covered demographics, asset ownership and well-being. Answers were recorded by enumerators on tablets using SurveyCTO software. Phone numbers for the 2020 phone survey were drawn from mobile phone transaction logs and the sample is representative of subscribers inferred based on their mobile phone data to be living in rural cantons eligible for Novissi aid (see appendix A.2). Note that because the sample is drawn based on inferred location, not all interviewees necessarily reside in an aid-eligible canton. The survey includes a question on canton of residence, and 68% of observations report living in a Novissi-eligible canton.

Of the phone numbers drawn, 35% responded, consented to the survey, and completed the entire survey. In total, after removing low-quality surveys and those missing poverty outcomes, the dataset contains 8,915 observations corresponding to individual subscribers. We reweight the survey for nonresponse using the same mobile phone features and machine-learning methods described in subsection 2.2.5. Our sample weights consist of the inverse of the draw probability and the inverse of the predicted probability of response. More details on the content of the 2020 phone survey, the sampling procedure, and the reweighting procedure are available in appendix A.2.4.

Construction of poverty outcomes

We construct four poverty outcomes from the survey data: consumption expenditure (captured in the 2018–2019 field survey only), an asset-based wealth index, a poverty probability index (PPI), and a PMT.

Consumption expenditure. The consumption expenditure outcome is only available in the dataset from the 2018–2019 field survey. Disaggregated expenditures for more than 200 food and non-food items are elicited in each household interview. The consumption aggregate is then adjusted for a price index calculated at the prefecture level. The final outcome measure is per capita adult equivalent household consumption expenditure, which we transform to US$ per day.

Asset index. We calculate a principal component analysis (PCA) asset index for households in the 2018–2019 field survey and for the households associated with

individuals interviewed in the 2020 phone survey. Asset indices are constructed with a PCA. The asset index is constructed from 24 underlying binary asset variables in the 2018–2019 field survey and 10 underlying binary asset variables in the 2020 phone survey. The asset indices for the two surveys are constructed independently, from different sets of assets, and therefore do not share a basis vector. The basis vector for each index is shown in Table S7.

The asset index explains 31.50% of the variance in asset ownership in the 2018–2019 field survey, and 53.45% of the variance in asset ownership in the 2020 phone survey. However, the variance explained in the two indices should not be directly compared since there are far fewer assets recorded in the 2020 phone survey than in the 2018–2019 field survey. We also note that the asset index for the 2020 phone survey dataset is dominated by variation in ownership of three assets (toilet, radio and motorcycle; see Table S7) and is therefore considerably less smooth than the asset index in the 2018–2019 phone survey dataset.

Poverty probability index. We use the scorecard for the current poverty probability index (PPI) used by Innovations for Poverty Action.[4] The index is calibrated based on a nationally representative survey conducted by INSEED in 2015 (N = 2,335). "Poverty probability" is scored based on ten household questions, including region of residence, education of adults and children, asset ownership, and consumption of sugar. We calculate the PPI only for households in the 2018–2019 field survey, as the data necessary for all components were not collected in the 2020 phone survey.

PMT. Using the data from the 2018–2019 field survey, we follow a stepwise forward selection process to select the 12 asset and demographic variables that are jointly most predictive of per capita household consumption (see Figure S4 and appendix A.1 for details). We use these variables to construct a consistent PMT for the 2018–2019 field survey and the 2020 phone survey. Following recent literature, we use a regularized linear model (Ridge regression) rather than a simple linear regression to maximize out-of-sample accuracy [116, 126]. For the 2018–2019 field survey, PMT consumption estimates are produced out-of-sample over tenfold cross validation. For the 2020 phone survey, we train the Ridge regression on the entire 2018–2019 field survey sample and use the fitted model to produce PMT consumption estimates for each phone survey observation. Over tenfold cross validation, the PMT explains 48.35% of the variance in log-transformed consumption expenditure in the 2018–2019 field survey. This explanatory power is similar to that of other national-scale PMTs reported in Indonesia, Peru and Jamaica(41%–66%) [85, 81]. The weights for the PMT are included in Table S8. As they are trained to predict consumption, PMT consumption estimates can be interpreted as estimated US$ per day.

[4]https://www.povertyindex.org/country/togo

Rural-specific PMT. We follow another stepwise forward selection process using the 2018–2019 field survey restricted to households in rural areas (N = 3,895) to create a PMT specific to rural areas with 12 components. The weights for the rural-specific PMT are shown in Table S9. Over ten-fold cross-validation the rural-specific PMT explains 17% of the variation in log-transformed consumption expenditure in the 2018–2019 field survey restricted to rural areas. We note that this explanatory power is substantially lower than that of other rural-specific PMTs evaluated in past work in Jamaica and Burkina Faso (36%–45%) [81, 83]. We produce out-of-sample values for the rural-specific PMT over cross validation for the 2018–2019 field survey, and use the fitted model to produce values for the 2020 phone survey. We mean-impute the rural-specific PMT for observations that do not have all necessary components in the 2020 phone survey dataset (N = 18). The correlation between the rural-specific PMT and general PMT is 0.75 in the 2018–2019 survey dataset restricted to rural areas, and 0.76 in the 2020 phone survey dataset.

Construction of occupation categories

We use self-reported occupation (of the household head for the 2018–2019 field survey, and of the respondent for the 2020 phone survey) to categorize occupations and later simulate occupation-based targeting. We first classify each of the self-reported occupations according to the occupation categories in the Novissi registry. We identify which of these categories are informal (in the Novissi registry, more than 2,000 unique occupations are considered informal—some of the most common ones are vendors, hairdressers, taxi drivers, tailors, construction workers and the unemployed). We further classify occupations in 10 broad categories according to the Afrostat system.[5] Table S10 records these categories, along with the proportion in each category in each of the two surveys and associated average consumption.

Summary statistics

Table S11 presents summary statistics on each of the two surveys; for the 2018–2019 household survey, results are presented separately for households who provide phone numbers (further broken down into those with phones numbers that match to the mobile phone metadata and those whose phone numbers do not match), and those without phone numbers. Note that since phone numbers for the 2018–2019 household survey were collected for a recontact protocol, a household without a phone number could represent a household without a phone or one that refused to be contacted for further surveys. We find that households providing phone numbers (average consumption = US$2.56 per day) are less poor than households not providing them (average consumption = US$1.75 per day); among those associated with a phone number, households that do not match to mobile phone metadata (average consumption = US$2.21 per day) are poorer than those that do (average consumption = US$2.59 per day). These

[5]https://www.afristat.org/nomenclatures/

patterns are consistent with related work in Afghanistan in which phone numbers were collected for the purpose of matching to mobile phone metadata. That study found that households with phones were wealthier than those without, and households associated with a matched phone number were wealthier than those that did not match [6].

Comparing summary statistics from the 2020 phone survey and 2018–2019 household survey, respondents to the 2020 survey tend to be poorer (average PMT 1.62 vs. 2.10), younger (average age = 33 vs. 44), and more predominantly male (23% women vs 28% women). These differences are not surprising given that the 2020 survey was conducted in rural areas whereas the 2018–2019 household survey was designed to be nationally representative.

2.2.2 Poverty maps

To simulate geographic targeting, we rely on poverty maps of Togo's prefectures (admin-2 level, 40 prefectures) and cantons (admin-3 level, 397 cantons). In the 2018–2019 field survey, the latitude and longitude of each household were recorded by enumerators as part of the interview, so we map each observation to a prefecture and canton using the geographic coordinates. For the 2020 phone survey, we ask each respondent to report their prefecture and canton of residence.

Prefecture poverty map

INSEED completed a survey-based poverty mapping exercise in 2017. Specifically, a PMT was calibrated on a small consumption sample survey conducted in 2015 (N = 2,335). 26,902 households were then surveyed in the field over three weeks in 530 EAs, sampled to be representative at the prefecture level. The interview included questions on demographics, education, asset ownership, and household characteristics that made up the PMT. The calibrated PMT was then used to infer the "consumption" of each household, and observations were aggregated to estimate the percentage of the population living under the Togo-specific poverty line of US$1.79 per day in each prefecture. Figure S5 shows the resulting poverty map. For validation, we evaluate the correlation between prefecture-level poverty rates from the poverty mapping exercise and average consumption in the 2018–2019 field survey. The Pearson correlation coefficient is -0.78, and the Spearman correlation coefficient is -0.70.

Canton poverty map

When COVID-19 first appeared in Togo in early 2020, it had been at least ten years since a household survey had been conducted in Togo that was representative at the canton level. Togo's last census was conducted in 2011, but did not include information on income, consumption, or asset ownership. We therefore rely on recently-produced publicly available satellite-based estimates of poverty which use deep learning models

trained on Demographic and Health Surveys (DHS) data from neighbouring countries to estimate the average relative wealth of each 2.4km tile in Togo [50]. We overlay the resulting tile-level wealth estimates with high-resolution estimates of population density inferred from satellite imagery Tiecke et al. (2017) [152] to obtain population-weighted average wealth estimates for each canton, shown in Figure S5. As noted in Chi et al. (2022) [50], the relative wealth measures are estimated with uncertainty. Thus, for validation, we evaluate the canton-level correlation between average wealth from the satellite-based poverty map and average consumption in the 2018–2019 field survey (though note that the latter survey is not representative at the canton level). The Pearson correlation coefficient is 0.57, and the Spearman correlation coefficient is 0.52.

2.2.3 Mobile phone metadata

We obtain mobile phone metadata (call detail records, or CDR) from Togo's two mobile network operators for certain time periods in 2018–2021. We focus on three slices of mobile network data: October–December 2018, April–June 2019 and March–September 2020. The three-month periods in 2018 and 2019 are matched to households interviewed in the first and second wave of the field survey, respectively. The seven-month period in 2020 is matched to outcomes for individuals interviewed in the phone survey in September 2020. Summary statistics on network activity in these periods are shown in Figure S6.

Our CDR data contain the following information. Calls: caller phone number, recipient phone number, date and time of call, duration of call, ID of the cell tower through which the call is placed; SMS messages: sender phone number, recipient phone number, date and time of the message, ID of the antenna through which the message is sent; mobile data usage: phone number, date and time of transaction, amount of data consumed (upload and download combined); mobile money transactions: Sender phone number, recipient phone number (if peer-to-peer), date and time of the transaction, amount of transaction, and broad category of transaction type (cash in, cash out, peer-to-peer or bill pay).

October–December 2018 and April–June 2019 phone data

Between 1 October and 30 December 2018, there were a total of 4.84 million unique mobile network subscribers between the two mobile phone networks (where a subscriber is any phone number that places at least one call or SMS on a network). Between 1 April and 30 June 2019, there were a total of 4.89 million mobile network subscribers. We identify spammers on the network as any phone number that placed an average of over 100 calls or 100 SMS messages per day, and remove any transactions associated with these numbers from our dataset. We remove 232 spammers in the 2018 time period and 162 spammers in the 2019 time period. In the 2018–2019 CDR, we observe only calls, SMS messages, and mobile money transactions (we do not observe mobile data usage).

March–September 2020 phone data

For data between March 1 and September 30, 2020, we observe a total of 5.83 million mobile network subscribers (note that this subscriber population does not necessarily reflect a 19% increase in subscribers from 2018–2019, since the slice is seven months rather than three months and there is significant month-to-month churn in subscribers; during the 3-month period from July–September 2020 we observe 5.20 million unique subscribers, a 6% increase from the 2019 period). We identify spammers as described above, resulting in the removal of transactions associated with 107 spammers from the 2020 CDR dataset. In the 2020 CDR, we observe calls, SMS messages, mobile data usage, and mobile money transactions.

Featurization

For each subscriber observed on the network in each of the three time periods, we calculate a set of 857–1,042 "CDR features" that describe aspects of the subscriber's mobile phone behaviour. These include:

Call and SMS features. We use open-source library bandicoot [120] to produce around 700 features relating to the calls and SMS messages each subscriber places and receives. These range from general statistics (for example, number of calls or SMS messages, or balance of incoming versus outgoing transactions), to social network characteristics (for example, number and diversity of contacts), to measures of mobility based on cell tower locations (for example, number of unique towers and radius of gyration).

Location features. Based on the locations of each of the cell towers in Togo, we calculate information about where each subscriber places their transactions. Specifically, we calculate the number and percentage of calls placed in each of Togo's 40 prefectures, and the number of unique antennas, cantons, prefectures, and regions that each subscriber visits.

International transaction features. Using country codes associated with phone numbers, we calculate the number of outgoing international transactions, separately for calls and SMS messages. We also calculate the total time spent on outgoing international calls.

Mobile money features. For each of four variables relating to transaction size–transaction amount, percent of balance, balance before transaction, and balance after transaction–we calculate the mean, median, minimum, and maximum, separately for incoming and outgoing mobile money transactions. We also calculate the total transaction count for each subscriber (separately for incoming and outgoing) and the total number of unique mobile money contacts (separately for incoming and outgoing). We perform these

calculations for all transactions together, as well as separately by transaction type (cash in, cash out, peer-to-peer, bill payments and other transactions).

Mobile data features. We calculate the total, mean, median, minimum, and maximum mobile data transaction for each subscriber, as well as the standard deviation in transaction size. We also calculate the total number of mobile data transactions and the number of unique days on which data is consumed. Note that mobile data features are only calculated for the 2020 CDR period, as our 2018–2019 CDR does not include mobile data records.

Operator. In our feature dataset we include a dummy variable for which of the two mobile network operators each subscriber is associated with.

2.2.4 Matching survey and phone datasets

Using phone numbers collected in surveys, we match survey observations to CDR features. As noted in 2.2.1, there are 4,618 households in the 2018–2019 field survey that provide a phone number, of which 4,171 match to CDR (90% of households with phone numbers, and 68% of households overall). We match households surveyed in the first survey wave to features generated in the October–December 2018 CDR period, and households surveyed in the second survey wave to features generated in the April–June 2019 CDR period. To build intuition on the relationships between phone-related features and poverty, Figure S7 compares four CDR features for those above and below the poverty line in the 2018–2019 household survey. As the 2020 survey was sampled based on the CDR dataset, all 8,915 observations in the 2020 survey dataset are matched to CDR.

Data privacy concerns

The CDR data we obtained for each subscriber contain personally identifying information (PII) in the form of the subscriber's phone number (it does not contain the individual's name, address or other PII), as well as other potentially sensitive information such as data about the subscriber's network and cell tower locations. To protect the confidentiality of these data, we pseudonymized the CDR prior to analysis by hash-encoding each phone number into a unique ID. The data are stored on secure university servers to which access is limited based on a data management plan approved by UC Berkeley's Committee for the Protection of Human Subjects.

We obtained informed consent from all research subjects in the phone survey prior to matching CDR records to survey responses. However, there are still open concerns around the use of CDR by bad actors, particularly as even pseudonymized datasets can frequently be de-anonymized for a subset of observations [56, 48]. Active research on

applying the guarantees of differential privacy to CDR datasets and associated machine-learning models holds promise for balancing the utility of CDR data with privacy concerns [8, 118]. For additional discussion of these considerations, see subsection 2.4.1.

2.2.5 Predicting poverty from phone data

We follow the machine learning methods described in prior work [36, 38, 6] to train models that predict poverty from CDR features. Specifically, we train a gradient boosting regressor with Microsoft's LightGBM for the two matched survey-CDR datasets separately. We tune hyperparameters for the model over threefold cross validation, with parameters chosen from the following grid:

- Winsorization of features: No winsorization, 1% limit
- Minimum data in leaf: 10, 20, 50
- Number of leaves: 5, 10, 20
- Number of estimators: 20, 50, 100
- Learning rate: 0.05, 0.075, 0.1

We train and evaluate the model over five-fold cross validation, with hyperparameters tuned independently on each fold, to obtain out-of-sample estimates of accuracy and out-of-sample predictions of poverty for each observation in our matched survey datasets. We then re-train the model on all survey data (for each of the two datasets separately), record feature importances (the total number of times a feature is split on over the entire forest), and use the final model to generate wealth predictions for every subscriber on the mobile phone network during the relevant time period.

We experiment with training models in this way for each of the relevant poverty outcomes: consumption expenditure, PMT, and asset index for the 2018–2019 field survey dataset and PMT and asset index for the 2020 phone survey dataset. Evaluations of model accuracy are found in Table S12. The correlation between the phone-based poverty predictions and a traditional PMT is 0.41, as trained and evaluated on the 2020 phone survey dataset (Table S12 Panel C). When trained and evaluated using the national 2018–2019 household survey with consumption data, the correlation between the phone-based poverty predictions and consumption is 0.46 (Table S12 Panel A).

Feature importances

Feature importances for each model are presented in Table S4. We note that in examining the feature importances, location-related features (number and percent of calls placed in each prefecture of the country) are very important. The correlation between phone-based poverty predictions using only these location features and a standard PMT

is 0.35 when trained and evaluated with the 2020 phone survey (versus 0.41 using all features). When trained and evaluated with the 2018–2019 field survey, the correlation between location-only phone-based poverty predictions and consumption is 0.42 (versus 0.46 when using all features). Given the relative importance of location features, we provide more in-depth analysis of the role of geography in phone-based targeting approaches in Table S15. Other important features in the full phone-based poverty scores relate to nighttime calling behaviour, mobile data usage and mobile money usage.

Aggregate validation of phone-based poverty estimates

Our machine learning models use cross-validation to help limit the potential that the predictions are overfit to the specific surveys on which they are trained (and on which they are later evaluated in the targeting simulations). To provide a more independent test of the validity of the CDR-based estimates, we compare regional aggregates of wealth based on the CDR model to regional estimates of wealth based on household survey data. In this exercise, we predict the consumption of roughly 5 million subscribers in Togo using the machine-learning model trained to predict consumption using the 2018–2019 national household survey, then calculate the average consumption of each prefecture and canton (where each subscribers' home location is inferred from CDR [155]).

Results, shown in Figure S8, indicate that the CDR-based estimates of regional poverty correlate with survey-based estimates of regional poverty. At the prefecture level, the Pearson and Spearman correlations of CDR-based consumption with survey-based consumption are 0.92 and 0.83, respectively; the correlations with the proportion of each prefecture living in poverty are -0.76 and -0.74. At the canton level, comparing the CDR-based estimates to the satellite-inferred canton poverty map from Figure S5, we find Pearson correlation = 0.84 and Spearman correlation = 0.68; compared to the average canton consumption in the 2018–19 field survey, Pearson correlation = 0.57 and Spearman correlation = 0.59. These correlations are toward the lower end of the range of correlations observed in prior efforts to estimate regional poverty with CDR [36, 38, 6].

Parsimonious phone expenditure method

In addition to the machine learning method for wealth prediction described above, we are interested in the performance of an intuitive, parsimonious method for approximating poverty with CDR. We focus on a measure of "phone expenditure" on the basis of costs of all calls placed and SMS messages sent by each subscriber. We apply standard rates for calls and SMS messages in Togo: 30 CFA (US$0.06) to send an SMS message and 50 CFA (US$0.09) per minute of call time.[6] We use these prices to infer the (approximate) amount spent by each subscriber from their outgoing mobile phone transaction logs. We find that

[6]These prices represent a typical Togolese phone plan, though there is considerable diversity in special promotions and friends-and-family plans available from Togo's two mobile phone operators, Moov and Togocom.

the phone expenditures method is substantially less accurate than the machine-learning-based method, with a correlation of 0.13 with both the 2020 phone survey PMT and the 2018–2019 household survey's consumption measure (Table S12).

2.2.6 Targeting evaluations

We simulate phone-based and counterfactual targeting methods for reaching the poorest individuals in Togo, using the two survey datasets described in subsection 2.2.1. Specifically, for each dataset, we simulate providing benefits to the poorest 29% of observations in the dataset based on a suite of counterfactual targeting options (with sample weights applied), and compare the population targeted to the population that is "truly poor," where ground truth poverty is determined using two different measurements. With the 2018–2019 in-person survey dataset, our main ground-truth wealth measure is based on consumption expenditure: we evaluate how well proxy measures of poverty reach those with the lowest consumption. For the 2020 phone survey dataset, our main ground-truth wealth measure is based on the PMT described in the Section 2.2.1 (this is necessary because consumption information was not collected in the phone survey).

Our main targeting evaluations simulate targeting 29% of individuals because the Novissi program had sufficient funds to target 29% of registrants in eligible cantons. The 29th percentile corresponds to a consumption threshold of US$1.17 per day in the 2018–2019 field survey dataset, and a PMT threshold of US$1.18 per day in the 2020 phone survey dataset. Our analysis shows how accurately each targeting method reaches the 29% truly poorest (Figure 2.2 and Table 2.1), those below the extreme poverty line, defined as three-quarters of the poverty line, or US$1.43 per day (Table S1), and those below the international poverty line of US$1.90 per day (Table S2).

Our evaluations are designed to measure how effectively several different targeting methods, described below, are at reaching the poorest individual mobile phone owners in each of the two survey populations. We focus on individuals rather than households because the Novissi program was designed and paid as an individual benefit. While social assistance program in other countries typically consider the household to be the unit of analysis that determines program eligibility, there is no notion of a household unit in the Novissi program (in part because the government does not possess data that links individuals to households). See subsection 2.4.1 for additional discussion of the implications of individual versus household-level analysis.

Likewise, our focus on mobile phone owners reflects the fact that the Novissi system in Togo distributed payments via mobile money; as such, anyone without access to a phone could not receive benefits irrespective of the targeting method — see 2.3.5 for a discussion of exclusion errors resulting from this constraint. In practice, this constraint only affects the analysis using the 2018–2019 in-person survey, where 4,171 of 6,171 respondents provided an active phone number. For analysis using the 2020 phone survey, we include all respondents, as every respondent had access to a phone. Future work could compare

phone-based targeting to counterfactual targeting methods that could be implemented in-person, and thus account for exclusion errors resulting from phone ownership.

Targeting methods and counterfactuals

Our evaluations use the two survey datasets to measure the performance of three targeting methods that were feasible when implementing the Novissi program: geographic blanketing (targeting everyone in certain geographies), occupation-based targeting (targeting everyone in certain occupation categories), and phone-based targeting. The location of subscribers targeted by each of these methods, in both the rural Novissi program and the hypothetical national program, are shown in Figure S12. Note that in the 2020 phone survey the unit of observation is the individual, while in the 2018–2019 field survey the unit of observation is the household: in practice, this means that our simulations with the 2018–2019 field survey dataset reflect a program that would provide benefits only to heads of household, and we do not account for household size in considering exclusion errors or social welfare. Future work could model phone-based targeting on a household basis by collecting phone numbers for all household members and calculating aggregate benefits assigned to each household; given survey data limitations we cannot perform this analysis.

With geographic targeting, the primary counterfactual approach considered by the government of Togo in implementing its rural assistance program, we assume that the program would target geographic units in order from poorest to wealthiest, and that all individuals in targeted units would be eligible for benefits. We report results from two different approaches to geographic targeting: (1) a program that targets the poorest prefectures (admin-2 region), defined as those prefectures with the lowest average predicted consumption based on a 2017 INSEED survey PMT; and (2) a program that targets the poorest cantons (admin-3 region), defined as those cantons with the lowest average wealth based on high-resolution micro-estimates of wealth inferred from satellite imagery. When targeting the N poorest geographic regions would result in more than 29% of individual receiving benefits, then N - 1 regions are targeted fully, and individuals from the nth poorest region are selected randomly until the 29% threshold is reached. See S5 and 2.2.2 for the poverty maps used for geographic targeting.[7]

In occupation-based targeting, we first evaluate the effectiveness of targeting informal workers, which is the eligibility criteria used by Novissi when it was first launched in April 2020, and which served as the basis for paying roughly 500,000 urban residents. In practice, this process involves categorizing the occupation of every individual respondent in both surveys as either formal or informal (including unemployed), applying the same definition of informality that was used by the Novissi program. In the simulations,

[7]While this purely geographic approach was considered carefully by the Government of Togo, it is less common in non-emergency settings, when other data can inform targeting decisions. For instance, it is common to combine some degree of geographic targeting with community-based targeting and/or proxy means tests.

informal workers are targeted first (in random order if there are more informal workers than can receive benefits) and formal workers are targeted last (also in random order, if the available benefits exceed the number of informal workers).

We also develop and test a hypothetical occupation-based approach, which we refer to as "optimal occupation-based targeting," which assumes that the policymaker had high-quality consumption data on the consumption of workers in each occupation and used that information to target the poorest occupations first. Although this approach was not considered in Togo's pandemic response, it was feasible with the data sources available in Togo at the time, and represents an upper-bound on the performance of a hypothetical occupation-based targeting system. We simulate this optimal occupation-based approach by calculating the average consumption of each occupation in the 2018–2019 field survey; occupations are then targeted in order of increasing average consumption. The average consumption of each occupation category is shown in Table S10. Note that because agricultural workers are the poorest category and make up 29% of the observations in the 2018–2019 field survey dataset and 41% of the observations in the 2020 phone survey dataset, in practice the precision and recall metrics reported in our targeting simulations reflect systems of occupation-based targeting that would prioritize agricultural workers only.

Of primary interest in the targeting evaluation is the performance of the targeting approaches based on mobile phone data. The phone-based (machine-learning) approach is the one described in the main text, which uses machine learning to construct a poverty score from rich data on mobile phone use and prioritizes the individuals with the lowest poverty scores (subsection 2.2.5). For reference, we also calculate the performance of a more parsimonious "phone (expenditures)" model, which prioritizes the individuals with the smallest total phone expenditures.

For completeness, our simulations also include results from targeting methods that were not feasible for the Novissi programme, as the data required to implement those methods were not available when Novissi was launched (though Togo plans to create a foundational unique ID system and comprehensive social registry in 2024).[8] In particular, we simulate targeting using an asset-based wealth index, constructed as described in subsection 2.2.1. For the hypothetical national simulations using the 2018–2019 field survey dataset, we also simulate targeting using a PPI and PMT. Finally, when simulating targeting the hypothetical national programme restricted to rural areas (Table S3), we also simulate targeting on a rural-specific PMT. We cannot simulate PPI or PMT-based targeting using the 2020 phone survey since the necessary data were not collected.

An important caveat is that the PMT that we use in the 2018–2019 survey is "perfectly calibrated" in the sense that it is both trained and evaluated on the same sample. In real-world settings, the predictive accuracy of a PMT declines as the time increases between the time of calibration and the time of application [41, 90]. As such, the performance of

[8] https://www.togofirst.com/en/public-management/0409-6177-togolese-deputies-approve-biometric-id-project

the PMT we report is likely an upper bound of the performance of a real-world PMT.

For the PMT in the 2018–2019 field survey dataset, as well as for CDR-based wealth estimates in both datasets, predictions are produced out-of-sample over cross validation so that they can be fairly evaluated in targeting simulations. Specifically, in each case, the training dataset is divided into ten cross validation folds; the machine learning model is trained on nine of the ten folds and used to produce predictions for the final fold. The training-and-prediction regime is repeated for all ten folds.

Measures of targeting quality

For each targeting method, we calculate two "threshold-agnostic" metrics of targeting accuracy — metrics that capture relationships between continuous measures of poverty rather than focusing on accuracy for targeting a specific portion of the population. These are:

Spearman correlation coefficient. Spearman's rank correlation coefficient is the Pearson correlation between the rank values of the true and proxy measures of poverty. We focus on the Spearman correlation rather than standard Pearson correlation as a measure of targeting quality because targeting concerns itself only with the ordering of observations according to poverty. Spearman's correlation coefficient is calculated as follows:

$$\rho = 1 - \frac{6\Sigma_{i=1}^{N}(r_i - \hat{r}_i)^2}{N(N^2-1)} \tag{2.1}$$

where N is the total number of observations, r_i is the rank of observation i according to the ground truth poverty measure, and $\hat{r}_i$ is the rank of observation i according to the proxy poverty measure.

ROC curves and area under the curve. Following Hanna and Olken (2018) [85], we trace receiver operator characteristic (ROC) curves that describe the quality of a targeting method at counterfactual targeting thresholds (Figure S9, left figures). At each counterfactual targeting threshold T we simulate targeting T% of observations according to the proxy poverty measure in question and calculate the true positive rate (TPR) and false positive rate (FPR) of the classifier with respect to reaching the T% poorest according to the ground-truth poverty measure. By varying T from 0% to 100%, we construct the ROC curves shown in Figure S9. The area under the curve (AUC) is used to summarize the targeting quality, with a random targeting method achieving an AUC of 0.5 and perfect targeting an AUC of 1. For convenience, we also include "coverage vs recall" figures (right figures of Figure S9) that show how program recall varies as the eligible percentage of the population increases. Note that since recall is another name for the true positive rate, Figure S9 panels B and D represent a rescaling of the ROC curves in Figure S9 panels A and C.

Targeting accuracy. Our analysis focuses on analysing the performance of a quota-based approach that ranks individuals from predicted poorest to predicted wealthiest, then targets the poorest 29% of individuals. We use the quota of 29% since the rural Novissi programme had sufficient funding to provide benefits to the poorest 29% of registrants in eligible cantons.[9] The 29th percentile corresponds to a consumption threshold of US$1.17 per day in the 2018–2019 field survey dataset, and a PMT threshold of US$1.18 per day in the 2020 phone survey dataset. We calculate the following metrics to describe how accurately targeting the poorest 29% according to each targeting method reaches (1) the 29% truly poorest, (2) those below the international poverty line of US$1.90 per day (57% of observations in the 2018–2019 field survey, and 76% of observations in the 2020 phone survey), and (3) those below the extreme poverty line, which was defined as three-quarters of the poverty line, or US$1.43 per day (41% of observations in the 2018–2019 field survey, and 53% of observations in the 2020 phone survey):

- Accuracy: Classification accuracy measures the proportion of observations that are identified correctly (targeted observations that are poor according to the ground-truth poverty measure, and non-targeted observations that are not poor according to the ground-truth wealth measure). Accuracy = $\frac{TP+TN}{TP+FP+TN+FN}$.
- Recall: Recall measures the proportion of all poor observations that are reached by a given targeting method. Recall = $\frac{TP}{TP+FN}$. Recall is closely related to the concept of exclusion errors (that is, the fraction of true poor who do not receive benefits, $\frac{FN}{TP+FN}$) since Recall = 1 - Exclusion error.
- Precision: Precision measures the proportion of targeted observations that are poor according to the ground-truth poverty measure. Precision = $\frac{TP}{TP+FP}$. Precision is closely related to the concept of inclusion errors (that is, the fraction beneficiaries who are non-poor, $\frac{FP}{TP+FP}$), since Precision = 1 - Inclusion error.
- Exclusion error: The proportion of true poor excluded from benefits. Defined as $\frac{FN}{TP+FN}$.
- Inclusion error: The proportion of beneficiaries who are not poor, that is, $\frac{FP}{TP+FP}$.

Note that the poverty lines are applied to consumption expenditure in the 2018–2019 field survey dataset, and to the PMT estimates in the 2020 phone survey dataset.

[9] This quota-based approach is not the only way that poverty scores could be used in targeting, though it is the only approach that we evaluate: for instance, a threshold-based approach might target everyone below a threshold poverty score; alternative approaches might provide cash transfers of different sizes depending on the poverty score of the beneficiary [41].

Social welfare

Using the two matched survey-phone datasets, we calculate aggregate utility under each of the targeting methods using a social welfare function. Following Hanna and Olken (2018) [85] we rely on CRRA utility, which models individual utility as a function of pre-transfer consumption and transfer size:

$$U = \frac{\Sigma_{i=0}^{N}(y_i + b_i)^{1-\rho}}{1-\rho} \tag{2.2}$$

Where N is the population size, y_i is the consumption of individual i, and b_i are the benefits assigned to the individual. Following Hanna and Olken (2018) [85] we use a coefficient of relative risk-aversion $\rho = 3$. To reflect the policy design of the Novissi program, we assume that all beneficiaries who receive a benefit receive the same value $b_i = b$.[10] To construct the social welfare curves, we:

1. Calculate a total budget available for each of the two datasets. We focus on programs that have a budget size analogous to that of rural Novissi, which aimed to distributed approximately US$4 million among the 154,238 program registrants, or US$25.93 per registrant. We therefore assign each dataset a total budget of US 25.93N$, where N is the total size of the dataset.

2. Simulate targeting T% of observations on the basis of each of our counterfactual targeting approaches.

3. Assign equal benefits to each of the targeted observations, with the budget divided evenly among targeted observations (so lower targeting thresholds T correspond to more benefits for targeted individuals).

4. Calculate aggregate utility by summing over benefits and consumption for each individual with the CRRA utility function. Note that non-targeted individuals are included in the welfare calculation; they are merely assigned 0 benefits. For the 2018–2019 field survey dataset we use consumption expenditure for y_i; for the 2020 phone survey dataset we use the PMT estimates.

5. By varying T between 0% and 100% of observations targeted, we trace out the social welfare curves shown in 2.3.

Fairness

We are interested in auditing our targeting methods for fairness across sensitive subgroups. Note that that notions of parity and fairness are debated in machine learning

[10] In principle, the benefit b_i paid to i could depend on characteristics of i, such as i's level of poverty. Although such an approach would substantially increase total welfare, in practice it is much more difficult to implement.

and policy communities: Barocas et al. (2017) describe how the three most popular parity criteria—demographic parity (benefits assigned to subgroups proportionally to their size), threshold parity (use of the same classification threshold for all subgroups), and error rate parity (equal classification error across subgroups)—are in tension with one another [27]. Moreover, Noreiga-Camparo et al. (2020) describe how tensions over parity criteria, prioritized subgroups, and positive discrimination lead to complicated prioritization compromises in the administration of targeted social protection programs [126].

Here we focus on two targeting-specific parity criteria:

Demographic parity. A targeting method satisfying demographic parity will assign benefits to a subgroup proportionally to the subgroup's presence in the population of interest. We evaluate demographic parity among the poor: that is, we compare the proportion of each subgroup living in poverty (below the 29th percentile in terms of consumption) to the proportion of each subgroup that is targeted (below the 29th percentile in terms of the proxy poverty measure used for targeting).

$DP = \frac{TP+FP}{N} - \frac{TP+FN}{N}$

Normalized rank residual. We are interested in whether certain subgroups are consistently ranked higher or consistently ranked lower than they "should" be by the counterfactual targeting approaches. We therefore compare the distributions of rank residuals across subgroups and targeting methods:

$RR_i = \frac{\hat{r}_i - r_i}{N}$

Here $\hat{r_i}$ is the poverty rank of individual i according to the proxy poverty measure and r_i is the poverty rank of individual i according to the ground-truth poverty measure.

We focus on seven dimensions for parity: gender, ethnicity, religion, age group, disability status, number of children, and marital status. We also evaluate parity across whether an individual is "vulnerable," where vulnerability is defined as one of the following traits: {female, over age 60, has a disability, has more than five children, is single}. We conduct this analysis using demographic information about the head of the household in the 2018–2019 field survey dataset, as these demographic variables were not all collected in the 2020 phone survey.

2.3 Results

Our main analysis evaluates the performance of Novissi's targeting approach that combines machine learning and mobile phone data — which we refer to more succinctly as the phone-based approach — by comparing targeting errors using this approach to targeting errors under three counterfactual approaches: a geographic targeting approach that the government piloted in summer 2020 (in which all individuals are eligible within the poorest prefectures (Togo's admin-2 level), or poorest cantons (Togo's admin-3 level);

occupation-based targeting (including Novissi's original approach to targeting informal workers, as well as an "optimal" approach to targeting the poorest occupation categories in the country); and a parsimonious method based on phone data without machine learning (that uses total expenditures on calling and texting as a proxy for wealth).

We present results that compare the effectiveness of these different targeting mechanisms in two different scenarios. First, we evaluate the actual policy scenario faced by the government of Togo in September of 2020, which involved distributing cash to 60,000 beneficiaries in Togo's 100 poorest cantons. This first scenario is evaluated using data collected in a large phone survey we designed for this purpose and conducted in September 2020. The "ground truth" measure of poverty in this first scenario is a PMT, as consumption data could not be feasibly collected in the phone survey. The PMT is based on a stepwise regression procedure, described in section A.1, which captures roughly 48% of the variation in consumption. Thus, for the first scenario focused on the rural Novissi programme, all targeting methods are evaluated with respect to this PMT. The phone-based machine-learning model is similarly trained using the PMT as ground truth. Second, we simulate and evaluate a more general and hypothetical policy scenario in which the government is interested in targeting the poorest individuals nationwide; this scenario is evaluated using national household survey data collected in person by the government in 2018 and 2019. The second simulation uses consumption as the ground truth measure of poverty. These data are described in subsection 2.2.1 and details on the evaluation are in subsection 2.2.6.

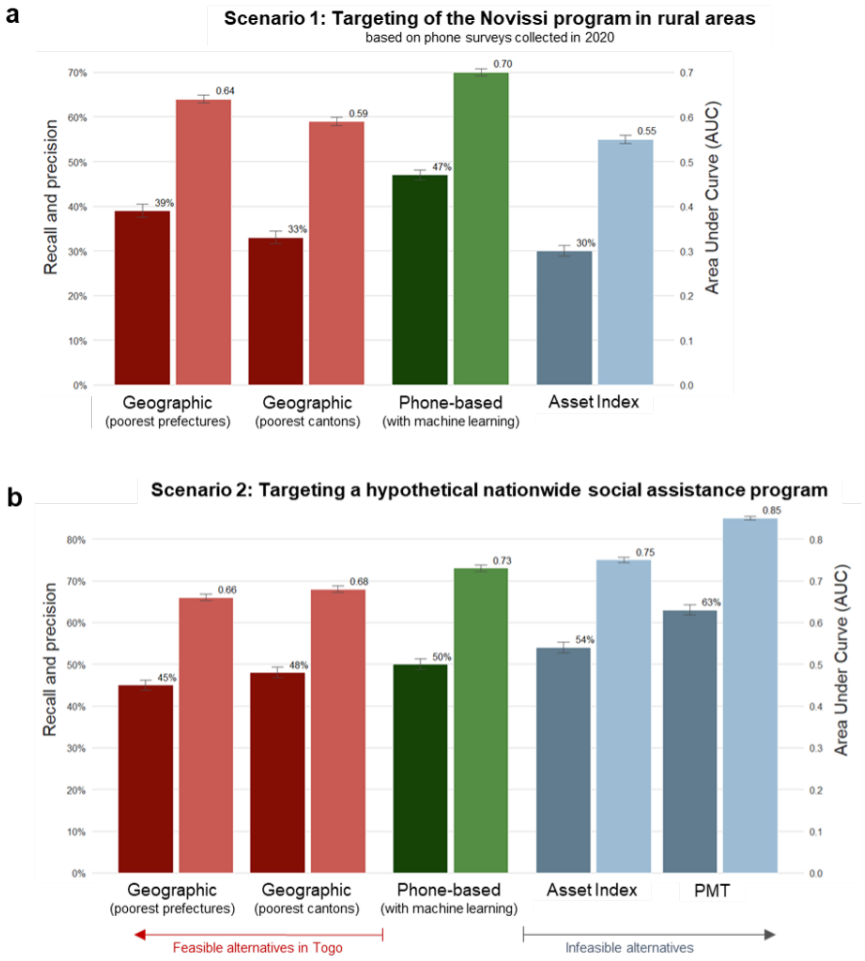

Figure 2.2: The performance of phone-based targeting (green) compared with alternative approaches that were feasible (red) and unfeasible (blue) in Togo in 2020. Targeting is evaluated for the actual rural Novissi programme (a), which focused on Togo's 100 poorest cantons (using a 2020 survey representative of mobile subscribers in the 100 cantons, where PMT is a ground truth for poverty since consumption data was not collected in the phone survey); and a hypothetical nationwide anti-poverty programme (using a national field survey conducted in 2018–2019, where consumption is a ground truth for poverty) (b). The darker bar in each pair indicates recall and precision (left axis), which is equivalent to 1 – exclusion error; the lighter bar in each pair indicates area under the curve (right axis). The bar height represents the point estimate from the full simulation; whiskers show s.d. produced from N = 1,000 bootstrap simulations. The figure highlights a subset of the results contained in Table 1.

Table 2.1: Performance of targeting mechanisms

	Targeting Novissi in rural Togo Based on 2020 Phone Survey (N = 8,915)			**Hypothetical nationwide program** Based on 2018-2019 Field Survey (N = 4,171)			
	Spearman	**AUC**	**Accuracy**	**Precision & Recall**	**Spearman**	**AUC**	**Accuracy**	**Precision & Recall**
Panel A: Targeting methods considered by the Government of Togo in 2020								
Prefecture	0.30	0.64	65%	39%	0.34	0.66	68%	45%
(Admin-2 regions)	(0.017)	(0.008)	(0.87%)	(1.51%)	(0.017)	(0.008)	(0.74%)	(1.27%)
Canton	0.19	0.59	61%	33%	0.39	0.68	70%	48%
(Admin-3 regions)	(0.019)	(0.009)	(0.78%)	(1.35%)	(0.016)	(0.008)	(0.71%)	(1.23%)
Phone	0.13	0.57	60%	32%	0.26	0.63	65%	40%
(Expenditures)	(0.020)	(0.010)	(0.71%)	(1.23%)	(0.017)	(0.009)	(0.81%)	(1.40%)
Phone	0.38	0.70	69%	47%	0.45	0.73	71%	50%
(Machine Learning)	(0.017)	(0.009)	(0.87%)	(1.18%)	(0.015)	(0.007)	(0.74%)	(1.28%)
Panel B: Common alternative targeting methods that could not be implemented in Togo in 2020								
Asset Index	0.10	0.55	60%	30%	0.51	0.75	74%	54%
	(0.018)	(0.009)	(0.48%)	(0.83%)	(0.014)	(0.007)	(0.69%)	(1.19%)
PPI	[data not available]				0.63	0.81	77%	60%
					(0.011)	(0.006)	(0.73%)	(1.25%)
PMT	[data not available]				0.72	0.85	78%	63%
					(0.009)	(0.005)	(0.70%)	(1.20%)
Panel C: Additional counterfactual targeting methods that were feasible in Togo in 2020								
Random	0.00	0.50	59%	30%	0.00	0.50	59%	29%
	(0.021)	(0.082)	(0.74%)	(0.26%)	(0.019)	(0.010)	(0.79%)	(1.36%)
Occupation	-0.11	0.45	55%	22%	-0.09	0.46	56%	24%
(As implemented)	(0.019)	(.007)	(0.62%)	(1.07%)	(0.019)	(0.095)	(0.53%)	(0.91%)
Occupation	0.25	0.61	66%	41%	0.41	0.69	72%	52%
(Optimally designed)	(0.016)	(0.008)	(0.58%)	(1.00%)	(0.016)	(0.008)	(0.72%)	(1.25%)

Notes: Targeting performance using mobile phone data and machine learning in comparison to counterfactual targeting strategies. The "true poor" are those who, according to survey data, are in the poorest 29% of the population. The first four columns evaluate targeting with a 2020 phone survey representative of subscribers in Togo's 100 poorest cantons, using a PMT as ground truth for poverty since consumption data were not collected. The last four columns evaluate targeting using nationally representative household survey data collected in 2018-2019, using consumption as a ground truth. Panel A compares the phone-based PMT (highlighted) to alternative targeting methods that the Government of Togo considered prior to expanding Novissi to rural areas. Panel B shows the performance of targeting methods that are commonly implemented but were infeasible in Togo at the time. Panel C indicates the performance of other targeting methods the government could have used. Accuracy, precision, and recall are evaluated by the extent to which they reach the poorest 29% (by construction, precision and recall are equal in this simulation, and are equal to 1 – exclusion error). Standard deviations, produced from 1,000 bootstrap simulations, shown in parentheses.

In the first scenario focused on reaching the poorest people in the 100 poorest cantons, we find that the phone-based approach to targeting substantially reduces errors of exclusion (true poor who are mistakenly deemed ineligible) and errors of inclusion (non-poor who are mistakenly deemed eligible) relative to the other feasible approaches to targeting available to the government of Togo (Figure 2.2 Panel A and Table 2.1, columns 3 to 6). We focus on the ability of each targeting method to reach the poorest 29% in each of the two survey datasets, as the rural expansion of Novissi only had sufficient funding to provide benefits to 29% of individuals in eligible geographies (Table S1 and Table S2 evaluate performance using alternative poverty thresholds). Using a PMT as a measure of "true" poverty status, phone-based targeting (area under the curve (AUC) = 0.70) outperforms the other feasible methods of targeting rural Novissi aid (for example, AUC = 0.59–0.64 for geographic blanket targeting). As a result, errors of exclusion (defined as 1 – Recall) are lower for the phone-based approach (53%) than for feasible alternatives (59%–78%).

Similarly, phone-based targeting outperforms most feasible methods when we simulate the targeting of a hypothetical national anti-poverty program (Figure 2.2 Panel B and Table 2.1, columns 7 to 10). Here, the phone-based approach is more effective at prioritizing the poor (AUC = 0.73) than geography-based alternatives (AUC = 0.66–0.68), and similarly leads to lower exclusion errors (50%) than most feasible alternatives (52%–76%). One exception in this hypothetical program is occupation-based targeting: whereas the Novissi program's original criteria of targeting informal workers would not scale well to a national program (76% exclusion errors), an alternative "optimal" occupation-based approach that we develop (subsection 2.2.6) — which assigns all transfers to the poorest occupational category (agricultural workers) — slightly outperforms phone-based targeting (48% exclusion errors).

Together, the results in Table 2.1 indicate that the phone-based targeting approach was more effective in the actual rural Novissi program than it would be in a hypothetical nationwide program. Our analysis suggests that the benefits of phone-based targeting are greatest when the population under consideration is more homogeneous, and when there is less variation in other factors (such as place of residence) that are used in more traditional approaches to targeting. For instance, when we restrict the simulation of the hypothetical national program to households in rural areas, the gains from phone-based targeting increase (Table S3).

We also find that the performance benefits of phone-based targeting increase as programs seek to target the most extreme poor. This increase can be seen by comparing Table 2.1, where targeting performance is measured by how many of the poorest 29% receive benefits, to Table S1, which measures whether households below the extreme poverty line (US$1.43 per capita daily consumption) receive benefits, and Table S2, which measures whether households below the poverty line (US$1.90 per capita daily consumption) receive benefits. Although all targeting methods perform better at targeting the extreme poor, the differential between the phone-based approach and other

methods is greater when the consumption threshold is lower.[11]

The phone-based approach that we develop relies heavily on machine learning to construct a poverty score for each mobile subscriber, where eligibility is a complex function of how the subscriber uses their phone. We also consider an alternative approach that does not use machine learning, but instead simply targets mobile phone subscribers with the lowest mobile phone expenditures over the preceding months. We find that this "phone expenditure" method (AUC = 0.57 for rural Novissi and 0.63 in for the hypothetical national program; Table 2.1) performs substantially worse than the machine-learning-based model (AUC = 0.70 for rural Novissi and 0.73 for the hypothetical national program). Although the phone expenditure model requires much less data and may be easier to implement, this parsimony increases targeting errors, and may also introduce scope for strategic "gaming" if used repeatedly over time.

An important factor in the success of the machine-learning model is the fact that it was trained on representative survey data collected immediately before the program's expansion. Since an individual's poverty status can change over time, and since the best phone-based predictors of wealth may also change, a model trained in one year or season may not perform well if applied in a different year or season. In Togo, we find that when the machine-learning model or the mobile phone data are roughly 18 months out of date, predictive accuracy decreases by 4–6% and precision drops by 10–14% (Table S5). These losses are nearly as large as the gains that phone-based targeting provides over geographic targeting—a finding that underscores the importance of training the model with current and representative data.

We also compare the phone-based approach to alternative targeting approaches that require a recent and comprehensive social registry. Although the Government of Togo did not have such a registry, this comparison helps situate this method relative to other methods commonly used by development researchers and policymakers. These results, shown in Table 2.1, can only be simulated using the national in-person survey, since the phone survey did not collect consumption data. The results are more ambiguous: the phone-based approach (AUC = 0.70–0.73) is approximately as accurate as targeting using an asset-based wealth index (AUC = 0.55–0.75), but less accurate than using a poverty probability index (AUC = 0.81) or a perfectly calibrated PMT (AUC = 0.85) (see subsection 2.2.1 for the differences between these indices). We note, however, that the performance of the "perfectly calibrated" PMT may substantially overestimate the performance of a real-world PMT, which is likely to decline steadily over time since calibration [41, 90].

[11] In this analysis, the wealth distribution of the underlying population is important: as more than half of the Togolese population is below the poverty line, the targeting methods are attempting to differentiate between different gradations of poverty. Just as precision increases as the target population grows — that is, from Table 2.1 to Table S1 to Table S2 — results may differ in contexts where the target population is much smaller.

2.3.1 Social welfare and fairness

Improvements in targeting performance translate to an increase in social welfare. Using the constant relative risk aversion (CRRA) utility function, we calculate aggregate welfare under the phone-based approach and each of the counterfactual targeting approaches. Under the CRRA assumptions, individual utility is a concave function of consumption. By assuming a fixed budget—which we fix at a size analogous to that of the Novissi rural aid programme, which had a budget of US$4 million to distribute among 154,238 programme registrants—and equal transfer sizes to all beneficiaries, we simulate the distribution of benefits among eligible individuals at counterfactual targeting thresholds to construct social welfare curves for each targeting method. This social welfare analysis also allows us to identify the optimal beneficiary share and corresponding transfer size. Figure 2.3 shows the utility curves for each of the targeting methods simulated, separately for the two populations. Note that phone-based targeting, geographic blanketing and an asset-based wealth index all achieve approximately the same maximum utility in the hypothetical national programme, but phone-based targeting dominates in the rural Novissi programme. Also note that all targeting methods outperform a universal basic income scheme if the beneficiary share and transfer size is well-calibrated.

These utilitarian welfare gains suggest that society as a whole will benefit from improved targeting, but do not imply that all subgroups of the population will benefit equally. Indeed, there is growing concern that algorithmic decision making can unfairly discriminate against vulnerable groups [67, 27, 126]. To address these concerns in the context of the Novissi programme, we audit the fairness of each targeting method across a set of potentially sensitive characteristics, while noting that notions of fairness and parity are contested and often in tension [106]. Figure 2.4 Panel A shows, as an example, that the phone-based approach does not cause women to be systematically more likely to be incorrectly excluded by the targeting mechanism from receiving benefits than men (see also subsection 2.3.1). Similarly, the phone-based approach does not create significant exclusion errors for specific ethnic groups (Figure 2.4 Panel B), religions, age groups or types of household, though there are small differences in targeting accuracy between groups (Figure S10). We also compare the fairness of the phone-based approach to several other targeting approaches by evaluating each method's demographic parity—that is, the extent to which each method under- or over-targets specific demographic subgroups relative to that group's true poverty rate (Figure 2.4 Panels C and D, Figure S11). Overall, we find that none of the targeting methods analysed naively achieves perfect parity across subgroups; a phenomenon referred to as "no fairness through unawareness" [62]. The largest parity differences occur with geographic targeting methods.

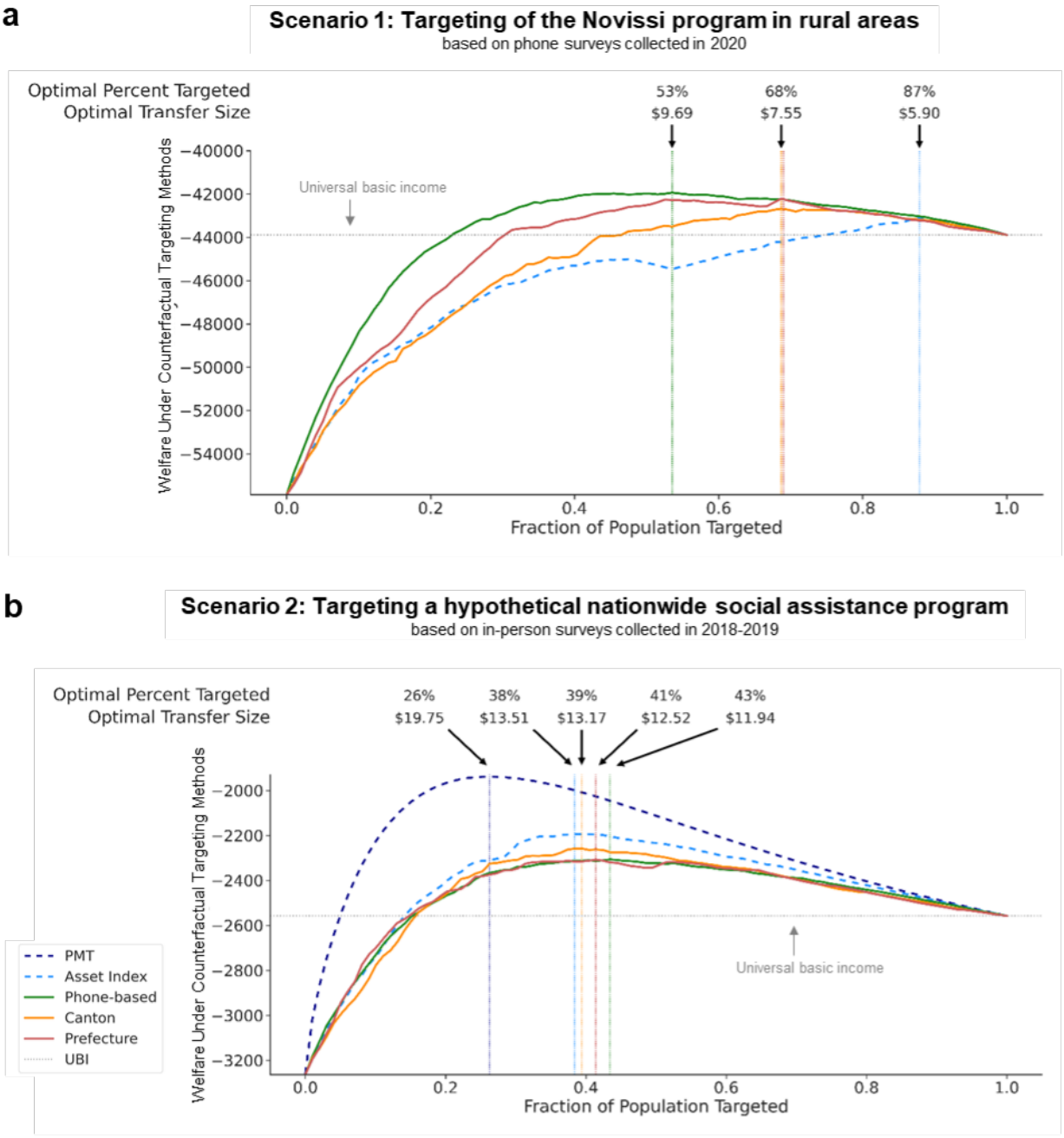

Figure 2.3: Aggregate social welfare is calculated (assuming CRRA utility) under counterfactual targeting approaches. We assume a fixed budget of US$4 million and a population of 154,238, with an equal transfer size for all beneficiaries. Utility curves for feasible targeting mechanisms are shown in solid lines; infeasible targeting mechanisms are shown in dashed lines. The horizontal dotted line indicates total social welfare for a universal basic income program that provides (very small) transfers to the entire population; vertical dotted lines indicate the targeting threshold and associated transfer size that maximizes social welfare for each targeting mechanism. Targeting is evaluated for the Novissi anti-poverty program in Togo's 100 poorest cantons (Panel A) and a hypothetical nationwide anti-poverty program (Panel B).

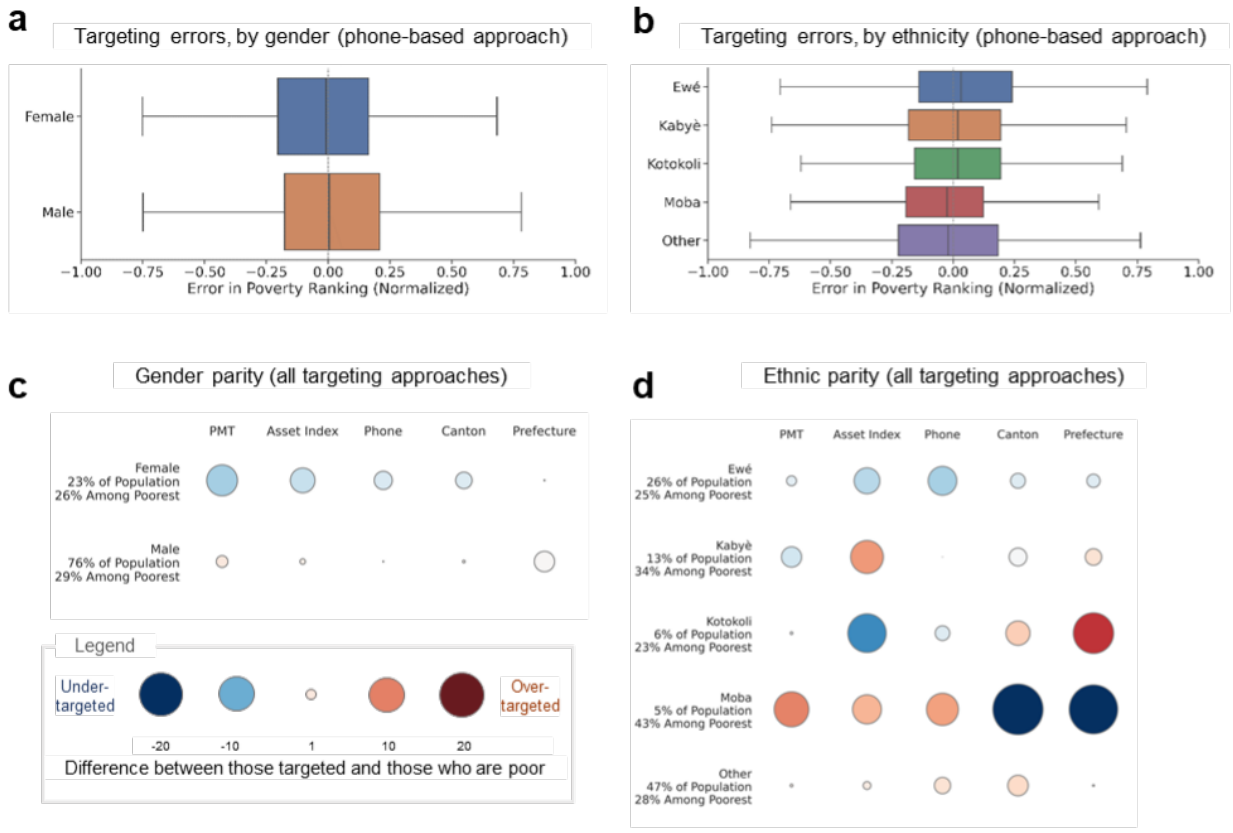

Figure 2.4: Distributions of differences between ranking according to predicted wealth from the phone-based approach and ranking according to true wealth (using the 2018–2019 field survey; N = 4,171), dissagregated by gender (a) and ethnicity (b). Boxes show the 25th and 75th percentiles, whiskers show the minimum and maximum, and the centre line shows the median of the distribution. Left-skewed bars indicate groups that are consistently under-ranked; right-skewed bars indicate groups that are consistently over-ranked. c, d, Evaluation of demographic parity across subgroups by comparing the proportion of a subgroup targeted under counterfactual approaches to the proportion of the subgroup that falls into the poorest 29% of the population (using the 2018–2019 field survey; N = 4,171), disaggregated by gender (c) and ethnicity (d). Bubbles show the percentage point difference between the proportion of the subgroup that is targeted and the proportion that is poor according to ground-truth data. Large red bubbles indicate groups that are over-targeted; large blue bubbles indicate groups that are under-targeted.

2.3.2 Differences Between Rural and National Evaluations

The results in Table 2.1 indicate that the phone-based targeting approach — as well as the counterfactual targeting approaches — was more effective in the actual rural Novissi program (first four columns of Table 2.1) than it would have been in a hypothetical nationwide program (last four columns of Table 2.1). There are several factors that may account for these differences. Some of these factors are difficult for us to test empirically, for instance the fact that the surveys were conducted at different points in time, used different teams of enumerators, and different data collection modalities (phone vs. in person). We investigate two factors that we can explore empirically: the geographic concentration of each survey and the ground truth measure of poverty (consumption vs. PMT). We additionally explore whether targeting results are sensitive to the use of a nationwide PMT vs. a rural-specific PMT.

Geographic concentration

Whereas the rural Novissi evaluation focuses on Togo's 100 poorest cantons, the hypothetical national program is evaluated nationwide (397 cantons). We therefore present results in Table S3 that restrict the simulation of the hypothetical national program to the 2,306 households in rural areas (out of 4,171 total). Comparing the results in Table S3 to the last four columns of Table 2.1, we find that the performance of all methods drops, as would be expected when the beneficiary population is more homogeneous. Importantly, we also observe that the relative performance of phone-based targeting increases: whereas the CDR-based method performed worse than the asset index and only slightly better than canton-based targeting in the full nationwide evaluation (last four columns of Table 2.1), the CDR-based method is on par with the asset index and substantially better than canton-based targeting when the nationwide survey is limited to rural areas (Table S3).

Consumption vs. PMT

Whereas the national evaluation uses a measure of consumption as ground truth, the rural Novissi evaluation uses a PMT as ground truth. Table S13 therefore simulates the hypothetical national program using a PMT as ground truth. Comparing the results in Table S13 to the last four columns in Table 2.1, we find that using a PMT rather than consumption as ground truth increases targeting accuracy across all of the targeting methods. However, switching from consumption to the PMT does not substantially improve the performance of the phone-based method relative to the counterfactual approaches. This latter finding suggests that the use of the PMT is likely not a major source of the difference between the relative performance of the CDR-based method in the rural Novissi program (first four columns of Table 2.1) and the hypothetical nationwide program (last four columns of Table 2.1).

National PMT vs. Rural PMT

Since the best predictors of welfare differ for rural and urban populations, we explore whether targeting results change when the PMT is calibrated using a rural rather than national population. Specifically, we construct a rural-specific PMT using the same methodology described in subsection 2.2.6, but restricting the training data to observations in the 2018-2019 field survey that are in rural areas. This rural PMT explains 17% of the variation in log-transformed consumption in rural areas, and is highly correlated (Pearson correlation = 0.75) with the general PMT. We then produce rural PMT estimates for respondents to the 2020 phone survey, and retrain the phone-based poverty prediction model to predict the rural-specific PMT in that population. Table S14 then presents results from simulating with the rural PMT as ground truth. Comparing Table S14to the first four columns of Table 2.1, we observe a noticeable improvement in the performance of the asset index, but other results are largely unchanged.

Relatedly, Table S4 shows the feature importances for different phone-based prediction models. Panels A and B show the top-10 features for the main models presented in Table 2.1, i.e., for predicting a PMT in the 2020 rural phone survey, and predicting consumption in the 2018-19 nationwide household survey. Panels C and D shows the top-10 features for predicting a PMT in the 2018-19 survey, and predicting a PMT in the 2018-2019 household survey, restricted to rural areas. The feature importances for the two national-scale models are similar, suggesting the role of the ground truth poverty measure may not be as important as the role of geography in creating the poverty prediction models. The feature importances for the two rural-focused models are less similar, which may be due to the fact that the 2020 phone survey is concentrated in the 100 poorest cantons, while in Panel D we restrict to rural areas, but these rural areas still cover the entire country.

Taken together, the results in this subsection suggest that the benefits of phone-based targeting are likely to be greatest when the population under consideration is more homogeneous, and when there is less variation in other factors (such as place of residence) that are used in more traditional approaches to targeting.

2.3.3 Location-based targeting

Several results emphasize the importance of geographic information in effective targeting. In particular, we observe that basic geographic targeting performs nearly as well as phone-based targeting in specific simulations — in particular, in simulations of a nationwide program that can afford to target a large proportion of the total population (e.g., Table S2). We also found that location-related features from the CDR are important in the phone-based prediction model (Table S4).

For these reasons, Table S15 explores the extent to which targeting could be based on a CDR-location model that only uses the CDR to infer an individual's home location (see subsection A.2.1 and Warren et al. (2022) [155]). As with the phone (expenditures) model,

the CDR-location model may be attractive to implementers since the data and technical requirements are reduced [122]. In Table S15, we observe that geographic targeting using phone-inferred home location is of slightly lower quality than geographic targeting using survey-recorded home location, and substantially worse than targeting using the machine learning approach.

We also investigate the correlation between different sources of information on an individual's location. Table S16 compares three different methods for identifying an individual's location, using roughly 4,500 respondents to the 2020 phone survey. At the prefecture (admin-2) level, most people (90%) self-declare living in the same canton in which they are registered to vote; there is also strong overlap between the individual's CDR-inferred location and self-declared location (70%). The accuracy is substantially lower at the canton level, which is likely due to error in the CDR-inference algorithm when spatial units are small, as well as to confusion among respondents as to which canton they live in (e.g., most respondents were confident in naming their village, but did not always know their canton).

Table S17 presents additional analysis to compare the mobile phone activity of each subscriber with their home location, as recorded in the survey and as inferred from their CDR. We find that 62-85% of the average subscriber's activity occurs in their home prefecture, and that all of the modal subscriber's activity occurs in their home prefecture. These results are consistent with the importance of location-related features in the prediction algorithm (and the relatively low mobility of the rural Togolese population).

This analysis may also provide some context for the difference in the accuracy of the geographic targeting methods between the rural evaluation and the national evaluation in Table 2.1. While canton-based targeting performs better in the national evaluation, which is consistent with past work showing that finer-resolution geographic targeting is preferred to lower-resolution geographic targeting [19, 146], prefecture-based targeting counter-intuitively performs better in the rural evaluation. We suspect this discrepancy is caused by three main factors: First, we expect that the estimates of average canton wealth are likely to be noisier than the estimates of average prefecture wealth, since the prefecture estimates aggregate over a larger population and the canton estimates rely on satellite-based inferences. Second, in the rural evaluation the prefecture is an important component of the PMT that is used as the ground truth measure of poverty (Table S8), so prefecture targeting relies on information that is structurally incorporated into the ground truth outcome (unlike in the national evaluation, where the ground truth outcome is consumption). Third, locations in the rural phone survey were self-reported, whereas locations were recorded on GPS devices by enumerators in the national survey; as noted, many respondents expressed confusion about their home canton.

2.3.4 Temporal stability of phone-based targeting

When simulating the performance of phone-based targeting, our main analysis uses each survey dataset to both train the machine learning model and, via cross-validation,

to evaluate its performance. These measures of targeting performance thus indicate what should be expected when training data (i.e., the ground truth measures of poverty and the matched CDR) are collected immediately prior to a program's deployment. This best-case scenario is what occurred in Togo in 2020: the phone survey was completed in October 2020 and Novissi was expanded beginning in November 2020. In other settings, however, it may not be possible to conduct a survey before launching a new program; it may likewise not be possible to access up-to-date mobile phone data.

To provide an indication of how long phone-based models and predictions remain accurate, Table S5 compares (i) the best-case scenario to alternative regimes where (ii) the training data are old but the CDR are current, and (ii) the training data are old and the CDR are also old. In these simulations, the "old" data are from the 2018-19 national household survey and corresponding 2019 phone dataset; the "current" data are the subset of 2020 phone survey respondents for whom CDR are available in 2019 and 2020 (N = 7,064). In all simulations, the 2020 PMT is used as the ground truth measure of poverty. Predictions for (i) are generated over 10-fold cross validation; predictions for (ii) and (iii) are out-of-sample with respect to the training data (since the models are trained on the 2018-19 field survey).

The results in Table S5 indicate that predictive performance decreases when the model is out of date, and decreases even further when the CDR are out of date. This is to be expected, since roughly two years elapsed between the "old" and "current" periods: in addition to changes in how people use their phones (which would disrupt the accuracy of the predictive model), the actual economic status of some individuals may have changed – for instance, due to the COVID-19 pandemic. There are also other important differences between the 2018-19 national household survey and the 2020 phone survey that could affect the extent to which a model trained on the former could accurately predict outcomes in the latter (such as the mode of data collection, the geographic concentration of the sample, and so forth – see subsection 2.3.2).

For the main simulations focused on reaching the poorest 29%, Table S5 suggests that accuracy decreases by 3-4 percentage points (4-6%) and precision decreases by 5-7 percentage points (10-14%) when out of date models and CDR are used for targeting. These losses are nearly as large as the gains of phone-based targeting over geographic targeting observed in Table 2.1, which emphasizes the importance of having current and representative training data for real-world deployment of phone-based targeting. However, in absolute levels, the phone-based predictions remain reasonably accurate despite the two-year gap between the training and test environments (i.e., the Spearman correlation with ground truth is ρ=0.35-0.36).

2.3.5 Program exclusions beyond the targeting algorithm

This novel approach to targeting requires careful consideration of the ways in which individuals can be incorrectly excluded from receiving programme benefits. Our analysis highlights six main sources of exclusion errors for the expansion of Novissi (Table 2.2):

(1) beneficiaries must have a SIM card and access to a mobile phone (field survey data from 2018–2019 indicate that 65% of adults and 85% of households have a phone; see also Supplementary Figure S3); (2) they must have used their SIM card recently, in order to generate a poverty score (between 72% and 97% of programme registrants); (3) they must be a registered voter (roughly 87% of adults); (4) they must self-target and attempt to register (roughly 40% of eligible individuals attempted); (5) they must succeed in registering, which requires basic reading and digital literacy (72% succeed); and (6) they must be successfully identified as eligible by the machine-learning algorithm (47% recall; Table 2.1). Many of these sources of possible exclusion overlap; Table S6 thus estimates, on the basis of the 2020 phone survey, the extent to which each successive step in registration creates additional exclusions. Appendix A.3 provides more details on the analysis of program exclusions. These results highlight the fact that algorithmic targeting errors are an important source of program exclusion, but that real-world program also face structural and environmental constraints to inclusion.

Table 2.2: Sources of exclusion from Novissi benefits

Exclusion Source	Proportion Included	Data and Calculations
Voter ID possession	83% - 98%	According to administrative data, 3,633,898 individuals are registered to vote in Togo. The electoral commission of Togo reports that this corresponds to 86.6% of eligible adults. The total adult population in Togo is not certain (the last census was in 2011), but Togo's national statistical agency estimates that there are 3,715,318 adults in Togo; the United Nations estimates 4.4 million adults [123]. These imply a voter ID penetration rate of either 82.6% and 97.8%, respectively.
SIM card and mobile phone access	50% - 85%	65% of individuals interviewed in the 2018-2019 field survey (N = 6,171) reported owning a phone; 85% of individuals were in a household with one or more phones. Rural penetration is lower (50% of individuals and 77% of households), as is penetration among women (53% for women vs. 79% for men; in rural areas, it is 33% for women and 71% for men) – see Figure S6. Phone penetration in Togo likely increased between the field survey (2018-2019) and the Novissi expansion (October 2020).
Past mobile phone use	72% - 97%	Poverty estimates were only constructed for subscribers who placed at least one outgoing transaction between March and September 2020. In a typical month, 2.5% of all phone numbers are newly registered (Figure S9), so with a one month gap between poverty inference and program registration we would expect 95-97% of registrations to be associated with a poverty score. However, 27% of all Novissi registrations (November-December 2020) did not match to CDR, likely due to new SIM purchases or registration on infrequently used SIMs (see Methods Section 6).
Program awareness	35% - 46%	245,454 unique subscribers attempted to register for the rural Novissi program. The total voting population of eligible areas is 528,562, implying a maximum registration rate of 46.44%. However, not all 245,454 registration attempts were made by people living in eligible areas; examining administrative data on home location from successful registrations we estimate that 87% of registration attempts came from eligible areas, implying an attempted registration rate of 40.40%. An alternative way to estimate attempted registration rates involves comparing the number of registration attempts made by phones below the poverty threshold (69,753) to our estimate of the number of voters in eligible cantons below the poverty threshold based on inferred home locations from mobile phone data (174,425, see Appendix B for details), which implies an attempted registration rate of 34.79% after scaling by 87% (to account for registrations that came from outside of eligible areas).
Registration challenges	72%	Registration for the Novissi program requires entering basic information into a USSD (phone-based) platform. According to program administrative data, of the 245,454 subscribers who attempted registration, 176,517 (71.95%) eventually succeeded. The average registration required four attempts.
Targeting errors	47%	Based on the estimates from our targeting simulations using the 2020 phone survey (Table 1), the exclusion error rate of the phone-based targeting algorithm is 53%.

Notes: We use multiple sources of administrative data, survey data, and government sources to estimate the extent to which different elements of the Novissi program's design may have led to errors of exclusion (Appendix A.3). Novissi eligibility requirements included: a valid voter ID (as a unique identifier and for home location), access to a mobile phone (to fill register using the USSD platform), past mobile network transactions (to estimate poverty from mobile network behavior), program awareness (to know that the program exists and to attempt to register), ability to register via the USSD platform (which requires basic digital literacy), as well as targeting errors from the phone-based machine learning algorithm. While this table calculates sources of exclusion as though they were all independent, Table S6 uses survey data to calculate overlaps in exclusions.

2.4 Discussion

Our analysis shows how non-traditional big data and machine learning can improve the targeting of humanitarian assistance. Beyond the gains in targeting performance, a key advantage of this approach is that it can be deployed quickly and responsively. In Togo, the government's objective was to deliver benefits to the poorest people in the country, so our efforts focused on training a machine-learning model to target the poor. In other settings, such as following natural disasters, the people most impacted by adverse events may not be the poorest [144]. With high-frequency phone data available in near real-time, related techniques might be used to more dynamically prioritize the people with the greatest need. For example, it may be possible to train a machine learning algorithm to identify people whose consumption fell by the greatest amount, based on changes in patterns of phone use following a crisis. Another possibility would be to simply use location information from mobile phone data to prioritize people who are likely to live in impacted regions (Table S15).

2.4.1 Limitations

It is important to emphasize that our phone-based approach is far from perfect, and may lead to important errors of both exclusion and inclusion. There are also practical limitations to this approach, for instance regarding data access and privacy [56, 112, 121, 1, 150, 127, 35]. Several such considerations are discussed in more detail below:

Phone ownership and access

As discussed in subsection 2.3.5, many individuals in LMICs do not own mobile phones. Thus, any targeting system based on mobile phone data may exclude those without phones from receiving program benefits. In the case of the Novissi program, the government used the mobile money system to disburse the cash transfers as a way to minimize human contact during the pandemic. Thus, in Togo, the use of phone data for targeting only created additional exclusions by requiring that program registrants had made at least one transaction on their SIM card in the months prior to registration. In general, incomplete mobile phone access highlights the need to allow for alternative pathways for individuals to register and receive benefits, and to create additional mechanisms for appeals, grievance redress mechanisms, and manual enrollment.

Data privacy

Mobile phone metadata, even when pseudonymized, contains sensitive information. subsection 2.2.4 describes several steps taken to protect the confidentiality of the data used in this project. More generally, special considerations arise when using personal

data from vulnerable populations [150], and human rights doctrine emphasizes that any form of communications surveillance should be "necessary and proportionate" [74].

In implementing the approach described in this chapter, we developed an IRB protocol, as well as a data management plan, that was approved by U.C. Berkeley's Committee for the Protection of Human Subjects. We followed principles of data minimization to limit the data collected and stored, and implemented organizational safeguards to restrict access to data. As an example, only IRB-approved researchers ever received access to CDR; data from the phone companies were shared with neither the Government of Togo nor GiveDirectly. Even the poverty scores derived from the phone data were restricted to IRB-approved researchers; the only data the government received was the list of SIM cards belonging to eligible beneficiaries below the targeted poverty threshold.

Future projects using mobile phone data for targeting should ensure that principles of data minimization and data sunsetting restrict the use of sensitive data to social protection objectives and limit the potential for "function creep" [150]. Further research on applying the guarantees of differential privacy to mobile phone metadata [8, 118] or implementing federated learning systems [108] could reduce the risk of data misuse or central data breaches.

Data access and consent

The fact that our approach requires access to mobile phone data owned by private companies poses an obstacle to the immediate and widespread use of such data for targeting humanitarian aid. There now exist several general frameworks and recommendations to facilitate the use of CDR in humanitarian applications [121, 127]. Yet such frameworks are still nascent, and without careful consideration may exclude important stakeholders and perspectives [1]; they also widen the scope for private companies to influence humanitarian and development decisions [151]. There also exist many ethical frameworks that rely on informed consent from participants for the use of personal data, including digital data such as CDR [47, 95]. Future programs should consider how consent pathways can be integrated with phone-based targeting, including opt-in (calculating poverty scores only after consent is provided) and opt-out (scrubbing data if consent is not provided at the time of registration) options.

Data representativity

To train the machine learning models, ground truth measures of consumption and wealth were collected using in-person and phone surveys. Since response rates were imperfect in the phone survey, we reweighted survey observations to make the training data more representative of all mobile subscribers (subsection 2.2.1). However, there are limits to the representativity of our training data, as dynamics of phone ownership

and phone sharing vary across population subgroups (Figure S3), and reweighting is an imperfect proxy.

To test for systematic bias based on data representativity, we perform ex-post audits to limit the likelihood that the trained models systematically disadvantage specific subgroups of the population (subsection 2.3.1), and find that the phone-based targeting method is no more biased than counterfactual targeting approaches. We believe such audits are essential to future work on wealth prediction and targeting based on nontraditional data. Audits could be improved with additional context-specific research about which sub-populations are at the greatest risk for systematic exclusion (for example, in this chapter we test for bias across age groups, genders, ethnicities, and more), and on considering alternative definitions for bias and fairness [126, 106, 27].

Unit of analysis

Our analysis focuses on individuals rather than households as the unit of analysis, partly reflecting the design of the Novissi program, and partly because there are no data in Togo that associate individuals with households. This limitation is important, since many real-world programs are targeted at the household level, but CDR are more naturally linked to individual subscribers. An important area for future work will thus be to explore the extent to which CDR can facilitate household-level targeting. Such work must account for the fact that a single SIM card is often shared across multiple members of the same household (and occasionally between households), and that some individuals use multiple SIM cards. Ideally, such an analysis would leverage authoritative data that uniquely identifies and links households, individuals, SIM cards, and phones.

Method of evaluation

Our main results are based on simulations of targeting methodologies using survey data collected prior to expansion of Novissi. An alternative approach to evaluating targeting performance would rely on survey data after program implementation, which would make it possible to more directly verify who did and did not receive program benefits, address issues related to the unit of analysis described above, and better attribute exclusion errors to different aspects of program design. While public health considerations in Togo prevented us from conducting a post-program survey, we hope future implementations of phone-based targeting can use post-program surveys to provide complementary evidence to what is described in this chapter.

Poverty dynamics

The phone-based approach we describe uses machine learning algorithms to predict which individuals are "poor," based on ground-truth assessment of poverty collected in surveys prior to program implementation. In the actual rural Novissi program, the ground truth measure of poverty was based on a proxy means test; in the hypothetical

national program, ground truth is based on consumption (subsection 2.2.1). However, particularly in the context of a crisis, an individual's poverty status can change; in such settings, pre-program poverty assessments may not accurately capture the population with the greatest need for support. Our data do not permit us to test whether phone data and machine learning can be used to determine if an individual has experienced a sudden fall in income or consumption, but we believe this is a promising area for future work.

Manipulation and gaming

When mobile phone data are used to determine eligibility for social benefits, individuals have incentives to strategically alter their behavior in order to "game" the system. This dilemma is not unique to phone-based targeting; it is a key consideration in the design of any targeting mechanism [79, 7], and one that affects traditional proxy means tests and poverty scorecards [113, 46]. However, recent evidence suggests that such distortionary effects may be limited [23], and complex eligibility criteria (such as the gradient boosting procedure described in subsection 2.2.5) should limit the scope for such gaming [72]. With Novissi in Togo, the one-off nature of the program likely eliminated most scope for strategic behavior; however, if such an approach were used continuously over time, alternative "manipulation-proof" approaches to machine learning may be more appropriate [34].

General equilibrium considerations

Our analysis of targeting effectiveness assumes there are no general equilibrium effects of the program on prices, wages, or interactions with informal transfers or insurance. For example, geographic targeting of transfers might lead to localized inflows of cash transfers that are large relative to the local economy, leading to changes in local demand for goods or supply of labor and therefore prices, wages or profits of local businesses [64, 55]. Similarly, since individuals are embedded in family and broader networks of informal transfers for redistribution, patronage and insurance and different targeting choices could have different effects on these existing informal arrangements [76, 119]. Equilibrium effects such as these may have important implications for the eventual distribution of impacts from the transfers. However, to cause a reversal of the policy implication of our analysis, general equilibrium effects would need to be more nuanced than merely present – for example, it would need to be that the false negatives under one method are more likely to share resources than the false negatives on another method.

2.4.2 Conclusion and future work

Most importantly, our results do not imply that mobile-phone-based targeting should replace traditional approaches reliant on proxy means tests or community-based targeting. Rather, these methods provide a rapid and cost-effective supplement that

may be most useful in crisis settings or in contexts where traditional data sources are incomplete or out of date. We believe that future work should explore how real-time data sources, such as the phone data used by Novissi, can be best combined with more traditional field-based measurements, so that these complementary data sources can be best integrated in the design of inclusive systems for social protection.

Chapter 3

Ultra-poverty targeting with machine learning and phone data in Afghanistan

This chapter is based on the paper "Targeting development aid with machine learning and mobile phone data: Evidence from an anti-poverty intervention in Afghanistan" [6], written in collaboration with Guadalupe Bedoya, Joshua Blumenstock, and Aidan Coville.

Abstract

Can mobile phone data identify ultra-poor households? By combining rich survey data from a "big push" anti-poverty program in Afghanistan with detailed mobile phone logs from program beneficiaries, we study the extent to which machine learning methods can accurately differentiate ultra-poor households eligible for program benefits from ineligible households. We show that machine learning methods leveraging mobile phone data can identify ultra-poor households nearly as accurately as survey-based measures of consumption and wealth; and that combining survey-based measures with mobile phone data produces classifications more accurate than those based on a single data source.

3.1 Introduction and context

The previous chapter introduces the use of mobile phone data for poverty targeting in broad national and regional-scale social protection programs in Togo. This chapter studies the use of mobile phone data for *ultra-poverty* targeting, identifying the poorest of the poor in one region of rural Afghanistan.

In this chapter, we match mobile phone metadata (call detail records, or CDR) from a large mobile phone operator in Afghanistan to household survey data from the Afghan

government's Targeting the Ultra-Poor (TUP) anti-poverty program. Eligibility for the TUP program was determined through a *hybrid targeting method*, combining a community wealth ranking (CWR) and a short follow-up survey. Our analysis assesses the accuracy of three counterfactual targeting approaches at identifying the actual beneficiaries of the TUP program: (i) our *CDR-based method*, which applies machine learning to data from the mobile phone company; (ii) an *asset-based wealth index*, which uses asset ownership to approximate poverty; and (iii) *consumption*, a common benchmark for measuring poverty in LMICs.

Our analysis produces three main results. First, by comparing errors of inclusion and exclusion using the program's hybrid method as a benchmark, we find that the CDR-based method is nearly as accurate as the commonly employed asset and consumption-based methods for identifying the phone-owning ultra-poor households. Second, we find that methods combining CDR data with measures of assets and consumption are more accurate than methods using any single data source. Third, we find that when non-phone-owning households are included in the analysis, the CDR-based method remains accurate if non-phone-owning households are classified as ultra-poor; however, targeting performance is quite poor if households without phones are ineligible for benefits. After presenting these main results, we compile data from several existing targeting programs to give an indication of the substantial reduction in marginal costs associated with CDR-based targeting.

The context of our empirical analysis – identifying ultra-poor households in Afghanistan – is a particularly challenging environment for data collection and program targeting, as 62% of the households classified as not *ultra*-poor still fall below the national poverty line. In such environments, when traditional options for targeting are not feasible, these methods may provide a viable alternative for identifying households with the greatest need. Given the policy relevance of these results, we conclude our analysis by discussing important ethical and logistical considerations that may influence how CDR methods are used to support targeting efforts in practice.

3.2 Methods

3.2.1 Targeting the "ultra-poor" in Afghanistan

Our empirical analysis relies on survey data collected as part of the Targeting the Ultra-Poor (TUP) program implemented by the government of Afghanistan with support from the World Bank. The TUP program was a "big push," providing multi-faceted benefits to 7,500 ultra-poor households in six provinces of Afghanistan between 2015 and 2018 [32]. Our analysis uses data from the baseline and targeting surveys from an impact evaluation of the TUP program conducted in Balkh province.

Ultra-poor designation. Eligibility for the TUP program was determined based on geographic criteria,[1] followed by a two-step process including a community wealth ranking (CWR) and a follow-up in-person survey. CWRs were conducted separately in each village, coordinated by a local NGO and village leaders, in collaboration with the government team. The CWR was followed by an in-person survey to determine whether nominated households met a set of qualifying criteria, coordinated by the NGO and government representatives, and based on a measure of multiple deprivation.

For a household to be designated as *ultra-poor*, and therefore eligible for program benefits, it had to be considered extreme-poor in the CWR (43% of households), and also meet at least three of six criteria:

1. Financially dependent on women's domestic work or begging
2. Owns less than 800 square meters of land or living in a cave
3. Primary woman under 50 years old
4. No adult men income earners
5. School-age children working for pay
6. No productive assets

Ultimately, 11% of the households classified as extreme-poor in the community wealth ranking step — 6% of the total population in the study villages — were classified as ultra-poor and were thus eligible for TUP benefits.

3.2.2 Household surveys

To facilitate Bedoya et al. (2019)'s [32] impact evaluation of the TUP program, household surveys were conducted in 80 of the poorest villages of Balkh province. A total of 2,852 households were surveyed, with ultra-poor households (N=1,173) oversampled relative to non-ultra-poor households (N=1,679).[2] Surveys were conducted between February and April 2016, following the CWR and eligibility verification. This survey window was timed to occur in the late winter and early spring, a few months before the harvesting season for wheat in Balkh.

The household survey was a long-form in-person survey that took approximately 3 hours for each household to complete. The survey covered a wide range of topics, including several modules related to household poverty and deprivation that feature in our analysis.

[1] The poorest villages were identified by the availability of veterinary services, financial institutions, and social services, and being relatively accessible [32].

[2] In our analysis, we restrict to the 2,814 households for which asset and consumption data are nonmissing.

Consumption. The consumption module of the TUP survey captured information on household food consumption for the week prior to the interview and non-food expenditures for the month or year prior to the interview. These are used to construct monthly *per capita* consumption values, as detailed in [32]. Based on these data, we measure the logarithm of *per capita* monthly consumption, using the same approach that the Afghan government used to determine the national poverty line. This monthly consumption aggregate thus captures a short-term (weekly) measure of food consumption during one of the planting seasons, as well as a medium-term (monthly and annual) measure of non-food expenditures [58, 133].

Asset index. We use survey data on household assets to construct a wealth index for each household, which provides an indication of each household's wealth relative to others in the survey. Specifically, we calculate the first principal component of variation in household asset ownership based on the sixteen items listed in Table S1, across the 2,814 households with complete asset data, after standardizing each asset variable to zero mean and unit variance. This wealth index explains 25.3% of the variation in asset ownership. Figure S1 shows the distribution of the underlying asset index components and Table S1 shows the direction of the first principal component. Broadly, we expect that the asset index will provide an indication of each household's long-term economic status, relative to other households in the survey.

Other variables. The TUP surveys collected several other covariates that we use in subsequent analysis. These include a food security index (composed of variables relating to the skipping and downsizing of meals, separately for adults and children), a financial inclusion index (composed of access to banking and credit, knowledge of banking and credit, and savings), and a psychological well-being index for the primary woman (standardized weighted scores on the Center for Epidemiological Studies Depression scale, the World Values Survey happiness and satisfaction questions, and Cohen's Stress Scale) — see Bedoya et al. (2019) [32]. The survey also collected data from each household on mobile phone ownership. Nearly all (99%) households with a cell phone provided their phone numbers and consented to the use of their call detail records for this study.

Sample representativity. Portions of our analysis are restricted to the 535 households from the TUP survey with phone numbers that match to our CDR (see subsection 3.2.3). Table 3.1 and Figure S2 compare characteristics of these households to the full survey population. There are some systematic differences: the 535-household sample is wealthier, which is consistent with households in the subsample being required to own at least one phone. For instance, while 88% of non-ultra-poor households in the TUP survey own at least one phone, only 72% of ultra-poor households own at least one phone.

Table 3.1: Summary statistics for different samples of survey respondents

Outcome	(1) Full sample (all observations)	(2) Matched Subsample	(3) Unmatched Owns Phone	(4) Unmatched No Phone
Panel A: Balance of Covariates				
Ultra-Poor	0.42 (0.49)	0.27 (0.45)	0.40 (0.49)	0.66 (0.47)
Asset Index	0.00 (2.01)	1.36 (2.60)	-0.05 (1.76)	-1.35 (0.79)
Log Expenditures	4.43 (0.71)	4.64 (0.70)	4.46 (0.70)	4.12 (0.65)
# Phones	1.35 (1.18)	1.72 (1.33)	1.59 (1.04)	0.00 (0.00)
Food Security Index	0.30 (0.90)	0.35 (0.74)	0.34 (0.93)	0.10 (0.89)
Financial Inclusion Index	0.15 (1.27)	0.34 (1.39)	0.15 (1.32)	-0.05 (0.79)
Psychological Well-being Index	0.35 (1.01)	0.38 (1.00)	0.43 (0.97)	-0.02 (1.07)
CWR Group	0.62 (0.90)	0.89 (1.02)	0.62 (0.88)	0.26 (0.66)
Panel B: Correlations Between Outcomes				
Ultra-Poor $\longleftrightarrow$ Asset Index	-0.32	-0.30	-0.27	-0.14
Ultra-Poor $\longleftrightarrow$ Consumption	-0.39	-0.30	-0.39	-0.26
Asset Index $\longleftrightarrow$ Consumption	0.37	0.34	0.34	0.15
N	2,814	535	1,807	472

Notes: Table reports average characteristics, with standard deviations in parentheses, of TUP survey respondents. Each column represents a different sample of respondents: (1) all respondents in the TUP survey; (2) Just those respondents who own a phone, where the phone number matches to the CDR obtained from the mobile phone operator; (3) Respondents who report owning a phone, but whose phone number does not match to the CDR obtained from the operator; (4) Respondents who report they do not own a phone.

Comparing survey-based measures of well-being and deprivation. As shown in Table 3.1 and Figure S3, the two survey-based measures of well-being are only weakly correlated. In the full sample, the correlation between the asset index and consumption is just 0.37; in the matched subsample, the correlation is 0.34. These modest correlations may be due in part to the fact that, as discussed above, the consumption data capture short- and medium-term deprivation, whereas the asset index is a better indicator of long-term wealth. Measurement error may also weaken these empirical correlations.

Also notable is the weak relationship between the two survey-based measures of deprivation and the ground truth ultra-poor designation: while the ultra-poor population makes up 27% of the overall sub-sample, less than half of the ultra-poor fall into the bottom 27% of the sample by wealth index or consumption. These differences may be partly attributable to measurement error, but they surely also arise from the fact that they are conceptually distinct constructs: while the consumption and asset indices focus primarily on economic flows and stocks, respectively, the ultra-poor designation was

designed to be more holistic and multidimensional, informed in part by community perceptions of vulnerability [141, 13].

The fact that the ultra-poor designation is not strongly correlated with the survey measures of consumption and wealth has important implications for the targeting analysis presented below. In particular, it suggests — and our later results affirm — that a policy targeted solely on assets or consumption data will do a poor job of differentiating between ultra-poor and non-ultra-poor. The relatively weak correlation between consumption and the asset index also hints at a later finding that targeting based on a combination of the two data sources performs better than targeting on a single source in isolation.

Sample weights. Since the TUP survey oversampled the ultra-poor (by a factor of roughly 12), portions of our analysis use sample weights to adjust for population representativeness. When sample weights are applied, it is explicitly noted; if not mentioned, no weights are applied. After sample weights are applied, the ultra-poor make up 5.98% of the overall population, and 4.63% of our matched subsample.

3.2.3 Mobile phone metadata

In a follow-up survey conducted in 2018, we requested informed consent from survey respondents to obtain their mobile phone CDR and match them to the survey data collected through the TUP project. CDR contain detailed information on:

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

For participants who consented, we match baseline survey data (collected November 2015 - April 2016) to CDR covering that same period, obtained from one of Afghanistan's main mobile phone operators. For households with multiple phones and a designated household head (N = 65), we match to CDR for the phone belonging to the household head. For households where the household head does not have a phone and someone else does (N = 17), we match to CDR for one of the households' phones selected at random. In total, for the 535 households in our sample, 629,543 transactions took place in the months of November 2015 to April 2016, broken down into 310,883 calls, 305,756 text messages, and 12,904 recharges.

From these CDR, we compute a set of 797 behavioral indicators that capture aggregate aspects of each individual's mobile phone use [120]. This set includes indicators relating to an individual's communications (for example, average call duration and percent initiated conversations), their network of contacts (for example, the entropy of their

contacts and the balance of interactions per contact), their spatial patterns based on cell tower locations (for example, the number of unique antennas visited and the radius of gyration), and their recharge patterns (including the average amount recharged and the time between recharges). The distributions of a sample of these indicators are shown in Figure S4.

3.2.4 Machine learning predictions

CDR-based method. Extending the approach described in Blumenstock (2015) [36], we test the extent to which ultra-poor status can be predicted from CDR. This analysis uses the 535 households who match to CDR to train a supervised machine learning algorithm to predict ultra-poverty status from the mobile phone features. The intuition — also highlighted in Figure S4 — is that ultra-poor individuals use their phones very differently than non-ultra-poor individuals, and machine learning algorithms can use those differences to predict ultra-poor status.

Our main analysis uses a gradient boosting model, which generally out-performs several other common machine learning algorithms for this task (see Table S3). The feature importances for the trained model are shown in Table S2. To limit the potential for overfitting, probabilistic predictions are generated via 10-fold cross-validation, with folds stratified to preserve class balance.[3] Additional details on the machine learning methods are provided in appendix B.1.

Combined methods. We also evaluate several approaches that use data from multiple sources to predict ultra-poor status. Our main *combined method* trains a logistic regression to classify the ultra-poor and non-ultra-poor households using the predicted ultra-poor probability from the CDR-based method (i.e., the output of the gradient boosting algorithm described above), as well as asset and consumption data collected in the TUP survey. For comparison, we similarly evaluate the performance of methods that combine only two of the available data sources (i.e., assets plus consumption, assets plus CDR, and consumption plus CDR). Predictions for each of the combined methods are pooled over 10-fold cross-validation.

3.2.5 Targeting accuracy evaluation

Evaluation on matched subsample. Our main analysis focuses on the 535 households for which we observe both CDR and survey data, and evaluates whether machine learning methods leveraging CDR data can accurately identify households designated as ultra-poor by the TUP program (using the two-step hybrid approach described in subsection 3.2.1). We compare the performance of the CDR-based method to the

[3]While cross validation is a standard evaluation strategy in the machine learning literature, for robustness we present results using a basic single train-test split in Table S6

performance of methods based on the wealth index, consumption data, and combinations of these data sources.[4] Each targeting method is evaluated based on classification accuracy, errors of exclusion (ultra-poor households misclassified as non-ultra-poor) and errors of inclusion (non-ultra-poor households misclassified as ultra-poor). We focus on the ultra-poor designation as the "ground truth" status of the household, against which other methods are evaluated, since it is the most carefully vetted measure of well-being for this population, and the proxy that the government used to target TUP benefits.

To evaluate the performance of the CDR-based and combined methods, we pool out-of-sample predictions across the ten cross-validation folds, so that every household in our dataset is associated with a CDR-based predicted probability of ultra-poor status that is produced out-of-sample.[5] To account for class imbalance, we evaluate model accuracy using a "quota method", by selecting a cut-off threshold for ultra-poor qualification such that each method identifies the proportion of ultra-poor households in our subsample; this cut-off also balances inclusion and exclusion errors. This quota-based approach reflects a scenario in which a program has a fixed budget constraint; it is also frequently used in the targeting literature [11, 140, 41]. In our 535-household matched dataset this threshold is 27%; in other samples (see following subsection), the percentage is different. We evaluate each method for precision (positive predictive value) and recall (sensitivity). To capture the trade-off between inclusion and exclusion errors for varying values of this threshold, we also construct receiver operating characteristic (ROC) and precision-recall curves for each method and consider the area under the ROC curve (AUC) as a measure of targeting quality. For each evaluation metric (precision, recall, and AUC), we bootstrap 1,000 samples from the original dataset to calculate the standard deviation of the mean of the accuracy metric. Each bootstrapped sample is the same size as the original dataset, drawn with replacement.

Accounting for households without phones. In order to focus our attention on how differences in the data used for program targeting affect targeting performance, our main results are based on the sample of 535 households for whom we have both survey data and mobile phone data. We also present results that show how performance is affected when the analysis includes TUP households for whom we do not have mobile phone data (typically because they do not have a phone or because they use a different phone network than the one who provided CDR). We provide analysis that targets such

[4]The CDR-based method uses supervised learning to model the ultra-poverty outcome, whereas the asset- and consumption-based approaches do not. To assess the importance of this difference, we experiment with applying machine learning methods to the asset and consumption data to model the ultra-poverty outcome. In results shown in Table S4, we find that a machine-learned asset predictor provides slight improvements on the standard asset-based wealth index and consumption measures. We continue to use the standard asset and consumption measures as benchmarks in the remainder of the chapter, however, as they are the targeting methods most frequently used in practice.

[5]In Table S6, we show that results are unchanged when we use a single train-test split, instead of 10-fold cross-validation.

households (1) before households with CDR, or (2) after households with CDR (see subsection 3.3.3). These results are evaluated on three different samples:

1. *Matched Sample:* The 535 households for whom could match survey responses to CDR.

2. *Balanced Sample:* This sample includes the 535 matched households as well as the 472 households in the TUP survey who report not owning any phone. It excludes households that own a phone on a different phone network than the one who provided CDR. The motivation for this sample is to provide an indication of targeting performance in a regime in which CDR can be used to target all phone-owning households. In addition to applying sample weights from the survey, households that do not own a phone are downweighted so that the balance of phone owners to non-phone-owners (with sample weights applied) is the same as in the baseline survey as a whole (with sample weights applied, 84% phone owners).

3. *Full Sample:* All 2,814 households in the TUP baseline survey for which asset and consumption data are available, with sample weights applied.

Note that the quota used to evaluate targeting changes for each sample, based on the number of households that are ultra-poor in the sample. For the matched sample, the targeting quota is 27.29%; for the balanced sample and full sample the quotas are 5.47% and 6.02%, respectively.

3.3 Results

Our first set of results evaluate the extent to which different targeting methods can correctly identify ultra-poor households. This analysis compares the performance of CDR-based targeting methods to asset-based and consumption-based targeting, using the sample of 535 households for which survey data and CDR data are both available.

An overview of these results is provided in Figure 3.1. Figure 3.1a shows the distribution of assets and consumption, as well as the distribution of predicted probabilities of being non-ultra-poor generated by the CDR-based and combined methods, separately for the ultra-poor and non-ultra-poor. The dashed vertical line indicates the threshold at which point 27% of households are classified as ultra-poor; we use this quota because 27% of households in this sample were designed as ultra-poor by TUP. Figure 3.1b provides confusion matrices that compare the true status (rows) against the classification made by each method (columns). These confusion matrices are also used to calculate the measures of precision and recall reported in Table 3.2 Panel A.

We find that the CDR-based method (precision and recall of 42%) is close in accuracy to methods relying on assets (precision and recall of 49%) or consumption (precision and recall of 45%). To evaluate the trade-off between inclusion errors and exclusion errors

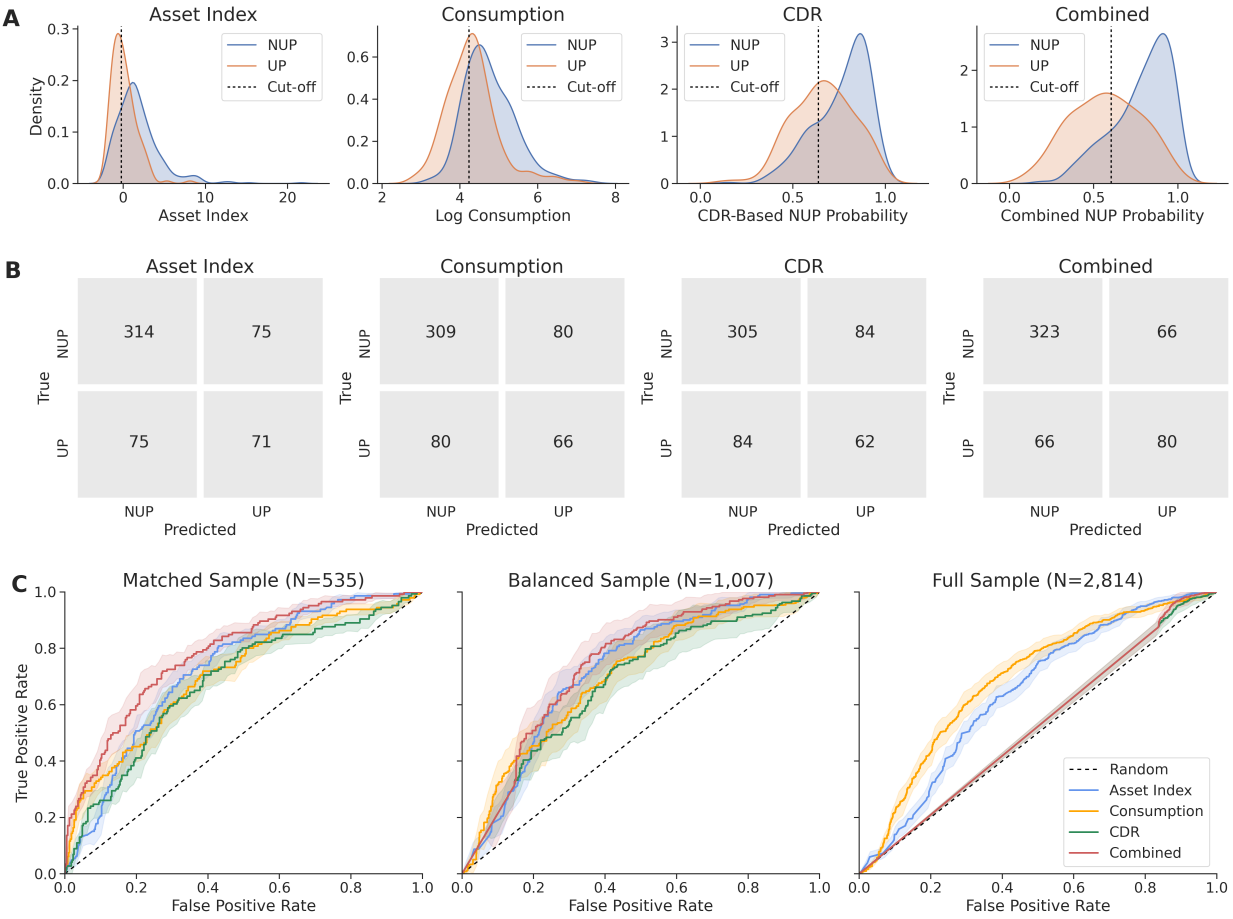

Figure 3.1: Predicting ultra-poor status from CDR. Panel A: Comparing the predictive accuracy of assets, consumption, and CDR-based methods for identifying the ultra-poor in our 535-household matched sample. To adjust for class balance, thresholds for classification (shown in dashed black vertical lines) are selected such that the correct number of households are identified as ultra-poor. Panel B: Confusion matrices showing the targeting accuracy of each method shown in Panel A. Panel C: ROC curves for each of the four targeting methods. In the third subplot, the CDR-based and combined methods target non-phone-owning households first as described in subsection 3.2.5.

resulting from selecting alternative cut-off thresholds, Figure Figure 3.1 shows the ROC curve associated with each classification method. The Area Under the Curve (AUC) scores for these curves, listed in Table 3.2, are comparable among methods, with assets (AUC=0.73) slightly superior to consumption (AUC=0.71) and the CDR-based method (AUC=0.68). The corresponding Precision-Recall curves are shown in Figure S5.

Table 3.2: Targeting simulation results

Targeting Method	(1) AUC	(2) Accuracy	(3) Precision	(4) Recall
Panel A: Matched Sample (N=535) - for whom we have survey and CDR data				
Random	0.50 (0.028)	0.60 (0.025)	0.27 (0.038)	0.27 (0.038)
Asset Index	0.73 (0.024)	0.72 (0.020)	0.49 (0.041)	0.49 (0.041)
Consumption	0.71 (0.026)	0.69 (0.023)	0.45 (0.038)	0.45 (0.038)
CDR	0.68 (0.027)	0.69 (0.021)	0.42 (0.042)	0.42 (0.042)
Combined	0.78 (0.022)	0.75 (0.020)	0.55 (0.039)	0.55 (0.039)
Panel B: Balanced Sample (N=1,007) - as above, plus households without phones				
Random	0.50 (0.017)	0.90 (0.006)	0.05 (0.010)	0.05 (0.010)
Asset Index	0.72 (0.026)	0.90 (0.006)	0.10 (0.013)	0.10 (0.013)
Consumption	0.70 (0.028)	0.90 (0.006)	0.15 (0.025)	0.15 (0.025)
CDR (Target Phoneless First)	0.68 (0.030)	0.90 (0.006)	0.11 (0.035)	0.11 (0.035)
CDR (Target Phoneless Last)	0.51 (0.028)	0.90 (0.006)	0.12 (0.033)	0.12 (0.033)
Combined (Target Phoneless First)	0.74 (0.026)	0.90 (0.006)	0.11 (0.046)	0.11 (0.046)
Combined (Target Phoneless Last)	0.57 (0.022)	0.90 (0.006)	0.18 (0.007)	0.18 (0.007)
Panel C: Full Sample (N=2,814) - as above, plus households with phones on other networks				
Random	0.50 (0.009)	0.89 (0.005)	0.06 (0.007)	0.06 (0.007)
Asset Index	0.65 (0.017)	0.89 (0.005)	0.07 (0.014)	0.07 (0.014)
Consumption	0.69 (0.015)	0.89 (0.006)	0.08 (0.031)	0.08 (0.031)
CDR (Target Phoneless First)	0.52 (0.008)	0.89 (0.005)	0.06 (0.008)	0.06 (0.008)
CDR (Target Phoneless Last)	0.48 (0.008)	0.89 (0.005)	0.08 (0.010)	0.08 (0.010)
Combined (Target Phoneless First)	0.52 (0.008)	0.89 (0.005)	0.06 (0.008)	0.06 (0.008)
Combined (Target Phoneless Last)	0.49 (0.008)	0.89 (0.005)	0.09 (0.009)	0.09 (0.009)

Notes: Four different measures of performance (columns) reported for different targeting methods (rows), using different samples of survey respondents (panels). Standard deviations, calculated using 1,000 bootstrap samples, in parentheses. Panel A: The 535-household subsample that is matched to CDR. Panel B: The 535-household matched sample, plus the 472 households that do not have a phone; this is meant to approximate targeting performance if CDR from all mobile networks were available. Sample weights are applied as described in subsection 3.2.5. Panel C: All 2,814 observations from the TUP survey, including households matched to CDR, households that own phones not matched to CDR, and households without phones, with sample weights applied. For Panels B and C, we simulate two types of CDR-based targeting: targeting households without phones first and targeting households without phones last.

3.3.1 Comparison of errors across methods

To better understand the nature of the mis-classification errors arising from the different datasets used for targeting, Table 3.3 compares the characteristics of correctly and incorrectly classified households for three different methods (targeting on assets, consumption, and CDR). Panel A highlights differences between ultra-poor households correctly classified as ultra-poor (True Positives) and ultra-poor households *mis*-classified as non-ultra-poor (False Negatives, also referred to as exclusion errors). Likewise, Panel B highlights differences between non-ultra-poor households correctly classified as non-ultra-poor (True Negatives), and non-ultra-poor households mis-classified as ultra-poor (False Positives, or inclusion errors). This analysis uses the *matched* sample (see Table 3.2) to highlight differences that arise when switching from one targeting dataset to another, on a population of households that are observed in all three datasets.[6]

[6]Similar analysis could also be performed using the balanced sample or the full sample; however, results would conflate differences caused by the targeting data (the current focus of Table 3.3) with the differences that arise from considering (or excluding) households without mobile phones (the current focus of Table 3.2).

Table 3.3: What types of households are misclassified?

Panel A: Ultra-Poor Households (Differences Between True Positives and False Negatives)

	Asset Index			**Consumption**			**CDR**		
	TP	FN	Diff.	TP	FN	Diff.	TP	FN	Diff.
Ultra-Poor	1.00	1.00	0.00	1.00	1.00	0.00	1.00	1.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Asset Index	-1.03	1.18	-2.21	-0.34	0.47	-0.81	-0.09	0.25	-0.34
	(0.49)	(1.34)	(0.17)	(1.09)	(1.69)	(0.23)	(1.16)	(1.70)	(0.24)
Consumption	4.21	4.40	-0.19	3.78	4.74	-0.96	4.29	4.32	-0.02
	(0.70)	(0.62)	(0.11)	(0.32)	(0.56)	(0.07)	(0.60)	(0.71)	(0.11)
# Phones	0.89	1.63	-0.74	1.02	1.48	-0.46	1.18	1.33	-0.16
	(0.68)	(1.12)	(0.15)	(0.73)	(1.14)	(0.16)	(0.61)	(1.21)	(0.15)
Food Security Index	-0.59	-0.51	-0.08	-0.83	-0.32	-0.51	-0.51	-0.58	0.07
	(1.13)	(1.10)	(0.18)	(1.19)	(0.99)	(0.18)	(1.14)	(1.09)	(0.19)
Financial Inclusion Index	-0.00	0.29	-0.29	0.10	0.19	-0.09	0.16	0.14	0.02
	(0.79)	(1.02)	(0.15)	(0.80)	(1.02)	(0.15)	(0.98)	(0.88)	(0.16)
Psychological Wellbeing Index	-0.35	-0.13	-0.22	-0.37	-0.12	-0.24	-0.31	-0.17	-0.14
	(0.92)	(0.94)	(0.15)	(0.86)	(0.98)	(0.15)	(0.81)	(1.02)	(0.15)
CWR Group	0.09	0.01	0.07	0.02	0.08	-0.06	0.06	0.04	0.03
	(0.44)	(0.12)	(0.05)	(0.12)	(0.41)	(0.05)	(0.40)	(0.24)	(0.06)

Panel B: Non-Ultra-Poor Households (Differences Between True Negatives and False Positives)

	Asset Index			**Consumption**			**CDR**		
	TN	FP	Diff.	TN	FP	Diff.	TN	FP	Diff.
Ultra-Poor	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Asset Index	2.53	-1.08	3.61	2.06	0.94	1.12	1.94	1.43	0.51
	(2.62)	(0.50)	(0.16)	(2.92)	(1.75)	(0.26)	(2.87)	(2.27)	(0.30)
Consumption	4.82	4.57	0.25	4.97	3.98	0.99	4.78	4.74	0.04
	(0.66)	(0.65)	(0.08)	(0.58)	(0.23)	(0.04)	(0.68)	(0.61)	(0.08)
# Phones	2.11	0.96	1.15	1.98	1.52	0.46	1.91	1.80	0.11
	(1.43)	(0.76)	(0.12)	(1.49)	(0.92)	(0.13)	(1.44)	(1.24)	(0.16)
Food Security Index	0.24	-0.16	0.40	0.24	-0.14	0.37	0.15	0.18	-0.02
	(0.87)	(1.03)	(0.13)	(0.88)	(0.99)	(0.12)	(0.91)	(0.94)	(0.12)
Financial Inclusion Index	0.80	-0.01	0.82	0.77	0.18	0.59	0.78	0.17	0.61
	(4.92)	(0.82)	(0.29)	(4.94)	(1.24)	(0.31)	(4.98)	(1.10)	(0.31)
Psychological Wellbeing Index	0.69	0.21	0.47	0.62	0.49	0.13	0.62	0.49	0.13
	(0.97)	(0.75)	(0.10)	(0.98)	(0.80)	(0.11)	(0.95)	(0.93)	(0.12)
CWR Group	1.30	0.84	0.46	1.23	1.13	0.10	1.26	1.01	0.25
	(1.00)	(0.96)	(0.12)	(1.03)	(0.94)	(0.12)	(1.01)	(0.98)	(0.12)

Notes: Table shows the average characteristics, with standard deviations in parentheses, of households that are correctly and incorrectly classified by each targeting approach, using the matched sample. Panel A highlights differences between between ultra-poor households correctly classified as ultra-poor (True Positives, TP) and ultra-poor households mis-classified as non-ultra-poor (False Negatives, FN; i.e., exclusion errors). Panel B highlights differences between non-ultra-poor households correctly classified as non-ultra-poor (True Negatives, TN), and non-ultra-poor households misclassified as ultra-poor (False Positives, FP; i.e., inclusion errors).

Across methods, false negatives (exclusion errors) have higher levels of food security, financial inclusion, and psychological well-being than true positives – that is, all three targeting methods misclassify ultra-poor households as non-ultra-poor when those ultra-poor households are better-off, according to other observable characteristics not used in the targeting. Likewise, false positives (inclusion errors) tend to fare worse than true negatives across these same measures. The exact pattern of differences depends on the targeting method; for instance, asset-based targeting (first set of columns) tends to misclassify ultra-poor as non-ultra poor when they have assets (the difference of -2.21 is large), but errors are not systematically correlated with consumption (the difference of -0.19 is relatively small). The CDR-based method in particular tends to prioritize households that score low on these alternative measures of well-being. These patterns suggest that the CDR-based targeting method may capture aspects of well-being that are not captured by standard survey-based measures of poverty such as wealth and consumption.

To test for systematic misclassification of certain types of households, Table 3.4 displays the overlap in errors of exclusion and inclusion between methods. Our results suggest that the three classifiers misidentify the same households at a rate only slightly above random.[7]

[7]The rates of overlap should be interpreted relative to the expected overlap in errors for random classifiers. Based on our selection of thresholds such that 27% of the sample is identified as ultra-poor, our three classifiers misidentify 15%-27% of the non-ultra-poor and 51%-65% of the ultra-poor. If these classifiers were random, we would expect approximately 20% overlap in inclusion errors and 55% overlap in exclusion errors.

Table 3.4: Overlap in targeting errors between methods

	Asset Index	Consumption	CDR	Combined
Panel A: Overlap in Errors of Exclusion				
Asset Index	100.00%	65.33%	57.33%	66.67%
Consumption	61.25%	100.00%	56.25%	62.50%
CDR	51.19%	53.57%	100.00%	63.10%
Combined	75.76%	75.76%	80.30%	100.00%
Panel B: Overlap in Errors of Inclusion				
Asset Index	100.00%	26.67%	22.67%	48.00%
Consumption	25.00%	100.00%	16.25%	37.50%
CDR	20.24%	15.48%	100.00%	46.43%
Combined	54.55%	45.45%	59.09%	100.00%

Notes: Table measures the extent to which the targeting errors produced by each pair of targeting methods overlap in the matched sample. Evaluation is performed on the matched sample of 535 TUP respondents. Panel A: Overlap between ultra-poor households that are misclassified as non-ultra-poor (errors of exclusion) for each targeting method. Panel B: Overlap between non-ultra-poor households that are misclassified as ultra-poor (errors of inclusion).

3.3.2 Combining targeting methods

Since the different targeting methods identify different populations as ultra-poor, there may be complementarities between asset, consumption, and CDR data. As shown in Panel A of Table 3.2, we find that a *combined method*, which takes as input the wealth index, total consumption, and the output of the CDR-based method, performs better (AUC = 0.78) than methods using any one data source (AUC = 0.68 - 0.73). As shown in Table S5, the full method also outperforms methods based on any two data sources (AUC = 0.75 - 0.76). The method that combines CDR and asset data (AUC = 0.76) may, however, be more practical than the combined method, since consumption data is difficult to collect for large populations.

3.3.3 Targeting households without phones

An important limitation of CDR-based targeting is that households without phones do not generate CDR. Here, we show how targeting performance is impacted when households without phones are included in the analysis. This analysis uses two additional samples of TUP households to evaluate targeting performance: (i) the *balanced sample*, which adds all of the 472 households without phones to the sample of 535 for whom we have matched CDR; the balanced sample is intended to illustrate the performance of CDR-based targeting if CDR were available from all operators in

Afghanistan — though it relies on the assumption that phone-owners observed on our mobile network are representative of all phone owners in Afghanistan (an assumption that is not fully satisfied, as shown in Table 3.1); and (ii) the *full sample*, which includes all 2,814 households surveyed in the TUP baseline with complete asset and consumption data; this sample includes an additional 1,807 households who report owning a phone, but whose number does not match to any number in the CDR provided to us by the single mobile operator.[8]

Results in Panels B and C of Table 3.2 show the performance of each targeting approach on the balanced and full sample, respectively. Note that as described in subsection 3.2.5, different targeting quotas are applied for each panel based on the proportion of each sample that is ultra-poor. In the CDR-based and combined approaches, we report performance when the households without CDR are targeted first (i.e. households without CDR are targeted in a random order and then the households predicted to be poorest are targeted until the quota is reached) as well as when households without CDR are targeted last (i.e., after the 535 households with phones are targeted, households without phones are included in a random order until the quota is reached).

Unsurprisingly, these results suggest that CDR-based targeting is not effective when a large portion of the target population does not own a phone (e.g. Panel C of Table 3.2, where only 16% of the sample has matching CDR). However, when we simulate more realistic levels of phone ownership in Panel B (84% of the households, based on our survey data), CDR-based targeting is once again comparable to asset- or expenditure-based targeting, particularly when households without phones are targeted first (AUC = 0.72, 0.70, 0.68 for assets, consumption, and CDR, respectively). On the other hand, if households without phones are targeted last (for example, if program administrators base targeting wholly on CDR and provide no benefits to any household without a phone), the CDR-based method only improves marginally on random targeting.[9]

3.3.4 Additional tests and simulations

Our main analysis considers the household head to be the unit of analysis. As described in subsection 3.2.3, this analysis is based on matching survey-based indices to phone data from the household head, which is consistent with the design of the TUP

[8]These 1,807 households include households that report owning a phone on a different network (this network is estimated to have around 30% market share in Afghanistan), as well as phones on our network that were not active during the six-month period of CDR that we analyze.

[9]A key nuance in this analysis is that for the CDR-based and combined methods where households without phones are targeted first, the precision and recall measures in Table 2 correspond to programs that only target households without phones (at random), as the number of households without phones exceeds the budget constraint of the program. The AUC score, on the other hand, is a summary statistic that represents targeting accuracy at all counterfactual targeting thresholds, and thus is not sensitive to the budget constraint — which explains the contrast between AUC and precision and recall in Table 2 Panels B and C. The ROC curves (Figure 3.1) and Precision-Recall curves (Figure S5) highlight how budget constraint affects precision and recall.

program and the TUP survey sample frame. An alternative approach matches survey data reported by the household head to all phone numbers associated with the household. As shown in Table S7, the predictive accuracy of these models is slightly attenuated relative to the benchmark results (Table S3).

We also explore the extent to which CDR can be used to predict other measures of socioeconomic status. Our main analysis focuses on the household's ultra-poor designation as the ground truth measure of poverty, since this label was both carefully curated and the actual criterion used to determine TUP eligibility. In Table S8, we report the accuracy with which CDR (obtained from the household head, who is typically male) can predict consumption and asset-based wealth (elicited from the primary woman of each household).[10] In general, these machine learning models trained to directly predict consumption or asset-based wealth do not perform well. This result contrasts with prior work documenting the predictive ability of CDR for measuring asset-based wealth [e.g. 36]. We suspect a key difference in our setting – aside from the fact that we are matching CDR to socioeconomic status at the *household* rather than the *individual* level – is the homogeneity of the beneficiary population: whereas [36] uses machine learning to predict the wealth of a nationally-representative sample of Rwandan phone owners, our sample consists of 535 individuals from the poorest villages of a single province in Afghanistan, where even the relatively wealthy households are quite poor.

3.4 Discussion

Our key finding is that, in a sample of 535 phone-owning households in poor villages in Afghanistan, machine learning methods leveraging phone data are nearly as accurate at identifying ultra-poor households as standard asset- and consumption-based methods. Further, we find that methods combining survey data with CDR perform better than methods using a single data source. However, as we demonstrate empirically, low rates of phone ownership — or the inability to access data from all operators — can undermine the value of CDR-based targeting. In our setting, the CDR-based approach still works well if households without phones are targeted before the CDR-based algorithm selects the poorest households with phones. However, this approach may not be appropriate in other contexts where phone ownership is less predictive of wealth, or where potential beneficiaries have the ability to strategically underreport phone ownership [34].

As mobile phone penetration rates continue to rise in LMICs [82], and as programs increasingly rely on mobile phones and money to distribute benefits [cf. 77], CDR-based targeting methods will likely play a more prominent role in the set of options considered by policy makers and program administrators — particularly in contexts like Afghanistan, where traditional targeting benchmarks are missing or unreliable. In just the past few years, for instance, data from mobile phone operators was used in the design of

[10] Due to the design of the TUP survey, which interviewed women in the household, we cannot avoid this mismatch between the survey respondent and the phone owner.

social assistance programs in Colombia, the Democratic Republic of Congo, Pakistan, and Togo [77, 122, 5].

Speed and cost. An advantage of CDR-based targeting is that it can be used in contexts where face-to-face contact is not feasible, dramatically reducing the time required to implement a targeted program. While it typically takes many months (or years) to implement a proxy-means test (PMT), community-based targeting (CBT), or consumption-based targeting, a CDR-based model can be trained in just a few weeks (see appendix B.3). Likewise, the marginal costs per household screened are substantially lower with CDR-based targeting than with CBT, PMT, or consumption-based targeting. For instance, Table S9 uses cost estimates obtained from the literature (and detailed in Table Table S10) to estimate targeting costs for the TUP program.[11] Whereas the marginal costs of screening an individual with a CBT or PMT are estimated at $2.20 and $4.00, respectively, the marginal cost of screening with CDR is negligible (see appendix B.3).[12] For the entire TUP program, which screened around 125,721 households in six provinces, CBT and PMT would add an additional estimated $276,586 and $502,884, respectively, corresponding to 2.18% and 3.97% of the total program budget.

To summarize, our results suggest that there is potential for using CDR-based methods to determine eligibility for economic aid or interventions, substantially reducing program targeting overhead and costs. Our results also indicate that CDR-based methods may complement and enhance existing survey-based methods. We note, however, that the practical and ethical limitations to CDR-based targeting are significant. We emphasize the need to consider these limitations and the constraints of specific local contexts alongside the efficiency gains offered by CDR-based targeting.

[11] In our cost calculations we obtain estimates for a CBT, rather than the hybrid approach used in the TUP program, as there is more information available on CBT-only costs in the literature. However, as the CBT cost can be interpreted as a lower bound for the cost of a hybrid approach, our qualitative results also apply to a hybrid approach.

[12] Marginal costs of CDR-based targeting are negligible because we assume no contact with screened individuals is required. In practice, it may be desirable to solicit informed consent to access CDR. If consent were collected in-person, the marginal costs would approach that of a PMT; if collected over the phone, there would still be significant cost savings, see appendix B.3).

Chapter 4

Comparing community-based and phone-based targeting in Bangladesh

This chapter is based on the paper "Comparing Community-Based Targeting to Big Data and Machine Learning in Bangladesh", written in collaboration with Anik Ashraf, Joshua Blumenstock, Raymond Guiteras, and Mushfiq Mobarak.

Abstract

The types of "big data"-driven targeting approaches described in the first two chapters of this thesis have enabled new paradigms for the targeting of social protections and humanitarian aid. However, these centralized and top-down approaches typically do not involve community participation or feedback, and thus may not capture nuanced conceptions of poverty as well as local, community-based approaches. This chapter uses a wealth of data from Bangladesh — including mobile phone records from the four major mobile network operators in Bangladesh, community-based wealth ranking data from 180 communities, a census of 100,000 households, and detailed survey data from 5,000 households — to directly compare "big data" targeting (based on mobile phone records) with community-based targeting. We find that phone-based targeting is more accurate than community-based targeting at identifying the consumption-poorest households, but both methods perform substantially worse than traditional proxy-means testing. We also explore the extent to which different data sources can be combined to more effectively identify poor and vulnerable households, and the extent to which different targeting methods work better for different types of communities and households.

4.1 Introduction and context

"Big data"-driven targeting approaches — like those introduced in the first two chapters of this thesis — are starting to be used in real-world social protection and humanitarian aid programs. Recently implemented aid programs have used measures of poverty inferred from mobile phone data [5, 6, 122] and satellite imagery [5, 146, 73] to target aid in settings where traditional social registries are incomplete, out-of-date, or missing altogether. While the results in the first two chapters of this thesis suggest that these approaches are less accurate than traditional survey-based targeting methods [5, 6], they have the advantages of speed, scale and low marginal costs. Particularly in settings where on-the-ground conditions require rapid, comprehensive, and targeted program roll-out (such as natural disasters, conflicts, and pandemics), "big data" targeting methods may be a policymaker's best option.

While this new targeting paradigm thus introduces a compelling and cost-effective option, it also raises important concerns. Chief among these is that top-down and centralized decision-making algorithms based on "big data" do not involve community participation or feedback. By contrast, Community-Based Targeting (CBT) — a very different paradigm for targeting that is used in many low and middle-income countries (LMICs)[1] — takes the community's input first and foremost, asking communities (or representatives of communities) to work together to select beneficiaries who are poor. CBTs can empower communities and increase community satisfaction with the process of allocating resources [9], and a number of donor agencies now make community input a requirement for targeted aid programs [159]. However, CBTs have been criticized for low targeting accuracy [51, 159], elite capture [130, 84, 52], and the exclusion of minorities and the socially isolated [104, 115]. The implementation of CBTs is also substantially more expensive and time-intensive than "big data approaches" (which can be implemented at a low cost and entirely remotely [6]).

This chapter presents the first direct comparison of "big data" targeting (in this setting, as in the first two chapters in this thesis, leveraging mobile phone metadata to infer poverty) to community-based targeting, and asks if and how the two approaches can be combined to most effectively identify poor and vulnerable households. This chapter also benchmarks the phone-based and community-based approaches in comparison to other established and emerging targeting approaches in LMICs, including proxy-means testing (PMT) [81], geographic targeting [19], targeting with categorical eligibility criteria, decentralized community-based targeting based on peer rankings [9, 31, 153], and random targeting via lotteries [20]. Thus, while our results speak most directly to recent work documenting the accuracy of targeting using machine learning and mobile phone data [5], this chapter also contributes to the larger literature comparing the accuracy of

[1] For example, over half of the cash transfer programs in Sub-Saharan Africa include a community-based targeting component [75]. CBT also plays a prominent role in the targeting of ultra-poor graduation programs [101], including BRAC's influential graduation program in Bangladesh [114].

suites of targeting approaches in LMICs [51, 101, 11, 130, 140].

Our analysis combines comprehensive mobile phone records from the four major operators in Bangladesh with community-based ranking data collected in 180 neighborhoods, roughly 5,000 household surveys, and poverty scorecard data from 100,000 households in the Cox's Bazar district of Bangladesh.[2] We highlight three main results. First, we show that phone-based targeting is more accurate than community-based targeting for identifying the consumption-poorest households. However, both methods are substantially less accurate than proxy-means testing. Second, we explore methods that leverage combinations of phone-based targeting, community-based targeting, and proxy-means testing. We show that the addition of phone- or community-based data does not improve the PMT, but combining community-based and phone-based targeting improves slightly upon using just one of these data sources. Finally, we assess heterogeneity in targeting accuracy across types of communities and types of households, finding that community-based targeting only outperforms phone-based targeting in urban communities and for very large and very small households.

4.2 Methods

In this chapter, we benchmark the accuracy of a number of approaches to targeting social protections in southern Bangladesh. We do this in the context of a cash transfer program that we developed in partnership with GiveDirectly and the Government of Bangladesh. The program received funding to make cash transfers to a total of 22,000 households in three sub-districts in southern Bangladesh — Ramu, Teknaf, and Ukhia.[3] Our main analysis compares proxy-means testing (PMT), community-based targeting (CBT), and phone-based targeting for identifying the consumption-poorest households in this setting, while ancillary analyses also study geographic targeting, categorical targeting, the progress-out-of-poverty index, an asset index, and peer rankings. Section 4.2.1 describes the data collected, Section 4.2.2 describes the construction of each of these targeting approaches, and Section 4.2.3 describes the metrics used to measure targeting accuracy.

4.2.1 Data

Our analysis relies on four main data sources:

- A **census** of all households in 200 randomly selected villages of three sub-districts in southern Bangladesh - Ramu, Teknaf, and Ukhia. The census was conducted

[2]This research was conducted in conjunction with a GiveDirectly cash transfer program conducted in Cox's Bazar in winter 2023. The program provided two monthly transfers of 7,500 Takas ($239 PPP) to households identified through phone-based targeting.

[3]The choice of the geographical region was based on preferences of the Government of Bangladesh and the donors of the program.

during February-March 2023 and created a listing of all households in the villages. It also collected information on, among others, household sizes, phone numbers of adult household members, and a few indicators of socio-economic status of the households that helps to calculate the Poverty Probability Index (PPI)[4] of the households. The census collected information for around 106,000 households.

- A **household survey** conducted in March 2023, which collected consumption expenditures[5] demographics, assets, and peer rankings.[6] The household survey was conducted with a representative random sample of 5,006 households from 180 randomly selected neighborhoods in the study area.

- Household wealth rankings from **community-based targeting exercises** conducted in November 2023 in each of the 180 neighborhoods. Our CBT exercises were conducted following a protocol that was developed and implemented by BRAC to determine beneficiaries for their own social safety net programs. In particular, the CBTs were conducted with participation from 12-25 households in each neighborhood.[7] To make the CBT exercises incentive compatible, participants were informed at the start of the meeting that the 20% poorest-ranked households would receive a one-time cash transfer of 1,000 Takas ($31.88 USD PPP) following the meeting.

- Complete **mobile phone metadata** from all consenting survey respondents from March to July 2023, including records of calls, texts, and mobile data usage.[8]

[4]https://www.povertyindex.org/country/bangladesh

[5]Consumption data were collected using the consumption expenditures module from the Bangladesh's 2016 Household Income and Expenditures Survey, excluding any items consumed by less than 1% of all households in the HIES *and* making up less than 1% of total expenditures for households that did consume them.

[6]In the peer rankings module, each household interviewed was asked about eight randomly selected households in their neighborhood. They were asked to report how well they knew the household and to rate the how well-off the household was on a scale of 1-5.

[7]The CBT protocol is summarized as follows. First, in neighborhoods of more than 100 households, enumerators split neighborhoods into contiguous segments of 50-100 households and conducted separate CBTs in each. Enumerators worked with senior community members to identify 12-25 households to join the meeting, inviting households from all walks of life and ensuring participation from women, students, farmers, businessmen, and laborers. Each meeting began with a "social mapping" exercise in which a community map was drawn with each household identified by name and occupation. Meeting attendees then worked together to rank the wealth of all households in the community by placing index cards representing each household on a string in the order of wealth.

[8]We analyzed mobile phone metadata from all four mobile network operators active in Cox's Bazar for all consenting survey respondents. Following the data protection procedures described in our IRB protocol, we pseudonymized or removed all personally identifying information, including phone numbers, prior to analyzing mobile phone metadata.

4.2.2 Targeting methods

The three main targeting methods we study are:

- The **phone-based targeting** approach uses machine learning methods to predict consumption expenditures from roughly 1,500 statistics on each subscribers' mobile phone use (including information about calls, texts, contact diversity, mobility, and mobile data usage).[9] Our machine learning methods are similar to those used in past work [5, 6, 4]: we use a gradient boosting model with hyperaparameters selected via three-fold cross validation.

- The **community-based targeting (CBT)** rankings from each community are used directly as a targeting approach. Rankings are normalized within each community to a 0-1 range for consistency across communities.[10]

- The **proxy-means test (PMT)** predicts poverty from survey-based covariates, including household characteristics (for example, the number of rooms and the material of the roof), demographic information (for example, the household size and gender of the household head), and asset ownership. 45 covariates are included in total. Following recent studies that show that machine learning approaches improve upon the traditional linear regression to construct PMTs [116, 126], we use a LASSO regression to predict consumption expenditures from PMT covariates, with the L1 penalty parameter chosen via cross validation.

We additionally replicate some less common targeting approaches that are also relevant counterfactuals:

- **Geographic targeting** at the union (admin-5) level, based on aggregating population-weighted wealth estimates from the relative wealth index [50]

- **Categorical targeting** based on eligibility criteria designed for urban Bangladesh by IFPRI [2]: a household is ranked as "poor" if it meets at least two of six categorical eligibility criteria[11]

- Other survey-based targeting approaches similar to the PMT, including Bangladesh's **poverty probability index (PPI)** and an **asset index** constructed with principal components analysis

[9] Subscriber-level statistics on mobile phone use are calculated using open source python library cider.

[10] This approach implicitly assumes that wealth ranges are consistent across neighborhoods; a more sophisticated approach could make use of data on neighborhood-level poverty to adjust rankings.

[11] IFPRI's original criteria are: Household head is a rickshaw/tricycle cart driver, household head is a non-agricultural day laborer, any member of the household ages 19-59 works as a maid servant, household does not own a fan, household does not own a TV, household walls are constructed from hemp/bamboo/other non-durable materials, and household is located in a slum or other poor area. We do not have data on slum locations in Cox's Bazar so we drop the final criterion.

- **Peer rankings**, based on taking the simple average of all wealth ratings for a given household by their neighbors

4.2.3 Evaluation metrics

Each targeting method is benchmarked to per capita household consumption expenditures collected in the household survey. Evaluation is conducted with a randomly selected 25% of the households in the survey, with the remaining 75% used to train the targeting methods that require machine learning. The evaluation is repeated 100 times on different random train-test splits, and we report the mean and standard deviation of each metric over the 100 runs.

We use the following metrics to evaluate targeting accuracy:[12]

- **Spearman correlation** between each targeting method's poverty ranking and ground truth consumption expenditures
- **Precision and recall** for identifying the poorest 21% of households[13]
- **Area under the ROC curve (AUC score)** integrating under the false positive rate - true positive rate curve by varying the targeting threshold from 0 to 100%

Some of the targeting methods we simulate do not produce rankings for all households. In such cases, households that are unranked are targeted last in our targeting simulations. In the phone-based targeting approach, 6% of households are not given a wealth ranking (2% of households in the survey do not provide a phone number or do not consent to matching survey data to mobile phone records; 4% of households in the survey provide at least one mobile phone number but no number is associated with transactions in our mobile phone metadata). 0.4% of households were not ranked in the CBT exercises and 2% of households had no peer rankings (because they were not known to the community).

[12]In our main analysis, targeting methods are evaluated for their ability to identify the poorest 21% of households in the entire sample (based in the budget constraint for the GiveDirectly program). However, community-based targeting is designed to identify the poorest households in each neighborhood, so in supplementary analysis we also study targeting accuracy for identifying the poorest 21% of households in each neighborhood.

[13]GiveDirectly's Cox's Bazar program had budget to provide transfers to 22,000 households of the approximately 106,000 households enrolled (21%). We evaluate targeting accuracy in this *quota setting* where precision is by definition equal to recall [41].

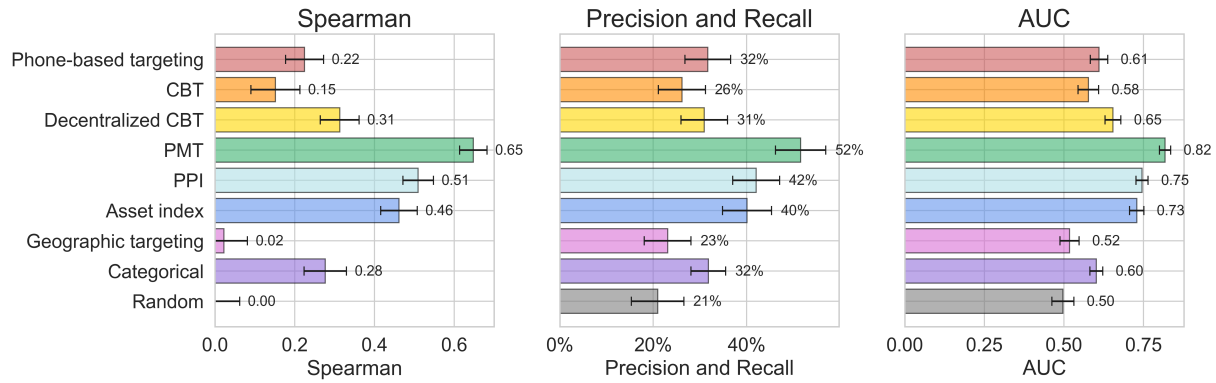

Figure 4.1: Targeting accuracy comparison, based on Spearman correlation with consumption (left), precision and recall for identifying the 21% consumption-poorest households (middle), and area under the ROC curve (right). Error bars show two standard deviations above and below the mean for each metric.

.

4.3 Results

4.3.1 Targeting accuracy of individual approaches

Our primary analysis compares the accuracy of phone-based targeting, community-based targeting (CBT), and proxy-means testing (PMT), as well as other benchmarks (Figure 4.1). We find that phone-based targeting (AUC = 0.61) is more accurate than CBT (AUC = 0.58). However, perhaps unsurprisingly, both approaches are substantially less accurate than PMT (AUC = 0.82). Other survey-based targeting variants (AUC = 0.72-0.74) also outperform phone-based targeting and CBT but are worse than PMT. Notably the peer ranking approach outperforms the CBT and phone-based targeting (AUC = 0.66).

These classification-based measures of targeting accuracy do not fully capture the quality of targeting methods: for example, two targeting approaches with the same targeting accuracy may perform differently in terms of the severity of incluion and exclusion errors (for example, one method may wrongly include households that are very wealthy, while another wrongly includes households that are just above the targeting threshold). To address this concern, Figure S2 plots the distributions of consumption expenditures for households that are included and excluded by each method (and for inclusion and exclusion errors specifically). This figure suggests that the best-performing targeting methods in term of accuracy are also the best-performing in terms of the poverty distribution of households included and excluded: the PMT tends to target the poorest households and exclude the richest ones; the CBT tends to target wealthier households and excluded poorer ones; and phone-based targeting performs in between the two.

Comparison to past work

The poor performance of a CBT in this setting may appear surprising based on past work. In particular, compared to the seminal study by Alatas et al. (2012) comparing proxy-means testing and community-based targeting in Indonesia [11] (which found only a small performance gap between CBT and PMT), the gap between CBT and PMT in this setting is much wider. However, there are two key differences between our study and the setting of Alatas et al. (2012): first, we use a *quota approach* to evaluation – with a quite narrow targeting quota at 21% – relative to the targeting objective in Alatas et al. (2012), which is to identify households below the poverty line. Second, our study focuses on targeting in a narrow georaphic area (120 neighborhoods in Cox's Bazar district in Bangladesh), in comparison to studies like Alatas et al. (2012) which study the targeting of poor households in entire countries are large areas within countries.

To better contextualize our results within past work, we compare our results on targeting accuracy of CBT and PMT approaches to empirical results from other settings that also rely on a quota approach to evaluation. We compare our results to targeting accuracy measures reported in three other papers: (1) our own work in Togo presented in Chapter 2 (which calculated targeting accuracy nationwide for a PMT at a 29% quota), (2) Schnitzer and Stoeffler (2022) [140], which evaluates the targeting accuracy of seven CBT-based and eight PMT-based social protection programs run in parts of Burkina Faso, Cameroon, Mali, Niger, and Senegal with targeting quoatas ranging from 21% to 67%, and (3) Brown et al. (2018) [41], which simulates PMT-based country-level targeting in eight African countries with 20% and 40% targeting quotas. Figure 4.2 plots the precision and recall of targeting approaches in each of these studies as a function of the targeting quota used. The fit is remarkably tight (in spite of large variations in data, program implementation, and study contexts), and our results appear to be well within the range of results reported in past work.

In addition to the salience of our quota approach to evaluation when comparing our results to past work, a further important difference between our study and many of the past studies on targeting accuracy of CBTs and PMTs is the narrow geographic scope of the program we study. Our study is limited to 120 neighborhood in Cox's Bazar district (which itself is around 1,000 square miles — slightly smaller than the state of Rhode Island). As a result, there is likely substantially less variation in poverty in our setting than in the settings of national-scale social protection programs. To partially test this hypothesis, Figure S5 simulates targeting more homogeneous subsets of our study population by poverty. The results confirm that targeting evaluations that for all methods except for random and geographic targeting, targeting simulations that are restricted to poorer subsets of the households in our survey result in lower targeting performance than evaluations conducted with the full set of households in our survey.

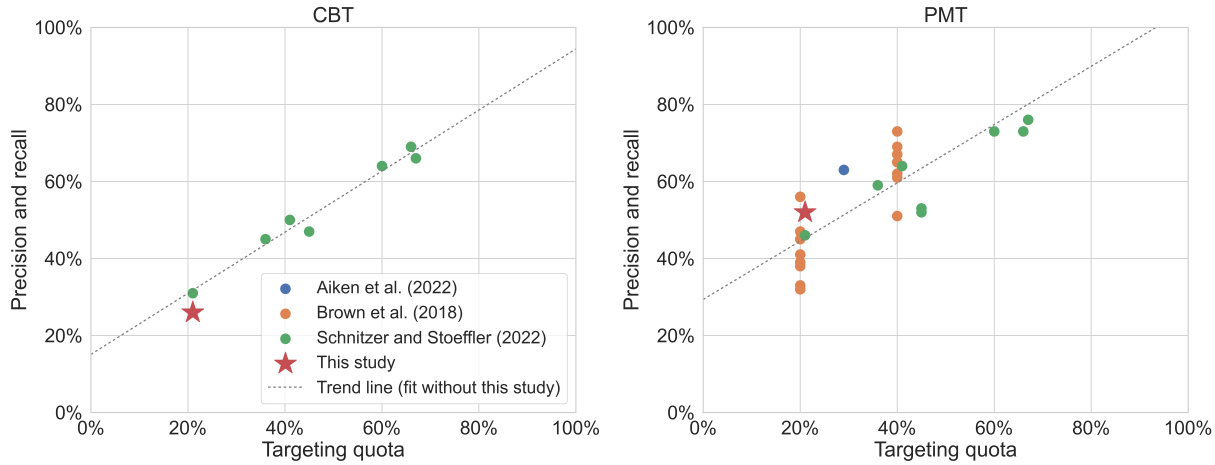

Figure 4.2: Comparison of our results on targeting accuracy (red stars) in comparison to past studies that also use a quota approach to targeting evaluation (green dots for Schnitzer and Stoeffler (2022) [140], blue dots for Brown et al. (2018) [41], and orange dots for Aiken et al. (2022) [5]. Targeting error rate is shown as a function of the targeting quota.

.

Targeting at the community level

While the typical goal of a targeted social protection program is to reach the poorest households among those screened, some may seek to identify the poorest households within each community, with a quota assigned at the community level. Moreover, this assignment of quotas more directly reflects the methods for CBT data collection: communities were asked to rank households from poorest to richest, and were told that the poorest 21% of households would receive a transfer. It is therefore possible that, while the CBT is weaker than phone-based targeting overall, it may perform better at identifying the poorest share of households *within* each community. To assess this possibility, we repeat the targeting evaluation, with the objective of identifying the poorest 21% of households within each neighborhood (or within each sub-neighborhood in settings where large neighborhoods were split due to size). In Figure S1, we show that while the absolute accuracy of each targeting method declines with this evaluation approach (unsurprising, since geographic variation between communities is no longer a useful signal for targeting), the quality of targeting approaches relaive to one another is unchanged: phone-based targeting is still more accurate than CBT, and less accurate than PMT.

Approaches to aggregating information from multiple phones

Our data also allow us to study additional variants of the phone-based targeting approach, in particular for households with multiple phones. While past work has generally used the phone of the household head to infer poverty for phone-based

targeting [5, 6], our detailed census data include phone numbers for all households members. While most (68%) of households in our census own only a single phone, 24% of households provided two phone numbers, and 6% provided three or more phone numbers (Figure S3). We experiment with four approaches to aggregating together poverty predictions from multiple phones within a household: taking the prediction of the most senior household member, taking the maximum prediction, taking the minimum prediction, and taking the mean of all predictions. Figure S4 shows that the aggregation approach has little impact on the overall accuracy of phone-based targeting, with taking the mean of predictions performing slightly better than other options. The small impact is unsurprising, as most households in our survey own only a single phone.

4.3.2 Combining targeting approaches

In this section, we evaluate the accuracy of targeting approaches that combine two or more of our three main targeting approaches (phone-based targeting, PMT, and CBT). We assess two methods for combining pairs of targeting approaches: the first based on combining rankings, and the second based on machine learning.

In the first approach, targeting approaches are combined as follows (taking the example of a combined targeting approach leveraging phone-based and CBT-based poverty rankings): start with full phone-based targeting (identifying the poorest 21% of households based on phone-based poverty predictions); then replace the household included by phone targeting that is ranked richest by CBT with the household ranked poorest by CBT but not included by phone-based targeting. Continue to replace households until the included households are all households that are included by CBT. This approach thus produces a set of combined approaches that represent degrees of merging of phone targeting and CBT (or any other pair of targeting methods).

Figure 4.3 plots the accuracy of these suites of targeting approaches. Overall, combining rankings from different targeting approaches does not appear to substantially increase targeting accuracy. The only slight exception is combining phone rankings with CBT rankings, which improves very slightly upon phone data alone (indicated by the maximum value of the blue line being slightly higher than its intersection with the y-axis).

In the second approach to combining data sources, we construct the combined targeting approach through the same machine learning pipeline as our primary phone-based approach: for example, the method combining phone-based targeting and PMT trains an ML model to predict consumption expenditures from phone features and PMT components; the method combining phone-based targeting and CBT trains an ML model to predict consumption expenditures from phone features and CBT rankings.

In Figure 4.4 we find that the PMT — by far the most accurate targeting approach on its own (AUC = 0.81) — is not meaningfully improved by adding information from phone-based targeting (AUC = 0.81) or CBT (AUC = 0.82). Combining phone-based targeting and CBT, however, improves targeting accuracy (AUC = 0.66) above phone-

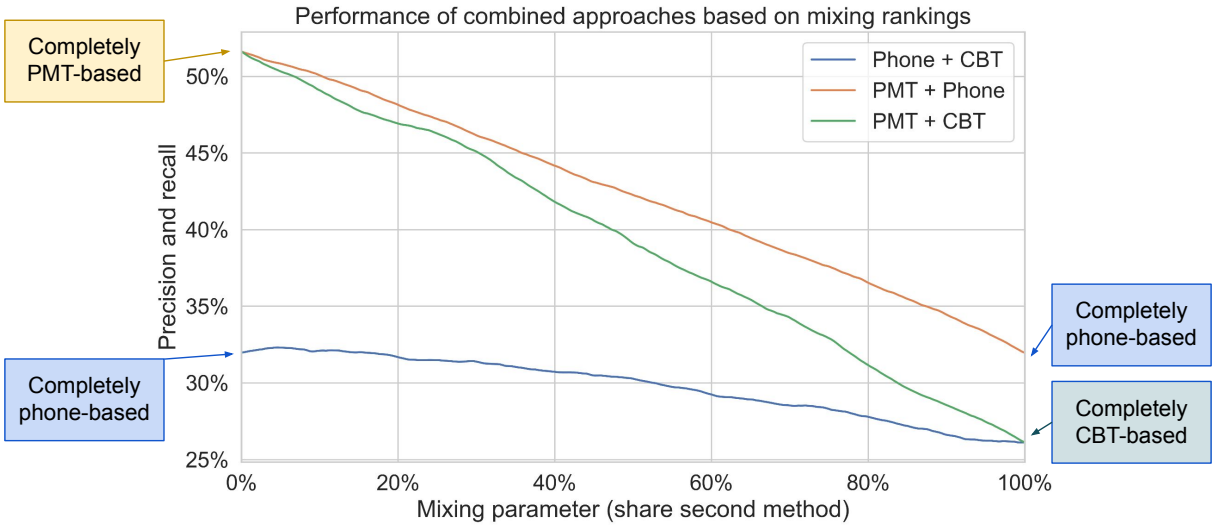

Figure 4.3: Accuracy for the suite of targeting approaches produced by combining two of the three main data sources. The "mixing parameter" on the x-axis shows what share of households identified by the first targeting approach (for example, phone-based for the blue line) have been replaced by households identified by the second targeting approach (for example, CBT for the blue line).

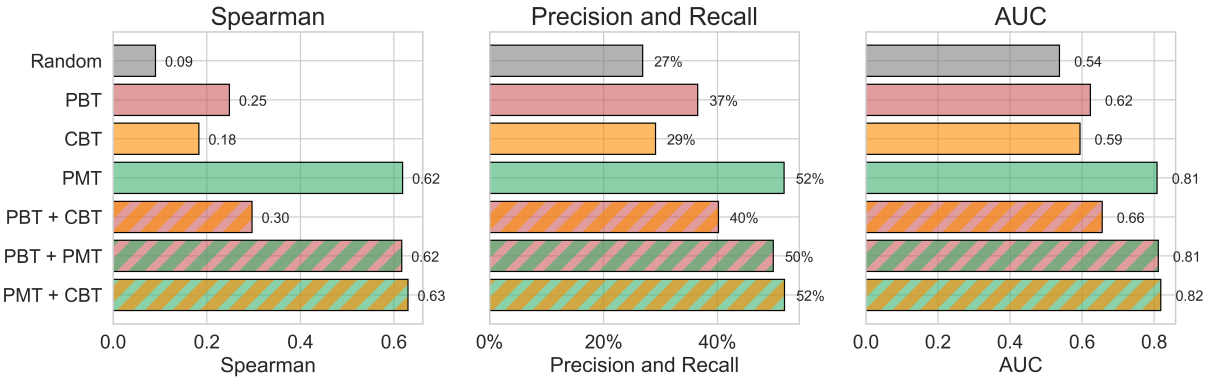

Figure 4.4: Targeting accuracy of methods combining one or more data sources using machine learning, in comparison to targeting using single data sources. Error bars show two standard deviations above and below the mean for each metric.

based targeting alone (AUC = 0.62) or community-based targeting alone (AUC = 0.59).[14] These results are consistent with the lack of improvement seen from combining data sources in the first ranking-based approach.

[14]Due to the time required to train models these results are calculated using just one train-test split from our 100 simulations, which explains why the accuracy results in Figure 4.4 differ slightly from Figure 4.1.

4.3.3 Heterogeneity in targeting accuracy

In our final set of results, we test for heterogeneity in targeting accuracy — measured in this section with Spearman correlation — by characteristics of neighborhoods and households. At the neighborhood level, we split neighborhoods into quartiles by size, share minority households,[15] and social connectedness.[16] We also contrast urban and rural neighborhoods. At the household level, we split households into quartiles by household size and social connectedness. We also compare minority vs. non-minority households and male vs. female-headed households.

The PMT performs best across the board in our heterogeneity analyses. There is nuance, however, in where phone-based targeting outperforms CBT. Phone-based targeting generally outperforms CBT, except in neighborhoods with a large share of minority households, in urban neighborhoods, and for very large households. Interestingly, there is little variation in CBT performance by the minority share, size, and connectedness of a neighborhood.[17]

[15]In this analysis households are considered minorities if they are either non-Bengali or non-Muslim.

[16]Connectedness is measured via the peer rankings module in the survey. Each household was asked, for eight randomly selected households in their neighborhood, how well they know the household on a scale of 1-4. Connectedness at the neighborhood level is defined as the average knowledge ranking for all households in the neighborhood. Connectedness at the household level is defined as the average knowledge ranking of all interviewees that were asked about the household in question.

[17]Table S1 formally tests for heterogeneity in targeting accuracy in a regression specification, and includes additional covariates beyond those shown in Figure 4.5.

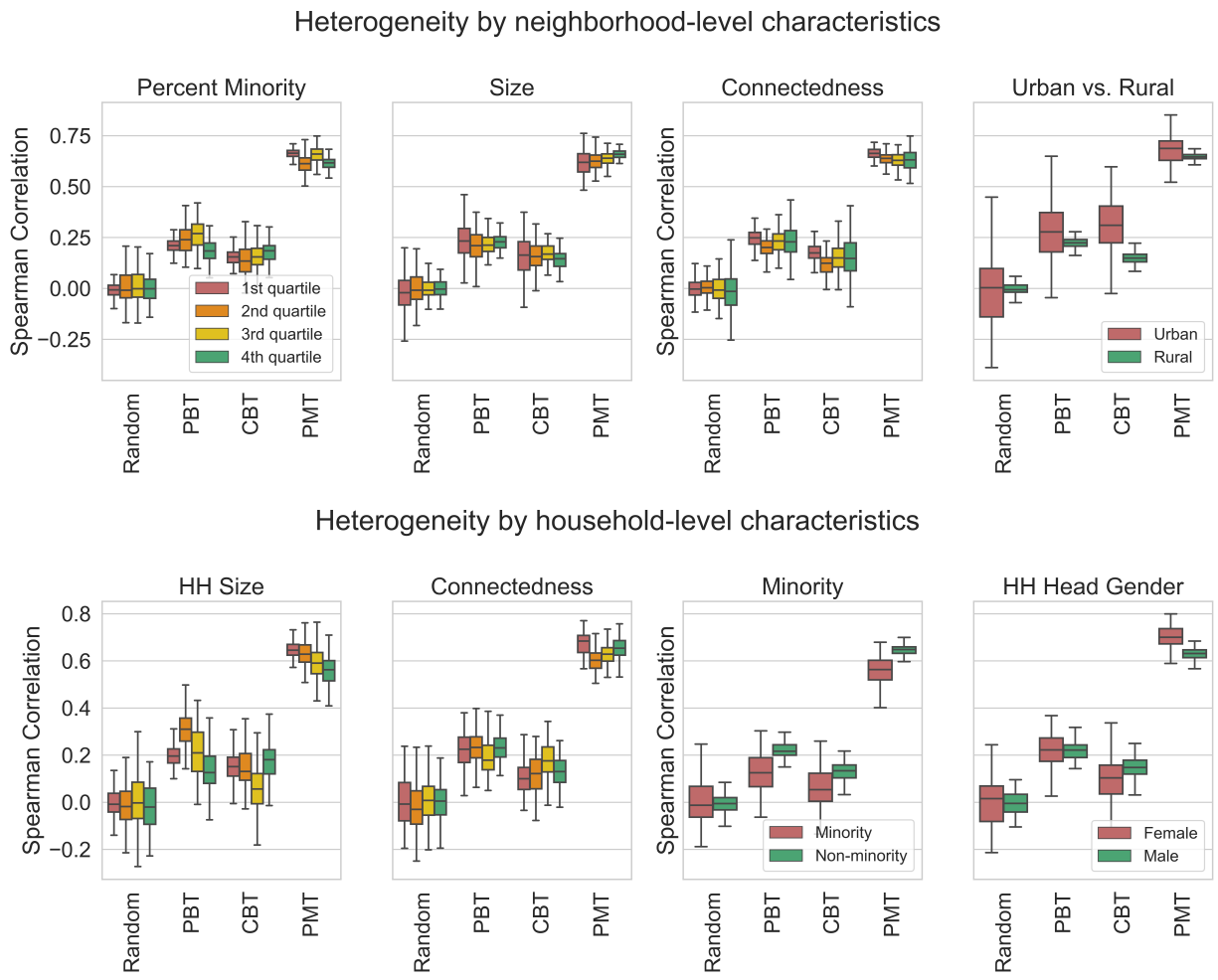

Figure 4.5: Heterogeneity in targeting accuracy by neighborhood-level characteristics (top row) and household-level characteristics (bottom row). Each plot shows the distribution of Spearman correlations (over the 100 random train-test splits) for each group.

4.4 Discussion

In summary, in the Cox's Bazar setting, we find that (1) phone-based targeting is more accurate than community-based targeting, but less accurate than proxy-means testing; (2) combining phone-based targeting and CBT yields a slightly more accurate approach than either method on its own, but combining CBT and PMT does not improve upon the targeting accuracy of the PMT, and (3) while there is heterogeneity across a number of dimensions in the performance of each targeting approach, CBT only outperforms phone-based targeting in urban areas, high minority neighborhoods, and for very large and very small households.

In terms of the accuracy of phone-based targeting, these results are consistent with the work presented in the second chapter of this thesis [5], which found that phone-based targeting is substantially less accurate than PMT. These results are also consistent with past work indicating the PMTs generally outperform CBTs in terms of targeting accuracy [51, 11, 130, 140]. One interesting point of deviation from past work is the comparison of peer rankings to the CBT: while past work has found slightly higher targeting accuracy for CBT than aggregation of peer rankings [153], here we find the opposite relationship.

This chapter abstracts away two important dimensions of targeting that are likely to influence policy choices around which targeting approach is best for a given program: cost and time. First, as described in the discussion section of Chapter 3, screening costs vary widely across programs: the median PMT-targeted program incurs a cost of around $4.00 per household screened, and the median CBT-targeted program incurs a cost of around $2.20 per household screened. Future work might in this space might consider fixed-budget programs in which budget spent on targeting reduces the share of budget available for cash transfers, and identify the welfare-optimizing targeting policy accounting for targeting costs.

The other key missing dimension is time: the PMT and CBT targeting methods studied in this chapter represent up-to-date PMT and CBT data (that is, data that were gathered at the same time as the consumption expenditures data that are used as ground truth). However, the average PMT-targeted social protection program updates its PMT only every 5-8 years [26]. Since poverty is dynamic [30], it is likely that an out-of-date PMT (or CBT) is substantially less accurate than the up-to-date methods studied in this chapter. Building on Hillebrecht et al. (2023) [90], future work might study the decay in accuracy of these targeting approaches as they become out-of-date and identify when targeting based on "big data", including mobile phone data, is more accurate than targeting using traditional PMT or CBT-based approaches.

Chapter 5

Measuring cash transfer impacts with surveys versus digital traces

This chapter is based on the paper "Estimating Impact with Surveys versus Digital Traces: Evidence from Randomized Cash Transfers in Togo" [4], written in collaboration with Suzanne Bellue, Joshua Blumenstock, Dean Karlan, and Christopher Udry.

Abstract

Do non-traditional digital trace data and traditional survey data yield similar estimates of the impact of a cash transfer program? In a randomized controlled trial of Togo's Novissi program, endline survey data indicate positive treatment effects on several welfare outcomes. However, impact estimates based on mobile phone data – processed with machine learning to predict welfare – do not yield similar results. This limitation likely arises from the underlying difficulty of using mobile phone data to predict short-term changes in wellbeing within a homogeneous, rural population. We discuss the implications of these results for using digital data in impact evaluation.

5.1 Introduction and context

This chapter returns to the setting of the second chapter — delivery of *Novissi* cash transfers during the COVID-19 pandemic — to study a new question distinct from targeting: can digital data sources like mobile phone data detect the *impacts* of anti-poverty programs?

Reliable estimates of post-program outcomes are essential to impact evaluation. In low- and middle-income countries (LMICs), such outcomes are traditionally measured

through surveys. However, the new paradigm for estimating living standards in LMICs based on the application of machine learning algorithms to nontraditional data (from mobile phones [e.g. 36, 38], satellites [e.g. 97, 158], and other digital sources [e.g. 68, 143]) opens up new opportunities for low-cost and large-scale impact evaluation. These estimates are attractive because they can be produced rapidly for large populations at a fraction of the cost of traditional surveys.

This chapter asks whether welfare outcomes estimated from "digital trace" data produce the same estimates of program impact as those obtained from traditional survey-based measures of welfare. We study these questions in the context of Togo's *Novissi* program. Recall that the *Novissi* program (Chapter 2) provided five monthly cash transfers of USD $13-15 to poor individuals living in rural Togo during the COVID-19 pandemic. *Novissi* cash transfers were targeted based on poverty predicted from mobile phone data, and they were rolled out in two phases (starting either in late 2020 or the middle of 2021). Through a randomized controlled trial (RCT) eligible individuals were randomly assigned to each phase. We conducted phone surveys between the phases, and obtained the complete mobile phone transaction logs of all consenting program participants.

Our first set of results uses the phone surveys to document the welfare impacts of the *Novissi* program, using traditional methods prespecified in a pre-analysis plan. Transfers increased food security (0.06 standard deviations (SD), standard error (se)=0.02), mental health (0.07 SD, se=0.02), and perceived economic status (0.04 SD, se=0.02). Effects on other outcomes are positive but not statistically significant. The effect on a composite index of welfare is 0.06 SD (se=0.02).

Second, we test whether cash transfers changed phone use. Receiving cash changed a number of aspects of beneficiary's phone use patterns (as measured from mobile phone transaction logs), such as increasing call volume and contact diversity. We examine 824 "features" of mobile phone use, and find that cash transfers statistically significantly impacted 35% of features with $p < 0.05$.

Last, we test whether *predictions* of welfare outcomes based on mobile phone data can be used to identify Novissi's treatment effects on welfare. We first observe that machine learning algorithms applied to phone data can predict a proxy means test (PMT) relatively well (and comparable to prior work), but the same approach is less effective at estimating the welfare measures impacted by the program (i.e., food security, mental health, and perceived economic status). Thus, we find that program impacts estimated based on these phone-predicted measures of welfare differ considerably from those based on surveys.

After presenting these results, we conduct analysis to understand why the phone data did not detect the impacts on food security, mental health, and other vulnerability measures; the analysis highlights several challenges associated with the homogeneity of the study population. We also show that food security and other vulnerability measures in Togo are less geographically concentrated than poverty, making it more difficult to predict them from phone data. Finally, we show that — even if the predictive models had been more accurate — impact evaluation using phone data could still be complicated

by issues of model drift, and by the difficulty of inferring impacts that were modest in magnitude. We conclude with a discussion of how these results can inform the broader conversation around the use of digital data for monitoring and impact evaluation.

This chapter contributes to two main literatures. The first documents the impacts of unconditional cash transfers on a range of welfare outcomes, including expenditures, food security, health, education, savings, and financial inclusion (for reviews, see Bastagli et al. (2016) [29] and Crosta et al. (2023) [54]). Several more recent papers document the welfare impacts of cash transfers distributed in response to the COVID-19 pandemic [22, 110, 102, 40]. Many of these studies are reviewed in Karlan et al. (2022) [102]. Broadly, this literature shows modest, positive, and statistically significant impacts of cash transfers on food security and mental health. The first portion of our analysis contributes to this literature by documenting the impacts of pandemic cash transfers in Togo, using an extensive cash transfer program where treatment was randomly assigned at the individual level. While the cash transfers we study are smaller ($13-15.50 per month) than most of the other programs studied ($15-50 per month), we document comparable effect sizes (0.04-0.07 SD).

The second, more nascent literature explores the use of digital data sources for measuring welfare and evaluating programs and policies. While early work in this space focused on documenting the potential for measuring welfare from mobile phone [36, 38, 5] and satellite data [97, 158, 50], more recent work has begun to ask whether program impacts can be estimated using these data sources. In particular, two recent papers find that estimates of the impact of large-scale development interventions, estimated using satellite imagery, are similar to but noisier than estimates based on surveys [93, 132]. The potential for using mobile phone data for impact evaluation remains unexplored.[1]

5.2 The GiveDirectly-Novissi program: Design and data

The GiveDirectly-Novissi program (GD-Novissi), implemented jointly by the Togolese Ministry of Digital Transformation and GiveDirectly, provided monthly cash transfers to 138,589 individuals in rural parts of Togo between November 2020 and August 2021. Eligible women received 8,620 FCFA (USD $15.50 = $38 PPP) per month, and eligible men received 7,450 FCFA (USD $13 = $33 PPP) per month for five months. GD-Novissi was one of several targeted transfer programs delivered under the Novissi umbrella during the pandemic.

Recall that registration and payment for GD-Novissi were entirely digital. Individuals registered for the program by dialing a toll-free mobile shortcode and filling out a brief USSD form. Registration required (i) a valid voter ID number, (ii) a valid SIM card, and

[1] An exception is ongoing work by Barriga-Cabanillas et al. (2023) [28], which uses a regression discontinuity design to study the impact of a cash transfer program in Haiti. Preliminary results, consistent with our own, suggest that phone-based estimates of food security are too noisy to detect the impact of cash transfers.

(iii) access to a mobile phone.[2] Cash was delivered monthly via mobile money, with mobile money accounts automatically opened for subscribers who did not already have them.

Recall that GD-Novissi used both geographic and poverty-related criteria for eligibility determination. First, beneficiaries had to be registered to vote in one of the 100 poorest cantons in the country. Second, poverty estimates for each registered subscriber were derived from their pre-program mobile phone records; only subscribers estimated to be living on less than \$1.25/day (the poorest 29% of subscribers) were eligible. Refer to Chapter 2 for a full description and evaluation of GD-Novissi's targeting approach and a discussion of the extent to which these eligibility criteria created systematic exclusions from the program.

GD-Novissi was advertised over several channels, including radio advertisements, communication with community leaders, SMS messages, and outreach by field teams. The program launched in November 2021; after three months, GD-Novissi had received 181,028 registrations, of which 49,083 met the eligibility criteria and received benefits.

We implemented a randomized controlled trial (RCT) among these 49,083 individuals who registered for GD-Novissi in its first three months and met the eligibility criteria. Prior to registration, eligible individuals were randomly assigned to treatment (N=27,673) and control (N=21,410) groups. Upon registration, subscribers in the treatment group immediately received the first of their five monthly cash transfers. Subscribers in the control group received the same total benefits as the treatment group, but their payments were distributed beginning in June 2021 and were bundled into three payments (N=21,410). Subscribers in the control group were not informed that they would receive transfers at a later date.

5.2.1 Endline Survey

To evaluate the impact of GD-Novissi cash transfers, we conducted an "endline" phone survey with both treatment and control individuals in May 2021, between zero and two months after members of the treatment group had received their final cash transfer (and before any of the control group subscribers had received a transfer — see Figure S1 for a visualization of the project timeline). Following our pre-analysis plan (American Economic Association Registry #7590), our survey collected information on food security, financial health, financial inclusion, mental health, perceived socioeconomic status, labor supply, health care access, and labor supply, along with a proxy-means test (PMT).[3]

The sample frame for the endline survey was subscribers who enrolled in the GD-Novissi RCT between November 2020 and January 2021 (N=49,083), stratified by treatment status and geography. The geographic stratification was implemented because

[2] A single SIM card could only register one voter ID. In Chapter 2, we estimate that 65% of individuals and 85% of households in Togo owned a mobile phone in 2019, and that 87% of Togolese adults possessed a voter ID.

[3] Table S3 reports the components of each outcome index.

one large region (Savanes) received payments unrelated to GD-Novissi during the period when GD-Novissi benefits were being delivered.[4] The final sample contains 9,511 observations, a completion rate of 39% relative to the 24,294 phone numbers called for the survey. Table S1 shows that attrition does not differ significantly between treatment and control groups.

Appendix D.1 provides more details on the endline survey, and Table S2 provides summary statistics and balance checks for the impact evaluation sample. We observe small and generally statistically insignificant differences in treatment assignment by gender, age, occupation, and place of residence.

5.2.2 Pre-treatment Survey

While our impact evaluation with survey data relies primarily on the endline survey conducted post-treatment, portions of our analysis use a pre-treatment phone survey conducted in September 2020, prior to the roll-out of the GD-Novissi program. This sample frame was defined as all active mobile subscribers whose primary home location was in those 100 poorest cantons, using geographic information available in the mobile phone data (see Chapter 2 and Appendix D.2 for details). In total, we completed 9,484 pre-treatment surveys.

As the primary objective of the pre-treatment survey was to collect PMT data that could be used to train the machine learning algorithms used to identify eligible GD-Novissi beneficiaries (see Chapter 2), it differed from the endline survey in two key respects. First, it was shorter and more focused on the PMT; did not contain a mental health module; and had fewer food security questions (Table S4). Second, the population was designed to be representative of *all* active mobile phone subscribers in Togo's 100 poorest cantons, not just those subscribers predicted to be below the poverty threshold. As shown in the first two columns of Table S2, the pre-treatment sample was still quite poor (average estimated daily per capita consumption of \$1.49, SD = \$0.74), but less homogeneously poor than in the endline survey (average consumption \$1.31, SD = \$0.49).

5.2.3 Mobile Phone Metadata

We obtained comprehensive mobile phone metadata from Togo's two mobile network operators for the duration of the GD-Novissi program. These data include detailed metadata about each phone call and text message sent or received on the mobile networks, including the phone number of the caller and recipient, the timestamp, the duration of calls, and the cell tower through which the call was placed. The data also include mobile data usage, including the phone number of the subscriber, the timestamp, and the amount of mobile data used for each mobile data transaction.[5]

[4] See Appendix D.3.2 for details on the Savanes program.

[5] Although the dataset shared by the mobile network operators also includes records of mobile money use, we do not use mobile money transactions in our main analysis since the treatment itself was delivered

We obtained informed consent from each respondent in the pre-treatment and endline surveys to match their survey responses to their mobile phone records.[6] We then generated sets of mobile phone *features* describing how each survey respondent used their mobile phone in the period preceding the survey. Features were generated using open source library `cider`[7] following the procedure described in Chapter 2. In total, we constructed 824 features relating to calling patterns, contact networks, mobility, location, data usage, international transactions, and more. Features for training models on the pre-treatment survey were generated using six months of mobile phone data preceding the survey (i.e., April - September 2020); features for training models on the endline survey were generated using mobile phone data from the six-month treatment period (November 2020 - April 2021).

5.3 Program impacts estimated using survey data

Our first set of results uses the endline survey to estimate the causal impact of GD-Novissi. These results are based on weighted regressions of each of the seven outcomes on treatment status and include strata, enumerator, and week of the survey fixed effects. To account for multiple hypotheses, we include p-values adjusted for the False Discovery Rate [14] for our seven pre-specified outcome indices.

Results in Table 5.1 Panel A indicate that GD-Novissi increased food security (by 0.06 SD, $p = 0.003$), mental health (by 0.07 SD, $p < 0.001$), and self-perceived socioeconomic status (by 0.04 SD, $p = 0.074$). These results are broadly consistent with studies of the effects of cash transfers during the COVID-19 pandemic in other contexts [22, 110, 102, 40].[8] GD-Novissi does not decrease individual labor supply (the coefficient is positive but not statistically significant, with a point estimate close to zero), consistent with evidence on the effect of cash transfers in other contexts [24, 21]. We observe no statistically significant effects on our indices of financial health, financial inclusion, or healthcare access, although the coefficient estimates are positive.[9]

The last column of Table 5.1 Panel A indicates that GD-Novissi increases an aggregate welfare index by 0.06 standard deviations ($p = 0.008$), where the aggregate index is constructed as an aggregated normalized index of the seven underlying outcome indices. The first column of Table 5.1 indicates no statistically significant impact on the proxy

via mobile money and thus mechanically (and dramatically) changed mobile money usage patterns for the treatment group. However, we explore the inclusion of mobile money data in Section 5.5.1.

[6]Following the data protection procedures described in our IRB protocol, we pseudonymized or removed all personally identifying information, including phone numbers, prior to linking these two datasets.

[7]https://global-policy-lab.github.io/cider-documentation/

[8]Appendix D.4 compares our survey-based results to impact evaluations of cash transfers in other settings during the COVID-19 pandemic.

[9]Our financial inclusion index measures the fraction of bank accounts and mobile money usage in households, excluding mobile money accounts of the respondents.

Table 5.1: Survey-based and phone-based treatment effects of the GD-Novissi program

	(1) PMT	(2) Food security	(3) Financial health	(4) Financial inclusion	(5) Mental health	(6) Perceived status	(7) Healthcare access	(8) Labor supply	(9) All seven indices
Panel A: Survey-based treatment effects									
Treatment	0.002	0.064***	0.026	0.007	0.072***	0.040*	0.010	0.009	0.061***
	(0.012)	(0.022)	(0.024)	(0.021)	(0.019)	(0.022)	(0.023)	(0.025)	(0.023)
Obs.	8,452	9,511	9,511	9,511	9,511	9,511	9,511	9,511	9,511
Control mean	1.31	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FDR q-value		0.014	0.428	0.732	0.002	0.149	0.732	0.732	0.021
Panel B: Predicting welfare outcomes using ML trained on pre-treatment survey									
R^2	0.143	0.002	0.013	—	—	0.034	—	0.046	—
Obs.	8,899	8,899	8,899	—	—	8,899	—	8,890	—
Panel C: Predicting welfare outcomes using ML trained on endline survey									
R^2	0.049	0.008	0.009	0.003	0.002	0.008	0.007	0.021	0.026
Obs.	8,448	9,507	9,507	9,134	9,507	9,507	9,522	9,507	9,507
Panel D: Phone-based treatment effects trained on the pre-treatment survey									
Treatment	-0.007***	-0.003	-0.013	—	—	-0.013	—	-0.000	—
	(0.002)	(0.016)	(0.014)	—	—	(0.013)	—	(0.012)	—
Obs.	48,759	48,759	48,759	—	—	48,759	—	48,759	—
Control Mean	1.409	0.000	0.000	—	—	0.000	—	0.000	—
Z-test p-value	0.731	0.016	0.159	—	—	0.483	—	0.002	—
Panel E: Phone-based treatment effects trained on the endline survey									
Treatment	-0.001	0.005	0.001	0.004	0.028*	0.001	0.021*	0.011	0.015
	(0.002)	(0.013)	(0.018)	(0.020)	(0.017)	(0.017)	(0.015)	(0.013)	(0.015)
Obs.	48,759	48,759	48,759	48,759	48,759	48,759	48,759	48,759	48,759
Control Mean	1.314	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Z-test p-value	0.800	0.021	0.404	0.910	0.088	0.162	0.677	0.955	0.100

Notes: Panel A shows treatment effects of GD-Novissi estimated using the endline survey. The dependent variable for each regression is indicated in the column title; see Appendix D.1 for variable construction. All regressions control for the enumerator, week of the survey, and strata fixed effects. All observations are weighted by sampling and response probabilities. Panels B and C show the performance of the machine learning models used to generate the predictions of welfare outcomes used to generate phone-based estimates of the treatment effects, measured by R^2 score (evaluated out-of-sample over five fold cross validation on the training set).Panels D and E report the treatment effects of GD-Novissi derived using the phone-based machine learning model's predictions. In Panel D, the pre-treatment survey is used to train the machine learning model; in Panel E, the endline survey is used to train the model. In Panels D and E treatment effects are estimated across all subscribers enrolled in the RCT by regressing the phone-based estimate of the outcome variable on treatment status, with standard errors determined with a Bayesian bootstrap procedure. The Z-test p-values in Panels D and E indicate the significance of the Z-test that the phone-based treatment effect and the survey-based treatment effect reported in Panel A are different. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

means test (PMT) measure of wealth. This is perhaps unsurprising given that the PMT is based on factors related to long-term poverty (see Chapter 2), which we do not expect would be influenced by a relatively modest cash transfer.

In Appendix D.3, we test for treatment effect heterogeneity on four pre-registered dimensions: gender, poverty, occupation, and region of residence (in or outside Togo's

northernmost region, Savanes). Treatment effects are not heterogeneous across any of these dimensions except for geography: treatment effects on food security and mental health were significantly larger for beneficiaries in the Savanes region in the far North of Togo (Figure S2 and Table S6).

Cash transfers were targeted and delivered at the individual level, but many of our outcomes are measured at the household level (Table S3), introducing two threats for our impact evaluation: (i) control individuals living in households with other members receiving GD-Novissi (which would downward-bias our estimates of treatment effects), and (ii) treated households with multiple members receiving GD-Novissi (which would upward-bias our estimates). 20% of treated individuals and 13% of control individuals in our survey reported that another member of their household received GD-Novissi. Restricting the impact evaluation analysis to the sample of individuals in single-transfer households yields estimates of impact similar in magnitude to our main results: food security increases by 0.06 SD, mental health by 0.07 SD, perceived economic status by 0.04 SD, and the composite welfare index by 0.06 SD.

5.4 Program impacts estimated using mobile phone data

Here we present our main, null result: although Novissi's treatment effects on both survey-based measures of welfare and several dimensions of phone use were positive and statistically significant, when using mobile phone data to estimate the welfare treatment effects of GD-Novissi, most estimates are close to zero and statistically insignificant.

An advantage of mobile phone data is that, in principle, they could help predict outcomes for a very large population (i.e., the full population of beneficiaries and non-beneficiaries with phones), using only a small sample survey to train the prediction model. If successful, such an approach could enable new paradigms for more rapid and lower cost impact evaluation using digital data. However, we find that in this context, mobile phone data do not capture the same treatment effects estimated with survey data.

5.4.1 Cash Impacts on Phone Use

Prior to estimating treatment effects on measures of welfare *predicted* from mobile phone data, we first measure the impacts of cash transfers on phone use itself. This analysis relies on the mobile phone transaction logs of all consenting RCT participants. We find that, across the 824 different metrics of phone use (calculated from phone transactions between November 2020 and April 2021, during the treatment group's transfers), there are statistically significant differences ($p < 0.05$) between treatment and control for 35% of metrics. Table 5.2 provides the standardized impacts of cash transfers on several easily interpretable dimensions of phone use: the cash transfer treatment increases calls by 0.02 SD (se = 0.009), contacts by 0.03 SD (se = 0.009), active days by 0.06 SD (se = 0.009), and unique prefectures (admin-2 units) visited by 0.04 SD (se = 0.009).

Table 5.2: Treatment Effects on Phone Use

	(1) Active days	(2) Calls	(3) Texts	(4) Contacts	(5) International Contacts	(6) % Initiated	(7) Regions	(8) Prefectures
Treatment	0.064***	0.021**	0.010	0.033***	0.015	-0.026***	0.049***	0.038***
	(0.009)	(0.009)	(0.010)	(0.009)	0.015 (0.013)	(0.010)	(0.009)	(0.009)
Control Mean	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Unstandardized Control Mean	98.892	625.654	70.979	77.098	3.164	0.902	4.236	9.867
Obs.	48,803	48,664	44,548	48664	22,410	44,548	48,803	48,803

Notes: Treatment effects on basic metrics of mobile phone use, selected from among the 823 metrics of mobile phone use used by our machine learning models. Metrics were selected by hand from the pool based on ease of interpretation: (1) active days of phone use, (2) total incoming and outgoing calls, (3) total incoming and outgoing texts, (4) unique contacts, (5) unique international contacts, (6) share of the individual's transactions initiated by them (rather than received from a contact), (7) unique regions visited (based on locations of mobile antennas), and (8) unique prefectures visited (based on locations of mobile antennas). All features are calculated over the entire six month treatment period (November 2020 - April 2022). All features are standardized to zero mean and unit variance in the control group (the unstandardized control mean is also provided for intuition). $^{*}p$ <0.1; $^{**}p$ <0.05; $^{***}p$ <0.01.

The cash transfers most consistently impact dimensions of phone use relating to calling: 55% of the 313 features related to calling patterns (such as number and duration of calls and diversity of call contacts) differ significantly between the treatment and control groups at the 0.05 level. Only 11% of text-related features are significantly different, 17% of mobile data usage-related features, and 38% of features related to mobility and location. No features relating to international transactions differ statistically significantly between treatment and control groups. The largest (in magnitude) treatment effects are on the share of contacts that make up 80% of an individual's calling time (-0.07 SD, se = 0.009), the share of contacts that make up 80% of an individual's calls (-0.07 SD, se = 0.009), and the number of unique days an individual uses their mobile phone (0.06 SD, se = 0.009).

5.4.2 Setup for Estimating Welfare Impacts from Phone Data

We now test whether welfare impacts can be estimated from mobile phone data. We test this approach under two different regimes. In the first regime, we assume that the only opportunity for survey data collection occurs before treatment is administered. In this scenario, pre-treatment survey data (collected in September 2020) are matched to pre-treatment phone data (March - September 2020), and machine learning is used to predict socioeconomic status from phone data.[10] Then, after the program has been implemented, an impact evaluation is conducted by comparing the *predicted* post-program outcomes of

[10]This is the same machine learning procedure that the GD-Novissi program used to determine program eligibility, as described by [5].

treated and control individuals, where the predicted outcomes are generated by applying the trained model (trained on pre-treatment data) to phone data collected post-program.[11]

In the second regime, we instead assume that the only opportunity to collect survey data is after administering treatment. In this approach, the machine learning algorithm is trained on post-treatment data (i.e., endline survey data from May 2021 that are matched to post-treatment phone data from November 2020 - April 2021), for a small sample of the actual beneficiary population. The trained model is then used to predict outcomes for all subscribers enrolled in the RCT, including those not surveyed.

5.4.3 Predicting Welfare Levels from Phone Data

We begin by testing the extent to which machine learning models can accurately predict individual welfare, separately in the first regime of analysis (using pre-treatment surveys) and the second regime (using endline surveys).

Panel B of Table 5.1 reports the accuracy with which pre-treatment survey outcomes can be predicted from mobile phone data. Specifically, for each outcome index captured in the pre-treatment survey, we calculate the five-fold cross-validated R^2 as follows. The full dataset that matches completed surveys to phone records (N=8,899) is divided randomly into five partitions ("folds"). A machine learning model is trained on four of the five folds and predictions are produced for observations in the remaining fold; the process is repeated for each of the remaining four folds. The percentage of variation explained by the predictions (R^2) is then calculated, pooling predictions across all the folds.[12] We use a gradient boosting model, with hyperparameters tuned using nested cross-validation on each fold.

Predictive accuracy is highest for the PMT (R^2 = 0.143, or a pearson correlation coefficient of $r = 0.381$) — a finding that replicates results in Chapter 2. We find that we cannot accurately predict any of the other welfare indices (food security, financial health, perceived status, and labor supply) from mobile phone features (R^2 = 0.002 - 0.046).[13]

Panel C of Table 5.1 shows results from analogous experiments, where the machine learning model is instead trained using the endline survey. With this survey, predictive accuracy for the PMT is lower (R^2 = 0.049, or a pearson correlation coefficient of 0.253). This difference is likely due in part to the more homogeneous population represented in the endline survey: while both surveys focus on the same rural areas, the endline survey was also restricted to *program-eligible* mobile subscribers, where eligibility was determined by predicted poverty (Section 5.2.1). We observe little predictive accuracy for

[11]This regime is similar in spirit to that used in the impact evaluations based on satellite imagery explored in [93] and [**ratledge2021using**].

[12]Survey weights and response weights are used in both training and calculating R^2 scores.

[13]We cannot include results on financial inclusion, mental health, healthcare access, nor the seven-index composite from Table 5.1, as the questions required to construct these indices were not included in the pre-treatment survey.

food security, mental health, perceived socioeconomic status, or any of the other welfare indices (R^2 = 0.002 - 0.026).

5.4.4 Estimating Treatment Effects from Phone Data

Our next set of tests investigates whether the phone-based estimates of welfare described in Section 5.4.3 can be used to estimate the treatment effects of GD-Novissi. In the first regime of analysis — where the only survey data available are collected pre-treatment — we train the machine learning model on the pre-treatment survey (as evaluated in Panel B of Table 5.1). We then compare *predicted endline outcomes* of treated and control individuals, where predicted endline outcomes are generated by passing post-treatment mobile phone data through the prediction model that was trained pre-treatment.[14]

Panel D of Table 5.1 shows the predicted average treatment effect of GD-Novissi, which is obtained by regressing the predicted outcome on the individual's treatment status. To estimate the variance of treatment effects derived from these phone-based predictions of welfare, we use a Bayesian bootstrap procedure [136] to incorporate both the first-stage uncertainty in our ML models' predictions and the variation in predictions across treatment and control subscribers.[15] The treatment effect is not statistically significant for any of the welfare indices besides the PMT, where it is negative. For food security and perceived socioeconomic status — which both had significant treatment effects in the survey — the point estimates of phone-based treatment effects are -0.003 and -0.013, respectively.

Panel E of Table 5.1 shows the results from the second regime where only endline survey data are available to train the machine learning model. Here, the models are trained using endline data (as evaluated in Panel C of Table 5.1); the models are then used to generate predicted welfare outcomes for all subscribers using post-treatment mobile phone data.[16] We estimate the predicted average treatment effects as before by comparing predictions between the treatment and control groups, and estimates of variance are again produced with a Bayesian bootstrap. We do not estimate statistically significant treatment effects on food security, financial inclusion, perceived socioeconomic status,

[14]Specifically, one model corresponding to each welfare outcome is trained using the pre-treatment survey and phone data (from March - September 2020), with hyperparameters tuned through 5-fold cross-validation specific to that outcome. Each model is then used to generate predicted welfare outcomes for treated and control individuals, using phone data from the treatment period (November 2020 - April 2021).

[15]Following Angrist et al. (2017) [15] and Dobbie and Song (2020) [61], we assign each observation that appears in the training and/or inference sets a "bootstrap weight" drawn from a Dirichlet distribution Dirichlet(1, ..., 1). These weights are used (in combination with the survey and response weights) in training the ML model, and in calculating treatment effects. We repeat this procedure with 100 random draws, and report the mean and standard deviation across these 100 bootstrap estimates.

[16]These subscribers include the surveyed subscribers on which the model is trained; results are unchanged if we restrict the inference set to subscribers who were not surveyed and, therefore, not used in training.

or the combined index; phone-based point estimates for these treatment effects range from 0.001 to 0.015 standard deviations. There is a positive and statistically significant treatment effect for the mental health index (0.028 SD, p = 0.053) and healthcare access index (0.021 SD, p = 0.076).

5.4.5 Additional Tests of Robustness

In Section 5.5, we examine several reasons why the positive treatment effects estimated using survey data were, in general, not observed in predictions generated with phone data. First, however, we present a few tests to ensure that the preceding results are robust to different variations of the machine learning methodology used to generate predicted treatment effects.

First, results are unchanged if we vary the duration of phone records used to generate welfare predictions. This experiment addresses the possibility that phone use may be most impacted by cash immediately following the transfer. For this test, we train and evaluate prediction models that use only two weeks of mobile phone data (instead of the six months used in our main analysis). When matching to the pre-treatment survey, we use mobile phone data from the two weeks during which the survey was conducted (September 17-30, 2020); for the post-treatment period, we use data from the two weeks immediately following the date on which each individual registered for GD-Novissi and received their first transfer. In Table S7, we do not observe improved predictive performance relative to Table 5.1 (R^2 = -0.002 - 0.112), and treatment effects are similar in magnitude and significance.

Second, we test whether using *changes* in mobile phone use between the pre-treatment period and the post-treatment period can improve predictive performance. Table S8 shows that using changes does not improve predictive performance for our machine learning models (R^2 = -0.004-0.036), and treatment effects remain insignificant.

Third, we test whether impact estimates based on mobile phone data are significant on subsets of the population where the survey-based treatment effects were largest. In particular, as discussed in Appendix D.3, the survey data indicate that treatment effects are larger for beneficiaries in the Savanes region in the north of Togo. However, Table S9 shows that when the machine learning model is trained using data from survey respondents in Savanes, and then evaluated on treatment and control individuals only in Savanes, predictive power remains low (R^2 = -0.012 - 0.120) the treatment effects estimated from the phone data remain statistically insignificant.

Finally, we run several tests of the machine learning models themselves. In addition to tuning the hyperparameters as described above, we take additional steps to ensure that data are not too sparse for the models being used [33]. Specifically, we introduce a feature selection step prior to model fitting, which eliminates all features that are not statistically significantly different between the treatment and control groups. Despite the substantial share of features that differ systematically between treatment and control subscribers

(Section 5.4.1), this feature selection step does not improve predictive accuracy (Table S10; R^2 = 0.000 - 0.139) or impact the significance of treatment effects.

5.5 Discussion

To summarize our main results, we find that (i) GD-Novissi cash transfers had positive and statistically significant impacts on food security, financial inclusion, mental health, perceived socioeconomic status, and an aggregate outcomes index in the endline survey; (ii) GD-Novissi transfers statistically significantly impacted many dimensions of mobile phone use, particularly around calling patterns and volume, but (iii) the welfare effects of GD-Novissi estimated using mobile phone data are not statistically significant, likely due to little predictive power of ML models to estimate welfare outcomes from mobile phone use.

The first result is broadly consistent with several studies finding positive impacts of cash transfers on food security and mental health during the COVID-19 pandemic [22, 40, 110, 102]. In comparison to other papers on COVID-19 cash transfers, the GD-Novissi transfer size is slightly smaller (monthly transfers USD 13-15.5 compared to USD 15-52 in other studies). However, effect sizes are of a similar magnitude to those observed in other studies.

The subsequent results are more nuanced and inform a rapidly evolving debate about if and how new digital data sources can be used to inform development research and policy. Where several recent studies have shown that phone data and machine learning can produce accurate estimates of consumption and asset-based wealth, we find that — at least in the rural Togolese context — a similar procedure does not produce reliable estimates of food security, mental health, or self-perceived economic status.

5.5.1 Challenges to estimating welfare from mobile phone data

Here, we explore four hypotheses that could explain why mobile phone data and machine learning can accurately predict wealth, but do not accurately predict food security or the other self-reported welfare outcomes.

Noise in survey data

Survey-based measures of food security and other vulnerability outcomes may be noisier than survey-based measures of economic poverty, and, therefore, more difficult to estimate from any underlying data source [91, 149]. However, the survey-based indices have sufficiently low measurement error that we estimate statistically significant survey-based treatment effects on a number of outcomes from the GD-Novissi cash transfers (Table 5.1). We further experiment with training classification models to predict an outcome with little to no measurement error: *treatment status*.

The results in Table S11 indicate that mobile phone data accurately predict GD-Novissi treatment status in neither the country as a whole (AUC = 0.515 - 0.522) nor the Savanes region specifically (AUC = 0.516 - 0.518). This result suggests measurement error in the survey is not the main reason for the null results in Table 5.1.[17]

Population homogeneity

A second possible explanation for the low predictive power of the ML models for non-PMT outcomes is the homogeneity of the study sample. Past work on predicting poverty from phone data has identified variation across an entire country or large regions [5, 36, 38]. In comparison, we seek to identify variation in poverty and vulnerability measures within the homogeneous subset of individuals identified to be living in poverty within Togo's poorest 100 cantons. In past work that has compared the accuracy of predicting poverty from mobile phone data in full-country evaluations vs. in rural areas only, predictive power is typically substantially lower when restricting to rural areas (r = 0.46 vs. 0.31 in Togo for poverty prediction at the individual level [5] and r = 0.64 vs. 0.50 in Rwanda for poverty prediction at the district level [36]). However, while the homogeneity of the population in our study helps explain why predictive power for the PMT is lower than in previous papers that evaluate nationally representative samples, it does not explain why predictive accuracy for vulnerability indices is substantially lower than for the PMT.

Relationship between phone use and vulnerability

A third hypothesis for why we are unable to predict any of our vulnerability indices from mobile phone data (when we are, to some extent, able to predict poverty) is that mobile phone use may be more closely related to long-term poverty outcomes than to short-term vulnerability metrics. For example, mobile money and mobile data usage are important predictors of wealth in Chapter 2, and are related to long-term investments in smartphones and financial services technologies. Short-term changes in food security, financial health, and mental health may not result in the types of investments in phone capabilities (such as buying a new smartphone or investing in a large airtime bundle) that would be clearly observable from mobile phone metadata. On the other hand, prior work has shown that phone use changes in response to short-term shocks [18, 39];

[17] To further confirm this result — and to test the validity of our pipelines for machine learning with mobile phone data — we replicate the experiment of predicting treatment status from six months of mobile phone data during the treatment period, this time including 'cheat code' features relating to mobile money use in the machine learning model. These features include information on the number and sizes of transactions placed and received by each subscriber, and thus directly reveal information about whether a subscriber has received a GD-Novissi cash transfer via mobile money. With the mobile money-related features included, the area under the curve score for predicting treatment status is 0.998. Table S13, which shows the feature importances for this machine learning model, further confirms that the key features used by the model relate to mobile money transactions.

we might therefore expect that other signals in the phone data, such as the timing or volume of outgoing calls, would reflect short-term changes in welfare. We cannot directly differentiate between these hypotheses in our context, but believe it is an exciting area for future work.

Spatial structure in outcome indices

A fourth and final hypothesis is that poverty may have more geographic structure than food security or the other vulnerability outcomes we examine. Spatial features obtained from the locations of cell towers through which subscribers place calls are critical features in the phone-based poverty prediction models described in Chapter 2. It is possible that food security and other vulnerability indices are less predictable from phone data because they are less related to geographic information. To test for whether the spatial structure could explain the difference in predictive power between the PMT and other outcomes, in Table S12, we calculate the within and between variance grouped by canton for both the pre-treatment and endline survey. We find that the ratio of between to within variance is substantially higher for the PMT (8.2 - 15.0 in the pre-treatment and endline surveys) than for any of the vulnerability indices (1.3 - 2.1). This result, combined with past documentation that spatial structure plays a key role in estimating poverty from mobile phone data [5, 89], suggests that spatial structure in an index may be an important determinant of whether it can be predicted from mobile phone data.

5.5.2 Additional challenges to estimating treatment effects from mobile phone data

Even if it were possible to accurately predict welfare outcomes from phone data, it might still prove difficult to use phone data to estimate the treatment effects of cash transfers on those same outcomes. Here, we provide suggestive evidence of two such issues: that the modest size of the GD-Novissi cash transfers generates only small changes in phone use, and that model drift in the relationship between phone use and vulnerability may complicate the repeated use of machine learning models over time.

Magnitude of impacts

A challenge for detecting treatment effects from mobile phone data in the context of Novissi is the program's modest transfer sizes and welfare impacts. Our survey-based impact evaluation results detect treatment effects of 0.04-0.07 standard deviations resulting from five monthly transfers of USD 13-15. Interventions of a larger magnitude would be expected to produce larger impacts (for example, Haushofer et al. (2016) [86] report a 0.26 SD increase in food security and mental health following a USD 404-1,525 PPP cash transfer in Kenya). The modest transfer sizes and impacts of the GD-Novissi program result in modest impacts on phone use (Table 5.2), which are difficult for an ML

model to detect. Cash transfer effects on welfare may be easier to recover from mobile phone data in the context of larger transfers.

Model drift

A specific challenge to identifying treatment effects in the first regime we study — training a model prior to program roll-out and deploying it later on to monitor impacts — is *model drift* in the relationship between phone use and vulnerability over time. Particularly in the context of shocks like the COVID-19 pandemic, a model trained well before a program's implementation may no longer be accurate when cash transfers are distributed. In Chapter 2 we empirically studied model drift in Togo, finding a substantial drop in accuracy when a model is trained two years prior to its deployment (Spearman correlation of 0.42 at the time of training vs. 0.35 at the time of deployment). To test the extent to which the same issues of model drift are present in this chapter, we evaluate the accuracy of predictions from our poverty prediction model trained on the pre-treatment survey for generating predictions using mobile phone data from the treatment period. In comparison to the R^2 score of 0.049 for the model trained on the endline survey, the predictions from the model trained on the pre-treatment survey achieve an R^2 of only 0.030, providing suggestive evidence of model drift in the nine months that elapsed between the pre-treatment and endline surveys.

5.5.3 Conclusion

In summary, our results show that in a context where survey data indicate small but statistically significant cash transfer impacts on several welfare outcomes and dimensions of phone use, estimates of welfare impacts derived using machine learning and mobile phone records are a tightly-estimated null. These results suggest that machine learning predictions of welfare derived from digital sources — while effective for estimating regional poverty [36] and targeting policies [5] — cannot naively replace traditional survey-based measurements in program monitoring and impact evaluation. Mobile phone data may perform better when treatment effects are larger, when there is more heterogeneity in key outcomes in the beneficiary population, and when phone-based estimates of outcomes are more accurate.

Chapter 6

Discussion

The chapters in this thesis introduce a new tool for the targeting of social protection programs based on machine learning and the analysis of digital data sources, particularly mobile phone data and satellite imagery. Chapters 2, 3, and 4 evaluate the accuracy of phone-based targeting (inferring poverty from metadata on mobile phone) for identifying households living in poverty in Togo, Afghanistan, and Bangladesh. Taken together, the results presented in these chapters suggest that targeting based on poverty inferred from mobile phone data is more accurate than geographic targeting or community-based targeting, but less accurate than targeting approaches that rely on high quality and up-to-date in-person household surveys, like proxy means tests (PMTs). The analysis presented in Chapter 5 builds on these cross-context results on targeting accuracy to assess whether poverty inferred from phone data can measure the *impacts* of targeted anti-poverty programs.

The results in this thesis have a number of policy implications that could inform the design of social protection and humanitarian aid delivery systems. The primary policy contribution is the introduction of phone-based targeting as a new tool in the social protection administrators "toolkit" for determining aid eligibility.[1] Phone-based targeting is most relevant in settings where standard administrative or survey-based registries of household poverty are unavailable or out-of-date, and where primary survey data collection is prohibitively expensive or logistically infeasible (particularly in instances of conflict, pandemics, natural disasters, and migration). In comparison to traditional targeting approaches like survey-based and community-based targeting, phone-based targeting has the advantage of rapid deployment at scale for a near-zero marginal cost of data collection (see Chapter 3 for a more detailed discussion of deployment costs).

The chapters of this thesis have also addressed a number of policy-relevant limitations of phone-based targeting. These include digital exclusion of households without mobile phones, privacy, algorithmic fairness, transparency of algorithmic eligibility criteria, and

[1] Satellite-based geographic "micro-targeting", as introduced in Chapter 2 and evaluated in more detail in Smythe and Blumenstock (2022) [146] could provide a related approach to fine-grained selection of geographies eligible aid based on passively collected remote sensing data.

the potential for manipulation and strategic behavior to "game" the decision threshold. Each of these concerns is discussed in more detail in Chapter 2, section 2.4. This thesis has also addressed data-centric limitations of phone-based targeting, including challenges to combining phone data and other data sources (Chapters 3 and 4) and limitations in the ability of phone data to measure changes in poverty over time (Chapter 5).

6.1 Directions for future work

Certain abstractions in this thesis limit the generalization of the results presented here to all policy scenarios facing the administrators of real-world social protection and humanitarian aid programs. Many of these abstractions represent interesting research directions that could provide evidence useful to policymakers in these settings. Most saliently, the primary metric of targeting success used in this thesis is accuracy (measured by exclusion errors, inclusion errors, precision, recall, area under the curve scores, and related metrics). However, as covered in the introduction (section 1.2), there are a number of other measures relevant to targeting success, including cost, speed, ability to adapt to poverty dynamics, acceptability to beneficiary communities, and impacts on poverty. Sections 6.1.1, 6.1.2, 6.1.3, and 6.1.4 explore potential future research directions in each of these areas. A further limitation of this thesis is that most of the machine learning aspects rely on standard approaches to supervised learning. Section 6.1.5 speculates at how newer advances in machine learning — particularly related to deep learning and optimal sampling — may improve the accuracy of ML-based poverty targeting.

6.1.1 Time: The role of poverty dynamics, distribution shift, and adaptivity

Much of this thesis — like much of the literature on targeting aid more broadly — has treated poverty as a static status, measured at at single point in time. However, in settings where household poverty is likely to change from month-to-month — particularly in the growing number of areas affected by climate shocks, conflict, and migration — poverty targeting approaches will need to adapt to changing conditions. The rich literature on poverty dynamics has documented that households move in and out poverty (as determined by consumption expenditures) on a fairly regular basis in low-income settings: for example, Baulch et al. (2000) [30] find that across panel studies in eight countries, the poverty status of 20-66% of households changes between survey waves.

The dynamic nature of poverty has several implications for poverty targeting that would be fruitful areas for future research. First, targeting poverty dynamics necessitates a shift towards *adaptive* targeting systems that identifying households moving in our out of poverty or experiencing economics shocks, rather than focusing on identifying households living in poverty at a single point in time. Traditional survey-based approaches are likely not a good fit for adaptive targeting, as surveys are typically

conducted infrequently. Real-time digital data sources like satellite imagery and mobile phone data are a more promising option for adaptive social protection, but the degree to which they can measure economic shocks or long-term poverty transitions is unknown. Future work could study panel data sources, matched to mobile phone data and satellite traces, to measure the extent to which these digital data streams can capture changes in poverty over time and target households experiencing economic shocks.

A second implication of the dynamic nature of poverty for aid targeting is the importance of temporal distribution shifts in ML-based poverty targeting models. Temporal distribution shifts have implications for both ML algorithms for poverty prediction based on survey data (like proxy-means tests) and those based on digital data (like phone-based targeting). PMTs provide a useful illustration of the challenges of distribution shift: The vast majority of the literature studying the performance of PMTs (including the chapters in this thesis) evaluate their targeting accuracy at a single point in time — the moment when the PMT data are collected and the PMT decision rule is implemented — which is also when the performance of the PMT is highest. In practice, most PMT-based poverty registries are updated infrequently: while many social protection administrations aspire to update the social registry and ML model regularly (e.g. every two years in Costa Rica; every three years in Colombia, Indonesia, and Mexico [26]), in reality updates typically occur roughly every 5-8 years [26, 96]. In the time between when the PMT data are collected and when policy decisions are made based on those data, the living conditions and poverty status of households may change, resulting in "model decay" in the extent to which the ML model reflects the up-to-date relationship between PMT covariates and poverty, and "data decay" in the extent to which the PMT covariates reflect the on-the-ground reality for households.

Future research could build on initial attempts [41, 90] to quantify the impacts of temporal distribution shifts on PMT accuracy. In particular, given the results in this thesis on the accuracy of phone-based targeting approaches, a question of substantial policy relevance is the length of time before the accuracy of an out-of-date PMT sinks below the accuracy of up-to-date phone based targeting. In all likelihood it is also the case that ML-based targeting methods that rely on digital data — including mobile phone data and satellite imagery — also suffer from temporal accuracy decay due to distribution shifts, so future work could also explore the extent of this decay and policy implications for how often such methods should be "refreshed". Finally, future research could assess new machine learning methods for domain adaptation and domain generalization [157, 16, 137] to design ML models that mitigate the accuracy harms of distribution shift in the poverty targeting setting.

6.1.2 Cost: Options for cost-benefit analysis of targeting approaches

A second aspect of poverty targeting not thoroughly explored in this thesis is the cost implications of different targeting approaches. An advantage of phone-based targeting and other approaches using passively collected digital data sources is that the marginal

cost of screening an additional household is close to zero (the fixed costs in terms of data collection and model calibration, however, are likely to be substantial). In comparison, and as summarized in Chapter 3, traditional survey-based and community-based targeting approaches have higher marginal costs per household screened (median costs in the literature of $4.00 and $2.20 for proxy-means testing and community-based targeting, respectively).

In spite of these substantial differences in costs, existing empirical studies have not typically incorporated screening costs in assessments of trade-offs between targeting approaches. In settings where targeting costs and benefits come from a single fixed budget for a social protection program — so every dollar spent on targeting is a dollar less that goes to benefits — it would be possible to identify the welfare-maximizing targeting approach based on utility functions relating household income or consumption expenditures to expected utility derived from a transfer [85]. With detailed data on costs of phone-based and alternative targeting approaches, future work adapt the social welfare framework introduced in Chapter 2 of this thesis to incorporate targeting costs, and identify the welfare-maximizing targeting approach for aid programs of different budgets. This line of research would provide guidance to policymakers on when cheaper but less accurate targeting approaches (like phone-based targeting) are likely to be better choices than more expensive but more accurate targeting approaches (like proxy-means testing).

6.1.3 Impacts: Targeting the poorest vs. targeting for treatment effects

Most work on poverty targeting — including the work presented in this thesis — designs and evaluates methods for identifying the poorest households in a community (typically measured based on income or consumption). However, policymakers may alternatively — or additionally — be interested in prioritizing households that are likely to be most *impacted* by receiving benefits. While standard utility functions posit a monotonic negative relationship between wealth and impact of receiving social benefits, this relationship may not be empirically supported: Haushofer et al. (2022) [87] find that, in the setting of a large cash transfer program in rural Kenya, the households identified as likely to benefit from a cash transfer program are very different from the households identified as poorest.

Future work could bridge between the literature on cash transfer impacts and the work presented in this thesis on targeting with digital data sources. A first step would be to build on randomized controlled trials comparing the poverty *impacts* of PMT-targeted and CBT-targeted social protection programs [11, 130] — which have generally not found differences in magnitudes of impacts between targeting approaches — to measure the poverty impacts of phone-based and satellite-based targeting methods in comparison to other targeting approaches. Further work could assess whether phone-based or satellite-based approaches can *predict* which households are likely to be most impacted by a cash

transfer or other anti-poverty intervention. These predictions could then be used to prioritize households most likely to be aided by an intervention for benefits at scale.

6.1.4 Perceptions: Measuring the acceptability of data-driven targeting approaches to beneficiary communities

No study to date has established whether beneficiary communities perceive digitally-targeted aid programs as legitimate and fair. Community concerns around fairness and social cohesion in social protection targeting [60] may be exacerbated in digital settings, since data-driven targeting relies on difficult-to-interpret ML estimators and passively collected data; conversely, concerns may be alleviated by the perceived objectivity of such an approach. Previous research on proxy-means test targeting social protection programs have shown that unintended negative dynamics — such as jealousy [60], stigma [134], and accusations of stealing [44] — can emerge when beneficiary concerns are not accounted for in data-driven targeting approaches. In extreme settings communities have rejected targeting approaches that are perceived as illegitimate [131], so it is critical to gather evidence on local conceptions of the legitimacy and fairness of data-driven targeting.

Future research could explore community perceptions of digital targeting methods (including phone-based and satellite-based targeting) through qualitative or observational studies, or through a randomized controlled trial. Building on two past randomized controlled trials that have compared beneficiary perceptions of fairness and legitimacy between proxy-means testing and community-based targeting [11, 130] (in Indoensia and Niger, respectively), a particularly interesting future study would randomize targeting methods at the community level between phone-based targeting, satellite-based geographic micro-targeting, proxy-means testing, and community-based targeting, and compare beneficiary assessments of perceived fairness, privacy, comprehensibility, and legitimacy of the targeting approaches.

6.1.5 Machine learning methods: Drawing on advances in ML to better leverage digital data sources

A final fruitful direction of future work lies in improved design of the machine learning methods used to infer poverty from digital data sources. Most of the work in this thesis has relied on standard supervised learning methods to train models to predict poverty from tabular features derived from digital data streams. However, digital trace data — and in particular phone data — have a great deal of structure that more advanced machine learning methods could leverage. For example, building on initial work on graph convolutional neural networks for multi-view network data like phone data [103], future work could aim to better leverage the network structure of mobile phone data and likely homophily of household poverty in the network for improved poverty prediction

from mobile phone data. Future work could also leverage the time-series nature of phone data and other digital trace data sources to better track temporal measures like poverty dynamics.

Given the limited training data typically available to train poverty prediction models based on mobile phone data and other digital data sources, a particularly relevant direction of future research on ML for poverty prediction lies in sampling approaches for training data. Most poverty prediction models — including the ones tested in this thesis — are trained on household surveys that were collected for a different purpose (typically, to provide nationally representative demographic and health statistics). These standardized surveys typically use (stratified) random sampling to reduce measurement error [92]. However, sampling training data according to these strategies may not optimize predictive power for a machine learning algorithm subject to a budget constraint for data collection. Adaptive sampling methods or strategies for one-step optimal experiment design could improve the performance of machine learning models trained to predict poverty from digital (or survey-based) data sources. Building on initial work I have contributed to in this space [147], future work could explore optimal adaptive data collection policies for both phone and field surveys for labeling training data for poverty prediction models.

6.2 Conclusion

This dissertation has introduced targeting methods for social protection and humanitarian aid programs in low-income and data-poor contexts based on passively collected digital data sources (mobile phone data and satellite imagery) and machine learning. Several chapters of this dissertation have rigorously evaluated the targeting accuracy of these new approaches in Togo (chapter 2), Afghanistan (chapter 3), and Bangladesh (chapter 4), in comparison to standard poverty targeting approaches, including survey-based approaches like proxy-means testing, community-based approaches, and geographic targeting. The work in this dissertation has also contributed to some of the first real-world implementations of targeting based on machine learning and digital data in the real world, including the aid programs described and evaluated in Togo and Bangladesh in Chapters 2 and 4.

Taken together, the studies presented here suggest that new targeting methods based on mobile phone data, satellite imagery, and machine learning are likely to become a useful tool to policymakers, but they are not a panacea. While these new algorithmic approaches are likely to be the best choice in some settings — particularly settings where traditional survey-based poverty data are unavailable or very out-of-date, and where primary data collection is infeasible due to conflict, pandemics or natural disasters — the results in this thesis show that they are generally not as accurate as traditional survey-based poverty targeting approaches, and governments and NGOs should continue to invest in high quality and comprehensive data collection for poverty targeting. Social

protection program administrators will also need to continue to weigh trade-offs in targeting criteria between accuracy, cost, speed, adpativity, acceptability, transparency, susceptibility to manipulation, and more. Future work in this space can help provide guidance on where digital and algorithmic targeting approaches fit into these trade-offs, and continue to push the machine learning frontier for improved targeting accuracy, ultimately helping channel limited social protection resources to those who need them the most.

Bibliography

[1] Rediet Abebe **andothers**. "Narratives and counternarratives on data sharing in Africa". **in***Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*: 2021, **pages** 329–341.

[2] Akhter Ahmed **and** M Mehrab Bakhtiar. *Proposed indicators for selecting needy participants for the Vulnerable Women's Benefit (VWB) Program in urban Bangladesh*. Intl Food Policy Res Inst, 2023.

[3] Emily Aiken **and** Tim Ohlenburg. *Novel digital data sources for social protection: Opportunities and challenges*. 2023.

[4] Emily Aiken **andothers**. *Estimating Impact with Surveys versus Digital Traces: Evidence from Randomized Cash Transfers in Togo*. techreport. National Bureau of Economic Research, 2023.

[5] Emily Aiken **andothers**. "Machine learning and phone data can improve targeting of humanitarian aid". **in***Nature*: 603.7903 (2022), **pages** 864–870.

[6] Emily L Aiken **andothers**. "Program targeting with machine learning and mobile phone data: Evidence from an anti-poverty intervention in Afghanistan". **in***Journal of Development Economics*: 161 (2023), **page** 103016.

[7] George A Akerlof. "The economics of" tagging" as applied to the optimal income tax, welfare programs, and manpower planning". **in***The American economic review*: 68.1 (1978), **pages** 8–19.

[8] Mohammad Alaggan **andothers**. "Sanitization of call detail records via differentially-private bloom filters". **in***Data and Applications Security and Privacy XXIX: 29th Annual IFIP WG 11.3 Working Conference, DBSec 2015, Fairfax, VA, USA, July 13-15, 2015, Proceedings 29*: Springer. 2015, **pages** 223–230.

[9] Vivi Alatas **andothers**. "Network structure and the aggregation of information: Theory and evidence from Indonesia". **in***American Economic Review*: 106.7 (2016), **pages** 1663–1704.

[10] Vivi Alatas **andothers**. "Self-targeting: Evidence from a field experiment in Indonesia". **in***Journal of Political Economy*: 124.2 (2016), **pages** 371–427.

[11] Vivi Alatas **andothers**. "Targeting the poor: evidence from a field experiment in Indonesia". **in***American Economic Review*: 102.4 (2012), **pages** 1206–1240.

[12] Harold Alderman. "Do local officials know something we don't? Decentralization of targeted transfers in Albania". **in***Journal of public Economics*: 83.3 (2002), **pages** 375–404.

[13] Sabina Alkire **andothers**. *Multidimensional poverty measurement and analysis*. Oxford University Press, USA, 2015.

[14] Michael L. Anderson. "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects". **in***Journal of the American Statistical Association*: 103.484 (2008), **pages** 1481–1495. ISSN: 01621459. URL: http://www.jstor.org/stable/27640197.

[15] Joshua D Angrist **andothers**. "Leveraging lotteries for school value-added: Testing and estimation". **in***The Quarterly Journal of Economics*: 132.2 (2017), **pages** 871–919.

[16] Martin Arjovsky **andothers**. "Invariant risk minimization". **in***arXiv preprint arXiv:1907.02893*: (2019).

[17] Kumar Ayush **andothers**. "Generating interpretable poverty maps using object detection in satellite images". **in***arXiv preprint arXiv:2002.01612*: (2020).

[18] James P. Bagrow, Dashun Wang **and** Albert-László Barabási. "Collective Response of Human Populations to Large-Scale Emergencies". **in***PLoS ONE*: 6.3 (**march** 2011), e17680. DOI: 10.1371/journal.pone.0017680. URL: http://dx.doi.org/10.1371/journal.pone.0017680 (**urlseen** 07/04/2013).

[19] Judy L Baker **and** Margaret E Grosh. "Poverty reduction through geographic targeting: How well does it work?" **in***World development*: 22.7 (1994), **pages** 983–995.

[20] Paul Bance **and** Pascale Schnitzer. "Can the Luck of the Draw Help Social Safety Nets?" **in**(2021).

[21] Abhijit Banerjee **andothers**. "Does Poverty Change Labor Supply? Evidence from Multiple Income Effects and 115,579 Bags". **in***National Bureau of Economic Research*: 27314 (2022).

[22] Abhijit Banerjee **andothers**. "Effects of a Universal Basic Income during the pandemic". **in***Innovations for Poverty Action Working Paper*: (2020).

[23] Abhijit Banerjee **andothers**. "The (lack of) distortionary effects of proxy-means tests: Results from a nationwide experiment in Indonesia". **in***Journal of Public Economics Plus*: 1 (2020), **page** 100001.

[24] Abhijit V Banerjee **andothers**. "Debunking the stereotype of the lazy welfare recipient: Evidence from cash transfer programs". **in***The World Bank Research Observer*: 32.2 (2017), **pages** 155–184.

[25] World Bank. *World development report 2021: Data for better lives*. 2021.

[26] V Barca **and** M Hebbar. "On-demand and up to date? Dynamic inclusion and data updating for social assistance". **in***GIZ (https://socialprotection. org/sites/default/files/publications_files/GIZ_DataUpdatingForSocialAssistance_3. pdf)*: (2020).

[27] Solon Barocas, Moritz Hardt **and** Arvind Narayanan. "Fairness in machine learning". **in***Nips tutorial*: 1 (2017), **page** 2017.

[28] Oscar Barriga-Cabanillas **andothers**. "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?" **in***Presentation at 2021 Pacific Development Conference*: (n.d.).

[29] Francesca Bastagli **andothers**. "Cash transfers: what does the evidence say". **in***A rigorous review of programme impact and the role of design and implementation features. London: ODI*: 1.7 (2016).

[30] Bob Baulch **and** John Hoddinott. "Economic mobility and poverty dynamics in developing countries". **in***The Journal of Development Studies*: 36.6 (2000), **pages** 1–24.

[31] Lori Beaman **andothers**. "Urban networks and targeting: Evidence from liberia". **in***AEA Papers and Proceedings*: **volume** 111. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203. 2021, **pages** 572–576.

[32] Guadalupe Bedoya **andothers**. *No household left behind: Afghanistan targeting the ultra poor impact evaluation*. techreport. National Bureau of Economic Research, 2019.

[33] Richard Bellman **and** Robert Kalaba. "A mathematical theory of adaptive control processes". **in***Proceedings of the National Academy of Sciences*: 45.8 (1959), **pages** 1288–1290.

[34] Daniel Björkegren, Joshua E Blumenstock **and** Samsun Knight. "Manipulation-proof machine learning". **in***arXiv preprint arXiv:2004.03865*: (2020).

[35] Joshua Blumenstock. *Don't forget people in the use of big data for development*. 2018.

[36] Joshua Blumenstock, Gabriel Cadamuro **and** Robert On. "Predicting poverty and wealth from mobile phone metadata". **in***Science*: 350.6264 (2015), **pages** 1073–1076.

[37] Joshua Blumenstock, Danaja Maldeniya **and** Sriganesh Lokanathan. "Understanding the impact of urban infrastructure: New insights from population-scale data". **in***Proceedings of the Ninth International Conference on Information and Communication Technologies and Development*: 2017, **pages** 1–12.

[38] Joshua E Blumenstock. "Estimating economic characteristics with phone data". **in***AEA papers and proceedings*: **volume** 108. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203. 2018, **pages** 72–76.

[39] Joshua Evan Blumenstock, Nathan Eagle **and** Marcel Fafchamps. "Airtime Transfers and Mobile Communications: Evidence in the Aftermath of Natural Disasters". **in***Journal of Development Economics*: 120 (**may** 2016), **pages** 157–181.

[40] Nicolas Bottan, Bridget Hoffmann **and** Diego A Vera-Cossio. "Stepping up during a crisis: The unintended effects of a noncontributory pension program during the Covid-19 pandemic". **in***Journal of Development Economics*: 150 (2021), **page** 102635.

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

[42] Stephanie Brunelin **andothers**. *Responding faster to droughts with satellites and adaptive social protection in Niger*. 2022.

[43] Gharad Bryan, James J Choi **and** Dean Karlan. "Randomizing religion: the impact of Protestant evangelism on economic outcomes". **in***The Quarterly Journal of Economics*: 136.1 (2021), **pages** 293–380.

[44] Francesco Burchi **and** Federico Roscioli. "Can integrated social protection programmes affect social cohesion? Mixed-methods evidence from Malawi". **in***The European Journal of Development Research*: 34.3 (2022), **pages** 1240–1263.

[45] Trent D Buskirk **and** Stanislav Kolenikov. "Finding respondents in the forest: A comparison of logistic regression and random forest models for response propensity weighting and stratification". **in***Survey Methods: Insights from the Field*: (2015), **pages** 1–17.

[46] Adriana Camacho **and** Emily Conover. "Manipulation of social program eligibility". **in***American Economic Journal: Economic Policy*: 3.2 (2011), **pages** 41–65.

[47] Fred H Cate **and** Viktor Mayer-Schönberger. "Notice and consent in a world of big data. International Data Privacy Law, 3 (2), 67-73". **in***International Data Privacy Law*: (2013).

[48] Alket Cecaj, Marco Mamei **and** Nicola Bicocchi. "Re-identification of anonymized CDR datasets using social network data". **in***2014 IEEE International Conference on Pervasive Computing and Communication Workshops (PERCOM WORKSHOPS)*: IEEE. 2014, **pages** 237–242.

[49] Raj Chetty, John N Friedman **and** Jonah E Rockoff. "Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates". **in***American economic review*: 104.9 (2014), **pages** 2593–2632.

[50] Guanghua Chi **andothers**. "Microestimates of wealth for all low-and middle-income countries". **in***Proceedings of the National Academy of Sciences*: 119.3 (2022), e2113658119.

[51] David Coady, Margaret Grosh **and** John Hoddinott. "Targeting outcomes redux". **in***The World Bank Research Observer*: 19.1 (2004), **pages** 61–85.

[52] Jonathan Conning **and** Michael Kevane. "Community-based targeting mechanisms for social safety nets: A critical review". **in***World development*: 30.3 (2002), **pages** 375–394.

[53] Paul Corral **andothers**. *Fragility and conflict: On the front lines of the fight against poverty*. World Bank Publications, 2020.

[54] Tommaso Crosta **andothers**. "Unconditional Cash Transfers: A Bayesian Meta-Analysis of 50 Randomized Evaluations in 26 Low and Middle Income Countries". **in***working paper*: (2023).

[55] Jesse M Cunha, Giacomo De Giorgi **and** Seema Jayachandran. "The price effects of cash versus in-kind transfers". **in***The Review of Economic Studies*: 86.1 (2019), **pages** 240–281.

[56] Yves-Alexandre De Montjoye **andothers**. "Unique in the crowd: The privacy bounds of human mobility". **in***Scientific reports*: 3.1 (2013), **pages** 1–5.

[57] Angus Deaton. "Measuring and understanding behavior, welfare, and poverty". **in***American Economic Review*: 106.6 (2016), **pages** 1221–1243.

[58] Angus Deaton. *The analysis of household surveys: a microeconometric approach to development policy*. World Bank Publications, 1997.

[59] Adeline Decuyper **andothers**. "Estimating food consumption and poverty indices with mobile phone data". **in***arXiv preprint arXiv:1412.2595*: (2014).

[60] Anne Della Guardia, Milli Lake **and** Pascale Schnitzer. "Selective inclusion in cash transfer programs: Unintended consequences for social cohesion". **in***World Development*: 157 (2022), **page** 105922.

[61] Will Dobbie **and** Jae Song. "Targeted debt relief and the origins of financial distress: Experimental evidence from distressed credit card borrowers". **in***American Economic Review*: 110.4 (2020), **pages** 984–1018.

[62] Cynthia Dwork **andothers**. "Fairness through awareness". **in***Proceedings of the 3rd innovations in theoretical computer science conference*: 2012, **pages** 214–226.

[63] Dennis Egger **andothers**. "Falling living standards during the COVID-19 crisis: Quantitative evidence from nine developing countries". **in***Science Advances*: 7.6 (2021), eabe0997.

[64] Dennis Egger **andothers**. "General equilibrium effects of cash transfers: experimental evidence from Kenya". **in***Econometrica*: 90.6 (2022), **pages** 2603–2643.

[65] Chris Elbers **andothers**. "Poverty alleviation through geographic targeting: How much does disaggregation help?" **in***Journal of Development Economics*: 83.1 (2007), **pages** 198–213.

[66] Ryan Engstrom, Jonathan Hersh **and** David Newhouse. "Poverty from space: Using high resolution satellite imagery for estimating economic well-being". **in***The World Bank Economic Review*: 36.2 (2022), **pages** 382–412.

[67] Virginia Eubanks. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press, 2018.

[68] Masoomali Fatehkia **andothers**. "Mapping socioeconomic indicators using social media advertising data". **in***EPJ Data Science*: 9.1 (2020), **page** 22.

[69] Masoomali Fatehkia **andothers**. "The relative value of facebook advertising data for poverty mapping". **in***Proceedings of the International AAAI Conference on Web and Social Media*: **volume** 14. 2020, **pages** 934–938.

[70] Deon Filmer **and** Lant H Pritchett. "Estimating wealth effects without expenditure data—or tears: an application to educational enrollments in states of India". **in***Demography*: 38.1 (2001), **pages** 115–132.

[71] Deon Filmer **and** Kinnon Scott. "Assessing asset indices". **in***Demography*: 49.1 (2012), **pages** 359–392.

[72] Amy Finkelstein. "E-ztax: Tax salience and tax rates". **in***The Quarterly Journal of Economics*: 124.3 (2009), **pages** 969–1010.

[73] Peter Fisker **andothers**. "Guiding Social Protection Targeting Through Satellite Data in São Tomé and Prıncipe". **in**(2022).

[74] Electronic Frontier Fundation. *Necessary & Proportionate: International Principles on the Application of Human Rights Law to Communications Surveillance, 2014*.

[75] Marito Garcia, Charity G Moore **and** Charity MT Moore. *The cash dividend: the rise of cash transfer programs in sub-Saharan Africa*. World Bank Publications, 2012.

[76] Xin Geng **andothers**. "Health insurance, a friend in need? Impacts of formal insurance and crowding out of informal insurance". **in***World Development*: 111 (2018), **pages** 196–210.

[77] Ugo Gentilini **andothers**. "Social protection and jobs responses to COVID-19". **in**(2022).

[78] Michael Gilraine, Jiaying Gu **and** Robert McMillan. *A new method for estimating teacher value-added*. techreport. National Bureau of Economic Research, 2020.

[79] Charles Goodhart. *Monetary relationships: a view from Threadneedle Street*. University of Warwick, 1976.

[80] Lena Gronbach, Jeremy Seekings **and** Vayda Megannon. "Social protection in the COVID-19 pandemic: lessons from South Africa". **in***Center for Global Development Policy Paper*: 252 (2022).

[81] Margaret Grosh **and** Judy L Baker. "Proxy means tests for targeting social programs". **in***Living standards measurement study working paper*: 118 (1995), **pages** 1–49.

[82] GSMA. *The mobile economy*. 2022.

[83] Julius Gunnemann. *PMT based targeting in Burkina Faso*. World Bank, 2016.

[84] Huawei Han **and** Qin Gao. "Community-based welfare targeting and political elite capture: Evidence from rural China". **in***World Development*: 115 (2019), **pages** 145–159.

[85] Rema Hanna **and** Benjamin A Olken. "Universal basic incomes versus targeted transfers: Anti-poverty programs in developing countries". **in***Journal of Economic Perspectives*: 32.4 (2018), **pages** 201–26.

[86] Johannes Haushofer **and** Jeremy Shapiro. "The short-term impact of unconditional cash transfers to the poor: experimental evidence from Kenya". **in***The Quarterly Journal of Economics*: 131.4 (2016), **pages** 1973–2042.

[87] Johannes Haushofer **andothers**. *Targeting impact versus deprivation*. techreport. National Bureau of Economic Research, 2022.

[88] Yutong He **andothers**. "Spatial-temporal super-resolution of satellite imagery via conditional pixel synthesis". **in***Advances in Neural Information Processing Systems*: 34 (2021), **pages** 27903–27915.

[89] Marco Hernandez **andothers**. "Estimating poverty using cell phone data: evidence from Guatemala". **in***World Bank Policy Research Working Paper*: 7969 (2017).

[90] Michael Hillebrecht, Stefan Klonner **and** Noraogo A Pacere. "The dynamics of poverty targeting". **in***Journal of Development Economics*: 161 (2023), **page** 103033.

[91] Lisa Hjelm, Astrid Mathiassen **and** Amit Wadhwa. "Measuring poverty for food security analysis: consumption-versus asset-based approaches". **in***Food and nutrition bulletin*: 37.3 (2016), **pages** 275–289.

[92] Stephen Howes **and** Jean Olson Lanjouw. "Does sample design matter for poverty rate comparisons?" **in***Review of Income and Wealth*: 44.1 (1998), **pages** 99–109.

[93] Luna Yue Huang, Solomon M Hsiang **and** Marco Gonzalez-Navarro. *Using satellite imagery and deep learning to evaluate the impact of anti-poverty programs*. techreport. National Bureau of Economic Research, 2021.

[94] ILO. *World Social Protection Report 2020–22: Social protection at the crossroads–in pursuit of a better future*. 2021.

[95] John PA Ioannidis. "Informed consent, big data, and the oxymoron of research that is not research". **in***The American Journal of Bioethics*: 13.4 (2013), **pages** 40–42.

[96] Ignacio Irarrázaval **andothers**. *Sole Information Systems on Beneficiaries in Latin America*. techreport. Inter-American Development Bank, 2011.

[97] Neal Jean **andothers**. "Combining satellite imagery and machine learning to predict poverty". **in***Science*: 353.6301 (2016), **pages** 790–794.

[98] Morten Jerven. *Poor numbers: how we are misled by African development statistics and what to do about it*. Cornell University Press, 2013.

[99] Graham Kalton **and** Ismael Flores-Cervantes. "Weighting methods". **in***Journal of official statistics*: 19.2 (2003), **page** 81.

[100] Thomas J Kane **and** Douglas O Staiger. *Estimating teacher impacts on student achievement: An experimental evaluation*. techreport. National Bureau of Economic Research, 2008.

[101] Dean Karlan **and** Bram Thuysbaert. "Targeting ultra-poor households in Honduras and Peru". **in***The World Bank Economic Review*: 33.1 (2019), **pages** 63–94.

[102] Dean Karlan **andothers**. "Social Protection and Social Distancing During the Pandemic: Mobile Money Transfers in Ghana". **in***National Bureau of Economic Research working paper*: (2022).

[103] Muhammad Raza Khan **and** Joshua E Blumenstock. "Multi-gcn: Graph convolutional networks for multi-view networks, with applications to global poverty". **in***Proceedings of the AAAI conference on artificial intelligence*: **volume** 33. 01. 2019, **pages** 606–613.

[104] Stephen Kidd. "Social exclusion and access to social protection schemes". **in***Journal of Development Effectiveness*: 9.2 (2017), **pages** 212–244.

[105] Stephen Kidd, Diloá Athias **and** Idil Mohamud. "Social registries: A short history of abject failure". **in***Development Pathways*: (2021).

[106] Jon Kleinberg, Sendhil Mullainathan **and** Manish Raghavan. "Inherent trade-offs in the fair determination of risk scores". **in***arXiv preprint arXiv:1609.05807*: (2016).

[107] Christoph Lakner **andothers**. *Updated estimates of the impact of COVID-19 on global poverty: Looking back at 2020 and the outlook for 2021*. 2021.

[108] Tian Li **andothers**. "Federated learning: Challenges, methods, and future directions". **in***IEEE signal processing magazine*: 37.3 (2020), **pages** 50–60.

[109] Kathy Lindert **andothers**. *Sourcebook on the foundations of social protection delivery systems*. World Bank Publications, 2020.

[110] Juliana Londoño-Vélez **and** Pablo Querubin. "The impact of emergency cash assistance in a pandemic: experimental evidence from Colombia". **in***Review of Economics and Statistics*: 104.1 (2022), **pages** 157–165.

[111] Joan Lopez. "Experimenting with poverty: The SISBEN and data analytics projects in Colombia". **in**(2020).

[112] Laura Mann. "Left to other peoples' devices? A political economy perspective on the big data revolution in development". **in***Development and Change*: 49.1 (2018), **pages** 3–36.

[113] César Martinelli **and** Susan Wendy Parker. "Deception and misreporting in a social program". **in***Journal of the European Economic Association*: 7.4 (2009), **pages** 886–908.

[114] Imran Matin, M Rabbani **and** M Sulaiman. "Crafting a graduation pathway for the ultra poor: Lessons and evidence from a BRAC programme". **in**(2008).

[115] N Maunder **andothers**. "Somalia: An evaluation of WFP's Portfolio (2012-2017)". **in***World Food Programme*: (2018).

[116] Linden McBride **and** Austin Nichols. "Retooling poverty targeting using out-of-sample validation and machine learning". **in***The World Bank Economic Review*: 32.3 (2018), **pages** 531–550.

[117] Sveta Milusheva **andothers**. "Challenges and opportunities in accessing mobile phone data for COVID-19 response in developing countries". **in***Data & Policy*: 3 (2021).

[118] Darakhshan J Mir **andothers**. "Dp-where: Differentially private modeling of human mobility". **in***2013 IEEE international conference on big data*: IEEE. 2013, **pages** 580–588.

[119] Ahmed Mushfiq Mobarak **and** Mark R Rosenzweig. "Selling formal insurance to the informally insured". **in**(2012).

[120] Y. de Montjoye, L. Rocher **and** A. Pentland. "bandicoot: a Python toolbox for mobile phone metadata." **in***Journal of Machine Learning Research*: 17 (2016), **pages** 1–5.

[121] Yves-Alexandre de Montjoye, Jake Kendall **and** Cameron F Kerry. "8 ENABLING HUMANITARIAN USE OF MOBILE PHONE DATA". **in***Trusted Data, revised and expanded edition: A New Framework for Identity and Data Sharing*: (2019), **page** 167.

[122] Anit Nath Mukherjee **andothers**. *Digital-first Approach to Emergency Cash Transfers: Step-kin in the DemocraticRepublic of Congo*. techreport. The World Bank, 2023.

[123] United Nations. "2019 revision of world population prospects". **in**(2019).

[124] United Nations. *Spending on social protection rose nearly 270% with the pandemic*. 2022.

[125] Albert L Nichols **and** Richard J Zeckhauser. "Targeting transfers through restrictions on recipients". **in***The American Economic Review*: 72.2 (1982), **pages** 372–377.

[126] Alejandro Noriega-Campero **andothers**. "Algorithmic targeting of social policies: fairness, accuracy, and distributed governance". **in***Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*: 2020, **pages** 241–251.

[127] Nuria Oliver **andothers**. *Mobile phone data for informing public health actions across the COVID-19 pandemic life cycle*. 2020.

[128] Sara Pavanello **andothers**. "Effects of cash transfers on community interactions: Emerging evidence". **in***The Journal of Development Studies*: 52.8 (2016), **pages** 1147–1161.

[129] Neeti Pokhriyal **and** Damien Christophe Jacques. "Combining disparate data sources for improved poverty prediction and mapping". **in***Proceedings of the National Academy of Sciences*: 114.46 (2017), E9783–E9792.

[130] Patrick Premand **and** Pascale Schnitzer. "Efficiency, legitimacy, and impacts of targeting methods: Evidence from an experiment in Niger". **in***The World Bank Economic Review*: 35.4 (2021), **pages** 892–920.

[131] Kate Pruce. "The politics of who gets what and why: learning from the targeting of social cash transfers in Zambia". **in***The European Journal of Development Research*: 35.4 (2023), **pages** 820–839.

[132] Nathan Ratledge **andothers**. "Using machine learning to assess the livelihood impact of electricity access". **in***Nature*: 611.7936 (2022), **pages** 491–495.

[133] Martin Ravallion. *Poverty lines in theory and practice*. **volume** 133. World Bank Publications, 1998.

[134] Keetie Roelen **andothers**. "Intra-household dynamics, social cohesion and women's empowerment: the effects of a graduation programme in Burundi". **in**(2019).

[135] Esther Rolf **andothers**. "A generalizable and accessible approach to machine learning with global satellite imagery". **in***Nature communications*: 12.1 (2021), **page** 4392.

[136] Donald B Rubin. "The bayesian bootstrap". **in***The annals of statistics*: (1981), **pages** 130–134.

[137] Shiori Sagawa **andothers**. "Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization". **in***arXiv preprint arXiv:1911.08731*: (2019).

[138] Norbert Rüdiger Schady **and** Maria Araujo. *Cash transfers, conditions, school enrollment, and child work: Evidence from a randomized experiment in Ecuador*. **volume** 3. World Bank Publications, 2006.

[139] Timo Schmid **andothers**. "Constructing sociodemographic indicators for national statistical institutes by using mobile phone data: estimating literacy rates in Senegal". **in***Journal of the Royal Statistical Society Series A: Statistics in Society*: 180.4 (2017), **pages** 1163–1190.

[140] Pascale Schnitzer **and** Quentin Stoeffler. "Targeting for Social Safety Nets: Evidence from Nine Programs in the Sahel". **in***Available at SSRN 4017172*: (2022).

[141] Amartya Sen. *The political economy of targeting*. World Bank Washington, DC, 1992.

[142] Umar Serajuddin **andothers**. "Data deprivation: another deprivation to end". **in***World Bank policy research working paper*: 7252 (2015).

[143] Evan Sheehan **andothers**. "Predicting economic development using geolocated wikipedia articles". **in***Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*: 2019, **pages** 2698–2706.

[144] Emmanuel Skoufias. "Economic crises and natural disasters: Coping strategies and policy implications". **in***World development*: 31.7 (2003), **pages** 1087–1102.

[145] Christopher Smith-Clarke, Afra Mashhadi **and** Licia Capra. "Poverty on the cheap: Estimating poverty maps using aggregated mobile communication networks". **in***Proceedings of the SIGCHI conference on human factors in computing systems*: 2014, **pages** 511–520.

[146] Isabella S Smythe **and** Joshua E Blumenstock. "Geographic microtargeting of social assistance with high-resolution poverty maps". **in***Proceedings of the National Academy of Sciences*: 119.32 (2022), e2120025119.

[147] Satej Soman **andothers**. "Can Strategic Data Collection Improve the Performance of Poverty Prediction Models?" **in***arXiv preprint arXiv:2211.08735*: (2022).

[148] Jessica E Steele **andothers**. "Mapping poverty using mobile phone and satellite data". **in***Journal of The Royal Society Interface*: 14.127 (2017), **page** 20160690.

[149] Getaw Tadesse, Gashaw T Abate **and** Tadiwos Zewdie. "Biases in self-reported food insecurity measurement: A list experiment approach". **in***Food Policy*: 92 (2020), **page** 101862.

[150] Linnet Taylor. "No place to hide? The ethics and analytics of tracking mobility using mobile phone data". **in***Environment and Planning D: Society and Space*: 34.2 (2016), **pages** 319–336.

[151] Linnet Taylor **and** Dennis Broeders. "In the name of development: Power, profit and the datafication of the global south". **in***Geoforum*: 64 (2015), **pages** 229–237.

[152] Tobias G Tiecke **andothers**. "Mapping the world population one building at a time". **in***arXiv preprint arXiv:1712.05839*: (2017).

[153] Carly Trachtman, Yudistira Hendra Permana **and** Gumilang Aryo Sahadewo. *How much do our neighbors really know? The limits of community-based targeting*. techreport. working paper, University of California, Berkeley, 2022.

[154] Burak Uzkent **andothers**. "Learning to interpret satellite images in global scale using wikipedia". **in***arXiv preprint arXiv:1905.02506*: (2019).

[155] Rachel Warren, Emily Aiken **and** Joshua Blumenstock. "Note: Home Location Detection from Mobile Phone Data: Evidence from Togo". **in***ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies (COMPASS)*: 2022, **pages** 685–692.

[156] World Bank. *Poverty and Shared Prosperity 2020 : Reversals of Fortune*. World Bank Publications. License: CC BY 3.0 IGO. The World Bank, 2020. URL: `https://openknowledge.worldbank.org/handle/10986/34496`.

[157] Huaxiu Yao **andothers**. "Wild-time: A benchmark of in-the-wild distribution shift over time". **in***Advances in Neural Information Processing Systems*: 35 (2022), **pages** 10309–10324.

[158] Christopher Yeh **andothers**. "Using publicly available satellite imagery and deep learning to understand economic well-being in Africa". **in***Nature communications*: 11.1 (2020), **page** 2583.

[159] Moeed Yusuf. "Community targeting for povery reduction: lessons from developing countries". **in**(2010).

[160] Manfred Zeller, Joseph Feulefack **and** Andreas Neef. *How accurate is participatory wealth ranking (PWR) in targeting the poor? A case study from Bangladesh*. techreport. 2006.

Appendix A

Supporting materials for Chapter 2

A.1 Selection of variables for proxy-means test

Our proxy-means test is used in analysis for both the 2018-2019 field survey (where we evaluate the PMT's accuracy as a targeting mechanism) and the 2020 phone survey (where we use the PMT as a measure of ground-truth poverty in the absence of a consumption measure). We construct the PMT using all observations from the 2018-2019 field survey (N = 6,171). We begin by identifying all information on demographics and asset ownership collected in the field survey that may correlated with poverty. In total, we identify 56 variables, including information on household assets and housing quality, education, marital status, age, ethnicity, health, location, and more.

Our goal is to identify a small subset of variables that are most predictive of household consumption. We use stepwise forward selection to identify the most predictive feature subsets of size K, for K ranging from 1 to 30. Specifically, we randomly divide our survey observations into a training set (75%) and test set (25%). For K = 1, we train a machine learning model to predict household consumption from each feature individually, and select the feature associated with the best model. For $K = 2$, we test adding each remaining feature to our model, and select the feature that adds the most predictive power. We continue the process for all K up to 30.

We perform the stepwise forward selection process first for a Ridge regression[1] and second for a random forest.[2] Supplementary Figure 4 plots the predictive accuracy (measured with R2) for each value of K for the two models.

We observe that the random forest is not significantly more accurate than the regression, and note a greater degree of overfitting with the random forest. We therefore select the Ridge regression, as the resulting coefficients are easier to interpret. We identify an "elbow" in the accuracy progression at K=12 features, so we use the feature subset

[1]The optimal L2 penalty is selected via a wide grid.

[2]The optimal ensemble size is chosen from 50, 100 via 3-fold cross validation and the optimal tree depth is chosen from 2, 4, 6, 8.

of size $K = 12$ in our PMT. These features and the weights associated with them are recorded in Table S8.

A.2 Design of the 2020 phone survey

This section describes the design and implementation of the 2020 phone survey, which took place in the last week of September and the first week of October 2020.

A.2.1 Sampling

The 2020 phone survey was designed to be representative of active mobile phone subscribers living in Togo's 100 poorest cantons. The sample frame for the survey was all mobile phone subscribers active on one of the two mobile networks in Togo between March 1 and September 30, 2020 (N = 5.83 million). Sampling was based on four metrics associated with each mobile phone subscriber: inferred probability of living in a rural Novissi-eligible area, registration to a previous Novissi program, inferred wealth based on phone data, and total mobile phone expenditure.

- Inferred probability of living in a rural Novissi-eligible canton: We used the machine learning model described in Warren et al. (2022) [155] to assign each subscriber a probability of living in a rural Novissi-eligible canton.
- Registration to a previous Novissi program: At the time of the survey, 22% of mobile network subscribers in Togo were already registered in the Novissi system, and therefore were associated with a ground-truth home canton based on the canton in which they are registered to vote. In our dataset of inferred home location likelihoods, we assigned any subscriber registered to vote in one of the 100 targeted cantons a 100% likelihood of geographical eligibility (N = 86,856). We assigned any subscriber registered to vote outside of these cantons a 0% likelihood of geographical eligibility (N = 1,046,905).
- Inferred poverty based on mobile phone data: We used ground-truth poverty data collected in a previous nationally-representative phone survey conducted in June 2020 to train a machine learning model to predict poverty from CDR. We followed the methods described in subsection 2.2.5 using the PMT as ground truth and CDR features from March 1 to September 30, 2020. We used the machine learning model to predict the poverty of each of the 5.83 million mobile phone subscribers in Togo.
- Mobile phone expenditure: We constructed the measure of total phone expenditure for each subscriber.

Based on the total number of voters registered in targeted cantons and individual mobile phone penetration in each canton (based on the 2018-2019 field survey, measured

at the prefecture level), we estimated that around 240,000 subscribers live in eligible cantons. We identified the 240,000 subscribers most likely to be living in a targeted canton (including all 86,856 subscribers registered in targeted cantons). Only these 240,000 subscribers were eligible to be surveyed.

We oversampled survey respondents based on two counterfactual targeting methods that we simulated pre-survey: predicted poverty based on phone data, and mobile phone expenditures, as described in 2.2.5. We divided the 240,000 subscribers into four quartiles based on phone-inferred poverty and mobile phone expenditures. We overlapped the quartiles to form eight "cells," based on the combination of the two targeting methods (for example, cell AA represents being in the lowest quartile by both targeting methods, while cell AD represents being in the lowest quartile by one method and the lowest quartile by the other, and cell BC represents being in the second-lowest quartile by one method and the second-highest quartile by the other). We assigned a cell weight of 0.20 to cells AD and BC (where the two methods disagree the most), a cell weight of 0.15 to cells AC and BD, a cell weight of 0.10 to cells AB and BC, and a cell weight of 0.05 to cells CD and DD (where the two methods disagree least).

Our sampling probabilities for the 240,000 survey-eligible subscribers were constructed as the product of a subscriber's cell weight and their probability of residing in a targeted canton (so subscribers likely to be living in targeted cantons are oversampled within each cell). The distributions of these draw probabilities are shown in Figure S13 Panel A. We use the inverse of these draw probabilities as sample weights in our downstream analysis, in combination with response weights - see Section (iv) below. We drew 40,000 phone numbers at random from the 240,000 survey-eligible subscribers, with assigned draw probabilities. We provided these 40,000 phone numbers in a random order to enumerators with the expectation that not all of them would be called in order to reach a goal interview quota of 10,000; indeed, only 30,244 phone numbers were called before the quota was reached – see appendix A.2.4 below.

A.2.2 Response Rates

In total, enumerators conducted 10,701 interviews out of 30,244 phone numbers that were called (overall response rate of 35.38%). Phone numbers were called in a random order, and were assigned to enumerators by language (with random assignment with groups of enumerators speaking the same language). While we have little information on subscribers pre-survey, we can examine differential nonresponse by (1) inferred geography based on CDR, (2) registration to a previous Novissi program, and (3) pre-survey mobile phone use (we focus on the phone-predicted measure of poverty and measure of daily expenditures on calls and texts that are used in the rest of the chapter). Table S18 displays response rates disaggregated along these dimensions. We find that response rates are higher for those registered to a prior Novissi program, those inferred to be living in the regions of Lomé Commune, Maritime, or Savanes, and those with

a high daily phone expenditure. Appendix A.2.4 describes how we reweight survey observations to account for differential nonresponse.

A.2.3 Removing Low-Quality Surveys

We identified unreliable enumerators by comparing the data collected in the survey with the information contained in the Novissi registry for the subset of survey respondents who had registered to a previous Novissi program. We begin our analysis by constructing "value-added" (VA) estimates for the enumerators in our data. We predict the VA of each enumerator on the basis of the correct answers to three questions for which we obtained ground-truth information from the Novissi database (canton, age and sex), and on the frequency of surveys with a single head of household (which avoids the roster part of the survey and simplifies the enumerator's work). We control for interviewee characteristics such as region and interview language to separate the enumerator's impact from observable interviewee selection.[3] Our approach to estimating enumerators' VA parallels the parametric empirical Bayes estimator of teacher's VA in past work [100, 49, 78].

We then normalize the VAs for each of the four dimensions (canton, age, gender, and number of surveys with only one adult), and take the average for enumerators who conducted more than twenty interviews. The bottom ten percent of enumerators have an average VA one standard deviation below the mean VA across all enumerators; we classify their surveys as "poor quality." The interviews of the three interviewers with an average VA lower two standard deviations below the total average VA are classified as "very poor quality."

1,180 surveys associated with enumerators who are ranked "poor quality" or "very poor quality" are removed from the dataset. We drop a further 606 surveys with missing data for the PMT or one or more of the counterfactual targeting methods, for a final survey dataset size of 8,915.

A.2.4 Reweighting for Nonresponse

As noted in section (ii), certain groups are more likely to respond to the survey than others. To make the final analysis representative of the initial sample frame (i.e., active mobile subscribers in the 100 poorest cantons) rather than just survey respondents, we reweight survey observations by likelihood of response based on pre-survey covariates [99, 45]. In our case, we train a machine learning model (using a gradient boosting model and the same set of hyperparameters used for wealth prediction from phone data) to predict response from our usual set of CDR features, along with whether a subscriber registered to a previous Novissi program. This model is trained on all 30,244 numbers that

[3] As the phone number list was randomized and then distributed to the enumerators, we believe there is little room for sorting.

were called, with "response" defined as responding to the survey, including all questions necessary to construct the PMT and counterfactual targeting outcomes, consenting to matching between survey responses and mobile phone data, and that survey passing the quality assessment step (see appendix A.2.3), for a total "responded" population of 8,915 (29%). As in other machine learning models described in this chapter, we tune hyperparameters over 5-fold cross validation and produce predictions for each observation over 10-fold cross validation. The model achieves a cross-validated AUC score of 0.71; feature importances for the model are shown in Supplementary Table 13. To assess the model's accuracy, Figure S14 compares binned estimates of response probability with true rates of response, and indicates that the response prediction model is well-calibrated. Figure S13 Panel B displays the distribution of response probabilities for observations included in the final survey dataset.

The final survey weights used in the chapter are the product of the inverse of the response probability and the inverse of the sampling probability described in appendix A.2.1; the distribution of survey weights are shown in Figure S13 Panel C.

A.2.5 Survey Content

Surveys lasted 30 minutes on average, and included questions on the demographics of the respondent and household members, assets owned by the household, subjective wellbeing of the respondent, the social services available to the household, and the impacts of COVID-19 on the household. The full survey instrument is publicly available online.[4]

A.3 Analyzing program exclusions

In Table 2.2, we present information on sources of exclusion from the Novissi program that are not inherently related to targeting. These estimates are drawn from diverse sources of administrative and survey data, specifically:

Voter ID penetration. According to government administrative datasets, 3,633,898 individuals were registered to vote in Togo by late 2019. The electoral commission of Togo reports that this corresponds to 86.6% of eligible adults. While the total adult population in Togo is hard to pin down (the last census was in 2011), Togo's national statistical agency (https://inseed.tg/) estimates that there are 3,715,318 adults in Togo, whereas the United Nations estimates 4.4 million adults in Togo [123], implying a voter ID penetration rates of 82.6% or 97.8%.

[4]https://jblumenstock.com/files/papers/TogoInstrument2020.pdf

Phone penetration. In the 2018-2019 field survey, 65% of individuals reported owning a mobile phone (Figure S3 Panel A) and 85% of households included at least one individual who owns a phone (Figure S3 Panel B). In rural areas, these rates drop to 50% of individuals and 77% of households. Rates of phone ownership are significantly lower among women (53%) than among men (79%), especially in rural areas (33% for women and 71% for men). These household survey-based estimates likely represent a lower bound, given the steady increase in phone penetration between 2018 and 2020. The Togolese government estimates 82% SIM card penetration in the country (though some people may have multiple SIM cards). Based on data from the mobile phone companies, we observe 5.83 million unique active SIMs in Togo between March and September 2020.

Past phone use. In order to construct a phone-based poverty estimate for a subscriber, they had to place at least one outgoing call or text on the mobile phone network in the period of mobile network observation prior to the program's launch (March – September 2020, with program registrations in November-December 2020). In Togo, a lower bound on this source of exclusion is the typical monthly rate of mobile phone turnover, which we estimate to be roughly 2.5% (see Figure S6). An upper bound is closer to 27%, which is the number of SIM cards that registered for Novissi November-December 2020 who did not make an outgoing transaction in the March-September. This discrepancy may be due to (i) individuals buying new SIM cards specifically to register for Novissi; or (ii) individuals registering for Novissi using existing SIM cards that were not in active use, for instance the SIM cards in multi-SIM phones.

Program awareness. Since individuals had to register for the Novissi program to receive benefits, program advertising and population awareness was a key goal. The program was advertised via radio, SMS, field teams, and direct communication with community leaders at the prefecture and canton level. In total, 245,454 subscribers attempted to register for the program. Although we do not observe the prefecture and canton of subscribers who attempt but do not succeed in registering in our administrative data, we know that 87% of successful registrants are in cantons eligible for benefits. Assuming the rate is approximately the same for attempters, we expect that around 213,545 of the attempters are in eligible cantons. The total voting population in eligible cantons is 528,562, for an estimated attempted registration rate of 40.40%.

Registration challenges. Registration for the Novissi program required the completion of a short (5 question) USSD survey. Of the 245,454 subscribers that attempted to register for the program, 176,517 succeed, for a 71.91% rate of registration success.

Overlaps among sources of exclusion. The above sources of exclusion are not independent and are therefore not cumulative. For instance, individuals who are not registered to vote may also be systematically less likely to have a mobile phone. For this reason, Table S6

uses the 2020 phone survey dataset — restricted to respondents who report living in an eligible canton — to calculate overlaps in sources of exclusion to the poor, including Voter ID possession, program awareness, registration challenges, and targeting errors using the phone-based targeting method. We cannot account for mobile phone ownership in this analysis since the 2020 survey was conducted over the phone, and sampled based on past CDR (see appendix A.2.1).

The final three columns of Table S6 show, based on the 2020 phone survey dataset, average characteristics of the population "succeeding" at each step: average PMT, percent women, and average age. The first panel shows successive exclusions for the entire population; the second panel focuses on just the poorest 29% (i.e., those who "should" be receiving aid, were everyone to register for the program and were the targeting algorithm perfect). In Panel A, we observe that to a certain extent the "right" types of people are dropping out at each step, which would be consistent with self-targeting observed in other contexts [10]: in particular, those who attempt to register are poorer than the overall population (average PMT = 1.45 vs. 1.62). There are little differences in the share of the successful population who are women or average age, except in the targeting stage.

Comparing Panels A and B of Table S6, we observe that the recall of the targeting algorithm is substantially higher among the population that owns a voter ID and succeeds in registration for the program (61%, as shown in Table S6, last row) than the overall population surveyed in the 2020 phone survey (47%, as shown in Table 2.1, row 4). This may be due to self-selection (i.e., the type of poor people who register for Novissi tend to also have low phone-based poverty scores). However, it could alternatively suggest that the phone-based targeting algorithm is best at identifying the poor among the types of subscribers who are aware of and register to the Novissi program.

A.4 Supplementary figures and tables

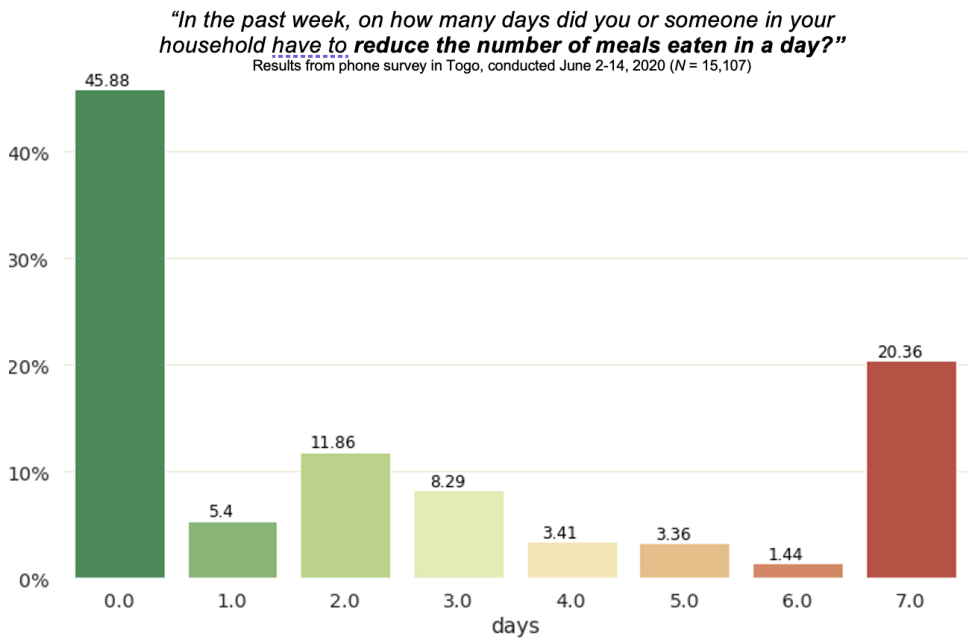

Figure S1: Food insecurity in Togo. In June 2020, we conducted a phone survey of 15,107 mobile phone owners in Togo. Survey weights are used to make responses representative of the population of mobile phone owners in Togo.

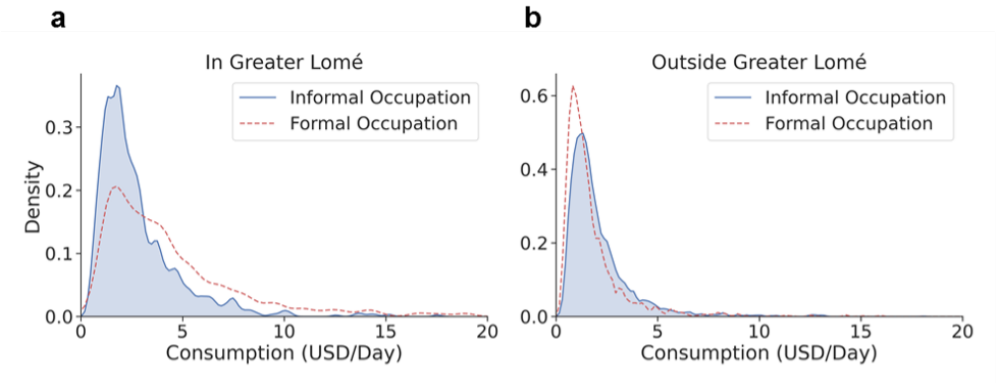

Figure S2: Wealth of formal vs. informal workers. Results based on analysis of nationally-representative household survey data collected by the Government of Togo in 2018-2019 (N = 6,171). Data is collected at the household-level, we assign a household-level informal occupation indicator if at least one of the adult household members is unemployed or employed in an informal occupation. See subsection 2.2.6.

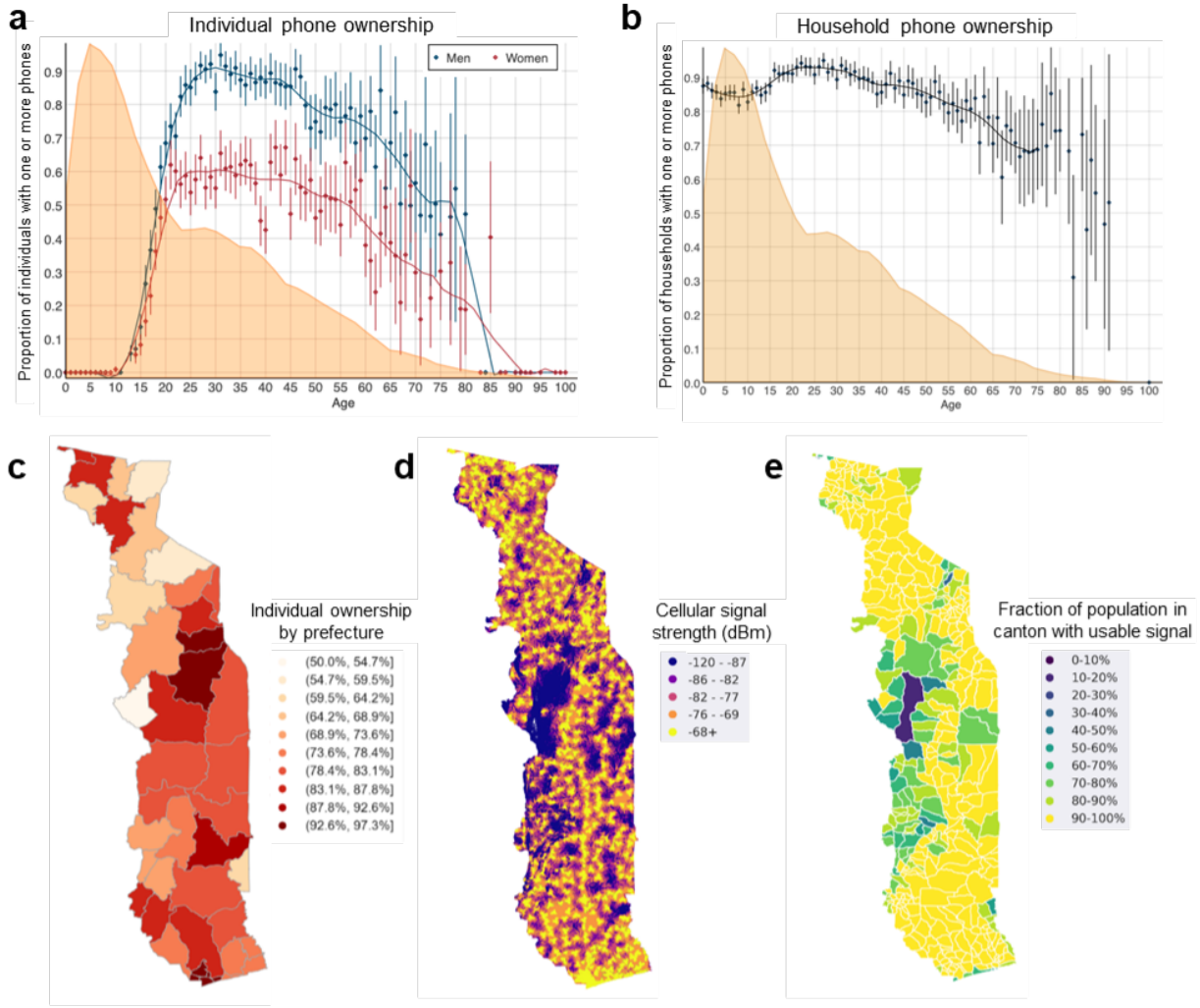

Figure S3: Mobile phone penetration and coverage in Togo. Based on nationally-representative household survey data collected in 2018-2019, we estimate a) the percentage of adults in Togo with one or more mobile phone, disaggregated by age and gender (the dots indicate the sample mean, while vertical bands indicate 95% confidence intervals derived from N = 27,483 total individual survey responses); b) the percentage of households in Togo with one or more mobile phones, disaggregated by the age of the head of household (the dots indicate the sample mean, while vertical bands indicate 95% confidence intervals derived from N = 27,483 total individual survey responses); and c) the percentage of individuals in each prefecture with one or more mobile phones. Using data on the location and signal strength of all cell towers in Togo, made available by Togocel (one of the two phone companies in Togo), we calculate d) the signal strength across Togo; and e) the fraction of the population in each canton with access to a usable signal, where signal greater than -86 dBm is generally considered usable, and sub-canton estimates of population density are derived from satellite imagery [152].

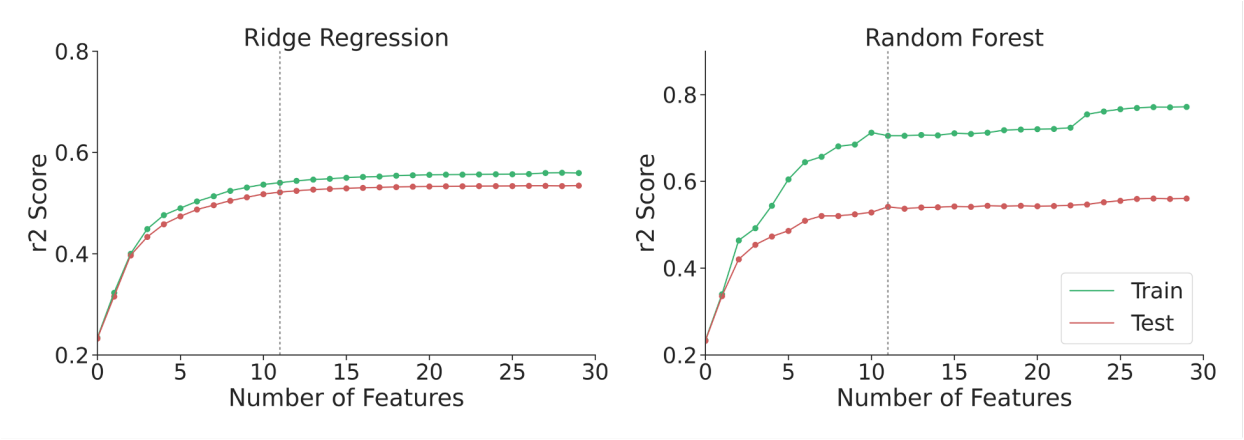

Figure S4: Selection of variables for proxy-means test. Each plot shows the accuracy (measured by r^2 score) of a proxy-means test using the most predictive feature subset of size K, where K is plotted on the x-axis. The left plot shows the accuracy obtained by a Ridge regression; the right plot shows the accuracy obtained by a random forest. Feature subsets are selected via stepwise forward selection.

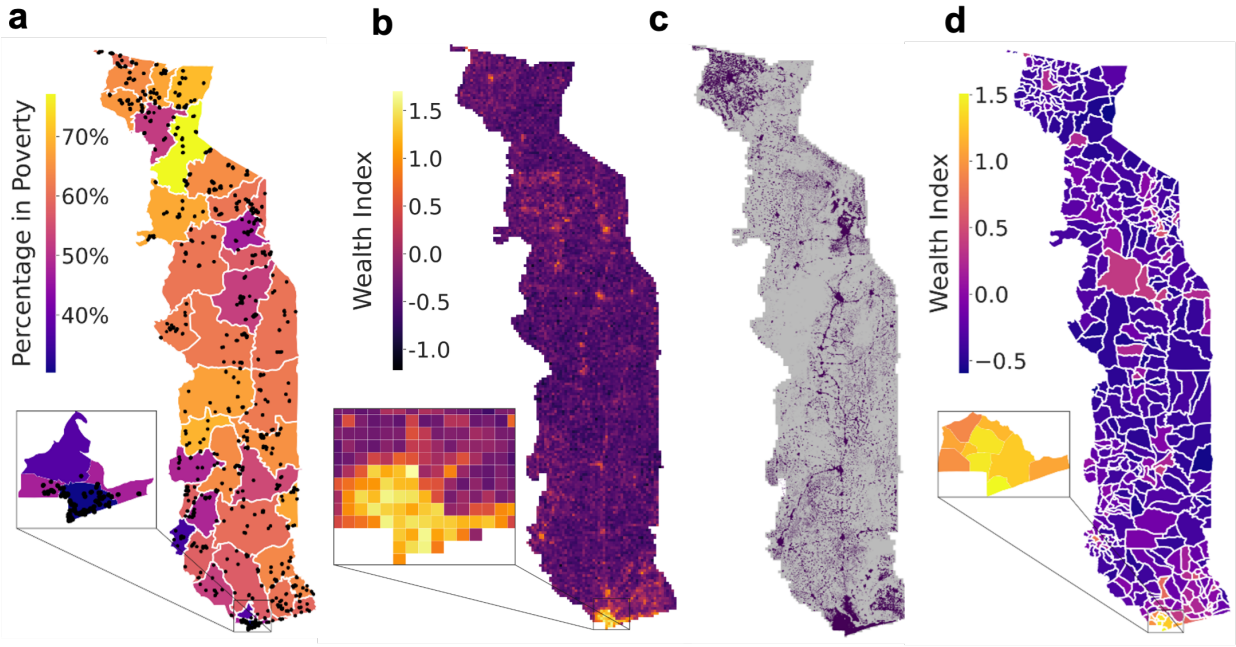

Figure S5: Poverty maps. (a) Prefecture (admin-2) poverty map inferred from 2017 field survey (N = 26,902), showing the percent of the population living under the poverty line by prefecture. Overlayed with locations of survey observations in black points. (b) High-resolution estimates of consumption derived from satellite imagery. (c) High-resolution estimates of population density derived from satellite imagery. (d) Canton (admin-3) poverty map inferred from satellite imagery by combining high-resolution consumption estimates and population density estimates to calculate weighted average consumption per canton.

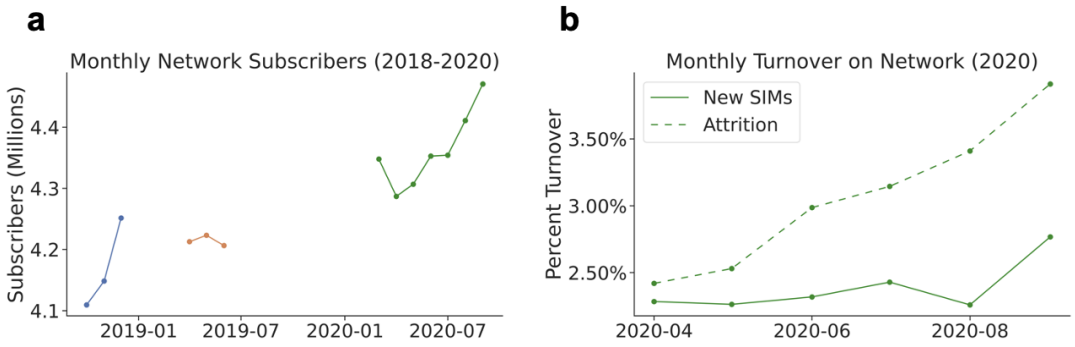

Figure S6: Mobile phone network activity. (a) Count of unique subscribers making at least one outgoing transaction (call or text) on the mobile network in each month. October-December 2018 shown in blue, April-June 2019 in orange, and March-September 2020 in green. (b) Monthly turnover from the network in April-September 2020. New SIMs are quantified as the proportion of subscribers in each month whose first observed transaction is in that month. Attrition is quantified as the proportion of subscribers in each month who make no further outgoing transactions after that month. Note that we do not observe CDR in the months prior to March 2020, so we show results starting in April 2020 in Panel B; nonetheless a small proportion of the new SIMs in Panel B are inevitably due to sparsity in the CDR (that is, subscribers who placed a transaction prior to March 2020 that is not recorded in our dataset). Likewise, we do not observe CDR past December 2020, so a small part of the attrition measured in Panel B is due to sparsity in CDR transactions.

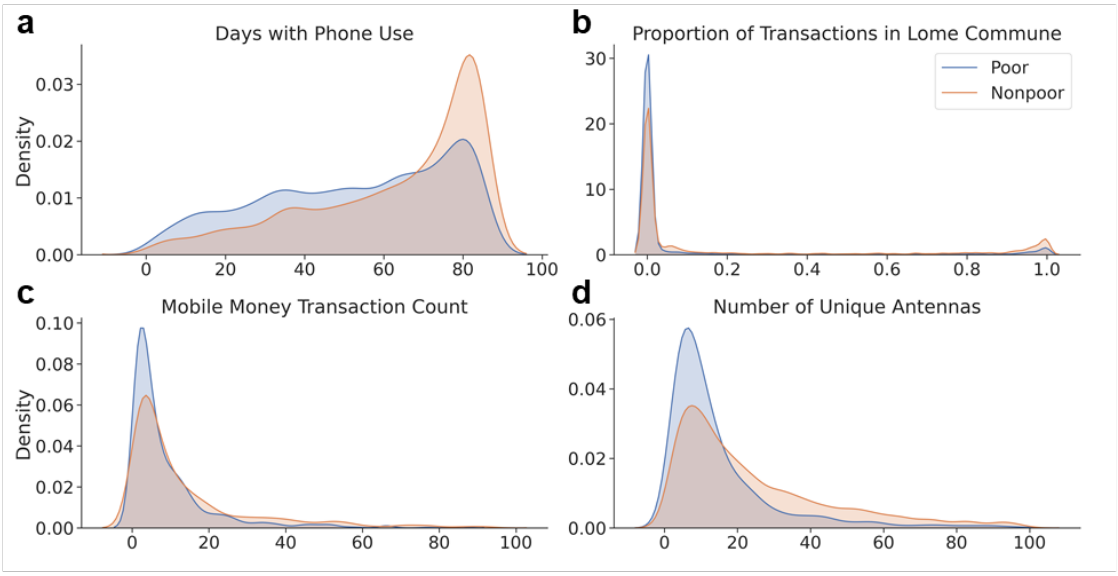

Figure S7: CDR features. Comparing the distribution for CDR features for those above and below the international poverty line (USD 1.90/day) in the 2018-2019 field survey dataset.

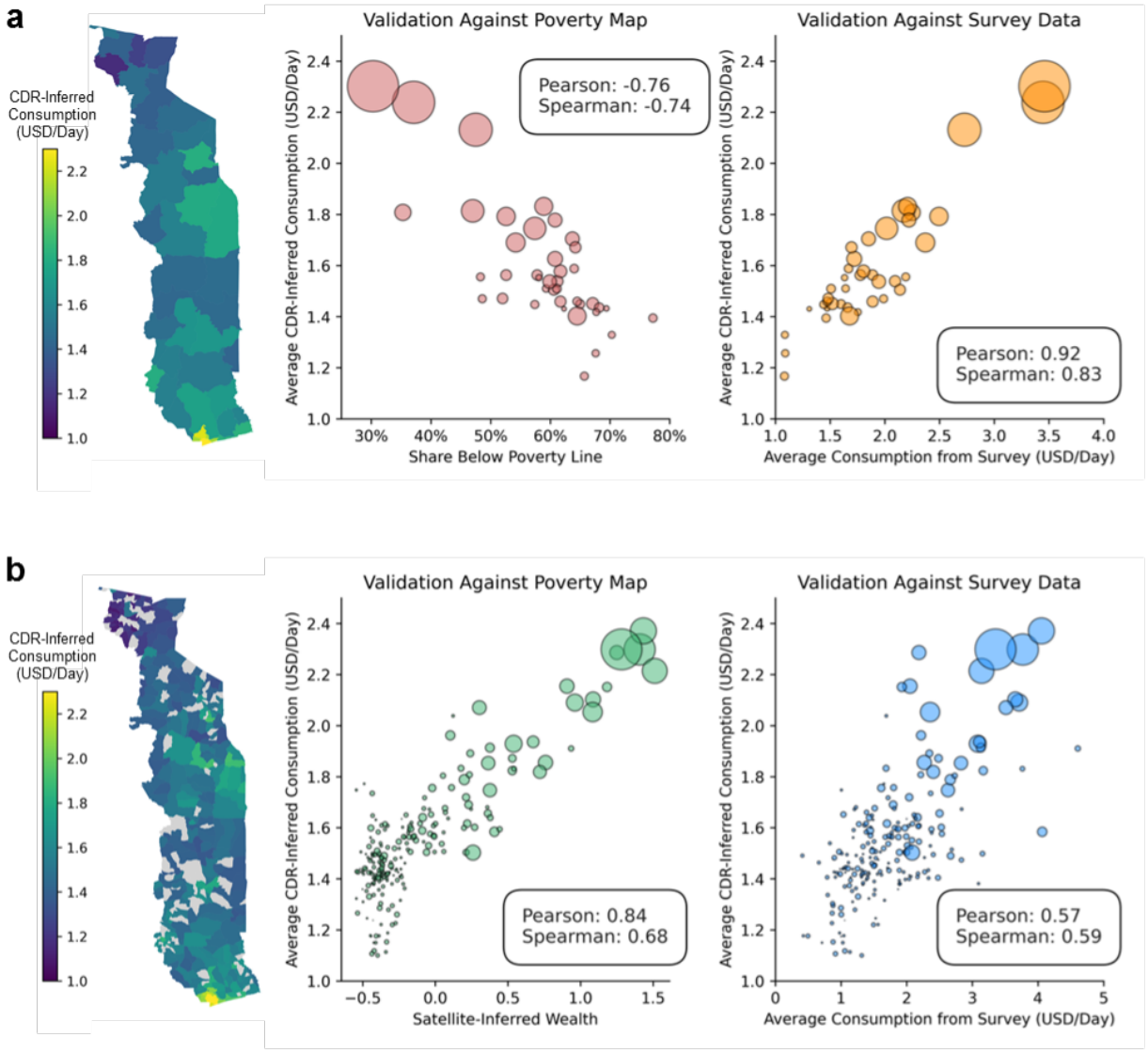

Figure S8: Spatial validation of phone-based poverty predictions. a) Map shows average phone-inferred consumption of subscribers in each prefecture (using CDR-based predictions trained on the 2018-19 in-person survey). Scatter plots compare average prefecture consumption, as derived from CDR (shown on y-axis), against two measures of poverty derived from the 2018-19 in-person survey (shown on x-axis): the share of people in the prefecture below the poverty line (middle plot), and the average consumption of households in the prefecture (right plot). b) Map shows average phone-inferred consumption of subscribers in each canton (cantons with no associated subscribers are shown in grey). Scatter plots compare average consumption per canton from the 2018-19 phone survey (evaluated across the 75% of all cantons in which there are observations in the 2018-19 field survey). Bubbles are sized by the number of subscribers assigned to each prefecture/canton.

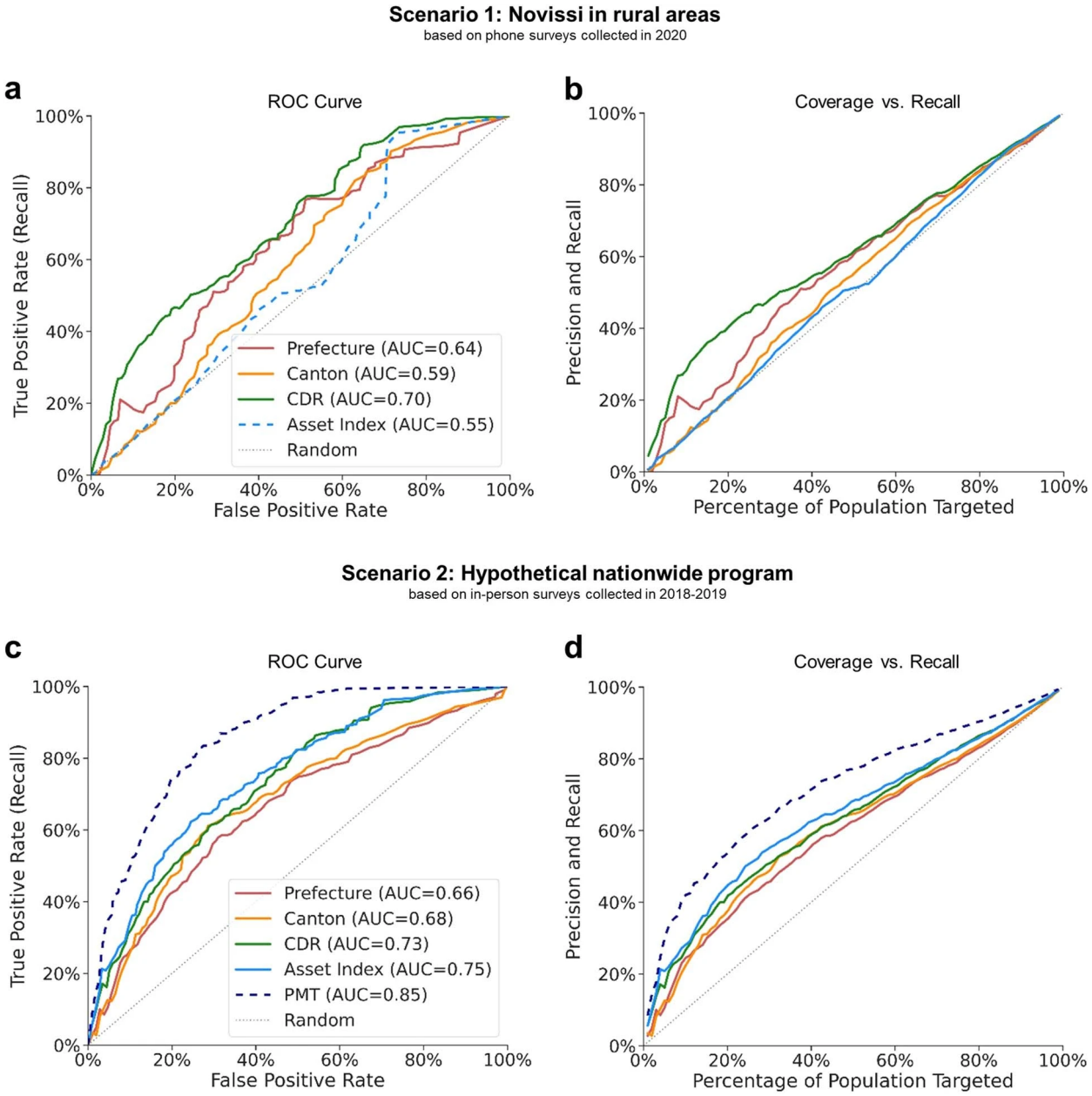

Figure S9: Top figures (a and b) show performance for the rural Novissi program, evaluated using 2020 phone survey. Bottom figures (c and d) correspond to the hypothetical national program, evaluated using the 2018–2019 field survey. ROC curves on left (a and c) indicate the true positive and false positive rates at different targeting thresholds. Coverage vs. Recall figures on right (b and d) show how precision and recall vary as the percentage of the population receiving benefits increases, i.e., they indicate the precision and recall for reaching the poorest K% of the population in programs that target the poorest K%. (Precision and recall are thus the same for each value of K by construction; see subsection 2.2.6).

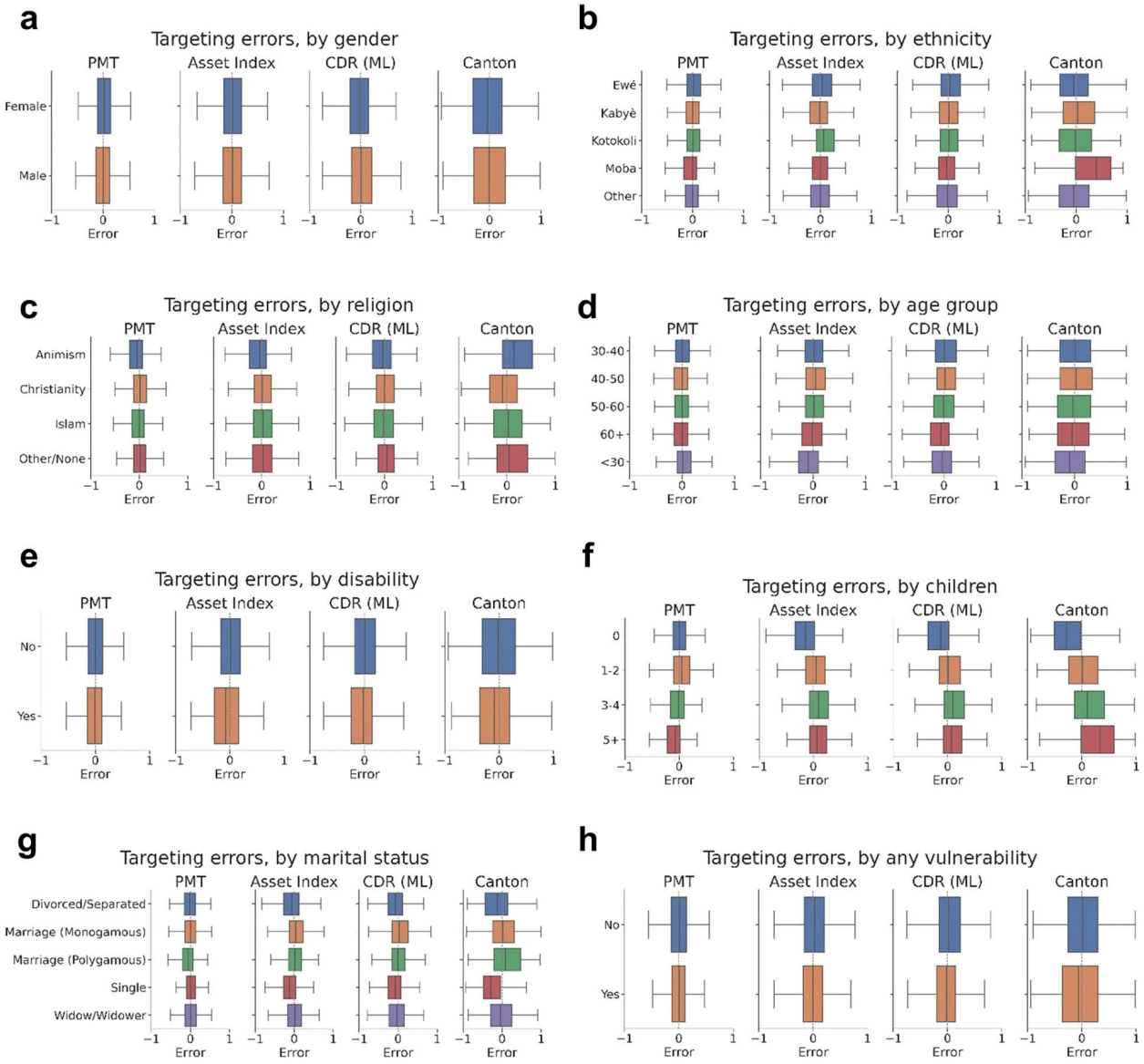

Figure S10: Boxplots showing distributions of normalized rank residuals (see subsection 2.3.1) aggregated by subgroup, using the 2018–2019 field survey dataset (N = 4,171). Boxes show the 25th and 75th percentiles, and the centre line shows the median of the distribution. Left-shifted boxes indicate groups that are consistently under-ranked by a given targeting mechanism, right-shifted boxes indicate groups that are consistently over-ranked by a given targeting mechanism.

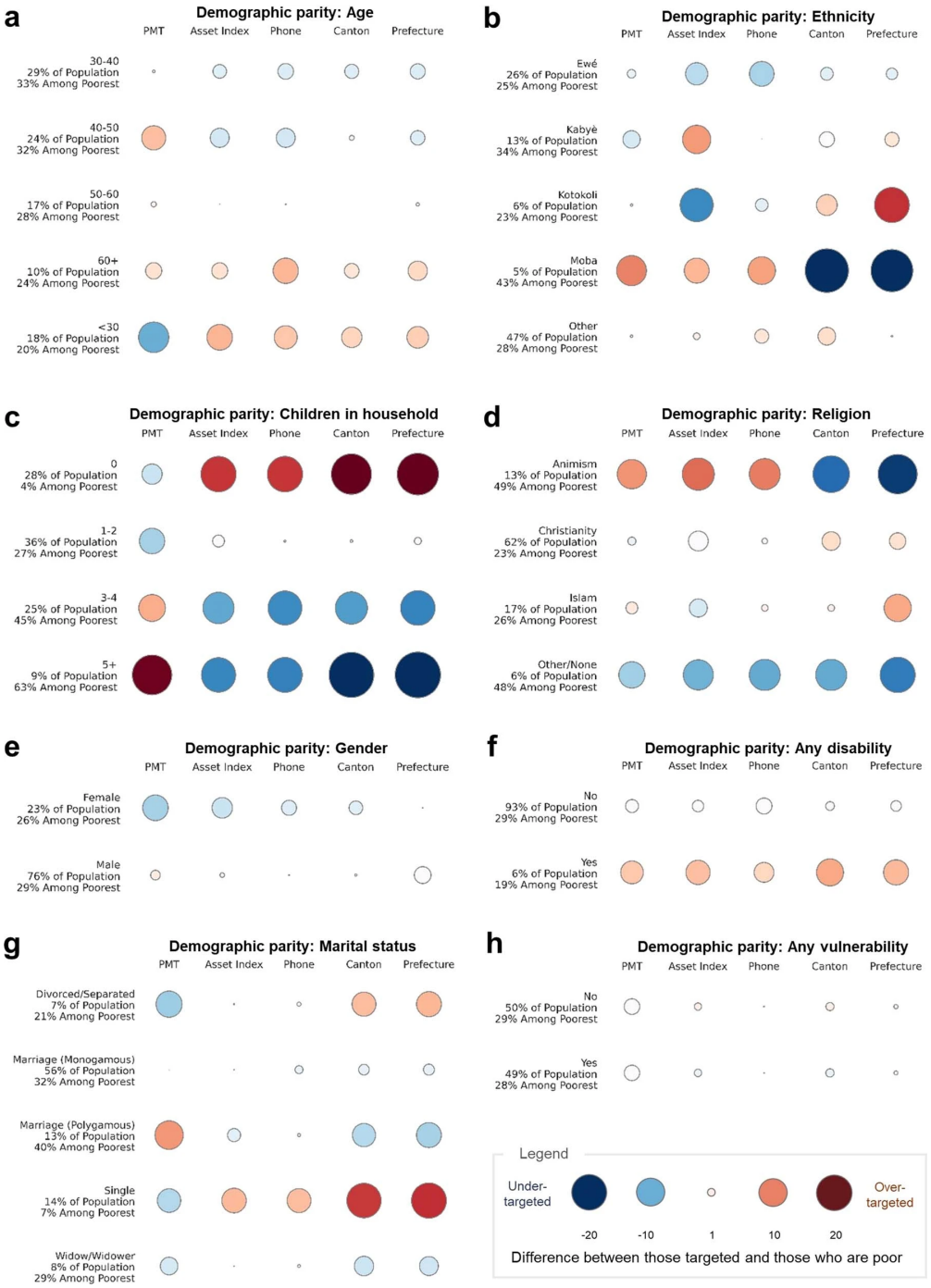

Figure S11: We evaluate demographic parity across subgroups by comparing the proportion of a subgroup targeted under counterfactual approaches to the proportion of the subgroup that falls into the poorest 29% of the population (using data from the 2018–2019 field survey matched to CDR, N = 4,171). Bubbles show the percentage point difference between the proportion of the subgroup that is targeted and the proportion that is poor according to ground-truth data. Large red bubbles indicate groups that are over-targeted; large blue bubbles indicate groups that are under-targeted.

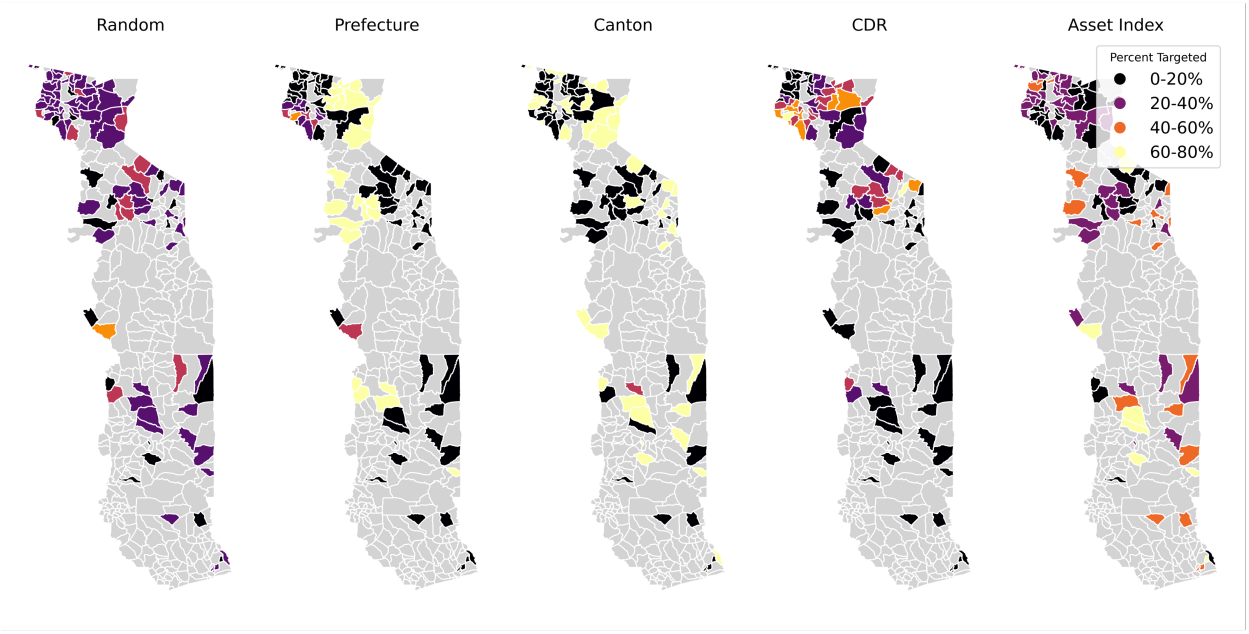

Figure S12: Share targeted by canton by different targeting methods. Panel A: Targeting share for the Novissi program in rural Togo, evaluated using individuals from the 2020 phone survey who report living in one of the 100 eligible cantons (N = 6,745). The respondent's self-reported canton and prefecture are used to color the map. Panel B: Targeting share for the hypothetical nationwide program, using data from the 2018-19 national household survey. Note that certain cantons have no observations in the 2018-2019 survey; these are shown in grey in Panel B. Cantons outside of the 100 poorest are shown in grey in Panel A.

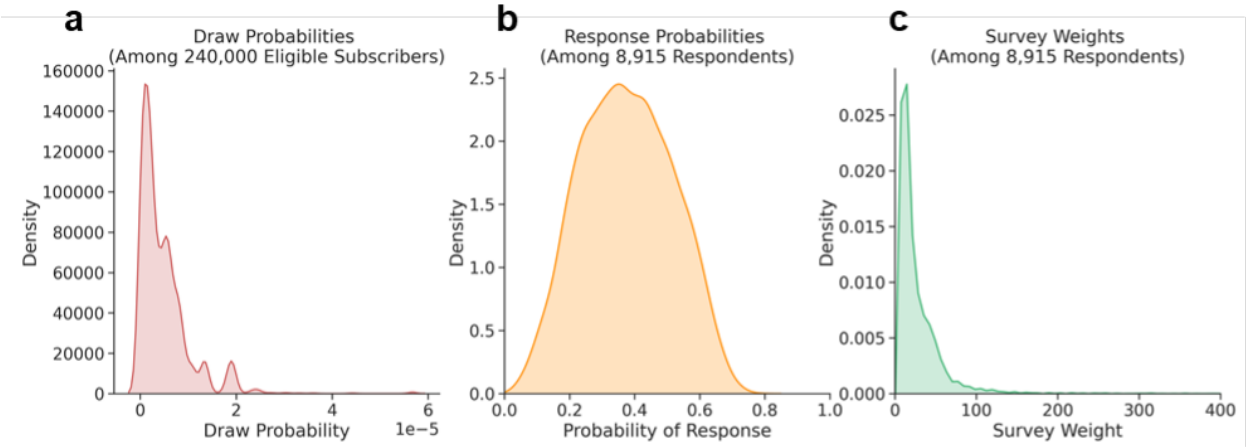

Figure S13: Distribution of sample weights for 2020 phone survey. Panel A: Distribution of draw probabilities among subscribers eligible for the survey. Panel B: Distribution of response probabilities for observations included in the final survey dataset, based on the response prediction model. Panel C: Distribution of sample weights (product of the inverse of the draw probability and the inverse of the response probability) for observations included in the final survey dataset.

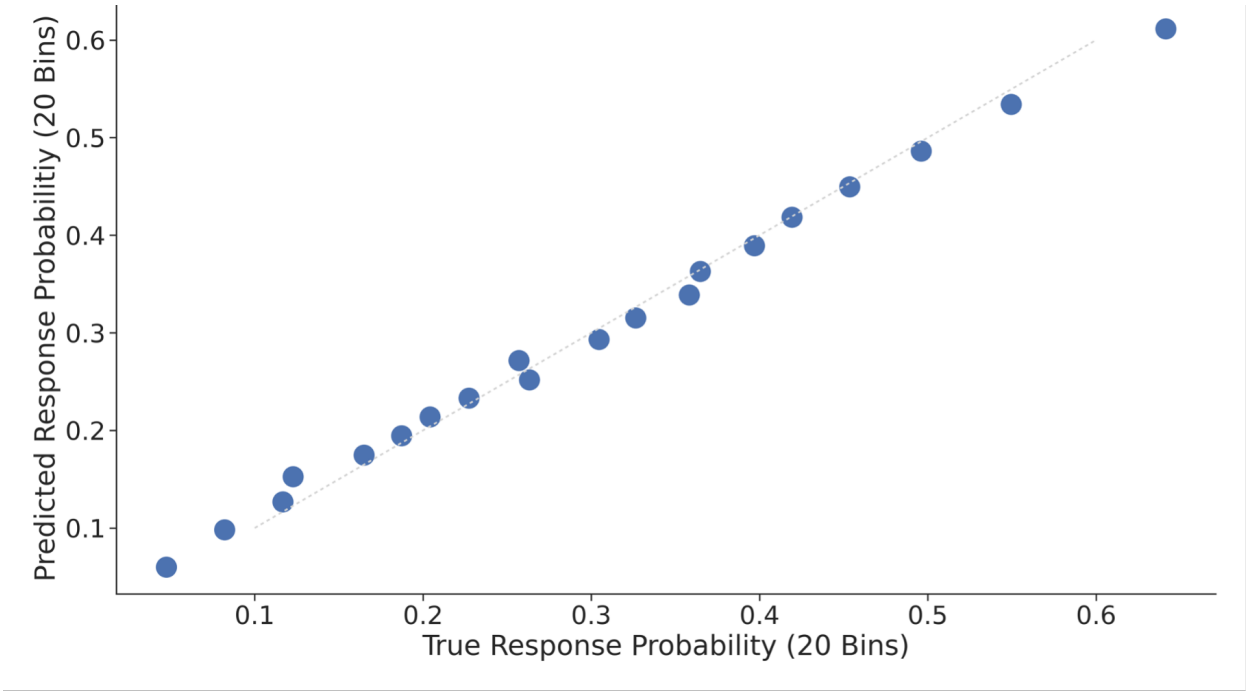

Figure S14: Calibration of response probabilities for 2020 phone survey. We compare the predicted probability of response (y-axis, binned into 20 quantiles) to the realized probability of response (x-axis, again binned into 20 quantiles) to confirm that the response prediction model is well-calibrated.

Table S1: Performance of targeting households below the *extreme* poverty line

	Targeting Novissi in rural Togo Based on 2020 Phone Survey (N = 8,915)			**Hypothetical nationwide program** Based on 2018-2019 Field Survey (N = 4,171)		
	Accuracy	**Precision**	**Recall**	**Accuracy**	**Precision**	**Recall**
Panel A: Targeting methods considered by the Government of Togo in 2020						
Prefecture	59%	61%	37%	67%	51%	44%
(Admin-2 regions)	(0.94%)	(1.49%)	(0.99%)	(0.73%)	(1.26%)	(1.09%)
Canton	54%	53%	32%	69%	54%	47%
(Admin-3 regions)	(0.86%)	(1.47%)	(0.91%)	(0.73%)	(1.26%)	(1.08%)
Phone	53%	50%	31%	64%	45%	39%
(Expenditures)	(0.85%)	(1.32%)	(0.90%)	(0.85%)	(1.46%)	(1.25%)
Phone	61%	64%	39%	69%	55%	48%
(Machine Learning)	(0.77%)	(0.94%)	(0.81%)	(0.73%)	(1.27%)	(1.09%)
Panel B: Common alternative targeting methods that could not be implemented in Togo in 2020						
Asset Index	53%	51%	31%	72%	60%	51%
	(0.54%)	(0.009)	(0.57%)	(0.71%)	(1.23%)	(1.05%)
PPI	[data not available]			76%	67%	57%
				(0.73%)	(1.26%)	(1.09%)
PMT	[data not available]			78%	70%	60%
				(0.70%)	(1.20%)	(1.03%)
Panel C: Additional counterfactual targeting methods that were feasible in Togo in 2020						
Random	53%	51%	31%	56%	33%	28%
	(0.84%)	(1.31%)	(0.88%)	(0.81%)	(1.39%)	(1.20%)
Occupation	47%	41%	25%	54%	29%	25%
(As implemented)	(0.76%)	(1.17%)	(0.80%)	(0.55%)	(0.96%)	(0.82%)
Occupation	59%	61%	37%	71%	58%	50%
(Optimal)	(0.68%)	(1.61%)	(0.71%)	(0.74%)	(1.28%)	(1.10%)

Notes: Analysis is similar to that presented in Table 2.1, but targeting is evaluated on the extent to which each method (still targeting the poorest 29%) provides benefits to individuals consuming less than the international extreme poverty line, set at 75% of the international poverty line or USD $1.43 per person per day (53% of observations in the 2020 phone survey dataset and 41% of observations in the 2018-2019 field survey). Spearman correlation and AUC are not reported here as they do not depend on the classification threshold, and are thus identical to the values reported in Table 2.1.

Table S2: Performance of targeting households below the poverty line

	Targeting Novissi in rural Togo Based on 2020 Phone Survey (N = 8,915)			**Hypothetical nationwide program** Based on 2018-2019 Field Survey (N = 4,171)		
	Accuracy	**Precision**	**Recall**	**Accuracy**	**Precision**	**Recall**
Panel A: Targeting methods considered by the Government of Togo in 2020						
Prefecture	47%	86%	34%	67%	51%	44%
(Admin-2 regions)	(0.67%)	(1.16%)	(0.46%)	(0.73%)	(1.26%)	(1.09%)
Canton	44%	80%	31%	69%	54%	47%
(Admin-3 regions)	(0.87%)	(1.51%)	(0.59%)	(0.73%)	(1.26%)	(1.08%)
Phone	41%	76%	30%	64%	45%	39%
(Expenditures)	(0.77%)	(1.32%)	(0.52%)	(0.85%)	(1.46%)	(1.25%)
Phone	48%	87%	34%	69%	55%	48%
(Machine Learning)	(0.76%)	(1.30%)	(0.51%)	(0.73%)	(1.27%)	(1.09%)
Panel B: Common alternative targeting methods that could not be implemented in Togo in 2020						
Asset Index	42%	77%	30%	72%	60%	51%
	(0.52%)	(0.89%)	(0.35%)	(0.71%)	(1.23%)	(1.05%)
PPI	[data not available]			76%	67%	57%
				(0.73%)	(1.26%)	(1.09%)
PMT	[data not available]			78%	70%	60%
				(0.70%)	(1.20%)	(1.03%)
Panel C: Additional counterfactual targeting methods that were feasible in Togo in 2020						
Random	39%	73%	29%	56%	33%	28%
	(0.76%)	(1.31%)	(0.51%)	(0.81%)	(1.39%)	(1.20%)
Occupation	38%	71%	28%	54%	29%	25%
(As implemented)	(0.77%)	(1.33%)	(0.52%)	(0.55%)	(0.96%)	(0.82%)
Occupation	46%	84%	33%	71%	58%	50%
	(0.61%)	(1.06%)	(0.42%)	(0.74%)	(1.28%)	(1.10%)

Notes: Analysis is similar to that presented in Table 2.1, but targeting is evaluated on the extent to which each method (still targeting the poorest 29%) provides benefits to individuals consuming less than the international poverty line of USD $1.90 per person per day (76% of observations in the 2020 phone survey dataset and 57% of observations in the 2018-2019 field survey). Spearman correlation and AUC are not reported here as they do not depend on the classification threshold, and are thus identical to the values reported in Table 2.1.

Table S3: Performance of targeting the hypothetical national program, when restricted to rural areas

Targeting a hypothetical nationwide program – but only in rural areas
Based on 2018 Phone Survey Restricted to Rural Areas (N = 2,306)

	Spearman	**AUC**	**Accuracy**	**Precision & Recall**
Panel A: Targeting methods considered by the Government of Togo in 2020				
Prefecture	0.16	0.57	64%	37%
(Admin-2 regions)	(0.023)	(0.011)	(0.97%)	(1.67%)
Canton	0.19	0.59	63%	36%
(Admin-3 regions)	(0.025)	(0.013)	(0.98%)	(1.69%)
Phone	0.15	0.59	63%	36%
(Expenditures)	(0.024)	(0.012)	(1.05%)	(1.81%)
Phone	0.30	0.65	67%	43%
(Machine Learning)	(0.023)	(0.012)	(1.00%)	(1.73%)
Panel B: Common alternative targeting methods that could not be implemented in Togo in 2020				
Asset Index	0.36	0.68	67%	44%
	(0.023)	(0.011)	(1.01%)	(1.74%)
PPI	0.55	0.77	72%	52%
	(0.017)	(0.009)	(1.07%)	(1.84%)
PMT	0.61	0.80	73%	54%
	(0.016)	(0.007)	(1.06%)	(1.84%)
Rural PMT	0.52	0.75	72%	51%
	(0.018)	(0.008)	(1.02%)	(1.75%)
Panel C: Additional counterfactual targeting methods that were feasible in Togo in 2020				
Random	0.00	0.50	59%	29%
	(0.024)	(0.012)	(1.04%)	(1.79%%)
Occupation	-0.13	0.44	54%	21%
(Novissi)	(0.024)	(0.011)	(0.89%)	(1.53%)
Occupation	0.31	0.63	63%	37%
(Optimal)	(0.023)	(0.010)	(0.62%)	(1.06%)

Notes: Analysis is similar to that presented in the last four columns of Table 2.1, but analysis is restricted to the 2,306 survey respondents (of the 4,171 total) who live in rural areas.

Table S4: Feature importances

Feature	Importance	Feature	Importance
Panel A: Predicting consumption, using 2018-2019 national household survey		*Panel B: Predicting a PMT, using 2020 phone survey*	
% in Tone	20	% in Tandjoare	31
% nocturnal calls	18	% in Doufelgou	21
# in Lome Commune	17	% in Cinkasse	18
% in Tandjoare	15	Mean data volume	15
% in Tchamba	14	% in Kpendjal-Ouest	14
% in Lome Commune	13	% in Agoe-Nyive	14
% in Agoe-Nyive	13	# in Kpendjal	14
SD call duration (weekends)	12	Median call duration (night)	13
Min time between calls (weekdays)	11	# in Golfe	11
Radius of gyration (night)	11	% in Keran	11
Panel C: Predicting a PMT, using 2018-2019 national household survey		*Panel D: Predicting a PMT, using 2018-2019 survey restricted to rural areas*	
% in Tchamba	25	% in Tchamba	29
% in Tandjoare	24	% in Tandjoare	22
% in Doufelgou	22	% in Doufelgou	22
% in Agoe-Nyive	20	% in Agoe-Nyive	22
% in Lome Commune	19	% in Kloto	21
# in Lome Commune	19	% in Tone	16
% in Tone	17	Radius of gyration	15
Radius of gyration (night)	16	% in Kpendjal-Ouest	12
Entropy of text contacts (day)	14	# in Dankpen	11
% in Tchaoudjo	13	SD churn rate	11

Notes: Feature importances for the 10 most important features selected by machine learning models trained to predict (a) Proxy Means Test from CDR, using a 2020 phone survey of mobile subscribers in Togo's 100 poorest cantons (N = 8,915); (b) consumption from CDR in the 2018-2019 field survey dataset (N = 4,171); (c) PMT from CDR in the 2018-2019 field survey dataset (N = 4,171), and (d) PMT from CDR in the 2018-2019 field survey dataset restricted to rural areas (N = 2,306). Feature importance is calculated based on the total number of times a feature is split upon in the prediction ensemble.

Table S5: How quickly does the accuracy of a phone-based targeting model degrade?

	Model	Phone data	Spearman	AUC	Accuracy	Precision	Recall
Panel A: Reaching the 29% Poorest							
(1) Best case	Current	Current	0.42	0.72	72%	51%	51%
			(0.019)	(0.010)	(0.73%)	(1.27%)	(1.27%)
(2) Old model	Old	Current	0.35	0.68	69%	46%	46%
			(0.019)	(0.010)	(0.75%)	(1.29%)	(1.29%)
(3) Old model and data	Old	Old	0.36	0.68	68%	44%	44%
			(0.019)	(0.010)	(0.74%)	(1.27%)	(1.27%)
(4) Geographic (Prefecture)			0.31	0.65	67%	43%	43%
			(0.020)	(0.009)	(0.89%)	(1.53%)	(1.53%)
(5) Geographic (Canton)			0.20	0.59	62%	34%	34%
			(0.023)	(0.011)	(0.83%)	(1.44%)	(1.44%)
Panel B: Reaching the extreme poor (48% of observations)							
(1) Best case	Current	Current	0.42	0.72	62%	68%	41%
			(0.019)	(0.01-)	(0.75%)	(1.30%)	(0.79%)
(2) Old model	Old	Current	0.35	0.68	60%	64%	38%
			(0.019)	(0.010)	(0.69%)	(1.30%)	(0.82%)
(3) Old model and data	Old	Old	0.36	0.68	60%	63%	38%
			(0.019)	(0.010)	(0.75%)	(1.29%)	(0.78%)
(4) Geographic (Prefecture)			0.31	0.65	59%	62%	38%
			(0.020)	(0.009)	(0.94%)	(1.61%)	(0.98%)
(5) Geographic (Canton)			0.20	0.59	54%	53%	32%
			(0.023)	(0.011)	(0.96%)	(1.65%)	(1.00%)
Panel C: Reaching the poor (74% of observations)							
(1) Best case	Current	Current	0.42	0.72	49%	90%	35%
			(0.019)	(0.01-)	(0.56%)	(0.97%)	(0.38%)
(2) Old model	Old	Current	0.35	0.68	47%	86%	34%
			(0.019)	(0.010)	(0.67%)	(1.16%)	(0.46%)
(3) Old model and data	Old	Old	0.36	0.68	47%	86%	34%
			(0.019)	(0.010)	(0.62%)	(1.08%)	(0.42%)
(4) Geographic (Prefecture)			0.31	0.65	48%	87%	34%
			(0.020)	(0.009)	(0.69%)	(1.18%)	(0.46%)
(5) Geographic (Canton)			0.20	0.59	44%	80%	31%
			(0.023)	(0.011)	(0.98%)	(1.69%)	(0.66%)

Notes: Table compares three scenarios: (1) "Best case": when the model is calibrated using survey data and phone data gathered just before deployment – these results are comparable to the paper's main analysis (slight differences are due to the sample restrictions described below); (2) "Old model": when the model is trained using a survey conducted two years before deployment, but the phone data are collected just before deployment; and (3) "Old model and data": when the phone-based wealth estimates are generated using survey and phone data from two years prior. Rows (4) and (5) show geographic targeting results using the same sample as in rows (1) – (3). In the simulations, the "old" data are from the 2018-19 national household survey and corresponding 2019 phone dataset; the 2020 phone survey PMT is used as the ground truth measure of poverty (restricted to respondents for whom CDR are available in 2019 and 2020, N = 7,064).

Table S6: Overlapping sources of exclusion from rural Novissi

Exclusion Source	*N*	Succeed	Drop Out	% Remaining	PMT	% Women	Age
Panel A: Attrition among overall population							
All	8,915	–	–	100.00%	1.62	23%	33.21
					(0.72)	(42%)	(11.91)
Own a voter ID	8,898	99.70%	0.30%	99.70%	1.62	23%	33.17
					(0.71)	(42%)	(11.81)
Attempt to register	5,145	45.48%	54.52%	45.34%	1.45	23%	33.30
					(0.57)	(42%)	(12.00)
Succeed in registration	4,092	76.84%	23.16%	34.84%	1.43	23%	33.05
					(0.54)	(42%)	(11.87)
Targeted by phone PMT	2,277	46.99%	53.01%	16.37%	1.28	21%	35.79
					(0.44)	(40%)	(11.96)
Panel B: Attrition among the poorest 29%							
All poor	3,209	–	–	100.00%	1.00	19%	36.22
					(0.15)	(39%)	(10.99)
Own a voter ID	3,207	99.77%	0.23%	99.77%	1.00	19%	36.16
					(0.15)	(39%)	(10.89)
Attempt to register	2,253	60.55%	39.45%	60.41%	0.99	20%	36.94
					(0.15)	(40%)	(11.19)
Succeed in registration	1,845	78.61%	21.39%	47.49%	0.99	19%	35.37
					(0.15)	(40%)	(11.03)
Targeted by phone PMT	1,257	60.56%	39.44%	28.76%	0.96	17%	36.67
					(0.15)	(37%)	(10.83)

Notes: Progressive sources of attrition from the rural Novissi program, where each row shows exclusion conditional on exclusions from preceding rows. The final three columns show characteristics of the population "succeeding" at each step. Panel A: Results estimated using the 2020 phone survey (N = 8,915). Panel B: Results estimated for just the poorest 29% from the 2020 survey (N = 3,209). There is no attrition based on mobile phone ownership or past phone use in this sample (in contrast to Table 2.2) since only active phone users were sampled for the phone survey. Values reweighted using sample weights.

Table S7: Asset-based wealth index

Asset	**Magnitude (2018-2019 Field Survey)**	**Magnitude (2020 Phone Survey)**
Electricity access	0.38	
Toilet	0.37	0.41
TV	0.35	
Electricity grid	0.35	
Garbage disposal	0.33	
Waste disposal	0.33	
Iron	0.26	0.06
Radio	0.20	0.23
Clean water (wet season)	0.16	
Clean water (dry season)	0.16	
Refrigerator	0.12	0.02
Walls	0.12	
Floor	0.11	
Mobile phone	0.11	
Water disposal	010	
Motorcycle	0.10	0.88
Computer	0.09	0.02
Roof	0.08	
Stove	0.07	0.06
Car	0.06	0.00
Tablet	0.01	0.00
Air conditioner	0.01	0.00
House	0.00	
Electricity (offgrid)	0.00	

Notes: Magnitude of first principal component for 2018-2019 field survey and 2020 phone survey.

Table S8: Proxy means test

Feature	β	**Feature (continued)**	β
Car	2.77	HHW Education 4	-0.18
Stove	1.77	Pref. Lacs	-0.18
Refrigerator	1.32	Pref. Sotouboua	-0.18
HHH Education 8	1.12	Pref. Kloto	-0.21
HHH Education 9	0.91	HHW Education 6	-0.21
HHH Hospitalization	0.81	Pref. Kpele	-0.21
Iron	0.63	Pref. Bas-Mono	-0.23
HHH Education 3	0.55	Pref. Lome Commune	-0.23
TV	0.50	Pref. Danyi	-0.24
All children in school	0.48	Pref. Yoto	-0.26
Pref. Cinkasse	0.39	Pref. Agoe-Nyive	-0.27
Pref. Tchamba	0.33	HHH Education 5	-0.27
Toilet	0.26	No children in school	-0.31
HHH Education 7	0.17	Pref. Assoli	-0.32
Pref. Est-Mono	0.14	Pref. Kpendjal-Ouest	-0.33
HHW Education 0	0.12	Pref. Zio	-0.33
Pref. Tchaoudjo	0.09	Pref. Amou	-0.34
Pref. Bassar	0.09	HHW Education 3	-0.34
Pref. Haho	0.07	Pref. Plaine du Mo	-0.34
Pref. Dankpen	0.04	Pref. Anie	-0.34
Pref. Moyen-Mono	-0.03	Pref. Tandjoare	-0.35
Pref. Oti-Sud	-0.06	Pref. Binah	-0.37
Pref. Oti	-0.08	Pref. Ave	-0.39
Pref. Wawa	-0.11	Pref. Keran	-0.41
Pref. Vo	-0.11	Pref. Kpendjal	-0.46
Pref. Ogou	-0.12	HHW Education 2	-0.50
Pref. Tone	-0.14	Pref. Kozah	-0.51
Pref. Agou	-0.15	HHH Education 2	-0.57
Pref. Akebou	-0.17	Pref. Blitta	-0.61
HHW Education 1	-0.17	HHH Education 1	-0.63
Some children in school	-0.17	Pref. Golfe	-0.68
Number of children	-0.17	Pref. Doufelgou	-0.75

Notes: Weights for linear model, trained on 2018-2019 phone survey (N = 6,171).

Table S9: Rural-specific proxy means test

Feature	β	**Feature (continued)**	β
Refrigerator	0.38	Pref. Akebou	0.03
HHH Hospitalization	0.32	Pref. Ogou	0.02
Motorcycle	0.31	Pref. Ave	-0.01
TV	0.28	Pref. Moyen-Mono	-0.03
Pref. Vo	0.26	Number of children	-0.08
Computer	0.24	Pref. Plaine du Mo	-0.08
Pref. Tchamba	0.21	Pref. Est-Mono	-0.10
Garbage removal	0.17	Pref. Dankpen	-0.12
Pref. Wawa	0.17	Pref. Binah	-0.13
Toilet	0.16	Pref. Tchaoudjo	-0.13
Pref. Kloto	0.16	Pref. Cinkasse	-0.14
Pref. Haho	0.16	Pref. Oti-Sud	-0.15
Pref. Yoto	0.14	Pref. Anie	-0.15
Pref. Bas-Mono	0.14	Pref. Oti	-0.18
Pref. Golfe	0.14	Pref. Kozah	-0.19
Pref. Kpele	0.14	Pref. Tone	-0.22
Pref. Lacs	0.13	Pref. Assoli	-0.22
Floor of solid materials	0.10	Pref. Blitta	-0.23
Pref. Zio	0.10	No children in school	-0.24
Pref. Lome Commune	0.09	Some children in school	-0.28
Pref. Agou	0.09	Pref. Doufelgou	-0.29
Roof of solid materials	0.08	Pref. Kpendjal-Ouest	-0.32
Pref. Bassar	0.08	Pref. Keran	-0.33
Pref. Amou	0.06	Pref. Kpendjal	-0.35
Pref. Danyi	0.06	Pref. Tandjoare	-0.40
Pref. Soutoubua	0.05		

Notes: Weights for linear model, trained on 2018-2019 phone survey restricted to rural areas (N = 3,895).

Table S10: Occupation categories

	2018-2019 Field Survey (N=6,171)			2020 Phone Survey (N=8,915)	
	Consumption	Proportion	N	Proportion	N
Intellectual Professions	$4.11 (3.55)	7%	277	7%	577
Intermediate Professions	$3.95 (3.56)	5%	197	3%	264
Administrators	$3.89 (3.57)	1%	32	0%	16
Managers and Directors	$3.70 (3.03)	3%	106	0%	36
Unemployed/Unknown	$3.19 (2.44)	8%	339	3%	275
Direct Services and Merchants	$2.75 (2.11)	23%	940	28%	2,111
Industry/Artisans	$2.47 (1.83)	15%	587	12%	1,026
Military Professions	$2.45 (1.25)	0%	17	1%	26
Elementary Professions	$2.21 (1.83)	2%	65	3%	249
Factory Workers	$2.17 (1.44)	7%	267	2%	165
Agricultural Professions	$1.53 (0.94)	29%	1,744	41%	4,170

Notes: Average daily per capita consumption per occupation category, with counts by category, separately for the 2018-2019 field survey and 2020 phone survey. Occupation categories for the 2018-2019 survey are for the household head, for the 2020 survey are for the individual respondent.

Table S11: Summary statistics for two survey datasets

	2018-2019 Household Survey					2020 Phone Survey
	Full Survey	Phone #	No Phone #	Phone #, Matched	Phone #, Unmatched	Full Survey
Consumption	2.39	2.56	1.75	2.59	2.21	[data not available]
	(2.41)	(2.38)	(2.41)	(2.42)	(1.78)	
PMT	2.10	2.22	1.65	2.23	2.03	1.62
	(1.43)	(1.47)	(1.16)	(1.47)	(1.38)	(0.72)
Occupation	56.42%	51.98%	72.94%	51.28%	59.63%	66.54%
(% Formal)	(49.59%)	(49.96%)	(44.43%)	(49.99%)	(49.08%)	(47.19%)
% Rural	51.93%	45.17%	77.12%	43.79%	60.17%	96.19%
	(49.96%)	(49.77%)	(42.01%)	(49.61%)	(48.97%)	(19.15%)
% Women	28.15%	23.61%	45.07%	23.43%	25.63%	23.27%
	(44.98%)	(42.47%)	(49.76%)	*42.36%)	(43.68%)	(42.25%)
Age	43.97	41.96	51.26	41.97	41.84	33.20
	(14.43)	(13.19)	(16.28%)	(13.15%)	*(13.71%)	(11.90)
N	6,089	4,571	1,518	4,171	400	8,915

Notes: Means and standard deviations for key outcomes in the 2018-2019 national household survey (N = 6,089) and 2020 phone survey concentrated in the 100 poorest cantons (N = 8,915). For the 2018-2019 national household survey, we break down the sample into two groups: households that provided enumerators with a phone numbers (N = 4,571) and those that do not (N = 1,518). We further break down the sample providing a phone number into two groups: households for which the phone number appears in data obtained from the mobile network operators (N = 4,171) and those for which it does not (N = 400). For the 2018-19 phone survey, occupation, gender, and age are assigned based on the head of household; for the 2020 phone survey they are assigned based on the respondent.

Table S12: Performance of phone-based approach to predicting wealth and consumption

	Consumption	**PMT**	**Asset Index**
Panel A: 2018-2019 Field Survey (N = 4,171)			
ML	0.46	0.62	0.74
Single Feature	0.13	0.16	0.11
Panel B: 2018-2019 Field Survey, Rural Only (N = 2,306)			
ML	0.31	0.44	0.51
Single Feature	0.09	0.12	0.08
Panel C: 2020 Phone Survey (N = 8,915)			
ML	–	0.41	0.40
Single Feature	–	0.13	0.14

Notes: Accuracy (Pearson correlation coefficients) for predicting poverty measures from CDR. ML predictions are produced over 5-fold cross validation and evaluated for pooled correlation. The "single feature" model estimates wealth and consumption based on the individual's total expenditures on calling and texting.

Table S13: Performance of targeting the hypothetical national program, with PMT as ground truth

Targeting a hypothetical nationwide program – with PMT as ground truth
Based on 2018-2019 National Household Survey (N = 4,171)

	Spearman	**AUC**	**Accuracy**	**Precision & Recall**
Panel A: Targeting methods considered by the Government of Togo in 2020				
Prefecture	0.50	0.73	72%	51%
(Admin-2 regions)	(0.014)	(0.006)	(0.73%)	(1.25%)
Canton	0.54	0.73	74%	55%
(Admin-3 regions)	(0.013)	(0.006)	(0.70%)	(1.22%)
Phone	0.31	0.64	66%	41%
(Expenditures)	(0.017)	(0.008)	(0.75%)	(1.30%)
Phone	0.56	0.78	73%	54%
(Machine Learning)	(0.014)	(0.006)	(0.73%)	(1.25%)
Panel B: Common alternative targeting methods that could not be implemented in Togo in 2020				
Asset Index	0.68	0.82	77%	60%
	(0.010)	(0.005)	(0.73%)	(1.26%)
PPI	0.74	0.86	81%	67%
	(0.009)	(0.005)	(0.68%)	(1.18%)
Panel C: Additional counterfactual targeting methods that were feasible in Togo in 2020				
Random	0.00	0.50	59%	30%
	(0.020)	(0.011)	(0.78%)	(1.34%%)
Occupation	-0.13	0.44	54%	21%
(Novissi)	(0.019)	(0.009)	(0.51%)	(0.88%%)
Occupation	0.50	0.73	76%	59%
(Optimal)	(0.015)	(0.006)	(0.71%)	(1.22%)

Notes: Analysis is similar to that presented in the last four columns of Table 2.1, but with the PMT as ground truth instead of consumption.

Table S14: Performance of targeting Novissi in rural Togo, when using the rural-specific PMT as ground truth

Targeting Novissi in rural Togo – with rural PMT as ground truth Based on 2020 Phone Survey (N = 8,915)				
	Spearman	AUC	Accuracy	Precision & Recall
Panel A: Targeting methods considered by the Government of Togo in 2020				
Prefecture	0.31	0.65	65%	40%
(Admin-2 regions)	(0.023)	(0.011)	(1.02%)	(1.76%)
Canton	0.19	0.60	62%	34%
(Admin-3 regions)	(0.025)	(0.012)	(1.03%)	(1.78%)
Phone	0.16	0.58	61%	33%
(Expenditures)	(0.023)	(0.012)	(1.02%)	(1.76%)
Phone	0.41	0.69	68%	46%
(Machine Learning)	(0.022)	(0.012)	(0.99%)	(1.70%)
Panel B: Common alternative targeting methods that could not be implemented in Togo in 2020				
Asset Index	0.46	0.71	68%	46%
	(0.021)	(0.011)	(0.99%)	(1.71%)
Panel C: Additional counterfactual targeting methods that were feasible in Togo in 2020				
Random	0.00	0.50	59%	29%
	(0.023)	(0.012)	(0.99%)	(1.70%)
Occupation	-0.12	0.45	55%	23%
(Novissi)	(0.024)	(0.011)	(0.93%)	(1.60%)
Occupation	0.26	0.61	65%	40%
(Optimal)	(0.022)	(0.010)	(0.66%)	(1.14%)

Notes: Analysis is similar to that presented in the first four columns of Table 2.1, but with the rural-specific PMT (as described in Methods §3.a) as ground truth.

Table S15: Geographic targeting with phone-inferred location.

	Targeting Novissi in rural Togo Based on 2020 Phone Survey (N = 8,915)				**Hypothetical nationwide program** Based on 2018-2019 Field Survey (N = 4,171)			
	Spearman	AUC	Accuracy	Precision & Recall	Spearman	AUC	Accuracy	Precision & Recall
Prefecture	0.30	0.64	65%	39%	0.34	0.66	68%	45%
(Survey-recorded)	(0.017)	(0.008)	(0.87%)	(1.51%)	(0.017)	(0.008)	(0.74%)	(1.27%)
Canton	0.19	0.59	61%	33%	0.39	0.68	70%	48%
(Survey-recorded)	(0.019)	(0.009)	(0.78%)	(1.35%)	(0.016)	(0.008)	(0.71%)	(1.23%)
CDR Prefecture	0.23	0.61	63%	36%	0.27	0.63	67%	44%
(Phone-inferred)	(0.016)	(0.008)	(0.76%)	(1.32%)	(0.017)	(0.008)	(0.74%)	(1.40%)
CDR Canton	0.12	0.56	58%	28%	0.31	0.65	69%	47%
(Phone-inferred)	(0.021)	(0.011)	(0.83%)	(1.43%)	(0.017)	(0.008)	(0.73%)	(1.27%)
Phone	0.38	0.70	69%	47%	0.45	0.73	71%	50%
(Machine Learning)	(0.017)	(0.009)	(0.87%)	(1.18%)	(0.015)	(0.007)	(0.74%)	(1.26%)

Notes: First two rows and final row replicate the results shown in Table 2.1. We add two additional counterfactual geographic targeting approaches based on location information derived from mobile phone data: targeting based on the average wealth of their home prefecture (row 3) or of their home canton (row 4). Home prefectures and cantons are inferred from outgoing mobile phone transactions [155]; the poverty of associated with each prefecture and canton is taken from the poverty maps shown in Figure S8.

Table S16: Correlation between sources of location data in 2020 phone survey

	Prefecture-level	**Canton-level**
Survey $\longleftrightarrow$ Voter	90.08%	69.77%
Survey $\longleftrightarrow$ Phone Data	70.08%	46.56%
Voter $\longleftrightarrow$ Phone Data	67.48%	44.89%

Notes: Correlation between the three sources of home location data available for observations in the 2020 phone survey: self-reported location collected in a survey, voter location recorded at the time of voter registration, and home location inferred from phone data. Each entry represents the percentage of observations (without sample weights applied) for which the two datasets agree on the respondent's location. Percentages are taken among the population (N = 4,515) for whom all three data sources are available (that is, individuals who were surveyed, whose phone numbers were registered for the rural Novissi program so that the canton and prefecture associated with their voter ID are included in Novissi administrative data, and who place at least one outgoing call between March to September 2020 so that their phone number is tied to a home prefecture and canton). This analysis cannot be carried out for the 2018-2019 field survey as fewer than 15% of the phone numbers collected in the survey registered for the rural Novissi program.

Table S17: Percentage of mobile phone activity initiated from a subscriber's home prefecture

	Share of phone transactions made from home prefecture (inferred from CDR)	**Share of phone transactions made from home prefecture (self-reported in survey)**
Panel A: 2018-2019 national household survey and April-June 2019 CDR		
Mean (and standard deviation)	75.46% (31.90%)	62.00% (40.05%)
Median	91.18%	81.84%
Mode	100.00%	100.00%
N	3,459,308	3,992
Panel B: 2020 phone survey and March-September 2020 CDR		
Mean (and standard deviation)	85.32% (18.78%)	68.00% (36.79%)
Median	94.00%	87.16%
Mode	100.00%	100.00%
N	5,615,393	8,183

Notes: Table indicates the fraction of outgoing calls and text messages that are routed through a cell tower in the subscriber's home prefecture. In the first column, "home location" is inferred from the subscriber's CDR [155]; in the second column, "home location" is recorded during a survey with the respondent. Panel A: results based on analysis from 2019, using CDR from three months in 2019 in the first column (N = 3,459,308), and survey respondents with known GPS coordinates from the 2018-2019 field survey in the second column (N = 3,992). Panel B: results based on analysis from 2020, using CDR from 7 months in 2020 in the left column (N = 5,615,393), and survey respondents with self-reported prefectures in the 2020 phone survey in the right column (N = 8,183).

Table S18: Response rates for 2020 phone survey

Group	Response Rate	N
Panel A: Previous Novissi registration		
Registered	37.82%	15,402
Unregistered	25.61%	14,085
Panel B: Phone-inferred region		
Lomé Commune	35.45%	189
Maritime	40.83%	1,254
Plateaux	30.17%	3,627
Centrale	31.91%	702
Kara	35.31%	6,582
Savanes	30.42%	17,034
Panel C: Phone-predicted poverty (USD/day)		
<$1.32	33.50%	7,372
1.32−1.42	33.55%	7,372
1.42−1.57	30.10%	7,371
$1.57+	30.79%	7,372
Panel D: Phone expenditures (USD/day)		
<$0.03	22.56%	7,372
0.03−0.08	28.15%	7,372
0.08−0.21	34.82%	7,371
$0.21+	42.66%	7,372

Notes: Response rate disaggregated by four dimensions: registration to a previous Novissi program (Panel A), region of Togo inferred from location of mobile phone transactions (Panel B), daily consumption inferred from mobile phone activity and machine learning (Panel C), and daily phone expenditures (Panel D).

Table S19: Feature importances for response reweighting model for 2020 phone survey.

Feature	**Importance**
Registered to previous Novissi program	15
Togocom subscriber	13
% nocturnal calls	10
% in Kpendjal	9
Active days	9
Mean balance of contacts	8
Median interactions per contact	7
Median time between calls (weekdays)	7
Active days (weekend)	6
Minimum time between calls	6

Notes: As described in appendix subsection A.2.4, the gradient boosting ensemble model is trained to predict the probability of response for a phone number drawn for the 2020 phone survey on the basis of pre-survey observable covariates (from CDR and previous Novissi registrations). Feature importance is calculated based on the total number of times a feature is split upon in the prediction ensemble.

Appendix B

Supporting materials for Chapter 3

B.1 Machine learning methods and hyperparameters

Although our work is focused on identifying the ultra-poor with CDR, we experiment with predicting four measures of ground-truth welfare with CDR features: ultra-poor status (binary), below the national poverty line (binary), asset index (continuous), and log consumption (continuous). For the binary measures, we experiment with four classification models: logistic regression (unregularized), logistic regression with L1 penalty, a random forest, and a gradient boosting model. For the continuous measures, we experiment with four regression models: linear regression, LASSO regression, a random forest, and a gradient boosting model. The linear models and random forest are implemented in Python's scikit-learn package. The gradient boosting model is implemented with Microsoft's LightGBM.

In each case, we produce predictions out-of-sample over 10-fold cross validation. We use nested cross-validation to tune the hyperparameters of each model over 5-fold cross-validation within each of the outer folds to avoid any information leakage between folds. We report both the mean score across the 10 folds as well as the overall score when data from all folds is pooled together. For the linear models and random forest, missing data is mean-imputed and each feature is scaled to zero mean and unit variance before fitting models (these transformations are done separately for each fold, with parameters fitted only on the training data for each fold). For the gradient boosting model missing values are left as-is and features are not scaled. We re-fit the model on the entire data, again tuning hyperparameters over 5-fold cross validation, to report selected hyperparameters and feature importances. We also report the top 5 features for each model, determined by the magnitude of the coefficient for the linear models, and by maximum impurity reductions for the tree-based models.

Hyperparameters are selected from the following grids for each model:

Linear/logistic regression

- Drop features where over X% of observations are missing data: X={50%, 80%, 100%}
- Drop features with variance under: {0, 0.01, 0.1}
- Winsorization limit: {0%, 1%, 5%}

LASSO regression

- Drop features where over X% of observations are missing data: X={50%, 80%, 100%}
- Drop features with variance under: {0, 0.01, 0.1}
- Winsorization limit: {0%, 1%, 5%}
- L1 penalty: {0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 10, 100}

Random Forest

- Drop features where over X% of observations are missing data: X={50%, 80%, 100%}
- Drop features with variance under: {0, 0.01, 0.1}
- Winsorization limit: {0%, 1%, 5%}
- Number of Trees: {20, 50, 100}
- Maximum Depth: {1, 2, 4, 6, 8, 10, 12}

Gradient boosting model

- Drop features where over X% of observations are missing data: X={50%, 80%, 100%}
- Drop features with variance under: {0, 0.01, 0.1}
- Winsorization limit: {0%, 1%, 5%}
- Number of Trees: {20, 50, 100}
- Minimum data in leaf: {5, 10}
- Number of leaves: {5, 10, 20}
- Learning rate: {0.05, 0.075}

B.2 Abbreviations in feature names

Figure S4, Table S8, Table S2, and Table S7 use a set of abbreviations in CDR feature names. This appendix lists the relevant abbreviations.

- BOC: Balance of contacts
- CD: Call duration
- IPC: Interactions per contact
- IT: Interevent time
- NOI: Number of interactions
- PPD: Percent pareto durations (percentage of call contacts accounting for 80% of call time)
- PPI: Percent pareto interactions (percentage of contacts accounting for 80% of subscriber's interactions)
- RD: Response delay
- RR: Response rate
- WD: Weekday
- WE: Weekend

B.3 Cost and speed calculations

In the discussion section we provide a cost and speed comparison between targeting methods, as some of the value-add of the phone-based targeting approach relies on how cheap and quick it is compared to asset, consumption- or CBT-based targeting approaches. Administrative data on targeting costs was not collected as part of the TUP program, so we turn to other studies of program targeting to estimate the costs of CBT and asset-based (or PMT) methods. We treat the costs of an asset index-based and PMT approach as equivalent in this section, as they both require comprehensive household surveys.[1] We identify three studies that provide variable targeting costs for PMT and CBT methods: [11] provide variable costs for CBT and PMT-based targeting of a single program in Indonesia; [101] provide variable costs for CBT and PMT-based targeting of an ultra-poor program in Honduras and one in Peru; and [140] provide variable costs

[1] In practice, an asset-based approach may be slightly cheaper than a PMT, as it does not require conducting a consumption module for a subset of surveys to train a PMT.

for three CBT-based programs and four PMT-based programs in seven countries in Sub-Saharan Africa.[2] Table S10 summarizes the cost estimates from each of these papers; we use the median per-household targeting cost for each method in our analysis ($2.20 per household for CBT and $4.00 per household for PMT). While using these global estimates to inform our model of targeting costs in Afghanistan is not ideal, since no data on targeting costs from the TUP program or other anti-poverty programs is available for the country, these values are the best available estimate on which to base our cost analysis.

We are unable to find any papers that document the targeting cost associated with consumption-based targeting, as consumption data is rarely used as a real-world targeting strategy. We therefore consider the costs of consumption to be strictly greater than the costs of targeting on a PMT, since consumption modules take longer to collect than PMT data in household surveys. In practice, we expect that the cost of targeting on consumption would be substantially greater than the cost of targeting on a PMT.

For phone-based targeting, we associate no cost with the collection and analysis of phone data. While in some cases phone data may require purchase from the operator, partnerships between mobile network operators and governments for social protection and public health applications have not, to date, involved payment [117]. The fixed cost of mobile data analysis is non-negligible but its contribution to marginal cost is close to zero as the number of screened households increases. A phone-based targeting method that collects informed consent from program applicants to analyze phone data would have nonzero marginal costs, though the cost of consent would depend on the modality of consent collection. If consent was collected in person, these costs would likely be only slightly lower than those of a PMT, as every household would need to be surveyed in person. If consent was collected over the phone via SMS or voice, these costs would likely be significantly lower.

It is worth noting that our benchmark in this chapter is the hybrid model with a CBT plus verification component, but due to limited estimates in the literature we leave this strategy out of our cost analysis. We consider the CBT a lower-bound estimate for the hybrid strategy, and therefore our results would be qualitatively unchanged if the hybrid strategy were also considered in cost comparison. [11] suggest that there are synergies in targeting approaches so that combining approaches is less costly than the sum of the costs of the two approaches individually, but costs are certainly greater than that of CBT targeting alone.

Our cost analysis finally relies upon administrative data from the TUP program. The TUP program in its entirety served 7,500 households across six provinces of Afghanistan. While there is no data available on the total number of households screened by the TUP program, the portion of the program in Balkh province that was enrolled in the RCT identified 1,235 ultra-poor households out of 20,702 households screened [32]. Assuming similar eligibility rates across Afghanistan, we estimate that the TUP program as a

[2]To our knowledge, no studies incorporate fixed targeting costs, as these are typically indistinguishable from fixed costs of other components of program set-up.

whole likely screened around 125,721 households. We use this value to estimate total targeting costs for the TUP program under counterfactual targeting approaches. Eligible households received benefits totaling $1,688, including a productive asset, cash transfers, a health voucher, training, biweekly social worker visits, and veterinarian visit once every two months during the year of intervention. The total benefits dispersed by the program were therefore on the order of 12.7 million (although the total program costs, including overhead, were closer to 15 million [32]); we use the total value of benefits to compare the costs of program targeting using our set of counterfactual targeting approaches to the direct costs of program benefits in Table S9. We find that targeting costs for a PMT or asset-based approach would represent approximately 3.97% of the total benefits delivered in the program; costs for a CBT approach would represent approximately 2.18% of the total benefits. In comparison, costs for the phone-based approach would be negligible.

When it comes to speed, in-person data collection for an asset-based (or PMT) targeting approach typically takes months or years to prepare and implement [156]. The CDR-based approach can be rolled out comparatively quickly — but there are still practical hurdles to implementation. First, training data for the CDR-based poverty prediction model must be collected, preferably shortly prior to program roll-out [5]. While in the TUP project training data was collected in-person in a household survey, in other contexts training data collection was expedited via a phone survey [36, 5]. Even if data is collected over the phone, it will typically take several weeks to design a survey instrument and collect data. Second, the CDR-based method requires data from mobile network operators. Data sharing agreements with mobile network operators take at minimum a few weeks to arrange, and substantially longer in the worst case [117]. Third, and finally, training a CDR-based poverty prediction model is expensive in terms of memory, computing power, and human capacity, and will likely take several weeks to implement.

B.4 Supplementary figures and tables

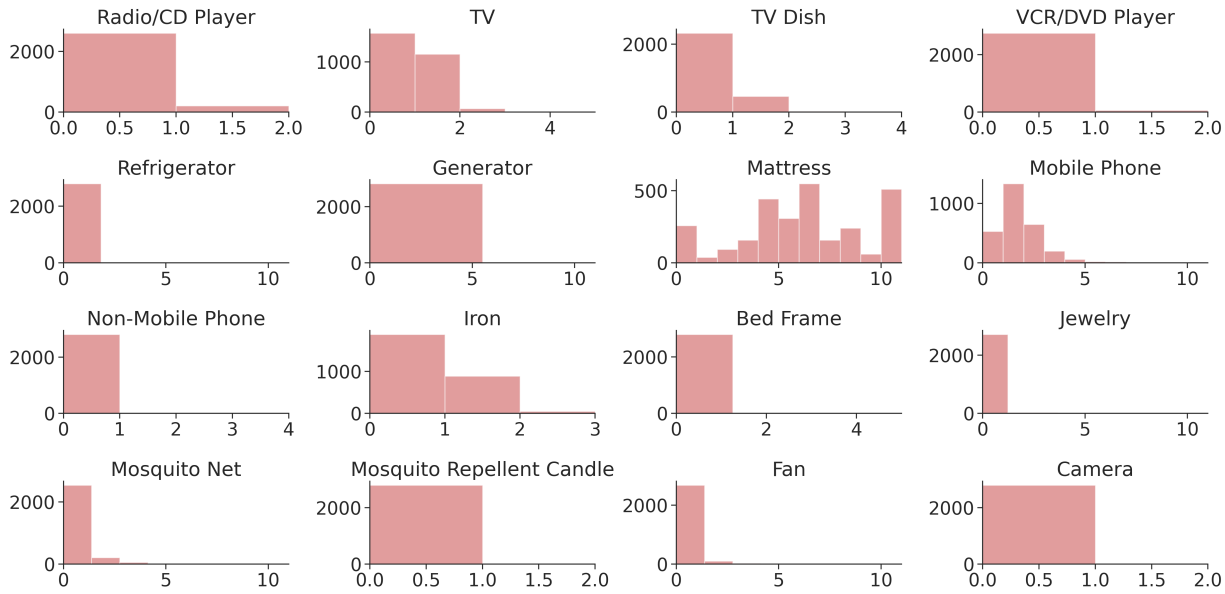

Figure S1: Histograms showing the distribution of each underlying asset used to construct the asset index.

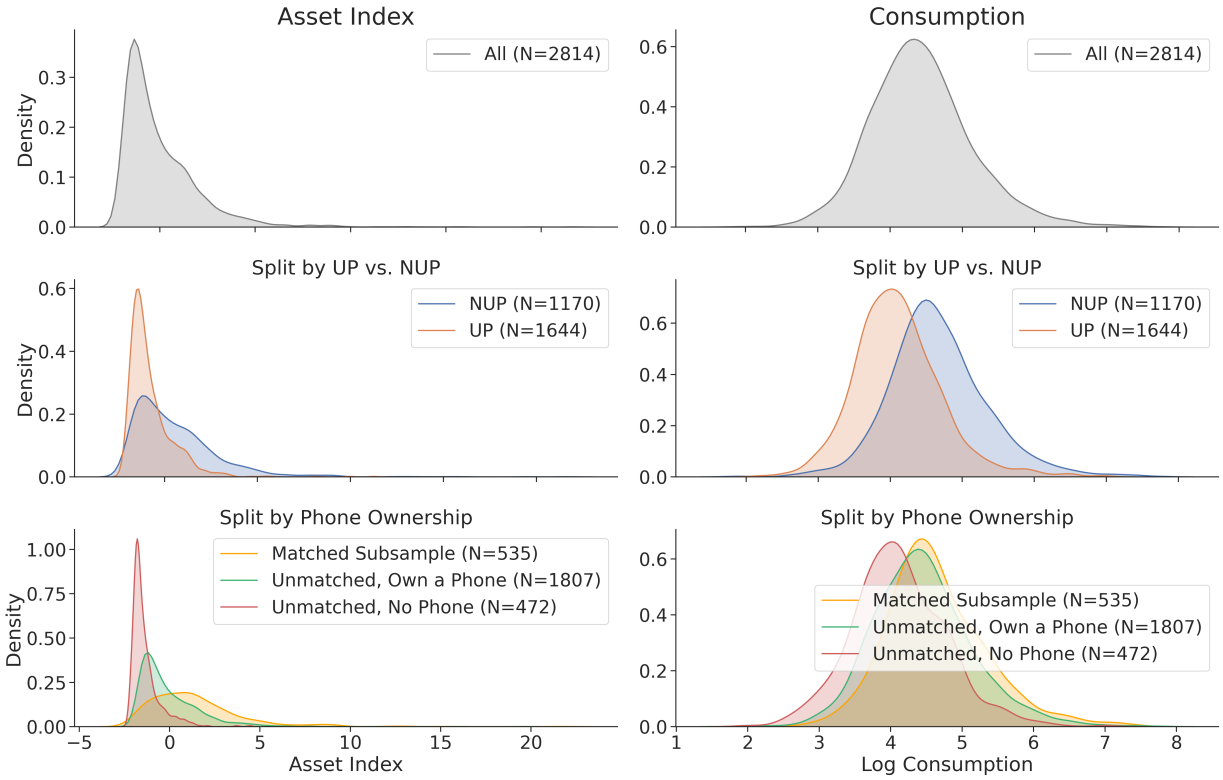

Figure S2: Distributions of asset index and log-transformed consumption, for the entire survey sample, separately for ultra-poor and non-ultra-poor households, and again separately for households in the subsample matched to CDR, households outside of the matched subsample that report owning at least one mobile phone, and households outside of the matched subsample that report not owning a mobile phone.

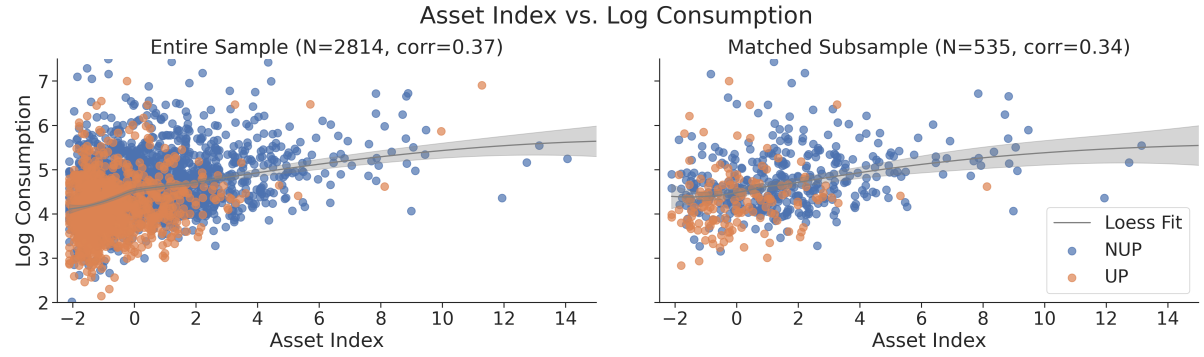

Figure S3: Correlation between asset index and log-transformed consumption, separately for the entire survey sample and the matched subsample. We include the LOESS fit, along with a 95% confidence interval.

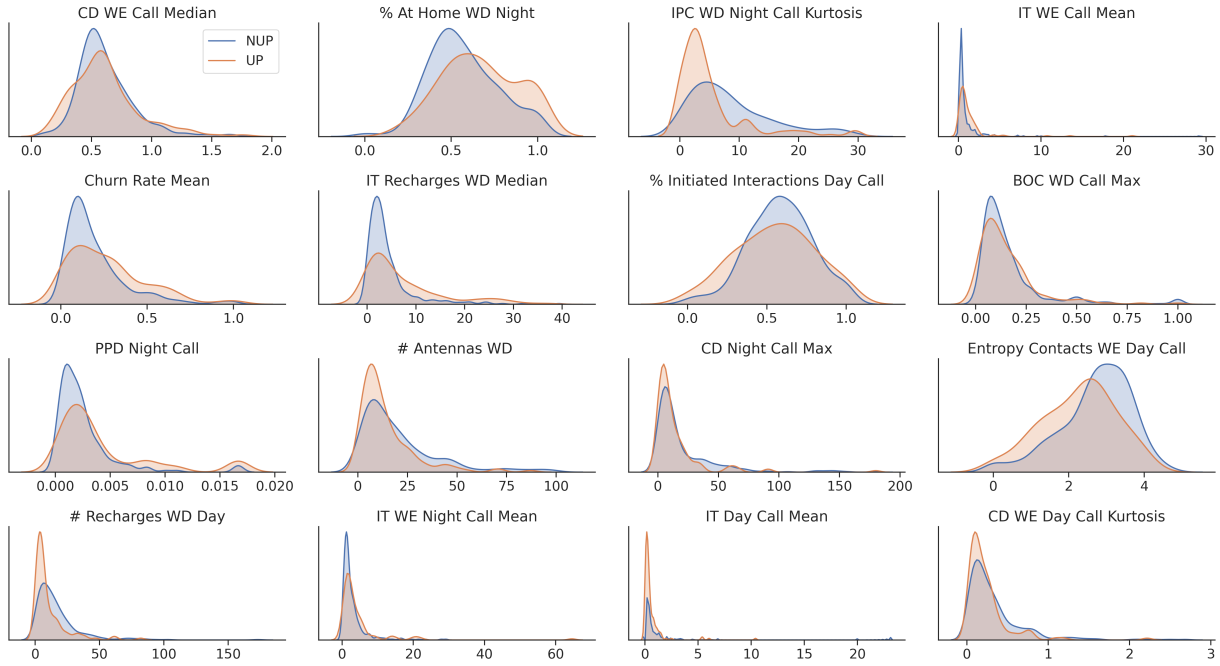

Figure S4: Kernel density estimates for 16 of the most important features for predicting ultra-poor status from CDR, with density estimates shown separately from UP and NUP households. Since many features are near-redundant, rather than showing the raw top 16 features from the table above, we show 16 selected features from the top 50. See appendix B.2 for abbreviations in feature names.

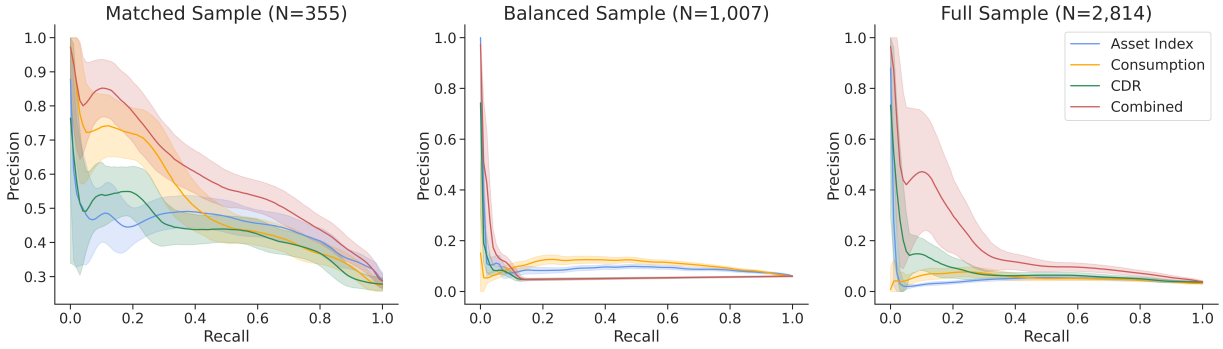

Figure S5: Precision-recall curves for each of the four targeting methods. In the third subplot, the CDR-based and combined methods target non-phone-owning households first as described in subsection 3.2.5.

Table S1: Direction of first principal component of asset ownership

Asset	**Magnitude**
Radio/CD Player	0.04
TV	0.37
TV Dish	0.29
VCR/DVD Player	0.15
Refrigerator	0.25
Generator	0.11
Mattress	0.24
Mobile Phone	0.31
Non-Mobile Phone	0.06
Iron	0.36
Bed Frame	0.29
Jewelry	0.27
Mosquito Net	0.26
Mosquito Repellent Candle	0.08
Fan	0.37
Camera	0.16

Notes: The asset index is calculated over the entire 2,814 household sample, without sample weights. We standardize each of the features to zero mean and unit variance before decomposition. The first principal component accounts for 25.28% of the variation in these standardized features.

Table S2: Feature importances (gradient boosting model)

Feature	Importance	Feature	Importance
CD WE Call Median	8	IT Recharges Night Min	3
% At Home WD Night	7	BOC WD Call Median	3
IPC WD Night Call Kurtosis	7	IT WE Call Min	2
CD Day Call Median	6	BOC WD Night Call Kurtosis	2
IT WE Call Mean	6	IPC Day Call Kurtosis	2
Churn Rate Mean	5	% Nocturnal WD Call	2
IPC Day Call Skew	5	IT WD Day Text Mean	2
IT Recharges WD Median	5	CD Night Call Max	2
% At Home Day	4	IT WE Call Skew	2
% Initiated Interactions Day Call	4	IPC WE Day Call Kurtosis	2
% Initiated Interactions WD Day Call	4	IT WE Text Median	2
BOC WD Call Max	4	% At Home WE Night	2
% Initiated Interactions WD Night Call	3	Entropy Contacts WE Day Call	2
PPD Night Call	3	# Recharges WD Day	2
IT Recharges WD Night Min	3	Entropy Antennas WD	2
IT Recharges Night Median	3	IPC Night Call Skew	2
IPC WD Night Text Mean	3	IT WE Night Call Mean	2
IT Recharges Day Kurtosis	3	# Contacts Day Call	2
IT Night Text Min	3	CD WD Call Max	2
IT WE Day Text Median	3	IT Day Call Mean	2
# Antennas WD	3	IT WD Night Text Min	2
CD WD Night Call Kurtosis	3	Entropy Antennas Day	2
IPC Night Call Kurtosis	3	% Initiated Interactions WE Day Call	2
IPC WE Night Call Kurtosis	3	CD WE Day Call Kurtosis	1
IPC WE Night Call Skew	3	IPC Day Call Std	1

Notes: For our selected machine learning model – the gradient boosting model used to predict ultra-poor status from CDR features – we display feature importances for the top 50 features. Feature importances for the gradient boosting model represent the total number of times the feature is used for a split in the entire ensemble of decision trees. We report feature importances when the model is trained on all 535 observations (rather than over cross validation). See appendix B.2 for abbreviations in feature names.

Table S3: Details of machine learning models

Model	AUC	Top Five Features
Logistic (No Penalty)	0.53	Reporting # Records, Active Days, Active Days Day, Active Days Night, Active Days WD
Logistic (L1 Penalty)	0.66	Reporting # Records, Active Days, Active Days Day, Active Days Night, Active Days WD
Random Forest	0.68	NOI Out Day Call, NOI Out WD Day Call, Nois Call, Entropy Contacts Night Call, NOI Out WE Call
Gradient Boosting	0.68	CD WE Call Median, % At Home WD Night, IPC WD Night Call Kurtosis, CD Day Call Median, IT WE Call Mean

Notes: Each row indicates performance (AUC) of a different machine learning algorithm, trained to predict ultra-poor status on the sample of 535 matched households. AUC is reported as the mean AUC score over 10-fold cross validation. See appendix B.2 for details of features.

Table S4: Machine learning an asset index

Model	AUC Score	Top Five Features
Logistic (L1 Penalty)	0.60	TV, TV Dish, Fridge, Mattress, Mobile Phone
Random Forest	0.73	Fridge, Iron, Bedframe, Mattress, TV Dish
Gradient Boosting	0.74	Mattress, Bedframe, Fridge, Mobile Phone, TV Dish

Notes: The asset index benchmark we used is constructed following standard procedures based on principal comnponent analysis (see Table S1). However, it is possible that an alternative asset-based predictor, trained using machine learning to predict ultra-poor status directly from the 16 underlying components, could perform better. We test this hypothesis by adapting our machine learning pipeline for identifying the ultra-poor from CDR to the task of identifying the ultra-poor from asset possession. As with the CDR-based prediction, we evaluate the model over nested cross validation in our 535-household matched sample: the model's predictions are evaluated out-of-sample over 10-fold cross validation, and within each fold hyperparameters are tuned over 5-fold cross validation. We retrain the model on the entire dataset to report hyperparameters and feature importances. Hyperparameters are chosen from the same grid as for the CDR-based models. We display the AUC score and top features for each model.

Table S5: Performance using one, two or three predictor datasets

Data Sources	**AUC**
Assets	0.73 (0.025)
Consumption	0.71 (0.000)
CDR	0.68 (0.028)
Assets + Consumption	0.76 (0.017)
Assets + CDR	0.76 (0.025)
Consumption + CDR	0.75 (0.016)
Assets + Consumption + CDR	0.78 (0.019)

Notes: AUC scores for targeting methods using a single data source, pair of data sources, and all three data sources together (in our 535-household matched sample). Standard deviations are calculated from 1,000 bootstrapped samples of the same size as the original sample, drawn with replacement.

Table S6: Targeting simulation results for one train-test split

	(1)	(2)	(3)	(4)
Targeting Method	**AUC**	**Accuracy**	**Precision**	**Recall**
Random	0.50	0.48	0.28	0.28
Asset Index	0.68	0.63	0.33	0.33
Consumption	0.74	0.74	0.53	0.53
CDR	0.75	0.70	0.47	0.47
Combined	0.82	0.74	0.53	0.53

Notes: Reproduction of main results (Table 3.2 Panel A) to using a single train-test split (for our 535-household matched sample, with 10% of the observations in the test set).

Table S7: Matching household to multiple phone numbers

Model	**AUC**	**Top Five Features**
Logistic (No Penalty)	0.50	Reporting # Records, Active Days, Active Days Day, Active Days Night, Active Days WD
Logistic (L1 Penalty)	0.65	Reporting # Records, Active Days, Active Days Day, Active Days Night, Active Days WD
Random Forest	0.67	NOI call, NOI Out WE Call, IPC WD Night Call Kurtosis, IPC Night Call Kurtosis, IT Recharges WD Day Min
Gradient Boosting	0.66	Churn Rate Std, CD WE Call Median, IPC WD Night Call Kurtosis, IPC Day Call Skew, % Initiated Interactions Day Call

Notes: In our main analysis, for multi-phone households we use only the phone number belonging to the household head (or to a random household member, where no household head is specified), leaving 535 household-level observations. Here we consider instead using machine learning methods to predict individual-level ultra-poverty, with a dataset of 634 individual phone numbers matched to the ground-truth wealth measures for the associated households. We find that the individual-level models are slightly less accurate than the household-level models presented in the main paper, but we focus on the household-level models in the main paper since the household was the unit of targeting in the TUP program. See appendix B.2 for abbreviations in feature names.

Table S8: Predicting other measures of poverty from CDR

Model	R^2 or AUC	Top Five Features
Panel A: Predicting below poverty line (binary)		
Logistic (No Penalty)	0.53	Reporting # Records, Active Days, Active Days Day, Active Days Night, Active Days WD
Logistic (L1 Penalty)	0.53	Reporting # Records, Active Days, Active Days Day, Active Days Night, Active Days WD
Random Forest	0.56	NOI Out Night Call, BOC Night Call Kurtosis, CD Day Call Skew, Nois Night Call, IT Night Call Kurtosis
Gradient Boosting	0.55	IT Night Call Kurtosis, IT Text Max, Radius Gyration WE Night, Entropy Antennas, NOI Out WD Call
Panel B: Predicting consumption (continuous)		
Linear Regression	-0.21	% Pareto Recharges WE Night, % Pareto Recharges WE, % Pareto Recharges Night, Entropy Contacts WD Day Text, PPI WE Night Text
LASSO Regression	-0.00	Reporting # Records, PPI Text, PPI Day Text, PPI Night Call, PPI Night Text
Random Forest	-0.02	Churn Rate Mean, IPC WE Night Call Kurtosis, IT Recharges WE Day Skew, IPC WE Night Call Skew, CD WE Call Median
Gradient Boosting	-0.03	CD WD Night Call Skew, IPC WD Day Text Skew, IT WD Night Call Min, IT WD Night Call Max, IT WE Night Call Max
Panel C: Predicting asset index (continuous)		
Linear Regression	-0.06	IPC Text Min, IPC WD Text Min, IPC WD Day Text Min, BOC WD Text Min, % Initiated Conversations WD
LASSO Regression	0.00	Active Days WE Day, Active Days WD, Active Days WE, Active Days, Active Days WD Day
Random Forest	0.00	IT Night Call Skew, IPC Text Min, IT WE Day Call Median, IT WE Call Median, Entropy Contacts WE Night Call
Gradient Boosting	-0.02	IT Text Median, Entropy Antennas WE, Entropy Antennas WD Night, Entropy Contacts WE Night Call, IT Recharges Night Min
Panel D: Predicting CWR group (continuous)		
Linear Regression	0.01	PPI Night Text, IT Recharges Day Skew, IPC WE Call Min, Active Days WE Night, IT Recharges WD Day Skew
LASSO Regression	0.05	PPI Night Text, Active Days WE Day, Active Days WE Night, IT Recharges WD Day Skew, IT Recharges Day Skew
Random Forest	0.04	# Contacts WE Day Call, Entropy Contacts WD Night Call, IPC Night Call Kurtosis, # Contacts WE Call, IT Call Kurtosis
Gradient Boosting	0.03	IT Call Kurtosis, IT Recharges Day Skew, # Contacts WE Day Call, IT Recharges Day Kurtosis, IPC WD Night Call Kurtosis

Notes: Machine learning results for predicting: (A) Below-poverty-line status, using consumption data and based on Afghanistan's national poverty line; (B) Total consumption (log-scale); (C) Asset index; and (D) Community Wealth Ranking. Performance is evaluated on the sample of 535 matched households. Binary metrics (A) are evaluated using the mean AUC score over 10-fold cross validation; Continuous metrics (B-D) are evaluated using the mean R^2 score over 10-fold cross validation. See appendix B.2 for details of features.

Table S9: Variable costs of different targeting methods

Targeting Method	Cost per HH screened	Total cost of targeting	Fraction of program costs spent on targeting
CBT	\$2.20	\$276,586	2.18%
PMT	\$4.00	\$502,884	3.97%
Consumption	>\$4.00	>\$502,884	>3.97%
Phone	\$0.00	\$0	0.00%

Notes: Costs for the TUP program, based on costs estimated from the literature. The TUP program screened an estimated 125,721 households; benefits valued at \$1,668 were provided to each of the 7,500 beneficiary households for a total benefits distribution of approximately \$12.7 million. The total value of benefits is used to obtain the targeting costs as a percentage of total program costs. For the Phone option, we assume no contact with beneficiaries is required; if contact were required, for instance to collect informed consent, variable costs would increase accordingly.

Table S10: Costs for CBT and PMT targeting methods obtained from the literature

Source	Location	Cost per household
Panel A: CBT		
Alatas et al. (2012)	Indonesia	$1.20
Karlan and Thuysbaert (2019)	Honduras	$1.67
Karlan and Thuysbaert (2019)	Peru	$1.90
Schnitzer and Stoeffler (2021)	Burkina Faso	$5.60
Schnitzer and Stoeffler (2021)	Niger	$5.40
Schnitzer and Stoeffler (2021)	Senegal	$3.20
Median		$2.20
Panel B: PMT		
Alatas et al. (2012)	Indonesia	$2.70
Karlan and Thuysbaert (2019)	Honduras	$2.62
Karlan and Thuysbaert (2019)	Peru	$3.05
Schnitzer and Stoeffler (2021)	Burkina Faso	$5.69
Schnitzer and Stoeffler (2021)	Chad	$9.50
Schnitzer and Stoeffler (2021)	Mali	$4.00
Schnitzer and Stoeffler (2021)	Niger	$6.80
Median		$4.00

Notes: Costs per household screened for two targeting methods obtained from three papers in the targeting literature. Costs in [11] are provided per-village; we use the average of 54 households per village to obtain per-household targeting costs. Cost for the CBT in [101] is provided as part of the cost for a hybrid CBT and verification approach; although an individual cost for the cBT alone is provided, it is possible this cost excludes some of the mutual costs for the two exercises and is therefore an underestimate of costs of a CBT alone. We use the median of the distribution of targeting costs in our cost analysis.

Appendix C

Supporting materials for Chapter 4

C.1 Supplementary figures and tables

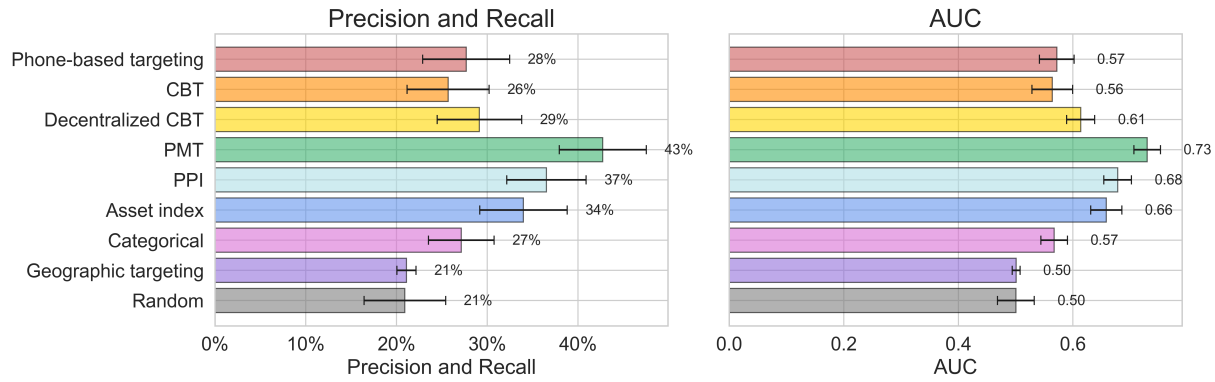

Figure S1: Targeting accuracy comparison for identifying the poorest households in each neighborhood. Accuracy based on precision and recall for identifying the 21% consumption-poorest households in each neighborhood (left), and area under the ROC curve (right). Error bars show two standard deviations above and below the mean for each metric.

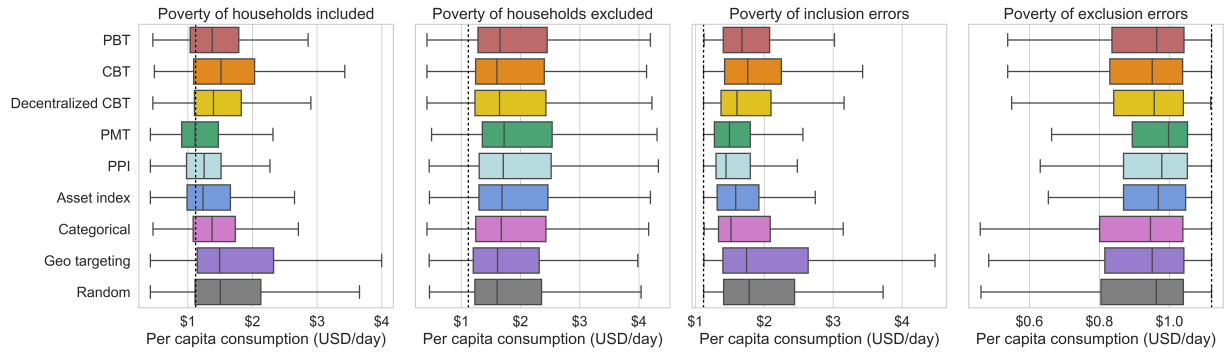

Figure S2: Consumption expenditure distributions for households included and excluded by each targeting method (left and center-left) and for inclusion errors and exclusion errors (center-right and right). The targeting threshold for the 21% poorest households is shown in a vertical dashes line in all figures.

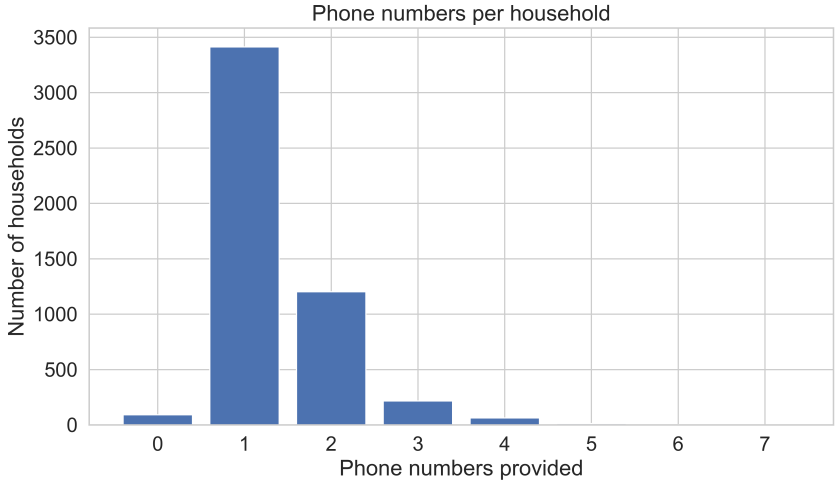

Figure S3: Histogram of household phone ownership in our survey.

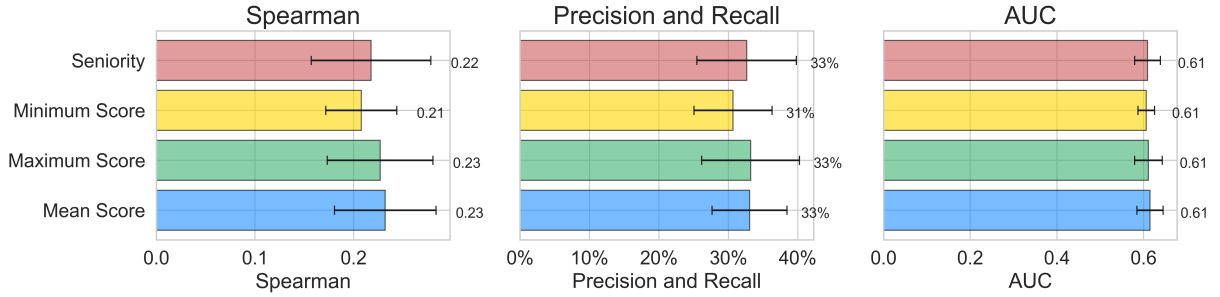

Figure S4: Targeting accuracy for different approaches to aggregating poverty predictions from multiple phones within a household. Accuracy based on precision and recall for identifying the 21% consumption-poorest households in each neighborhood (left), and area under the ROC curve (right). Error bars show two standard deviations above and below the mean for each metric.

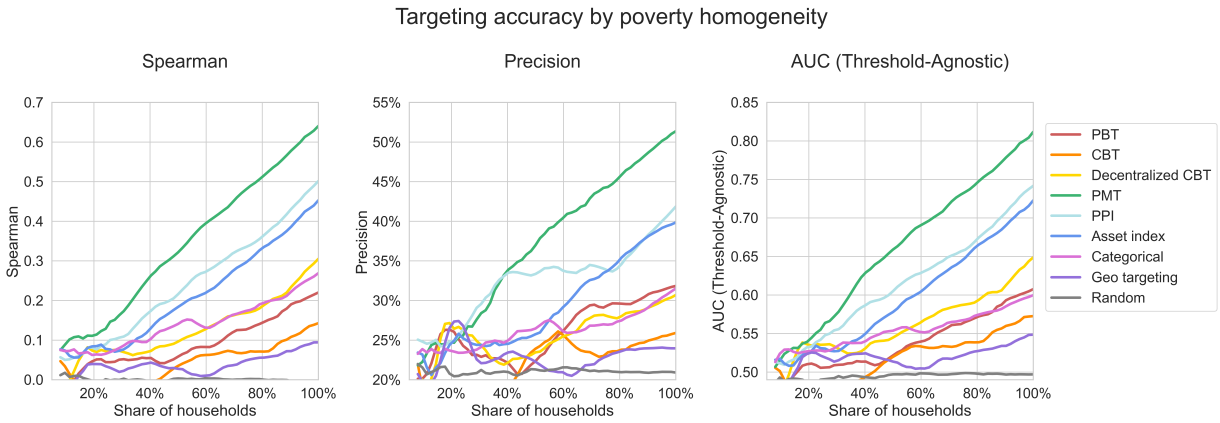

Figure S5: Targeting accuracy of each targeting method as a function of the poverty homogeneity of the population. The x-axis represents the share of households from our survey included, ranked by poverty: thus 20% incdicates restricting the targeting evaluation to the 20% poorest households in our survey.

Table S1: Heterogeneity in targeting accuracy

	Phone targeting	CBT	PMT
Panel A: Neighborhood characteristics			
Households	-0.136*** (0.030)	-0.022 (0.030)	-0.054 (0.032)
Size (square km)	0.055 (0.029)	-0.003 (0.029)	0.001 (0.031)
Density	-0.032 (0.034)	-0.007 (0.034)	0.034 (0.036)
Urban	-0.016 (0.201)	0.063 (0.203)	0.074 (0.215)
% Minority	0.014 (0.051)	0.012 (0.051)	0.066 (0.054)
Connectedness	-0.050 (0.039)	0.048 (0.039)	0.050 (0.041)
Average consumption	-0.097** (0.030)	-0.188*** (0.030)	0.035 (0.032)
Inequality (gini)	0.055 (0.030)	0.142*** (0.030)	0.048 (0.032)
Panel B: Household characteristics			
HH head gender	-0.054 (0.072)	-0.087 (0.073)	-0.102 (0.077)
HH head age	-0.099*** (0.026)	-0.059 (0.027)*	0.130*** (0.028)
HH head employed	-0.051 (0.067)	0.062 (0.067)	-0.054 (0.071)
HH head minority	0.204 (0.148)	0.005 (0.149)	0.202 (0.158)
HH size	0.116*** (0.026)	0.122*** (0.026)	-0.237*** (0.028)
Connectedness (in)	-0.021 (0.031)	-0.034 (0.031)	-0.057 (0.033)
Connectedness (out)	-0.014 (0.019)	-0.008 (0.019)	-0.045 (0.020)*
Phone transactions	0.177*** (0.024)	-0.013 (0.024)	0.062* (0.026)
Consumption	-0.468*** (0.027)	-0.486*** (0.027)	-0.401*** (0.029)

Notes: Kitchen sink regression for predicting targeting accuracy from neighborhood and household characteristics. Targeting accuracy is measured as the standardized signed difference between a household's rank according to a targeting method (phone-based targeting, CBT, or PMT) and the household's rank according to consumption. A coefficient of 1 in this table thus indicates that a one unit change in the associated variable increases a household's rank relative to consumption by 1 standard deviation. "Connectedness" in Panel A is measured as the probability that a surveyed household in the neighborhood reports knowing another randomly selected surveyed household in the neighborhood. "Connectedness (in)" in Panel B is measured as the probability that the household in question known another randomly selected household in the neighborhood, and "connectedness (out)" is the probabiliy that another randomly selected household knows this household.

Appendix D

Supporting materials for Chapter 5

D.1 Additional details of the impact evaluation survey

D.1.1 Sample Frame

The sample frame for the endline survey was drawn from subscribers who successfully enrolled in the GD-Novissi RCT between November 2020 and January 2021 (N=49,083). Thus, the sample was restricted to individuals who (a) had active mobile phone accounts; (b) were registered to vote in one of Togo's 100 poorest cantons; (c) completed the registration procedure for GD-Novissi; and (d) were predicted, based on their mobile phone data, to consume less than $1.25 per day.

The sample was stratified by treatment status and geography. The former was done to maximize statistical power in estimating treatment effects; the latter was done to account for the fact that one large region (Savanes) received payments unrelated to GD-Novissi during the period when GD-Novissi benefits were being delivered.[1]

D.1.2 Response Rate

The enumerators called all 24,294 phone numbers of our final sample in random order. We successfully surveyed 10,129 individuals (response rate of 42%). After removing low-quality surveys (see Appendix D.1.3), our final sample contained 9,511 observations (completion rate of 39%). This completion rate is similar to other phone surveying completed during COVID-19: for example [63] analyzes random-digit dialing to conduct surveys on well-being in nine countries during COVID-19 and reports completion rates ranging from 17% to 59%. Table S1 shows that attrition rates do not differ statistically significantly between the treatment and control groups.

[1]See Appendix D.3.2 for details on the other Savanes program. Of the 36,090 subscribers registered in the Savanes region, we sampled 36%; of the 12,993 subscribers registered outside of the Savanes region, we sampled 88%.

D.1.3 Data Collection Monitoring

We identified surveyors who performed poorly by comparing the data collected with the information contained in Novissi administrative data. We began our analysis by constructing "enumerator effects" (EE) estimates for every enumerator in our data. We predicted the EE on the basis of the correct answers to five questions for which we obtained the "truth" from the Novissi registry (prefecture, canton, age, gender, and Novissi status), and on the frequency of very short surveys (below 15 minutes) as well as surveys with no children reported (which avoids the roster part of the survey and simplifies the surveyor's work.). We controlled for interviewee characteristics such as region and interview language to separate the enumerator's impact from observable interviewee selection.[2] Our approach to estimating EE parallels the parametric empirical Bayes estimator of teacher effects ([100, 49, 78]).

We then normalized the EEs for each of the seven dimensions (prefecture, canton, age, gender, Novissi status, number of short surveys, number of surveys without kids), and took the sum of the coherently signed components for enumerators who conducted more than ten interviews. We classified the interviews of enumerators with an average EE lower than the sample mean minus two standard deviations as of "very poor quality" and remove them from the sample. 615 observations collected by five enumerators who were ranked "very poor quality" were removed from the dataset. In addition, on the second day of the survey, while monitoring data quality, we noticed an enumerator who was performing extremely poorly. After a warning from his supervisor, the quality of his data collection improved. We removed the data collected by this enumerator during the first two days (60 observations). Thus, we only use data from only 9,511 high-quality surveys in our main analysis.

D.1.4 Weights

We reweight observations in the endline survey by the inverse of the sampling probability and the inverse of the probability of response. Sampling probabilities are determined by the four sampling strata, as follows:

- Savanes, treatment: 30.48%
- Savanes, control: 41.26%
- Outside Savanes, treatment: 76.25%
- Outside Savanes, control: 100.00%

[2]The phone number list was randomized and then distributed to the enumerators, so we believed that there is little room for sorting.

To calculate response weights, we train a machine learning model to predict survey response from pre-survey covariates. In total, 9,511 phone numbers completed the survey out of 24,294 numbers sampled. We include the following pre-survey covariates as features in our model:

- 824 features relating to phone use in the six months pre-survey (November 2022 - April 2022).
- 6 features from the Novissi registry: Age, gender, canton of registration (one hot encoded), number of payments received up until the survey date, profession (one hot encoded for the 20 most common professions), and registration week (one hot encoded).
- An indicator for treatment vs. control group.

Using a similar pipeline to the machine learning methods described in Section 5.4.3, we train a LightGBM classifier to predict response, and produce an out-of-sample predicted probability of response for each phone number using five-fold cross-validation. We tune hyperparameters using three-fold cross-validation separately on each of the five folds. With all predictions pooled together, our model achieves an AUC of 0.69. In Figure S3, we confirm that the predicted probabilities of response are well-calibrated by binning the predicted probability of response into ten equal-sized bins, and plotting the average realized response rate for phone numbers in each bin.

The overall weight for each observation is the product of the inverse of the sampling probability and the inverse of the (predicted) probability of response.

D.1.5 Outcomes

The survey contained modules on household food security and consumption, health, access to social services, poverty, mental health, and experience with the Novissi program. Following our pre-analysis plan (American Economic Association Registry #7590), we constructed seven primary indexed outcomes using the index construction methodology described in [43]. These seven indices are: food security, financial health, financial inclusion, mental health, perceived socioeconomic status, health care access, and labor supply.[3]

We standardize all our outcomes so that, within our control group of eligible active mobile subscribers, each outcome had zero mean and unit variance, with the exception of the mental health measure for which we use the Kessler K6 distress scale methodology [**kessler2002short**]. Specifically, we first standardize each component — signed coherently beforehand — by subtracting its control group mean and dividing

[3]Table S3 reports the specific wording of each component of each outcome index.

by its control group standard deviation. We then calculate the sum of the standardized components and standardize the sum again by the control group standard deviation.[4]

In addition to these seven primary welfare outcomes, we collected a proxy-means test (PMT) for each subscriber that proxies for consumption. We used the proxy-means test developed in Chapter 2.

D.1.6 Attrition and Balance Checks

We test for differential nonresponse between the treatment group and the control group in the impact evaluation survey by regressing a binary indicator for response on treatment status, among all 24,294 phone numbers called. In Table S1, we find that there is no statistically significant difference in response rates between the treatment and control groups.

We also test for covariate balance between the treatment and control groups in our impact evaluation survey sample in Table S2. We find that the treatment and control groups in the impact evaluation survey are balanced on self-reported age, gender, and occupation (Panel A), with results robust to substituting administrative data from the Novissi program for self-reported survey data (Panel B).

D.2 Additional details of the pre-treatment survey

Details of the sampling and design for the pre-treatment survey — conducted pre-program in September 2020 (see Figure S1) — are available in Chapter 2.

The phone survey reached 9,484 mobile subscribers inferred to be living in parts of Togo eligible for the GD-Novissi program, and collected information on poverty, living conditions, food security, and health. Critically, the pre-treatment survey collected the components for our PMT (Appendix D.1) and the components of three of the four indices for which we observe significant treatment effects of GD-Novissi: food security, financial health and perceived socioeconomic status.

The financial health and perceived socioeconomic status indices are constructed identically to the indices in the impact evaluation survey; however, only certain components of the food security index were collected in the pre-treatment survey, so we construct a "reduced food security index" for the pre-treatment survey. The reduced food security index is less comprehensive than the food security index collected in the impact evaluation survey (Table S3) and, in particular, does not include questions on food consumption. The components for the reduced food security index are listed in Table S4.

[4]We impute missing components using the other components in an index unless the missing components are children-related and the family had no children, in which case we compute the index omitting those components.

As in the impact evaluation survey, each index is constructed following [43], by standardizing each component across the surveyed population, summing components, and then standardizing the resulting index.

Also, as in the impact evaluation survey, each observation is weighted by the inverse of the sampling probability and the inverse of the probability of response. Details on the estimation of weights are available in Chapter 2. We use weights throughout our analysis involving the pre-treatment survey, except where otherwise noted.

D.3 Treatment effect heterogeneity

We test for treatment effect heterogeneity on four pre-registered dimensions: gender, poverty, occupation, and region of residence (in or outside of the Togo's northernmost region, Savanes). For each dimension, we test for heterogeneous treatment effects on our seven outcomes and the aggregate welfare index in Table 5.1 Panel A.

D.3.1 Heterogeneity by gender, wealth, and occupation

We find little evidence that treatment effects were heterogeneous across the socioeconomic and demographic subgroups that we pre-specified in our pre-analysis plan. In particular, while GD-Novissi had an important gender component, whereby women received roughly 15% more money per month than men, the welfare impacts on women were not significantly larger than for men. These results can be seen in Panel A of Table S6: while women, in general, are worse off than men (the third row of coefficients), the coefficient on the interaction between treatment and female is never significantly different from zero.[5]

Panels B and C of Table S6 likewise indicate that treatment effects did not differ by pre-treatment wealth or occupation. In Panel B, we compare treatment effects for people with PMT scores above and below the sample median, and, with the exception of healthcare access, do not observe significant differences for any outcome. Panel C indicates that treatment effects were not different for farmers – the most common occupation in the rural areas of Togo where GD-Novissi was implemented, and the occupation reported by 60% of endline survey respondents.

D.3.2 Geographic heterogeneity

There is, however, one dimension where we find evidence of substantial heterogeneity in treatment effects, which is by the location of the beneficiary. In particular, Panel D

[5]According to administrative data from the Novissi program, women represent half of GD-Novissi beneficiaries, and 45% of our surveyed sample. However, in suggestive evidence of strategic behavior at the household level in GD-Novissi registration, the share of women among survey respondents is only 27%. That is, 40% of the phone numbers registered with female voter ID cards were answered by men.

of Table S6 highlights how treatment effects on food security and mental health were significantly larger for beneficiaries in the Savanes region in the far North of Togo (Figure S2). Indeed, with the exception of healthcare access, the treatment effects for beneficiaries outside Savanes are all close to zero and no longer statistically significant once we account for the differential effect of treatment in Savanes.

Savanes is unique in several respects: most GD-Novissi beneficiaries (70%) reside in the Savanes region, it is generally poorer than other regions eligible for GD-Novissi, it had higher rates of COVID-19 and related curfews than other regions, and the government provided an independent round of cash transfers called *Savanes-Novissi* (unconnected to GD-Novissi) to all residents in Savanes in February 2021 (two months before our endline surveys were conducted). While understanding the reasons for the substantial geographic heterogeneity between Savanes and non-Savanes beneficiaries is not the focus of this paper, we explore briefly below three possible hypotheses that could explain why the treatment effects of GD-Novissi are observed mainly in the Savanes region: (i) that differential registration for Savanes-Novissi between the GD-Novissi treatment and control groups resulted in additional cash impacts for the treatment group, (ii) that mobility reductions resulting from curfews in the Savanes region made cash transfers more impactful in Savanes than the rest of the country, and (iii) that price differences between Savanes and the rest of the country gave cash transfers more purchasing power in Savanes.

Interaction Between GD-Novissi and Savanes-Novissi

In addition to the GD-Novissi program considered here, the Government of Togo implemented three other targeted cash transfer programs under the Novissi umbrella during the pandemic period. One of these, "Savanes-Novissi", provided one-time cash transfers of USD 8-10 to all residents of Savanes who registered for Novissi in a two-week period beginning on February 22, 2021. Women received a one-time transfer of CFA 6,125 (USD 9.80), and men received a transfer of CFA 5,250 (USD 8.40). A total of 244,302 Savanes residents registered for and received Savanes-Novissi, of whom 114,311 (46.79%) were already registered for GD-Novissi.

We observe an approximately 20 percentage point difference in registration rates for Savanes-Novissi between the treatment and control groups in GD-Novissi, with the control group substantially more likely to register for the Savanes-Novissi program. 41% of the treatment group registered for Savanes-Novissi, while 63% of the control group registered.

There are two plausible explanations for the difference in enrollment: first, GD-Novissi provided enough assistance for the treatment group, so they were less in need of further cash transfers, and second, that confusion in communications around the two programs resulted in members of the treatment group believing they were ineligible for Savanes-Novissi. The second explanation is particularly plausible for two reasons. First, people located in Savanes who were registered for GD-Novissi were eligible for

Savanes-Novissi, but were required to register separately for Savanes-Novissi, which could be confusing. Second, because the treatment group was receiving cash transfers from GD-Novissi in February 2021, the government of Togo initially excluded treated people of GD-Novissi from the Savanes-Novissi program. While the Savanes-Novissi amount was transferred at the time of registration for everyone else, people in the GD-Novissi treatment group received Savanes-Novissi cash transfers in the second week of the period of registration.

We included specific questions in the impact evaluation survey to distinguish between the two explanations. We first asked people if they registered with Savanes-Novissi. If not, we asked them an open-ended question why not (Table S14). The enumerators were told to classify the answers in one of the eight pre-defined categories, including "other" and "I don't know". Three of the possible categories are related to the confusion hypothesis ("I did not think I was eligible", "I did not think I needed to register, since I already registered with GD-Novissi", and "I heard about the program after the end of the registration period"). Three others are related to GD-Novissi impact hypothesis ("I receive GD-Novissi, I don't need extra money", "I have enough money, I don't need extra money", and "Other people are more in need than me, I prefer them to get the money").

Qualitatively, the treatment group is more likely to be confused about the eligibility criteria than the control group. The first main reason why people did not register is the lack of information, and there is a ten percentage points difference between the treatment and the control group: 36.5% of the control group versus 49.6% of the treatment group was confused about the eligibility criteria. Less than 3% of people in both treatment arms reported a lack of need for Savanes-Novissi as the main reason, supporting the second hypothesis for the enrollment differences (confusion in eligibility criteria).

However, in comparing the welfare outcomes of people who did and did not register with Savanes-Novissi by treatment arm (Table S15), we do observe that people from the treatment group who did not register with Savanes-Novissi have a higher food security and financial health index than those who did register. There are no such differences between registrants and non-registrants in the control group. The fact that the treatment group self-selected in Savanes-Novissi supports the first hypothesis, suggesting that GD-Novissi contributed to the low enrollment rates for Savanes-Novissi in the treatment group.

We conclude, based on this evidence, that both hypotheses (the welfare impact of GD-Novissi and confusion around eligibility criteria) likely contributed to lower registration rates for GD-Novissi in the treatment group in Savanes. Importantly, however, the difference in registration rates does not explain the larger welfare impacts of GD-Novissi in Savanes in comparison to the rest of the country: if anything, we would expect the lower registration rates for Savanes-Novissi in the treatment group to attenuate treatment effects relative to other regions of the country.

Impacts of the Savanes Curfew on Mobility

After a surge in the number of COVID-19 cases, the entire region of Savanes was placed under curfew from January 17 to February 21, 2021. It is the only part of Togo where there was a strict shutdown due to COVID-19 during the implementation of GD-Novissi. A plausible explanation for the welfare impacts of GD-Novissi in Savanes is that mobility restrictions led to an increased need for assistance in the Savanes region relative to other parts of the country.

We test this hypothesis using mobility indicators derived from mobile phone data from subscribers inferred to be living in the Savanes region before and during the period of curfew. Using the frequency-based home location detection methods described in [155] we infer which subscribers are likely to be residing in the Savanes period during the curfew. We proxy mobility with the number of unique towers and the number of unique cantons visited over the course of 21 days in November, December, January, and February.[6]

First, we use the phone-based mobility indicators to verify whether the curfew in the Savanes region induced people to move less. We use a difference-in-differences strategy to identify the effect of the curfew on our sample mobility. Our estimating OLS equation is

$$y_{it} = \rho_i + \sum_{\tau=-3,\tau\neq-1}^{0} \beta_\tau (S_i \times C_\tau) + \gamma_t + \alpha X_{it} + \varepsilon_{it}$$

where ρ_i is an individual fixed-effect that captures all observable and unobservable time-invariant individual characteristics, γ_t is a period fixed-effect, and ε_{it} is an individual-period shock. S_i is a dummy for living in Savanes, and C_τ ($-3 \leq \tau \leq 0$) are dummies indicating the number of periods relative to the curfew (February 2021). We control for the number of transactions X_{it} to make sure that the potential change in our mobility metric is not mechanically driven by a change in the number of phone calls and texts. The parameter of interest is β_0, which measures the effect of the curfew on our metric for mobility y_{it} relative to the previous period (the omitted category β_{-1}). The coefficients β_{-3} and β_{-2} are pre-trends coefficients that capture the difference in mobility between Savanes and Not-savanes inhabitants relative to the omitted variable before the curfew.

Table S16 shows the curfew's effect on mobility. Mobility in the Savanes region decreases by 0.3 towers per month on average (compared to a control mean of 5.76 towers per month), and the pre-trend coefficients are not statistically different from zero, indicating that the curfew put in place in the Savanes region did significantly decrease mobility of the region's residents.

[6]Whenever someone makes a call or text, that transaction is associated with the tower closest to where she is. However, many people in our sample do not make a transaction every day. The number of unique towers or cantons observed in the data is the best information we have on actual mobility, but remains a noisy metric for mobility.

To test whether mobility reductions may have driven the positive GD-Novissi treatment effects in Savanes, we test for differential impacts of GD-Novissi among Savanes subscribers by baseline mobility quintiles (with baseline mobility derived pre-treatment in the months of September and October 2021). Mobility is again derived from mobile phone data and proxied by the number of unique cell towers a subscriber visits in the months of September and October. While we observe substantial mobility reductions across quintiles (Figure S4), we do not observe a differential impact of GD-Novissi by baseline mobility (Figure S5). This result suggests that it is unlikely that mobility impacts drove the GD-Novissi treatment effects in Savanes, as there is no differential impact for the most mobile pre-curfew within the Savanes region.

Price Differences

A final possible explanation for the GD-Novissi treatment effects in Savanes (in comparison to the rest of the country) is that price differences between the Savanes region and the rest of the country give GD-Novissi transfers more purchasing power in Savanes. We collected price information for staple goods in the consumption module of the impact evaluation survey; in our analysis, we restrict to goods for which at least 50% of the respondents provided a price. Among these seven goods, we observe statistically significant differences in prices between Savanes and the rest of the country for only three goods (Table S17). Palm oil and milk are more expensive in Savanes, while Niebe is cheaper. Given that there are no systematic price differences in a consistent direction between Savanes and the rest of the country, we conclude that price differences are not a major driver of the GD-Novissi treatment effects in Savanes.

D.4 Comparison with related work on COVID-19 cash transfers

Since the COVID-19 pandemic, a growing body of research has emerged to document the welfare impacts of cash transfers distributed in response to the pandemic. Many of these studies are reviewed in [102]. Broadly, this literature shows modest, positive, and statistically significant impacts of cash transfers on a wide range of welfare metrics, including food security and mental health.

Specifically, [22] use phone surveys and an RCT design to show that universal basic income transfers of USD 22.5 nominal per month to households under lockdown in Kenya reduced the probability of households experiencing hunger (by 5-11 percentage points, relative to a control mean of 68%), and had modest positive impacts on mental health. Similarly, [110] use an RCT and phone surveys to measure impacts of a monthly VAT refund of USD 19 in Colombia, finding a 4.4 percentage point increase in the probability of treated households purchasing food in the week preceding the survey (relative to a control mean of 72%), but no statistically significant impacts on food security. The paper

also reports positive and statistically significant impacts on mental health indices (1.2-2.1 percentage points) and a financial health index (0.055 standard deviations). [102] follow a similar experimental design, using an RCT and several rounds of phone surveys to evaluate the impact of eight monthly cash transfers of $15, recording an 8% increase in food consumption among treated households. In a non-randomized approach, [40] use online surveys and a regression discontinuity design to show that pension payments of USD 43-50 per month in Bolivia decreased the probability of households going hungry by 8-12 percentage points, relative to a comparison mean of 22%.

The first portion of our analysis contributes to this literature by documenting the impacts of pandemic cash transfers in Togo, using an extensive cash transfer program where treatment was randomly assigned at the individual level. While the cash transfers we study are smaller ($13-15.50 per month) than most of the other programs studied ($15-50 per month), we document comparable effect sizes (0.04-0.07 standard deviations).

Our results on heterogeneous treatment effects (Table S6) are also broadly consistent with the other papers studying the impacts of COVID-19 cash transfers on well-being, which for the most part do not find significant heterogeneity across dimensions studied [110, 102]. However, two results stand in contrast: [110] finds that treatment effects are driven primarily by households in urban areas, while we find that treatment effects are driven primarily by households in Savanes, which is the most rural region of Togo; and [102] finds that treatment effects on food security are larger for female-headed households than male-headed households, whereas we find no heterogeneous treatment effects by gender of the recipient.

D.5 Additional tests for estimating treatment effects from phone data

In this section, we use alternative specifications to test whether it is possible to recover treatment effects from GD-Novissi mobile phone records. In table S7 we test using a two-week period to derive features from mobile phone data rather than a six-month feature period. In Table S8, we try using changes in features between the pre-treatment and during-treatment periods to predict each of our outcomes (using the endline survey as ground truth). In Table S9 we try predicting outcomes and inferring treatment effects in the Savanes region only, since the survey-based treatment effects were only observable in Savanes. In Table S10 we try the same specifications using only features that are statistically significantly different between the treatment and control groups (22% of all features). Finally, to test for whether noise in survey data is the cause of low predictive power, in Table S11 we train and evaluate a model to predict treatment status from the mobile phone feature set. The poor performance of each of these models suggests that it is the inability of phone data to identify differences between the treatment and control groups — rather than an issue of noisy survey data — that drives the low predictive

power of the phone-based models and thus the null effects in downstream inference tasks.

D.6 Supplementary figures and tables

Figure S1: GD-Novissi timeline

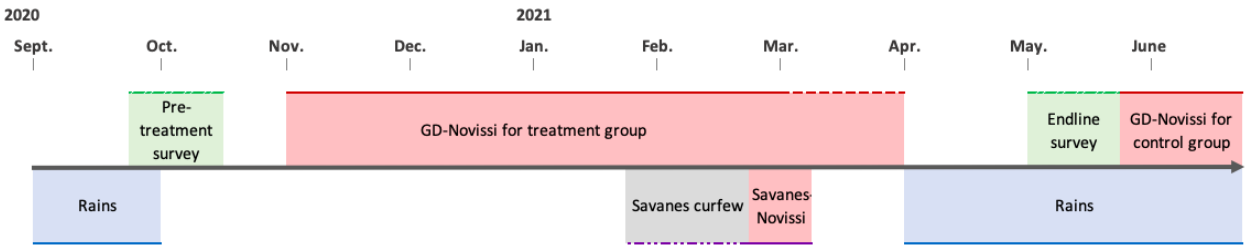

Figure S2: Beneficiaries per canton (admin-3 units)

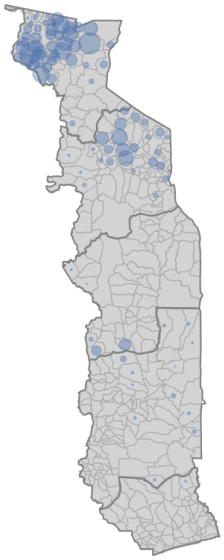

Figure S3: Confirming the calibration of response weights

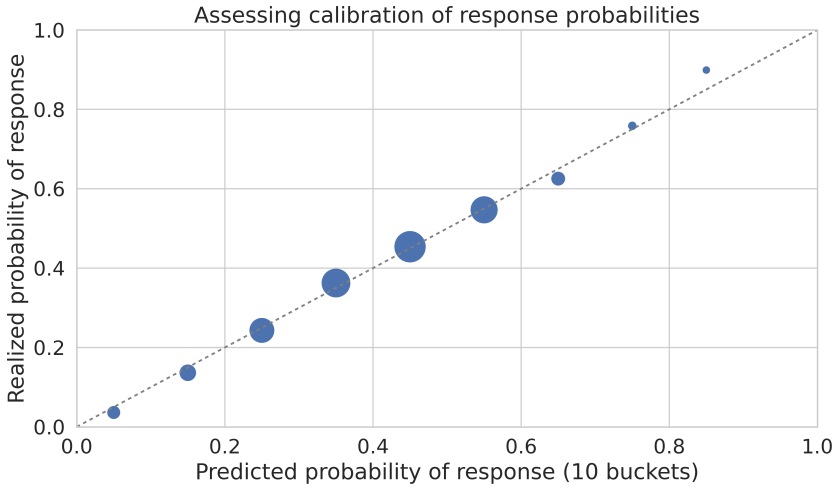

Figure S4: Mobility reductions in Savanes by baseline mobility quintile

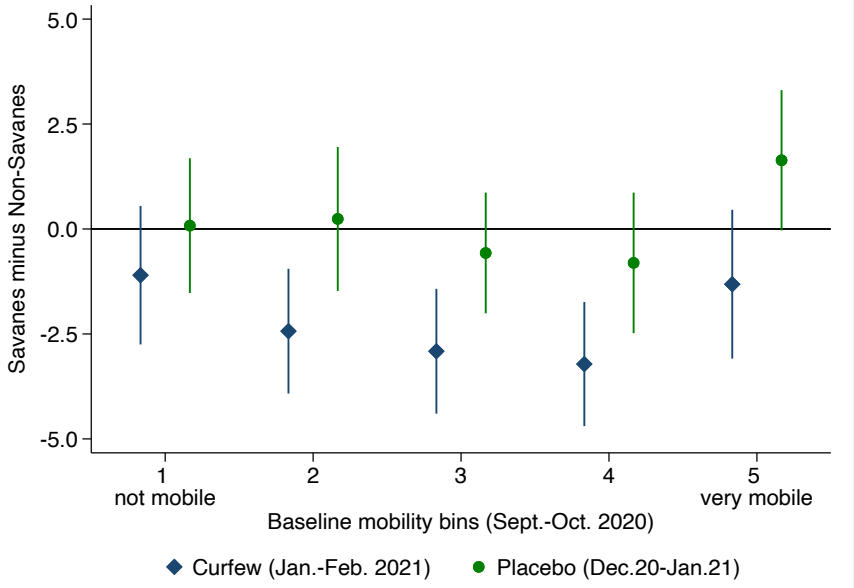

Figure S5: Impacts of GD-Novissi on the aggregated index by baseline mobility quintile

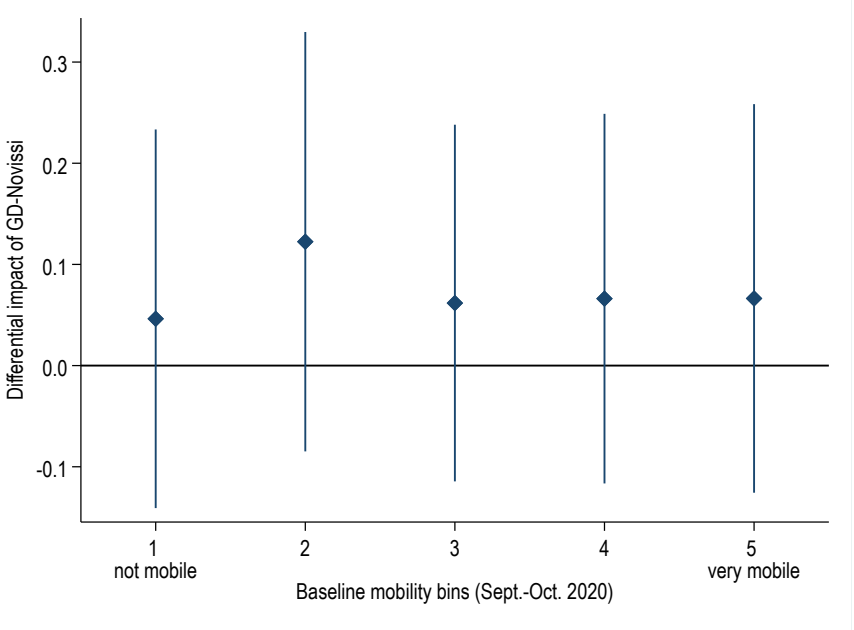

Table S1: Differential attrition

	Probability of non-response
Treatment	-0.01 (0.01)
N	24,294
Control Mean	0.61

Notes: Differential attrition from impact evaluation survey sample for treatment and control groups. Effect of treatment on attrition is estimated with a simple linear regression specification with non-response as the dependent variable. Regression is conducted without fixed effects. $^{*}p$ <0.1; $^{**}p$ <0.05; $^{***}p$ <0.01.

Table S2: Summary Statistics and Balance Checks

	Baseline Sample		*Endline Sample*		
	N	Mean	N	Mean	Diff. T-C
Panel A. Survey data					
PMT	8,821	\$1.49	8,452	\$1.31	\$0.00
		(0.74)		(0.49)	(0.01)
Female	8,821	0.23	9,511	0.31	0.03**
		(0.42)		(0.46)	(0.01)
Age	8,716	33.37	9,310	36.03	-0.30
		(11.98)		(11.44)	(0.30)
Farmers	8,819	0.41	9,511	0.59	-0.02
		(0.49)		(0.49)	(0.01)
Savanes	8,821	0.51	9,443	0.72	-0.01
		(0.50)		(0.45)	(0.01)
Panel B. Novissi registry data					
Female	5,493	0.50	9,511	0.49	0.02*
		(0.50)		(0.50)	(0.01)
Age	5,493	36.02	9,429	37.63	0.09
		(13.96)		(12.70)	(0.33)
Farmers	5,402	0.23	9,375	0.38	-0.02*
		(0.42)		(0.49)	(0.01)
Savanes	5,493	0.52	9,511	0.74	-0.00
		(0.50)		(0.44)	(0.01)

Notes: This Table presents summary statistics for the pre-treatment and the endline samples. Column "N" indicates the number of respondents, "Mean" indicates the sample mean, with the standard deviation in parenthesis. Column "Diff. T-C" contains the balance checks that are conducted by regressing the demographic variable of interest on treatment status (balance checks are conducted for the endline survey only). All observations are weighted by sampling probabilities. All regressions control for the enumerator, week of the survey, and strata fixed effects. Robust standard errors are in parenthesis. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

Table S3: Components for impact evaluation outcomes

Question	Possible Answers
Panel A. Food security	
Yesterday, how many meals did you eat?	0-3
In the past 7 days, how often were you unable to eat preferred foods because of a lack of money or other resources?	Two weeks
In the past 7 days, how often have you had to limit portion size at meal times?	0 Never - 4 Every day
In the past 7 days, how often did you have to reduce the number of meals eaten in a day?	0 Never - 4 Every day
In the past 7 days, how often have the children over 3 in your household had to reduce the number of meals eaten in a day?	0 Never - 4 Every day
Yesterday, how many meals did the children over 3 in your household eat?	0-3
When was the last time your household had each of the following items: Powdered milk, sugar, smoked anchovy, fresh onion, dried fish, sesame, red palm oil, traditional bread, orange, cowpea/dried beans.	0 Never - 5 Less than a week
How much did you spend on purchasing each of the items above, the last time?	Integer
Panel B. Financial health	
Were you able to save money last month? If so, how much?	Integer
God forbid, if your household stopped getting income from any source, how long could your household easily continue to meet your basic needs for food and housing? (Winsorized 95th percentile.)	Integer
God forbid, if there was a major emergency and your household needed money, how much money could you easily obtain within the next seven days? (Winsorized 95th percentile.)	Integer
Panel C. Financial inclusion	
Fraction of the adults in the household with a bank account.	Float
Fraction of the adults in the household (excluding the respondent) with a mobile money account.	Float
Panel D. Mental health (Kessler K6 nonspecific distress scale)	
During the past 7 days, about how often did you feel nervous?	Integer
During the past 7 days, about how often did you feel hopeless?	Integer
During the past 7 days, about how often did you feel restless or fidgety?	Integer
During the past 7 days, about how often did you feel that everything was an effort?	Integer
During the past 7 days, about how often did you feel so sad that nothing could cheer you up?	Integer
During the past 7 days, about how often did you feel worthless?	Integer
Panel E. Self perception of socioeconomic status	
In general, relative to other people in Togo, would you say that you are...	1 very poor - 5 very well off
How do you think other communities perceive the wealth of your household?	1 very poor - 5 very well off
Panel F. Labor supply	
Hours worked last week (winsorized 99th percentile)	Integer
During the past 7 days, how much income/pay did you receive?	Integer
Panel G. Healthcare access	
The last time you or someone else in your household needed healthcare, did you get healthcare?	Yes/no
When you last needed health care, did you get it at the hospital?	Yes/no
God forbid, if a child in your household needed to go the hospital, would you be able to bring him or her?	Yes/no

Notes: Components for each of the outcomes in the endline survey. All indices are produced using the index construction methadology from Bryan et al. (2021) [43] except for the mental health index, which is based on simple addition of the components.

Table S4: Reduced food security index

Question	**Possible Answers**
Yesterday, how many meals did you eat?	0-3
In the past 7 days, how often were you unable to eat preferred foods because of lack of money or other resources?	0 Never - 4 Every day
In the past 7 days, how often have you had to limit portion size at meal times?	0 Never - 4 Every day
In the past 7 days, how often have you had to reduce the number of meals eaten in a day?	0 Never - 4 Every day
In the past 7 days, how often have the children in your household over age three had to reduce the number of meals eaten in a day?	0 Never - 4 Every day
Yesterday, how many meals did the children in your household over age three eat?	0-3
In the past 7 days, were you able to buy the amount of food you usually buy?	Yes/no

Notes: Components for the reduced food security index in the pre-treatment survey.

Table S5: Variation Between and Within Canton

Outcome	(1) Between Variance	(2) Within Variance	(3) Ratio
Panel A: Pre-treatment survey			
PMT	4.30	0.29	15.00
Food Security	1.438	0.96	1.44
Financial Health	1.25	0.98	1.27
Perceived Socioeconomic Status	1.80	0.95	1.89
Labor Supply	1.58	0.94	1.67
Panel B: Endline survey			
PMT	2.08	0.25	8.21
Food Security	1.85	0.99	1.87
Financial Health	1.65	0.99	1.87
Financial Inclusion	1.59	0.92	1.72
Mental Health	1.89	0.98	1.92
Perceived Socioeconomic Status	1.26	0.99	1.28
Healthcare Access	1.46	1.00	1.45
Labor Supply	1.57	0.96	1.63
Aggregate Welfare Index	2.03	0.98	2.08

Notes: Between vs. within variance, with groups defined by canton (self-reported in the baseline survey, determined by location of Novissi registration in the endline survey). Only individuals in cantons with at least 10 individuals surveyed are included in the analysis. All outcomes except for the PMT are standardized to 0 mean and unit variance in the control group.

Table S6: Survey-based treatment effect heterogeneity

	(1) Food security	(2) Financial health	(3) Financial inclusion	(4) Mental health	(5) Perceived status	(6) Health care access	(7) Labor supply	(8) All seven indices
Panel A: Gender								
Treatment * Female	-0.037	-0.015	-0.070	0.006	0.003	-0.075	-0.011	-0.053
	(0.049)	(0.051)	(0.047)	(0.041)	(0.048)	(0.054)	(0.050)	(0.049)
Treatment	0.077***	0.034	0.025	0.072***	0.042	0.036	0.018	0.081***
	(0.025)	(0.029)	(0.025)	(0.023)	(0.027)	(0.026)	(0.033)	(0.027)
Female	-0.050	-0.121***	0.201***	-0.074**	-0.116***	-0.079**	-0.221***	-0.122***
	(0.035)	(0.039)	(0.034)	(0.030)	(0.037)	(0.037)	(0.037)	(0.036)
Panel B: Poverty								
Treatment * Poor	0.039	-0.021	-0.068	-0.049	-0.021	0.098**	-0.049	-0.019
	(0.045)	(0.052)	(0.046)	(0.039)	(0.047)	(0.048)	(0.054)	(0.048)
Treatment	0.053	0.048	0.040	0.105***	0.051	-0.048	0.024	0.072**
	(0.033)	(0.036)	(0.034)	(0.028)	(0.034)	(0.035)	(0.034)	(0.035)
Poor	-0.120***	-0.024	-0.111***	0.012	-0.080**	0.062*	0.058	-0.054**
	(0.030)	(0.037)	(0.033)	(0.029)	(0.034)	(0.033)	(0.040)	(0.035)
Panel C: Occupation								
Treatment * Farmer	0.016	0.032	0.024	-0.032	-0.027	0.049	0.028	0.024
	(0.046)	(0.050)	(0.044)	(0.039)	(0.046)	(0.048)	(0.052)	(0.048)
Treatment	0.052	0.006	-0.011	0.090***	0.053	-0.018	-0.010	0.043
	(0.037)	(0.039)	(0.035)	(0.031)	(0.036)	(0.039)	(0.043)	(0.039)
Farmers	-0.152***	-0.114***	-0.277***	-0.031	-0.172***	-0.002	-0.135***	-0.234***
	(0.032)	(0.037)	(0.032)	(0.028)	(0.035)	(0.033)	(0.040)	(0.035)
Panel D: Region								
Treatment * Savanes	0.087**	0.041	0.055	0.064*	0.065	-0.076*	0.015	0.066
	(0.042)	(0.044)	(0.041)	(0.037)	(0.044)	(0.045)	(0.043)	(0.043)
Treatment	0.000	-0.003	-0.033	0.025	-0.007	0.066*	-0.002	0.012
	(0.032)	(0.031)	(0.031)	(0.029)	(0.034)	(0.034)	(0.029)	(0.032)
Savanes	-0.076**	0.024	-0.037	0.014	-0.010	0.144***	0.017	0.020
	(0.032)	(0.033)	(0.030)	(0.029)	(0.035)	(0.032)	(0.033)	(0.033)
Obs	9,511	9,511	9,511	9,511	9,511	9,511	9,511	9,511

Notes: Heterogeneous treatment effects for outcomes for which we detect a statistically significant survey-based treatment effect in Table 5.1 Panel A. The dependent variable for each regression is indicated in the column title; see Appendix D.1 for variable construction. In Panels A, C, and D, gender, occupation, and region of residence are determined by information provided by the respondents in the survey. In Panel B, poverty is determined by having a below-median PMT score. All regressions control for the enumerator, week of the survey, and strata fixed effects. All observations are weighted by sampling probabilities and response probabilities, and observations are restricted to subscribers who were active prior to the program's launch. Robust standard errors are in parentheses. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

Table S7: Estimating Treatment Effects from Two Weeks of Mobile Phone Data

	(1) PMT	(2) Food security	(3) Financial health	(4) Financial inclusion	(5) Mental health	(6) Perceived status	(7) Healthcare access	(8) Labor supply	(9) All seven indices
Panel A: Predicting welfare outcomes using ML trained on pre-treatment survey									
R^2	0.112	0.003	0.014	—	—	0.017	—	0.028	—
Obs.	8,593	8,593	8,593	—	—	8,593	—	8,584	—
Panel B: Predicting welfare outcomes using ML trained on endline survey									
R^2	0.031	0.004	0.006	-0.002	0.004	0.001	0.007	0.010	0.011
Obs.	8,238	9,261	9,261	8,898	9,261	9,261	9,276	9,261	9,261
Panel C: Phone-based treatment effects trained on the pre-treatment survey									
Treatment	-0.004***	0.001	0.014	—	—	-0.003	—	0.007	—
	(0.001)	(0.014)	(0.014)	—	—	(0.013)	—	(0.012)	—
Obs.	46,327	46,327	46,327	—	—	46,327	—	46,327	—
Control Mean	1.467	0.000	0.000	—	—	0.000	—	0.000	—
Z-test p-value	0.822	0.018	0.673	—	—	0.745	—	0.005	—
Panel D: Phone-based treatment effects trained on the endline survey									
Treatment	-0.001	0.014	0.008	-0.012	-0.001	0.012	0.004	0.011	0.010
	(0.001)	(0.012)	(0.013)	(0.015)	(0.014)	(0.012)	(0.012)	(0.012)	(0.012)
Obs.	46,327	46,327	46,327	46,327	46,327	46,327	46,327	46,327	46,327
Control Mean	1.306	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Z-test p-value	0.793	0.049	0.524	0.464	0.002	0.267	0.819	0.943	0.049

Notes: Replication of Table 5.1 using two weeks of phone data to derive features (rather than six months). In the first regime — in which the ML models are trained using data from the impact evaluation survey — the two weeks of phone data for model training are obtained from the two weeks during which the pre-treatment survey took place in September 2021. The mobile phone data used to train the ML model in second regime — in which the ML models are trained using data from the endline survey — is taken from the two weeks immediately after each subscriber registered for GD-Novissi. The immediate post-treatment two weeks are used to generate well-being predictions in both regimes. Standard errors from Bayesian bootstrap procedure in parentheses. $^{*}p$ <0.1; $^{**}p$ <0.05; $^{***}p$ <0.01.

Table S8: Estimating Treatment Effects from Mobile Phone Data using Changes in Features

	(1) PMT	(2) Food security	(3) Financial health	(4) Financial inclusion	(5) Mental health	(6) Perceived status	(7) Healthcare access	(8) Labor supply	(9) All seven indices
Panel A: Predicting welfare outcomes									
R^2	0.036	0.005	0.004	0.004	-0.004	0.006	0.000	0.015	0.020
Obs.	8,446	9,504	9,504	9,131	9,504	9,504	9,519	9,504	9,504
Panel B: Phone-based treatment effects									
Treatment	-0.003**	0.000	-0.015	-0.007	0.025*	-0.011	0.005	-0.004	-0.005
	(0.002)	(0.018)	(0.019)	(0.016)	(0.017)	(0.018)	(0.019)	(0.015)	(0.014)
Obs.	48,726	48,726	48,726	48,726	48,726	48,726	48,726	48,726	48,726
Control Mean	1.318	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Z-test p-value	0.699	0.025	0.177	0.591	0.062	0.068	0.861	0.657	0.015

Notes: Replication of Table 5.1 Panel B using changes in phone-derived features between the pre-treatment and during-treatment periods as inputs to the model (once for the six month time period, and once for the two week time period). Standard errors from Bayesian bootstrap procedure in parentheses. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

Table S9: Estimating Treatment Effects from Mobile Phone Data in Savanes

	(1) PMT	(2) Food security	(3) Financial health	(4) Financial inclusion	(5) Mental health	(6) Perceived status	(7) Healthcare access	(8) Labor supply	(9) All seven indices
Panel A: Predicting welfare outcomes using ML trained on pre-treatment survey									
R^2	0.120	-0.012	0.009	—	—	0.042	—	0.038	—
Obs.	3,701	3,701	3,701	—	—	3,701	—	3,698	—
Panel B: Predicting welfare outcomes using ML trained on endline survey									
R^2	0.034	0.007	-0.003	-0.006	0.001	-0.002	0.001	0.016	0.023
Obs.	3,089	3,478	3,478	3,368	3,478	3,478	3,481	3,478	3,478
Panel C: Phone-based treatment effects trained on the pre-treatment survey									
Treatment	-0.005***	-0.002	-0.009	—	—	-0.011	—	-0.005	—
	(0.002)	(0.020)	(0.014)	—	—	(0.015)	—	(0.016)	—
Obs.	35,889	35,889	35,889	—	—	35,889	—	35,889	—
Control Mean	1.412	0.000	0.000	—	—	0.000	—	0.000	—
Z-test p-value	0.778	0.028	0.208	—	—	0.548	—	0.002	—
Panel D: Phone-based treatment effects trained on the endline survey									
Treatment	-0.003**	0.003	0.004	0.010	0.022	-0.008	0.000	0.013	0.008
	(0.002)	(0.018)	(0.018)	(0.024)	(0.020)	(0.022)	(0.017)	(0.017)	(0.016)
Obs.	35,889	35,889	35,889	35,889	35,889	35,889	35,889	35,889	35,889
Control Mean	1.307	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Z-test p-value	0.652	0.034	0.477	0.929	0.068	0.127	0.738	0.900	0.059

Notes: Replication of Table 5.1 with only subscribers located in Savanes. Standard errors from Bayesian bootstrap procedure in parentheses. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

Table S10: Estimating Treatment Effects from Mobile Phone Data using Only Significant Features

	(1) PMT	(2) Food security	(3) Financial health	(4) Financial inclusion	(5) Mental health	(6) Perceived status	(7) Healthcare access	(8) Labor supply	(9) All seven indices
Panel A: Predicting welfare outcomes using ML trained on pre-treatment survey									
R^2	0.139	0.000	0.007	—	—	0.027	—	0.044	—
Obs.	8,899	8,899	8,899	—	—	8,899	—	8,890	—
Panel B: Predicting welfare outcomes using ML trained on endline survey									
R^2	0.042	0.008	0.005	0.002	0.004	0.006	0.008	0.021	0.024
Obs.	8,448	9,507	9,507	9,134	9,507	9,507	9,522	9,507	9,507
Panel C: Phone-based treatment effects trained on the pre-treatment survey									
Treatment	-0.006***	-0.010	-0.013	—	—	-0.013	—	0.001	—
	(0.002)	(0.018)	(0.014)	—	—	(0.013)	—	(0.012)	—
Obs.	48,759	48,759	48,759	—	—	48,759	—	48,759	—
Control Mean	1.431	0.000	0.000	—	—	0.000	—	0.000	—
Z-test p-value	0.755	0.011	0.159	—	—	0.491	—	0.002	—
Panel D: Phone-based treatment effects trained on the endline survey									
Treatment	-0.001	0.001	0.002	0.004	0.028*	-0.001	0.021	0.012	0.016
	(0.002)	(0.015)	(0.018)	(0.019)	(0.017)	(0.017)	(0.016)	(0.013)	(0.015)
Obs.	48,759	48,759	48,759	48,759	48,759	48,759	48,759	48,759	48,759
Control Mean	1.313	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Z-test p-value	0.802	0.016	0.425	0.920	0.088	0.141	0.709	0.927	0.101

Notes: Replication of Table 5.1 with only features that are statistically significantly different between the treatment and control groups. Standard errors from Bayesian bootstrap procedure in parentheses. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

Table S11: Estimating treatment status from mobile phone data

Phone Data Period	**AUC**	*N*
Panel A: All subscribers in RCT		
Six months	0.5151	49,079
Two weeks	0.5219	46,370
Panel B: Savanes only		
Six months	0.5254	36,086
Two weeks	0.5153	34,049

Notes: Predictive performance of a gradient boosting model for predicting treatment status from mobile phone records from the treatment period. Predictions are obtained over 5 fold cross validation, and the pooled area under the curve (AUC) score is reported.

Table S12: Variation Between and Within Canton

Outcome	(1) Between Variance	(2) Within Variance	(3) Ratio
Panel A: Pre-treatment survey			
PMT	4.30	0.29	15.00
Food Security	1.438	0.96	1.44
Financial Health	1.25	0.98	1.27
Perceived Socioeconomic Status	1.80	0.95	1.89
Labor Supply	1.58	0.94	1.67
Panel B: Endline survey			
PMT	2.08	0.25	8.21
Food Security	1.85	0.99	1.87
Financial Health	1.65	0.99	1.87
Financial Inclusion	1.59	0.92	1.72
Mental Health	1.89	0.98	1.92
Perceived Socioeconomic Status	1.26	0.99	1.28
Healthcare Access	1.46	1.00	1.45
Labor Supply	1.57	0.96	1.63
Aggregate Welfare Index	2.03	0.98	2.08

Notes: Between vs. within variance, with groups defined by canton (self-reported in the baseline survey, determined by location of Novissi registration in the endline survey). Only individuals in cantons with at least 10 individuals surveyed are included in the analysis. All outcomes except for the PMT are standardized to 0 mean and unit variance in the control group.

Table S13: Feature importances in a machine learning model including mobile money data

Feature	**Importance**
Maximum amount of transactions in category "other"	281
Maximum balance before outgoing transactions	246
Mean balance before outgoing transactions	211
Maximum balance before outgoing transactions in category "other"	172
Mean amount of transactions in category "other"	145
Number of outgoing transactions	140
Number of outgoing transactions in category "other"	123
Mean balance before outgoing transactions in category "other"	114
Mean balance after outoging transactions in category "other"	105
Maximum balance before outgoing transactions in category "other"	102

Notes: Feature importances for machine learning model predicting treatment status from mobile phone data *including data on mobile money transactions* using six months of phone data from during the treatment period (see Section 5.5.1). Feature importances are derived from the gradient boosting model as the total number of times a feature is split upon in the entire ensemble of regression trees. Only the top 10 most important features are shown.

Table S14: Reasons for non-registration to Savanes-Novissi

	Control	Treatment
(1) I did not think I was eligible.	16.9% [46]	23.8% [115]
(2) I did not think I needed to register since I registered with GD-Novissi.	5.9% [16]	9.7% [55]
(3) I heard about the program after the end of the registration period.	13.7% [37]	16.1% [80]
(4) I receive GD-Novissi, I don't need extra money.	0.6% [2]	2.2% [13]
(5) I have enough money, I don't need extra money.	1.2% [3]	0% [0]
(6) Other people are more in need than me, I prefer them to get the money.	0.9% [2]	0.2% [1]
(7) Other.	27.8% [68]	16.8% [87]
(8) I don't know or refuse.	33% [78]	30.8% [156]
Total	100% [252]	100% [507]

Notes: Answers to the impact evaluation survey question "Why didn't you attempt to register for Savanes-Novissi in March of this year?", separately between the treatment and control groups (restricted to subscribers who earlier answered that they had not attempted to register for Savanes-Novissi). Observations are weighted by sampling probabilities. Counts are shown in square brackets.

Table S15: Registration with Savanes-Novissi

	(1) Food security	(2) Financial health	(3) Financial inclusion	(4) Mental health	(5) Perceived status	(6) Health care access	(7) Labor supply	(8) **All seven indices**
T, non-SN	0.148***	0.063	0.041	0.091***	0.110***	0.023	0.017	0.131***
	(0.037)	(0.042)	(0.034)	(0.030)	(0.037)	(0.038)	(0.044)	(0.039)
T, SN	0.039	-0.036	0.053	0.070*	0.074*	-0.010	-0.049	0.037
	(0.040)	(0.045)	(0.039)	(0.036)	(0.040)	(0.045)	(0.051)	(0.044)
C, non-SN	0.042	-0.023	0.038	-0.017	0.085**	0.052	-0.036	0.038
	(0.038)	(0.044)	(0.039)	(0.033)	(0.040)	(0.038)	(0.046)	(0.041)
C, SN Mean	-0.01	0.02	-0.01	0.04	0.00	0.04	0.01	0.03
F-test 1-2	7.43***	5.55**	0.10	0.39	0.85	0.57	2.39	5.37**
Obs.	4,755	4,755	4,755	4,755	4,755	4,755	4,755	4,755

Notes: Results for regressing the main survey outcomes on the interaction of GD-Novissi treatment status and Savanes-Novissi registration status. *T* indicates treatment, *C* indicates control, and *SN* and *non-SN* indicates beneficiaries and non-beneficiaries Savanes-Novissi, respectively. *F-test 1-2* row provides the p-value of the statistical comparison of the coefficients for "Treatment, not Savanes-Novissi" and "Treatment, Savanes-Novissi". All regressions control for the enumerator, week of the survey, and strata fixed effects. All observations are weighted by sampling probabilities. Robust standard errors are in parentheses. $^{*}p$ <0.1; $^{**}p$ <0.05; $^{***}p$ <0.01.

Table S16: Effects of the Savanes curfew on mobility

	Number of unique towers
β_0	-0.305*** (0.120)
β_{-2}	-0.037 (0.117)
β_{-3}	-0.010 (0.119)
N	18,398
Sample Mean	5.76

Notes: Differences-in-differences specification for identifying the impacts of the Savanes curfew on mobility proxied from mobile phone records. β_0 is the coefficient of interest, representing the effect of February (the month of the curfew) relative to January (pre-curfew). β_2 and β_3 represent the effects of November and December (relative to the month of January). *p<0.1; **p<0.05; ***p<0.01.

Table S17: Prices across regions

	(1) Milk (sachet)	(2) Sugar (sachet)	(3) Onion (pile)	(4) Palm oil (liter)	(5) Bread (unit)	(6) Orange (pile)	(7) Niebe (bowl)
Savanes region	50.3*	2.1	-7.7	68.4***	5.3	6.7	-74.9***
	(28.1)	(4.4)	(5.6)	(19.9)	(6.7)	(9.6)	(19.9)
Not-Savanes Mean	352.4	105.3	197.2	875.2	229.5	199.9	1461.5
Obs	1,097	3,200	4,822	2,120	2,524	1,473	2,316

Notes: This table provides price differences between Savanes and non-Savanes regions, based on prices of goods reported by impact evaluation survey respondents. The dependent variables are indicated in the column title. All regressions control for enumerator, week of survey, and treatment status fixed effects. All observations are weighted by sampling probabilities. Robust standard errors are in parentheses. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.