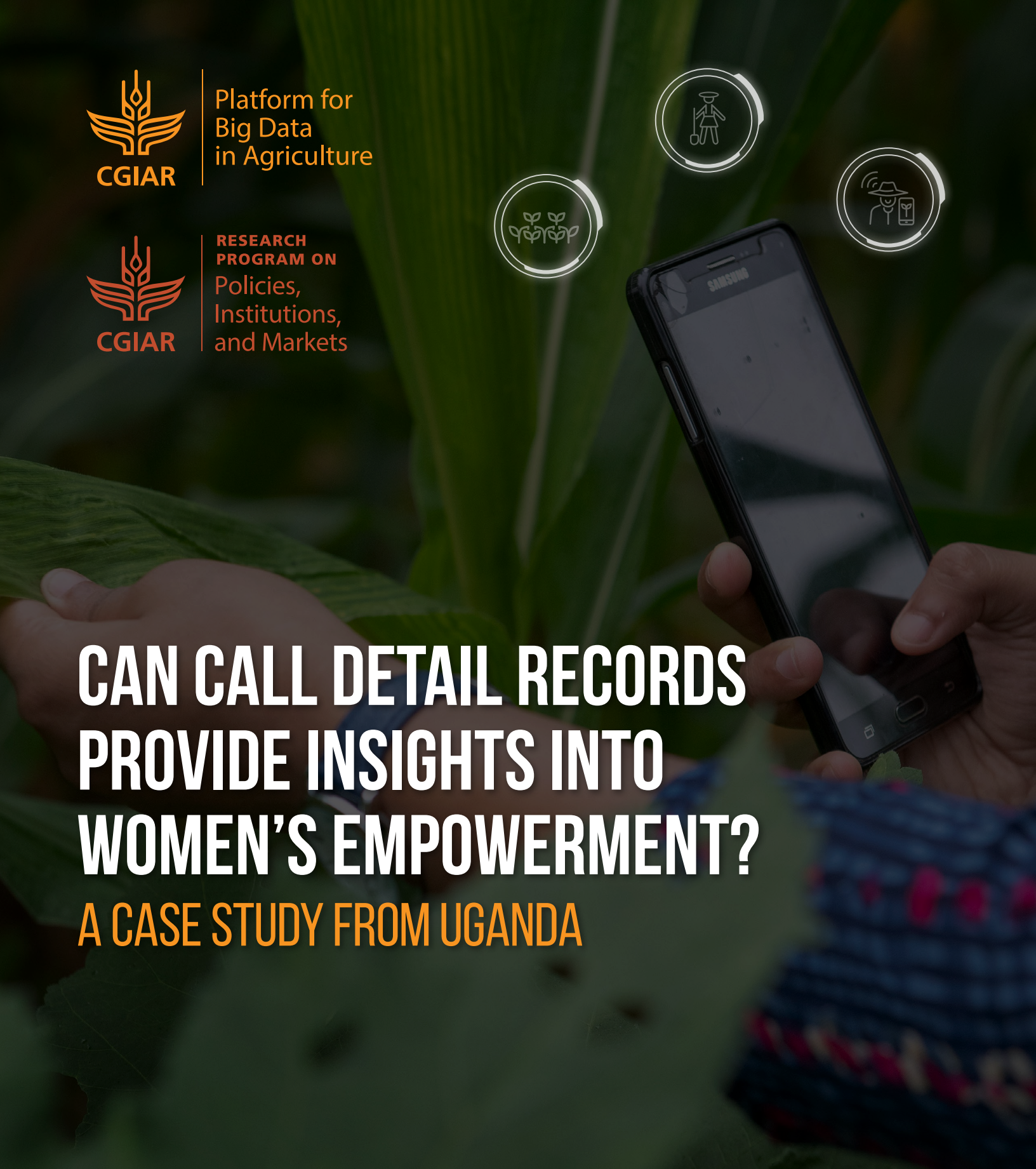
CGIAR Platform for Big Data in Agriculture

CGIAR RESEARCH PROGRAM ON Policies, Institutions, and Markets



CAN CALL DETAIL RECORDS PROVIDE INSIGHTS INTO WOMEN'S EMPOWERMENT?

A CASE STUDY FROM UGANDA

Dalberg

Alliance

Bioversity International

CIAT International Center for Tropical Agriculture

KIT Royal Tropical Institute

IFPRI INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE

The **CGIAR Platform for Big Data in Agriculture** is a cross-cutting program of the global CGIAR consortium of non-profit research institutes looking into virtually every aspect of food security spanning: genomics, breeding, agroecology, climate science, and the socioeconomic drivers and context of food systems change. The Platform tends to data standards and data sharing, digital innovation strategy and technology transfer, and research into the intersection of digital technologies and agricultural development in emerging regions.
CGIAR is a global research partnership for a food secure future dedicated to reducing poverty, enhancing food and nutrition security, and improving natural resources.

https://bigdata.cgiar.org/

The **CGIAR Research Program on Policies, Institutions, and Markets (PIM)** leads action-oriented research to equip decisionmakers with the evidence required to develop food and agricultural policies that better serve the interests of poor producers and consumers, both men and women. PIM combines the resources of CGIAR centers and numerous international, regional, and national partners. The program is led by the International Food Policy Research Institute (IFPRI).

www.pim.cgiar.org

Citation:
Slavchevska V; Tyszler M; Burra DD; Seymour G; Sementsov D; Van Lierde A; King B. 2021. Can call detail records provide insights into women's empowerment? A case study from Uganda. CGIAR Platform for Big Data in Agriculture; CGIAR Research Program on Policies, Institutions, and Markets (PIM). 26 p.

Design and layout: Ximena Hiles

Photo credits: International Center for Tropical Agriculture (CIAT) and CGIAR System Organization

Photos available on Flickr: https://www.flickr.com/photos/ciat/ and https://www.flickr.com/photos/cgiarconsortium/

Conflict of interest: The authors declare that there are no known conflicts of interest associated with this research; the financial support for this work was provided for research purposes only and did not influence the findings.

January 2021

CGIAR Platform for Big Data in Agriculture

CGIAR RESEARCH PROGRAM ON Policies, Institutions, and Markets

CAN CALL DETAIL RECORDS PROVIDE INSIGHTS INTO WOMEN'S EMPOWERMENT?

A CASE STUDY FROM UGANDA

Vanya Slavchevska,[1] Marcelo Tyszler,[2] Dharani Dhar Burra,[1] Greg Seymour,[3] Denys Sementsov,[4] Astrid Van Lierde,[4] and Brian King[5]

[1] Alliance of Bioversity International and the International Center for Tropical Agriculture (CIAT), Cali, Colombia
Corresponding author: v.slavchevska@cgiar.org
[2] KIT Royal Tropical Institute, Amsterdam, the Netherlands
[3] International Food Policy Research Institute (IFPRI), Washington, DC, USA
[4] Dalberg Data Insights, Brussels, Belgium
[5] CGIAR Platform for Big Data in Agriculture, Cali, Colombia

Dalberg

Alliance
Bioversity International
CIAT International Center for Tropical Agriculture Since 1967 Science to cultivate change

KIT Royal Tropical Institute

INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE
IFPRI

ACKNOWLEDGEMENTS

This work was undertaken as part of the CGIAR Platform for Big Data in Agriculture, and the CGIAR Collaborative Platform for Gender Research (2017–2019), which was a part of the CGIAR Research Program on Policies, Institutions, and Markets (PIM). The team of scientists were from the Alliance of Bioversity International and CIAT, Dalberg Data Insights, the International Food Policy Research Institute (IFPRI), and KIT Royal Tropical Institute. Funding support for this study was provided by the CGIAR Platform for Big Data in Agriculture and the CGIAR Collaborative Platform for Gender Research (PIM). This publication has not gone through IFPRI's standard peer-review procedure. The opinions expressed here belong to the authors, and do not necessarily reflect those of PIM, IFPRI, the CGIAR Platform for Big Data in Agriculture or CGIAR.

The authors would like to thank Cheryl Doss and Jessica Heckert for their helpful review and comments on the phone survey and Mayda Calderon for research assistance with the literature review.

ABSTRACT

Across the globe, sex-disaggregated data to track gender equality and women's empowerment remain scarce as they cover few countries and are collected irregularly. There has been a growing interest in identifying alternative data sources that are common across countries and can provide higher spatio-temporal coverage to measure and monitor progress on women's empowerment and gender equality. This study explores one such data source: mobile phone usage data, also called call detail records (CDRs). We use CDRs of mobile phone users in Uganda combined with data from a phone survey to train machine-learning models to predict the sex of the mobile phone user and several indicators of economic empowerment such as ownership of a house and land, occupation, and decision-making over household income. The most accurate of the models predicts the sex of the mobile phone user with 78% accuracy. The different indicators of economic empowerment are predicted with accuracies ranging from 57% to 61%. We also predict users' sex and economic empowerment jointly. When we first predict the sex of the user and then economic empowerment, no noticeable improvements occur in the predictive accuracies over the separate predictions for the five indicators. However, when we predict economic empowerment and then the sex of the user, we achieve high accuracy rates ranging from 81% to 87%. Mobile phone usage data hold potential for gender research although they are not without limitations.

KEYWORDS

CALL DETAIL RECORDS (CDRS); MOBILE PHONE METADATA; MACHINE LEARNING; WOMEN'S EMPOWERMENT; AFRICA

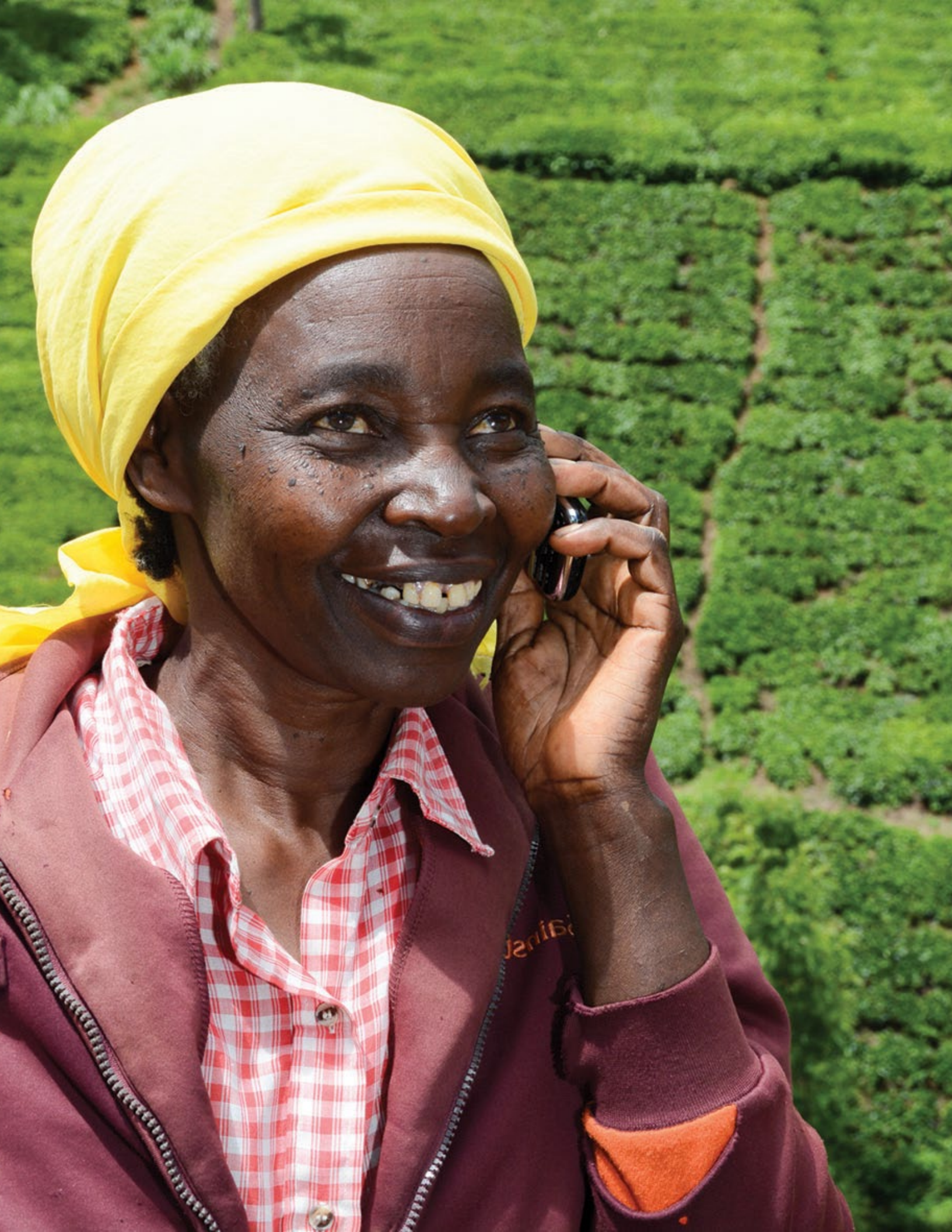
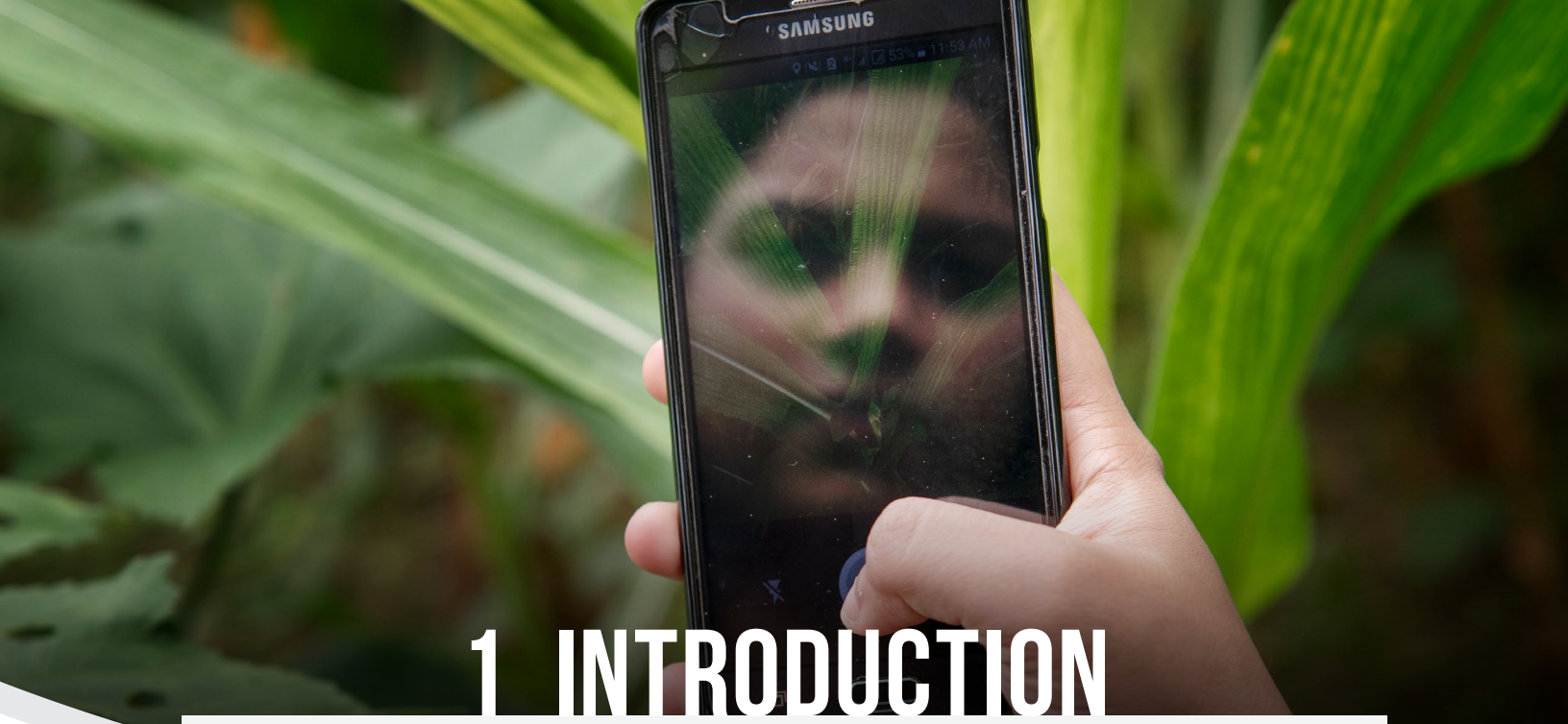


TABLE OF CONTENTS



1 INTRODUCTION

Despite significant improvements in the collection of data on gender equality and women's empowerment, significant gaps in the types of data collected, geographic coverage, and frequency of data collection remain. Most large-scale household and agricultural surveys rarely collect sex-disaggregated data to enable governments to track indicators of gender equality (Hillesland et al., 2020). The few that do, such as the Living Standards Measurement Study – Integrated Survey on Agriculture (LSMS-ISA), Demographic Health Surveys (DHS), and Women's Empowerment in Agriculture Index (WEAI) (Alkire et al., 2013), are implemented infrequently and cover few countries, thus impairing efforts to monitor progress on gender equality and women's empowerment globally.

For these reasons, there is a growing need to identify new data sources and respective methodologies to fill gender data gaps. One relevant data source with significant potential is mobile phone usage data, also called **call detail records** (CDRs). CDRs are produced every time mobile phones are used to call, send short messages (SMS), or browse the internet. CDRs capture a wealth of information on calling patterns (e.g., time and duration of calls), mobility (captured by cell towers through which calls were routed), and social networks (e.g., number of unique contacts, etc.), which have been shown to be correlated significantly with the socioeconomic status of the users (Blumenstock et al., 2015; Khaefi et al., 2019). CDRs coupled with ground-truth data for training machine-learning models can be used to classify and predict the characteristics of large populations, potentially without the need for large-scale surveys. Past studies have used CDRs to assess various socioeconomic dimensions at both an aggregate geographic level and an individual level. Studies have used CDRs to look at population estimates (Deville et al., 2014), poverty (Blumenstock et al., 2015; Steele et al., 2017), human migration (Lu et al., 2016), the spread of epidemics (Balcan et al., 2009; Meloni et al., 2011; Wesolowski et al., 2015), disaster response (Bengtsson et al., 2011); literacy (Montjoye et al., 2013), and even personality (Montjoye et al., 2013).

This study assesses the potential of using CDRs to predict the sex and economic empowerment of mobile phone users as well as whether sex and economic empowerment can be predicted jointly. Several earlier studies have used CDR data and developed models to predict the sex of mobile phone users (Al-Zuabi et al., 2019; Dalberg Data Insights, 2019; Felbo et al., 2017; Frías-Martínez et al., 2010; Jahani et al., 2017). Studies have also used CDRs to predict individual-level proxies for socioeconomic status (Gutierrez et al., 2013; Sundsøy et al., 2016a). This study contributes to the literature by assessing the feasibility of predicting unique indicators of economic empowerment such as employment, land and house ownership, and decision-making over income. Moreover, to the best of our knowledge, this is the first study to model and predict indicators of economic empowerment jointly with the sex of users, thus providing insights into the use of CDRs to monitor indicators of gender equality and women's empowerment.

This study uses two different types of data from subscribers of MTN Uganda, one of the largest mobile service providers in the country. First, we conducted a phone survey to collect ground-truth data from a representative random sample of subscribers of MTN Uganda. Then, using the respondents' phone numbers, we merged the ground-truth data with the respective CDRs. Building on the methodology developed by Dalberg Data Insights (2019) to predict the sex of mobile phone users, we train and test the predictive accuracy of five different models for predicting mobile phone users' sex, land ownership, house ownership, occupation, and decision-making power over household income. The sex of mobile phone users is predicted with 78% accuracy. Land and home ownership are predicted with 64% and 65% accuracy, respectively. Users' occupation, which comprises three categories (whether the user works in agriculture, in non-agriculture, or is not in the labor force), is predicted with 57% accuracy. The model for decision-making over household income does not perform any better than simply predicting the majority category, and this is likely because the indicator is highly unbalanced with more than 80% of the full sample reporting having some decision-making over household income.

In addition, we test whether it is possible to improve the predictive accuracy of the models by sequentially predicting users' sex and economic empowerment status. When we first predict the sex of the user and then economic empowerment, no noticeable improvements occur in the predictive accuracies over the separate predictions for the five indicators. However, when we predict economic empowerment and then the sex of the user, we achieve high accuracy rates ranging from 81% to 87% depending on the respective indicators.

The results are encouraging and illustrate that CDRs could be operationalized for targeting of development projects and programs as well as for tracking progress on gender equality and women's empowerment at the national and sub-national levels. There are, however, some important biases and ethical concerns regarding CDRs, which we touch upon in the discussion.

The article is organized in the following manner. In section 2, we review the related literature. In section 3, we discuss the approach as well as the data sources. Section 4 presents the results, and section 5 concludes with a discussion of the results and potential concerns with using CDRs.

2 RELATED LITERATURE

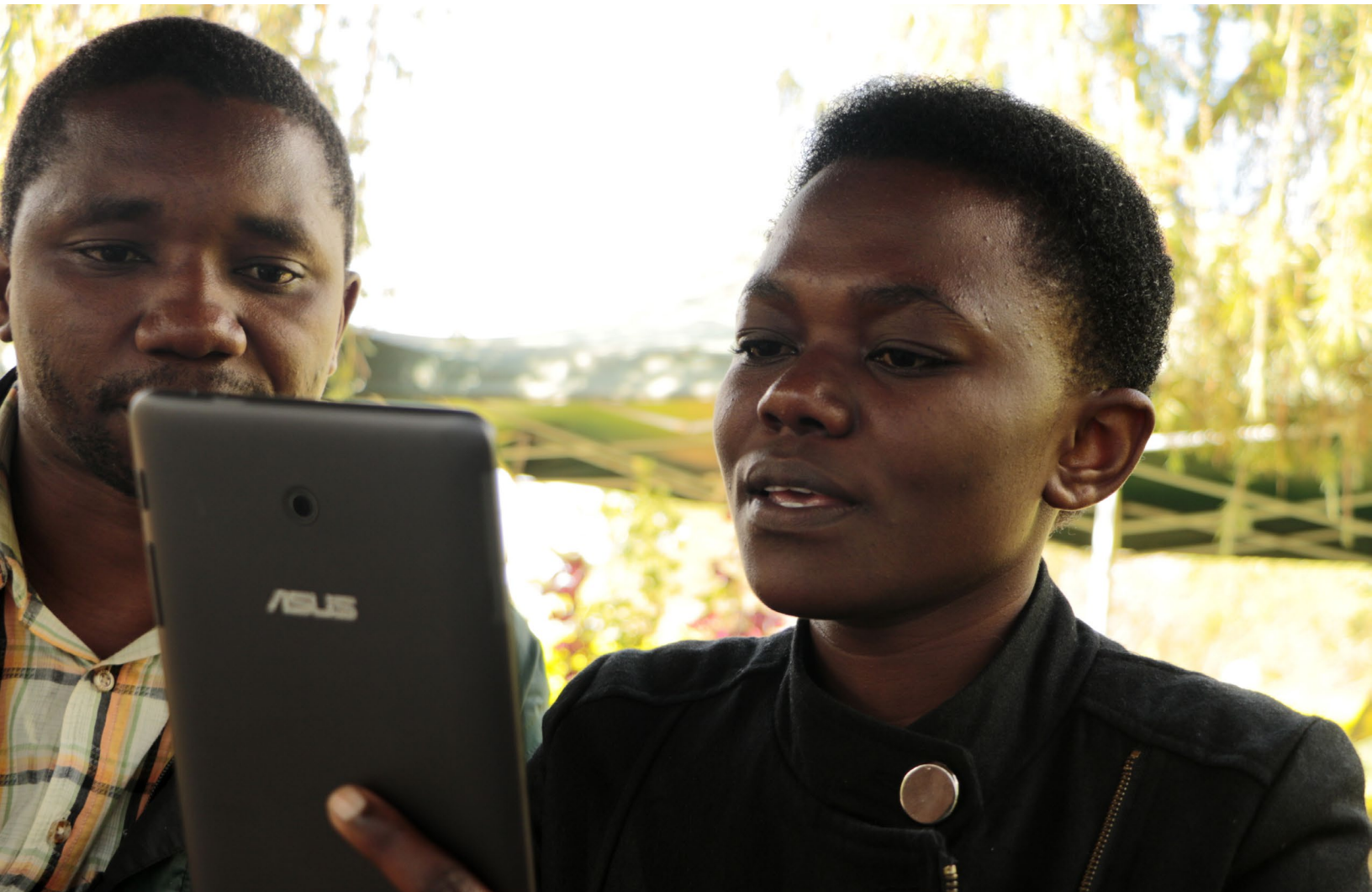
At the center of this study is the assumption that certain characteristics of mobile phone usage, social networks, mobility, and recharging patterns are correlated with empowerment and its intersection with the sex of the mobile phone user. Following Kabeer (1999), empowerment – the ability to exercise strategic choices – requires resources (including current access and future claims to material, human, and social resources), agency (the process of decision-making and negotiation) and achievements (well-being outcomes). Although no other study has explicitly tested CDRs for predicting the empowerment of users, several studies have examined the linkages between CDRs and socioeconomic well-being and between CDRs and the sex of mobile phone users, and these studies are related to our work.

Using a diverse set of indicators for socioeconomic status, several studies have examined the relationship between features of CDRs and socioeconomic status at both the community and individual level. This is motivated by the fact that mobile phone users of different socioeconomic status in terms of education, occupation, and wealth are likely to have distinct patterns of phone usage. For example, compared with well-off users, poor users can be expected to make fewer and shorter calls, have smaller networks of contacts, and travel shorter distances; they may also select different phone plans, including topping up phones more frequently but with lower amounts (Gutierrez et al., 2013). Hundreds of CDR indicators could be constructed and used to characterize users and to classify them in different socioeconomic groups. Using fixed-line CDRs from England, Eagle et al. (2010) find that social diversity, captured by the diversity of individuals' social networks, is a good proxy for the socioeconomic development of regions. A similar study from France finds that human mobility patterns, measured by the distance traveled by an individual (radius of gyration) and the diversity of movements over her locations (mobility entropy), are strongly correlated with socioeconomic indicators such as education, unemployment, income, and deprivation (Pappalardo et al., 2015). In Ecuador, the volume of mobile phone activity has also been linked to higher income and higher education with no differences for men and women (Castillo et al., 2018).

Besides the diversity of CDR features used to map socioeconomic status, a considerable diversity of socioeconomic indicators is used across studies. Several studies have looked at poverty. Smith-Clarke et al. (2014) propose a methodology that aggregates mobile phone usage features at a cell tower level to estimate poverty rates in Côte d'Ivoire at granular geographic levels. In Sri Lanka, Fernando et al. (2018) show that CDRs can be useful in measuring indicators of socioeconomic development, proxied by the characteristics of the dwelling, and this can be done in both post-conflict regions as well as in fast-developing urban

regions. Several studies have gone a step further and proposed machine-learning models for estimating socioeconomic development. Using aggregate cell-phone data, Soto et al. (2011) predicted indicators of the socioeconomic status of a population, achieving correct prediction rates of more than 80% for an urban population. More recently, Blumenstock et al. (2015) developed a model to predict wealth using a composite wealth index at an individual level. They used both CDRs and a phone survey, and used the model to predict the wealth of the out-of-sample population at both the national and district level. Steele et al. (2017) showed that combining CDRs with satellite data was better positioned to model traditional measures of poverty at disaggregated geographic levels.

CDR indicators capturing social, economic, and mobility patterns of users have been used to predict users' employment status and even their profession. By deriving economic, social, and mobility features for each mobile user, Sundsøy et al. (2016b) predicted individual employment status with up to 70.4% accuracy. The study goes a step further by showing that the data could be aggregated at cell tower resolution, suggesting a potential opportunity to provide a mapping of labor market conditions, including labor market vulnerabilities, across geographic areas. In fact, Toole et al. (2015) developed a method to detect mass layoffs to predict changes in aggregate unemployment rates using CDRs. Focusing on the closure of a large manufacturing plant, they describe a structural break model to detect the date and the size of the mass layoff. They use a Bayesian classification model to identify affected individuals by observing changes in calling behavior following the plant's closure. Job loss is associated with a significant reduction in social and mobility behavior and, because these behavioral changes are captured in CDRs, CDRs could be used in monitoring unemployment trends in near real time, especially around significant economic shocks or policy changes.



In addition to socioeconomic status, studies have examined how men and women may use phones differently. Differences in usage patterns could be used for predicting the sex of the user, which, along with other demographic information about the user, is often unavailable or unreliable, especially in developing countries where the use of pre-paid phones is common. Mehrotra et al. (2012) use transaction data from Rwanda to show that men are significantly more active phone users; but, although they are more active during the day, women are active in the evenings. The study also finds significant gender differences in phone activity around holidays and politically important days. Dalberg Data Insights (2019) find that women tend to have fewer calls, and most of the calls are incoming but with longer average duration. Women also have fewer contacts, travel less on average, and top up their phones less frequently and with smaller amounts than men (Dalberg Data Insights, 2019).

Several recent studies have proposed machine-learning models to predict the sex of mobile phone users. Jahani et al. (2017) developed a framework with more than 1,400 features derived from CDRs that could be used to predict individual characteristics. They showed that models, trained on 10,000 users, predicted users' sex with 74.3% to 88.4% accuracy in a developed country and with 74.5% to 79.7% accuracy in a developing country. Rather than focusing on developing advanced features from the data, Felbo et al. (2017) exploited the raw CDRs focusing on the temporal modality. Their methods reached 78% predictive accuracy for sex and 62% accuracy for age. More recently, Al-Zuabi et al. (2019) applied different machine-learning models to predict mobile phone users' sex and age based on CDRs as well as customer service and billing data. They used data on 18,000 users provided by SyriaTel Telecom Company. The model achieved 85.6% accuracy in predicting the sex of users and 65.5% accuracy in predicting their age.

Dalberg Data Insights, in partnership with GSMA, developed the Gender Analysis and Identification Toolkit (GAIT) (Dalberg Data Insights, 2019), which calculates more than 150 different indicators summarizing mobile phone usage for each subscriber, and feeds the indicators into a machine-learning model, which are used to predict the sex of subscribers. In the case of Uganda, the GAIT method was able to predict sex with 72% accuracy. Building exactly on this work, we extend the analysis to predict indicators of economic empowerment, including users' main occupation, ownership of key assets, and control over income as well as the intersection of these indicators with the sex of the users. To the best of our knowledge, no previous studies have examined the feasibility of predicting women's (and men's) economic empowerment from CDRs.

3 DATA AND METHODS

As mentioned in the introduction, this study uses two sources of data: a phone-based survey with MTN subscribers in Uganda and de-identified CDRs of the MTN subscribers who participated in the phone survey. The CDRs were used to train machine-learning classifiers to predict indicators collected from the survey.

3.1 PHONE SURVEY

In October 2019, we conducted a phone survey of randomly selected MTN subscribers in Uganda. Phone numbers were randomly selected from a database of more than 12 million subscribers using a sampling probability proportional to the population defined in the MTN database, stratified by rural-urban status for each of 11 administrative sub-regions (Southwest, Karamoja, East, Central 1, Central 2, Teso, Lango, Acholi, Elgon, Western, and West Nile). The objective was to reach at least 10,000 successful interviews. This was deemed a minimal threshold to have a representative sample suitable for training the machine-learning model. All ethical protocols were followed during data collection, management, and analysis, and the Institutional Review Board (IRB) of the International Center for Tropical Agriculture (CIAT) (now part of the Alliance of Bioversity International and CIAT) provided the clearance letter.

The phone survey was designed to minimize response burden and refusals to participate. It included seven questions (see Annex on p. 24). The questions were informed and aligned in phrasing with questions used in national surveys, such as Living Standards Measurement Studies (LSMS) and Demographic and Health Surveys (DHS). The survey included questions about the respondents' age, sex, main occupation, land and house ownership, and decision-making about household income. Questions were asked to verify whether the person who answered the phone was the main user of the current subscriber identification module (SIM) card (i.e., the person who uses the phone more than anyone else).[1] If the person who answered the phone was not the main user or if the person was younger than 18 years old, the interview was stopped; enumerators were instructed to attempt (up to three times) to arrange a time to speak with the main user.

A total sample of 10,417 respondents (main users of the SIM card) was successfully interviewed. The sample comprised 3,946 (38%) women and 6,471 (62%) men. Table 1 provides descriptive statistics. Almost half of the respondents are married (43%), followed by single (32%). Slightly more than a quarter of the respondents are subsistence farmers, while 23% are self-employed outside of agriculture.

1 SIM cards are inserted into a mobile phone to enable customers to access a mobile operator's network. Every SIM card has a unique telephone number.

Significant gender gaps occur in occupation, decision-making, and asset ownership. Twenty-six (26%) percent of the men are self-employed outside of agriculture vis-à-vis 17% of the women; 19% of the men are non-agricultural wage workers in contrast to 13% of the women; and 12% of the men are commercial farmers vis-à-vis 5% of the women. Approximately 84% of the respondents report making decisions regarding the use of household income, either alone or jointly with their partner. Nearly half of the men indicate that they make decisions alone compared with only 40% of the women. Moreover, women are more likely than men to report that their spouse is the main decision maker (10% of the women versus 2% of the men). Forty-one percent of the men own both a house and land compared with 27% of the women. One half of the women do not own either a house or land compared with 32% of the men.

Based on the questionnaire, we construct four different proxies for economic empowerment, focusing on land ownership, house ownership, decision-making over household income, and occupation. Land ownership is an indicator equal to one if the respondent owns any land, alone or jointly with someone else, and zero otherwise. House ownership is equal to one if the respondent owns a house, either alone or jointly with someone else, and zero otherwise. Decision-making over income is also defined as a binary indicator and is equal to one if, over the last 12 months, the respondent, alone or with his or her partner, usually made the decisions about the use of the total household income. Finally, we grouped the information on respondents' occupation in the last 12 months into three categories. The first category is agriculture and includes subsistence farmers, commercial farmers, and agricultural wage workers. The second category is non-agriculture, which includes non-agriculture self-employment and wage employment with or without a contract. The third category comprises people who have been unemployed or not in the labor force such as retired and unemployed people who are not looking for work, housewives, and students.

3.2 CALL DETAIL RECORDS (CDRs)

The call detail records (CDRs) for the sampled subscribers cover the 2-month period from September to October 2019. This includes voice call metadata (timestamp, phone number calling, phone number called, duration), short message service (SMS) metadata (timestamp, phone number sending message, phone number being messaged), internet usage metadata (timestamp, phone number usage, and amount), top-up metadata (timestamp, phone number topping up, and amounts), and antenna location. We leveraged upon the GAIT approach to develop individual subscriber-level indicators from CDRs that capture 167 distinct features with respect to phone usage, social networks, mobility, and recharging patterns (Dalberg Data Insights, 2019). Examples of phone usage indicators include the number and duration of calls, share of outgoing activity, share of text messages, and volume of data used. Examples of social network indicators include the number of unique contacts and a social diversity score, which reflects how an individual divides his/her time among different contacts. Location data, based on cell phone tower, were used to classify each phone number as urban or rural. All indicators are normalized to have zero-mean and unit-variance. A full list of the indicators derived from the CDRs is available upon request.

3.3 MACHINE-LEARNING MODELS[2]

Machine learning (ML) is primarily focused on prediction. Given some observed data on y and x, how do we predict y given new values of x? ML models try to find the set of x-variables (or "features") and respective coefficients that achieve the highest degree of prediction accuracy when using never-before-seen data.

In building ML models, the data are typically divided into training and testing sets. The training set is used for estimating (or "training") a model. ML algorithms have different hyperparameters that control the model

2 This section draws on the excellent descriptions of machine-learning approaches provided by Varian (2014), Mullainathan and Spiess (2017), and Storm et al. (2020).

complexity and behavior. Choosing the best hyperparameter values is typically done using a process of *k*-fold cross-validation. First, the testing set is randomly partitioned into *k* subsamples (or "folds"). Then, the model is iteratively estimated for a range of hyperparameter values, holding out a single fold for validation. The model with the best average performance on the validation set is selected. Finally, the testing set is used to evaluate how well the chosen model performs.

In this study, we use the extracted CDR features and phone survey data to train four different supervised learning models to predict five characteristics of mobile phone users (referred to as "classification problems"): sex, land ownership, house ownership, occupation, and decision-making power over income. The ML algorithms we test are K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest, and Extreme Gradient Boosting (XGBoost).

Initially, we treat each classification problem as independent. We separately predict the sex of the mobile phone user, along with each of the four empowerment indicators: land ownership, house ownership, occupation, and decision-making power over household income.

In the next step of our analysis, we explore the predictive accuracy of the models wherein CDRs are used to predict both the sex of the mobile phone user and each of the empowerment indicators. Across many developing countries, the sex of a mobile phone user is often not collected and registered by the Mobile Network Operator (MNO). Even if it is collected, it is inaccurate because of phone sharing and/or the use of pre-paid phone cards. One recent study in Bangladesh found that 78% of female mobile phone users were, in fact, registered as male (GSMA, 2018). Second, no clear evidence exists on how men's and women's mobile phone usage patterns and the underlying CDR features relate to economic empowerment. To address these issues, we train a model to predict the sex of the user and then use the subsample of users predicted as women to retrain the models to classify phone users according to the economic empowerment indicators. Next, we reverse the order. We first train a model to classify phone users according to economic empowerment indicators and then retrain the model to predict the sex of the users within each classified group.

In the final step of our analysis, we use the best performing model to classify and predict across the entire user base of 12.7 million MTN subscribers according to economic empowerment status and sex.

4 RESULTS

Our primary metric of model performance is prediction accuracy on the testing set. As a baseline comparison for evaluating different models, we used a naïve classification based on the majority class reflected in the data. For example, given that 38% of mobile phone users in the sample are female, a model that always classifies the phone user as male would be expected to achieve a prediction accuracy of roughly 62% (the proportion of men in the sample). The better a model performs, the more it will exceed this baseline accuracy. The XGBoost model achieved the highest average testing prediction accuracy for sex, house ownership, decision-making, and occupation; the SVM model achieved the highest average accuracy for land ownership.

In addition to accuracy, we report sensitivity and specificity values for each model. Sensitivity is a statistical metric that assesses a model's ability to accurately predict positive results (i.e., the probability that an actual positive result is classified as such). Specificity measures how well a model accurately predicts negative results (i.e., the probability that an actual negative result is classified as such).[3] For example, given a binary indicator for the sex of the user, which is equal to one if the user is female and zero if the user is male, sensitivity refers to the proportion of correctly predicted female users and specificity refers to the proportion of correctly predicted male users.

4.1 PERFORMANCE OF THE MODEL WHEN PREDICTING EACH CLASSIFICATION PROBLEM INDEPENDENTLY

Table 2, column 2 presents the testing predictive accuracy achieved by independently predicting each of the five classification problems. The sex of the respondent is predicted with an accuracy of 78% vis-à-vis a baseline accuracy of 62%. However, 85% of the men are correctly predicted as men, while 66% of the women are correctly predicted as women, suggesting that the obtained accuracy can be ascribed to the identification of male subscribers. The most predictive features for distinguishing men from women are indicators of basic phone usage (such as the ratio of the duration of incoming over outgoing calls, duration of incoming voice calls) and social network indicators (such as the number of unique contacts) (see Table 3 for details).

3 https://bit.ly/3qORncp

The best machine-learning model predicts ownership of a house with 65% accuracy and ownership of land with 64% accuracy, improving meaningfully on the baseline accuracy of 51% and 52%, respectively. Approximately 84% of the survey respondents reported participating in the decisions around household income, either solely or jointly with the spouse/partner. The best machine-learning model improves this accuracy by a mere 5 percentage points, which is not surprising given how skewed the indicator is toward one of the response categories.

The final indicator of interest is the occupation of the respondent, which is aggregated into three categories: (i) agricultural work, (ii) non-agricultural work, and (iii) not in the labor force (housewives, students, and retired respondents). The model predicts these categories with an accuracy of 57%, which is a significant improvement over the 44% baseline accuracy.[4]

Across all the models used for independent predictions, the XGBoost model for decision-making variable obtained the lowest sensitivity value (6%), while the XGBoost model for predicting the sex of the mobile phone user obtained the highest sensitivity value (66%). Similarly, SVM with rbf Kernel for land ownership obtained the lowest specificity value (60%), while the decision-making variable prediction using XGBoost showed the highest specificity value (98%).

4.2 PERFORMANCE OF THE MODEL WHEN JOINTLY PREDICTING THE SEX OF THE USER AND EMPOWERMENT INDICATORS

Table 2, column 3 shows out-of-sample predictive accuracy when we first predict the sex of the user and then retrain the classifier to predict the empowerment indicators. We compare the performance of the current approach with another, wherein first the sex of the user is predicted followed by prediction of economic empowerment indicators, without re-training the classifier, that is, predictions are made with the model trained with both sexes. These are shown in Table 2, column "trained on both sexes, tested on predicted as women." The results are similar, suggesting that there is no benefit to re-training the classifier, as the accuracy of the model that is trained and tested on both sexes is similar to the one that is trained on both sexes, but tested only on those predicted as women (see Table 2 for additional details). Additionally, the predictions for both instances are significantly higher than the baseline predictions, except for the indicator for decision-making over household income (Column 1 of Table 2).

We also examine how the model performs when the indicators of economic empowerment are predicted first, followed by the prediction of whether the user is a man or a woman. These results are reported in Table 4. We observe that the sex of the user is predicted with 70% accuracy (83% sensitivity and 50% specificity) when the user makes decisions alone or jointly over income, and with 78% accuracy (60% sensitivity and 80% specificity) when the respondent does not make the decision. Similarly, the predictive accuracy for sex of the user is also high when ownership of assets (either land or a house) is predicted first. The predictive accuracy is 83% when the user owns either land or a house (81% sensitivity and 86% specificity for land ownership and 83% sensitivity and 87% specificity for home ownership), while it is 86% when the user does not own a house (67% sensitivity and 92% specificity), and is 83% (60% sensitivity and 94% specificity) when the user does not own land. The predictive accuracy for sex of the user is even higher when occupation/work is predicted first. For users whose primary occupation is predicted to be agriculture, their sex is predicted with 87% accuracy (68% sensitivity and 94% specificity). The accuracy of

4 We also test other aggregations of the occupation categories. In one, we distinguish between people with (i) paid employment (such as non agricultural self-employment, non-agricultural wage workers, and commercial farmers) and people out of work or in unpaid work (unemployed, subsistence farmers, and housewives). Marginal classes – those that constitute <1% of the sample and students – are excluded. The model's predictive accuracy for this categorization is 67.5%. Its performance relative to the baseline predictive accuracy of 57.5% is significant. Its disadvantage, however, is that extending the model to the complete database of users will likely obtain lower accuracy because it will not know how to classify the marginal classes. When we add back the excluded categories into a separate and third category, then the predictive accuracy of the model drops to 57.7% (baseline 55%).

predicting the sex for those engaged in non-agricultural work is 81% (97% sensitivity and 70% specificity). These results strongly suggest that training the classifier to predict economic empowerment and then the sex of the user provides higher prediction accuracies than the reverse approach, wherein prediction of sex is followed by prediction of empowerment indicators. This is a promising approach for obtaining more detailed information on women's and men's economic empowerment.

4.3 USING THE INDEPENDENT VARIABLE CLASSIFICATION MODELS TO PREDICT ACROSS THE ENTIRE MTN USER BASE

Using the best performing model for independent variable classification, we predict that women are 31% of the entire user base. With regard to empowerment indicators, we predict that nearly all have input in the decisions about household income, 70% own a house, 54% own land, and 43% work in agriculture. Among those predicted as women, roughly 50% are predicted to have input in decisions about household income, 50% own a house, 50% own land, and 21% work in agriculture (see Table 5).

5 DISCUSSION AND CONCLUSIONS

Data from mobile phones offer a wealth of new behavioral and spatial information and may hold potential to transform the way economic and gender analyses are carried out.

Women and men have distinct patterns of phone usage and, as a result, models to predict the sex of users achieve relatively high accuracy. The predictive accuracy for the sex of respondents is notably higher in this study than in the earlier study by Dalberg Data Insights (2019), which achieved a predictive accuracy of 72%. It is comparable to that of Felbo et al. (2017), whose model manages to predict the sex of users with 78% predictive accuracy. Our predictive accuracy is also within the range identified by Jahani et al. (2017), who estimate that standard machine-learning models using only CDRs and trained on 10,000 users are sufficient to predict users' sex with an accuracy of 74.5% to 79.7% in a developing-country context.

Novel to our study is the joint prediction of users' sex and economic empowerment. Most promising is the approach of first predicting the indicators of economic empowerment such as a house or land ownership and occupation, and then predicting the sex of the user. With this approach, we achieve predictive accuracy for the sex of users of more than 80% (except for decision-making over household income). This potentially indicates that the diversity in empowerment indicators is much larger than the diversity in the sex of users, which, once sorted, improves the accuracy of the successive sex prediction significantly, more than if we first predict sex and then empowerment indicators, with or without refitting of the classifier.

How can these findings be operationalized? To start with, approaches are needed to monitor progress on the SDGs for both men and women. Goal 5 aims at empowering all women and girls, and one key aspect is to improve access to and secure rights over land. Sex-disaggregated data on land ownership are not collected regularly by the national statistical offices (NSO) of developing countries (Doss et al., 2015; Hillesland et al., 2020; Kieran et al., 2015; Slavchevska et al., 2020), thus leaving significant gaps in the monitoring of progress toward improving women's land rights.

The approaches could also be used to allocate resources and target interventions with the individuals and at the places where they are most needed. This study shows that economic empowerment indicators could be predicted with high levels of accuracy. Moreover, it shows that, after predicting the indicators of economic empowerment, predicting the sex of the respondents achieves high accuracy too, giving governments and development partners near-real-time sex-disaggregated information about well-being and economic empowerment. These methods could also be deployed to measure changes in unemployment for both men and women in response to a wide range of shocks varying from global financial crises and epidemics to local closures of factories.

Although these new data and approaches demonstrate that they could be useful in monitoring indicators of economic empowerment, they also indicate areas where additional data and analyses are needed. First, in our study, the prediction of decision-making was difficult. Approximately 84% of the respondents indicated that they had a say in the decisions about household income; the machine-learning models could not improve over the naïve model of simply predicting the majority category. Although it is possible that most men and women in Uganda take part in the decisions about the use of household income, other issues such as phrasing of the question cannot be dismissed. The phrasing of the question on decision-making about household income was broad and was left to the respondents to report based on their own perceptions. Even if the question was phrased more narrowly, perceptions of one's decision-making role over household income may vary widely across individuals – from individuals who are sole decision-makers with complete autonomy over all household income to individuals with some but relatively limited control over some parts of the household income. The challenge of predicting this indicator raises questions as to whether this type of indicator could be meaningfully predicted from CDRs.

Second, the model learns from the data. Therefore, any biases incorporated in the data will be reflected in the model predictions. In their study of mobile phone users in Rwanda, Blumenstock and Eagle (2010) find that mobile phone owners are on average younger, wealthier, more educated, and predominantly male compared to the general population. They also find that women are more likely to use a shared phone than men, suggesting that the voices of a key group of mobile phone users are not heard. This is also likely the case in Uganda, but we did not collect sufficient socioeconomic data to compare how the mobile phone users differ from the general population. Besides being better off, women mobile phone users are likely to be more empowered than the women in the full population. Thus, findings obtained from mobile phone usage data cannot be extrapolated to the general population in cases where gaps in mobile phone penetration remain significant.

SIM sharing may further complicate the potential of CDRs for gender analyses. The current analyses are limited to mobile phone users who are the primary users of the SIM and usually do not share the phone with others. If women are more likely to share a phone with other family members, then women are also more likely to be excluded from the analyses. Therefore, more work is needed to describe patterns of phone sharing and how these data can be incorporated into rather than excluded from the analyses.

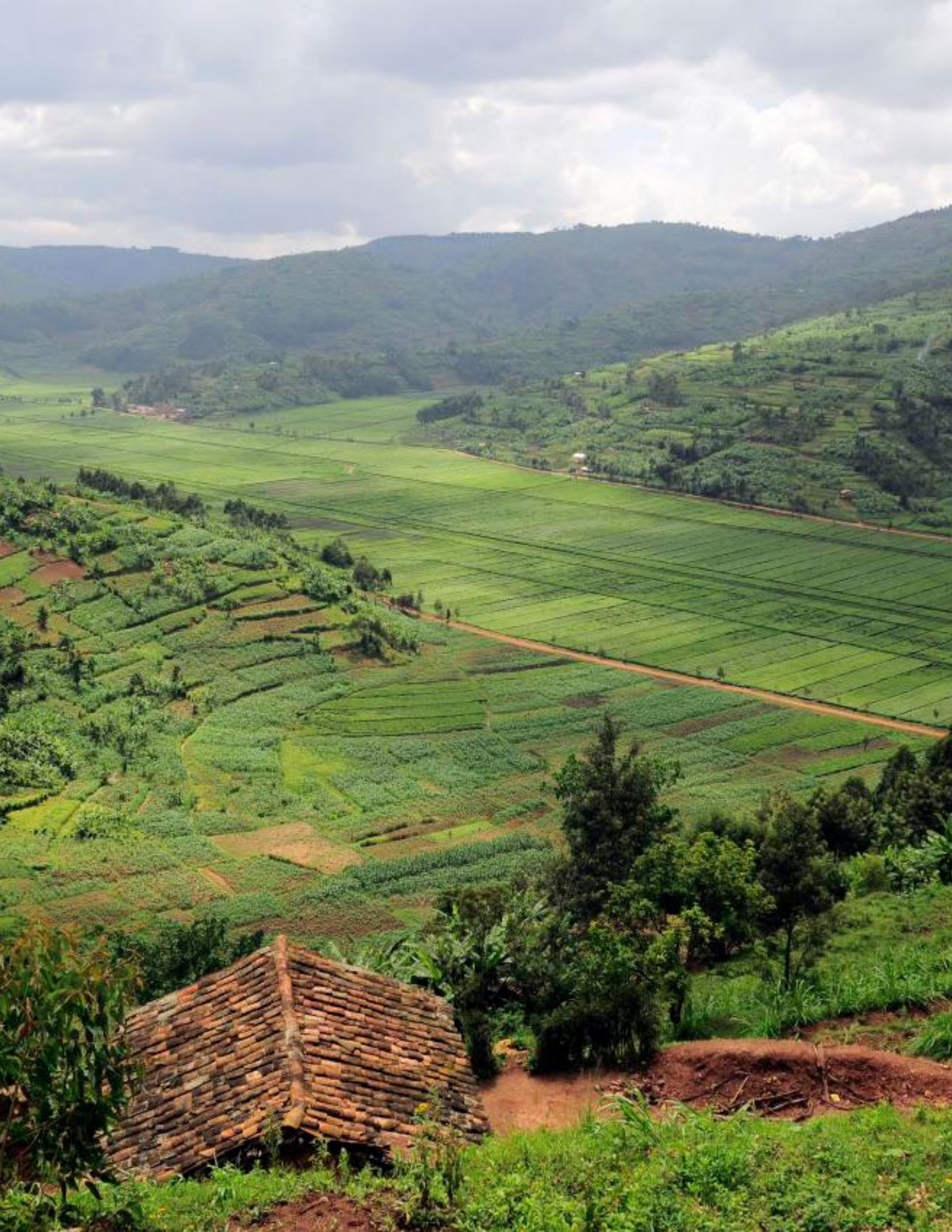
Third, the predictions of the model are static and are based on 2 months of CDR data. They reflect the correlation between self-reported empowerment indicators and phone usage behavior. If any changes occur meaningfully over time, the model will not be accurate any more.

Finally, although mobile phones have now penetrated even remote areas in many low- and middle-income countries (LMICs) (Silver et al., 2019), mobile phone data are not widely available due not only to privacy concerns but also to a reluctance by mobile phone providers to share the data. Further, a mobile gender phone gap persists in many LMICs. Women are 8% less likely than men to own a mobile phone and 20% less likely than men to access the internet on a mobile phone (GSMA, 2020).

These limitations and others should be tackled in future research. As this type of work is rather new, various avenues are open for future research. For instance, the proposed approach in combination with multiple national-level household surveys (e.g., LSMS-ISA or DHS) can be used to first cross-validate the findings of this study, and, if found suitable, can be used to monitor gender empowerment indicators in near real time. Additionally, we envision a scaling out opportunity across countries wherein national-level household surveys are already conducted for the combined use of CDR-based near-real-time tracking of empowerment with survey-based approaches to help track empowerment indicators across time. However, access to CDRs is complex, and open algorithm and data service platforms such as OPAL[5] and AIDA[6] would need to be leveraged to meet data access requirements.

Existing survey-based empowerment indicators (e.g., WEAI) capture empowerment at a particular time period, and these might not be available at specific instances such as during disaster response or while designing an intervention. With the proposed approach, we aim to incorporate time-sensitive dimensions into existing indicators so as to better understand complex concepts, and translate this into the design of gender-sensitive development strategies. Combining traditional survey methods of data collection with CDRs holds potential for capturing indicators of economic empowerment.

5 https://www.opalproject.org/about-opal
6 https://aida.dalbergdatainsights.com/



REFERENCES

Al-Zuabi IM; Jafar A; Aljoumaa K. (2019). Predicting customer's gender and age depending on mobile phone data. *Journal of Big Data,* 6(1), 18. https://doi.org/10.1186/s40537-019-0180-9

Alkire S; Meinzen-Dick R; Peterman A; Quisumbing A; Seymour G; Vaz A. (2013). The Women's Empowerment in Agriculture Index. *World Development*, 52, 71–91. https://doi.org/10.1016/j.worlddev.2013.06.007

Balcan D; Colizza V; Gonçalves B; Hud H; Ramasco JJ; Vespignani A. (2009). Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences of the United States of America,* 106(51), 21484–21489. https://doi.org/10.1073/pnas.0906910106

Bengtsson L; Lu X; Thorson A; Garfield R; von Schreeb J. (2011). Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: A post-earthquake geospatial study in Haiti. *PLoS Medicine,* 8(8), 1–9. https://doi.org/10.1371/journal.pmed.1001083

Blumenstock J; Eagle N. (2010). Mobile divides: Gender, socioeconomic status, and mobile phone use in Rwanda. *ACM International Conference Proceeding Series*, (August). https://doi.org/10.1145/2369220.2369225

Blumenstock J; Cadamuro G; On R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264), 1073–1076. https://doi.org/10.1126/science.aac4420

Castillo G; Layedra F; Guaranda MB; Lara P; Vaca C. (2018). The Silence of the Cantons: Estimating Villages Socioeconomic Status Through Mobile Phones Data. *2018 5th International Conference on EDemocracy and EGovernment, ICEDEG 2018,* (April), 172–178. https://doi.org/10.1109/ICEDEG.2018.8372308

Dalberg Data Insights. (2019). *Differences in mobile money and phone usage between men and women in Uganda: Case study.* Retrieved from https://bit.ly/2YhtE8q

Deville P; Linard C; Martin S; Gilbert M; Stevens FR; Gaughan AE; ... Tatem AJ. (2014). Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences of the United States of America,* 111(45), 15888–15893. https://doi.org/10.1073/pnas.1408439111

Doss CR; Kovarik C; Peterman A; Quisumbing A; Bold M. (2015). Gender inequalities in ownership and control of land in Africa: myth and reality. *Agricultural Economics,* 46(3), 403–434.

Eagle N; Macy M; Claxton R. (2010). Network diversity and economic development. *Science,* 328(5981), 1029–1031. https://doi.org/10.5040/9780755621101.0007

Felbo B; Sundsøy P; Pentland AS; Lehmann S; de Montjoye YA. (2017). Modeling the Temporal Nature of Human Behavior for Demographics Prediction. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. https://doi.org/10.1007/978-3-319-71273-4_12

Fernando L; Lokanathan S; Surendra A; Gomez T. (2018). Predicting population-level socio-economic characteristics using Call Detail Records (CDRs) in Sri Lanka. P*roceedings of the 4th International Workshop on Data Science for Macro-Modeling, DSMM 2018 - In Conjunction with the ACM SIGMOD/PODS Conference*. https://doi.org/10.1145/3220547.3220549

Frías-Martínez V; Frías-Martínez E; Oliver N. (2010). A Gender-centric Analysis of Calling Behavior in a Developing Economy Using Call Detail Records. *Papers from the AAAI Spring Symposium on Artificial Intelligence and Development,* March 22-24, 37–42.

GSMA. (2018). *The Gender Analysis & Identification Toolkit: Estimating subscriber gender using machine learning. London:* GSM Association.

GSMA. (2020). *Connected Women: The Mobile Gender Gap Report 2020*. Retrieved from https://www.gsma.com/r/gender-gap/

Gutierrez T; Krings G; Blondel VD. (2013). Evaluating socio-economic state of a country analyzing airtime credit and mobile phone datasets. ArXiv Preprint, 1–6. Retrieved from http://arxiv.org/abs/1309.4496

Hillesland M; Slavchevska V; Henderson H; Okello P; Oumo FN. (2020). Beyond the sex of the holder: understanding agricultural production decisions within household farms in Uganda. *Journal of Gender, Agriculture and Food Security*, 5(01), 14–27. https://doi.org/10.19268/JGAFS.512020.2

Jahani E; Sundsøy P; Bjelland J; Bengtsson L; Pentland AS; de Montjoye YA. (2017). Improving official statistics in emerging markets using machine learning and mobile phone data. *EPJ Data Science,* 6(1). https://doi.org/10.1140/epjds/s13688-017-0099-3

Kabeer N. (1999). Resources, agency, achievements: Reflections on the measurement of women's empowerment. *Development and Change*, 30(May), 435–464. https://doi.org/10.1111/1467-7660.00125

Khaefi MR; Burra DD; Dianco RF; Pradipta DM; Alkarisya M; Muztahid R; ... Idzalika R. (2019). *Modelling Wealth from Call Detail Records and Survey Data with Machine Learning: Evidence from Papua New Guinea.* https://doi.org/doi: 10.1109/BigData47090.2019.9005519

Kieran C; Sproule K; Doss CR; Quisumbing A; Kim SM. (2015). Examining gender inequalities in land rights indicators in Asia. *Agricultural Economics*, 46(S1), 119–138.

Lu X; Wrathall DJ; Sundsøy PR; Nadiruzzaman M; Wetter E; Iqbal A; ... Bengtsson L. (2016). Unveiling hidden migration and mobility patterns in climate stressed regions: A longitudinal study of six million anonymous mobile phone users in Bangladesh. *Global Environmental Change*, 38, 1–7. https://doi.org/10.1016/j.gloenvcha.2016.02.002

Mehrotra A; Nguyen A; Blumenstock J; Mohan V. (2012). Differences in phone use between men and women: Quantitative evidence from Rwanda. *Proceedings of the Fifth International Conference on Information and Communication Technologies and Development*.

Meloni S; Perra N; Arenas A; Gómez S; Moreno Y; Vespignani A. (2011). Modeling human mobility responses to the large-scale spreading of infectious diseases. *Scientific Reports,* 1, 1–7. https://doi.org/10.1038/srep00062

Montjoye Y-A de; Quoidbach J; Robic F; Pentland A. (2013). Predicting Personality Using Novel Mobile Phone-Based Metrics. In: Greenberg AM; Kennedy WG; Bos ND. (Eds.), *Lecture Notes in Computer Science* (7812, pp. 48–55). Berlin, Heidelberg: Springer-Verlag.

Mullainathan S; Spiess J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives,* 31(2), 87–106. https://doi.org/10.1257/jep.31.2.87

Pappalardo L; Pedreschi D; Smoreda Z; Giannotti F. (2015). Using big data to study the link between human mobility and socio-economic development. Proceedings - 2015 IEEE International Conference on Big Data, IEEE Big Data 2015, 871–878. https://doi.org/10.1109/BigData.2015.7363835

Silver L; Simth A; Johnson C; Taylor K; Jiang J; Anderson M; Rainie L. (2019). *Mobile Connectivity in Emerging Economies.* Retrieved from https://pewrsr.ch/369wiS0

Slavchevska V; Doss CR; de la O Campos AP; Brunelli C. (2020). Beyond ownership: women's and men's land rights in sub-Saharan Africa. *Oxford Development Studies.* https://doi.org/10.1080/13600818.2020.1818714

Smith-Clarke C; Mashhadi A; Capra L. (2014). Poverty on the cheap: Estimating poverty maps using aggregated mobile communication networks. *Conference on Human Factors in Computing Systems - Proceedings,* 511–520. https://doi.org/10.1145/2556288.2557358

Soto V; Frías-Martínez V; Virseda J; Frías-Martínez E. (2011). Prediction of socioeconomic levels using cell phone records. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),* 6787 LNCS(1), 377–388. https://doi.org/10.1007/978-3-642-22362-4_35

Steele JE; Sundsøy PR; Pezzulo C; Alegana VA; Bird TJ; Blumenstock J; ... Bengtsson L. (2017). Mapping poverty using mobile phone and satellite data. *Journal of the Royal Society Interface,* 14(127). https://doi.org/10.1098/rsif.2016.0690

Storm H; Baylis K; Heckelei T. (2020). Machine learning in agricultural and applied economics. *European Review of Agricultural Economics,* 47(3), 849–892. https://doi.org/10.1093/erae/jbz033

Sundsøy P; Bjelland J; Reme B-A; Iqbal AM; Jahani E. (2016a). Deep Learning Applied to Mobile Phone Data for Individual Income Classification. In *2016 International Conference on Artificial Intelligence: Technologies and Applications.* (pp. 96–99). Atlantis Press.

Sundsøy P; Bjelland J; Reme B-A; Jahani E; Wetter E; Bengtsson L. (2016b). *Estimating individual employment status using mobile phone network data.* (December). Retrieved from http://arxiv.org/abs/1612.03870

Toole JL; Lin YR; Muehlegger E; Shoag D; González MC; Lazer D. (2015). Tracking employment shocks using mobile phone data. *Journal of the Royal Society Interface,* 12(107). https://doi.org/10.1098/rsif.2015.0185

Varian HR. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2), 3–28. https://doi.org/10.1257/jep.28.2.3

Wesolowski A; Qureshi T; Boni MF; Sundsøy PR; Johansson MA; Rasheed SB; ... Singer BH. (2015). Impact of human mobility on the emergence of dengue epidemics in Pakistan. *Proceedings of the National Academy of Sciences of the United States of America,* 112(38), 11887–11892. https://doi.org/10.1073/pnas.1504964112

TABLES

TABLE 1. BASIC DEMOGRAPHIC CHARACTERISTICS OF SAMPLED MOBILE PHONE USERS.

RESPONDENTS' AGE (YEARS)	FEMALES	MALES	P VALUE	SIG.
Mean	33.77	35.20	0.00	***
Marital status	**FEMALES**	**MALES**	**P VALUE**	**SIG.**
Divorced	9%	6%	0.00	***
Living together	10%	10%		
Married monogamous	39%	46%		
Married polygamous	4%	6%		
Single	32%	31%		
Widowed	5%	0%		
Main occupation (main source of income)	**FEMALES**	**MALES**	**P VALUE**	**SIG.**
Agricultural wage worker	0%	1%	0.00	***
Commercial farmer – producing mainly for sale	5%	12%		
Housewife	13%	0%		
Non-agriculture self-employed	17%	26%		
Non-agriculture wage workers with contract	13%	19%		
Non-agriculture wage workers without a contract	4%	4%		
Retired or unemployed and not looking for work	1%	1%		
Student	8%	6%		
Subsistence farmer (production intended mainly for consumption by your household)?	26%	26%		
Unemployed	11%	5%		
Decision-making	**FEMALES**	**MALES**	**P VALUE**	**SIG.**
Respondent	40%	48%	0.00	***
Spouse/partner	10%	2%		
Respondent and spouse/partner jointly	34%	42%		
Someone else	16%	8%		
Ownership of assets	**FEMALES**	**MALES**	**P VALUE**	**SIG.**
House (only)	13%	13%	0.00	***
Land (only)	10%	14%		
Both	27%	41%		
Neither	50%	32%		
N	3946	6471		

Note: The table reports the p value from a one-way ANOVA test for respondents' age and p values from a Chi-squared test for the categorical variables marital status, main occupation, decision-making, and ownership of assets. ***, **, * indicate significance levels of 1%, 5%, and 10%, respectively.

TABLE 2. ACCURACY OF THE BEST CLASSIFIER OF EACH PROBLEM IN THE TESTING SET.

CLASSIFICATION PROBLEM	METRICS	BASELINE (%)	TRAINED ON BOTH SEXES, TESTED ON BOTH SEXES (%)	TRAINED ON BOTH SEXES, TESTED ON THE PREDICTED AS WOMEN (%)	TRAINED ON THE PREDICTED AS WOMEN, TESTED ON THE PREDICTED AS WOMEN (%)
		(1)	(2)	(3)	(4)
Sex (male/female)	Accuracy	62	78	NA	NA
	Sensitivity	NA	65.6	NA	NA
	Specificity	NA	85.5	NA	NA
House ownership (yes/no)	Accuracy	51	65	66	60
	Sensitivity	NA	63		58.7
	Specificity	NA	62		58.5
Land ownership (yes/no)	Accuracy	52	64	64	62
	Sensitivity	NA	25.4		71.4
	Specificity	NA	60.3		51.2
Decisions made alone or with spouse (yes/no)	Accuracy	84	84	78	76
	Sensitivity	NA	6		91.3
	Specificity	NA	98		14.6
Occupation (agriculture, non-agriculture, not in labor force)	Accuracy	44	57	55	54
	Sensitivity	NA	52		25
	Specificity	NA	76.6		88.9

Note: XGBoost is the best performing ML model in all classification problems, except "Owning land," for which SVM with rbf Kernel performs best.

TABLE 3. THE TOP 10 FEATURES EXPLAINING EACH INDICATOR

RANKING FROM HIGHEST TO LOWEST	GENDER	LAND	HOUSE	OCCUPATION	HOUSEHOLD DECISION POWER
1	Incoming/outgoing voice duration ratio	Central 1	Weekend evening data volume	Central 1	Unique contacts outgoing SMS count
2	Incoming voice duration	Data volume	Data volume	Distinct sites count	Incoming average voice duration
3	Weekday evening incoming voice duration	Urban	Data volume average	Urban	Weekday evening outgoing SMS count
4	Weekday evening outgoing voice duration	Weekday evening data volume	Urban	All week count	Top-up value average
5	Central 1	Data volume average	Unique contacts outgoing SMS count	Incoming voice duration average	Weekday evening incoming SMS count
6	Unique contacts count	Unique contacts incoming voice count	Central 1	Data volume	Momo sum amount received
7	Unique contacts incoming voice count	Weekday incoming voice duration	Weekday average voice duration	Weekday morning incoming SMS count	Data volume average
8	Unique contacts incoming SMS count	Outgoing voice duration	Data count	Top-up value average	Weekend evening incoming voice duration
9	Weekday incoming voice duration	Unique contacts outgoing voice count	Weekday evening outgoing SMS count	Top-up value sum	Unique contacts incoming SMS count
10	Outgoing/incoming ratio	East Central	Weekend evening outgoing SMS count	Southwest	Weekday evening data count

TABLE 4. ACCURACY OF THE BEST CLASSIFIER ON SEX PREDICTION, CONDITIONAL ON THE EMPOWERMENT INDICATOR PREDICTION

INDICATOR	CATEGORY	SEX PREDICTION ACCURACY (%)	SENSITIVITY (%)	SPECIFICITY (%)
Decision over income	Alone	70	83.3	50.0
	With spouse/none	78	63.5	86.2
House ownership	Owning	83	83.0	87.2
	Not owning	85	66.6	91.1
Land ownership	Owning	84	80.8	85.6
	Not owning	83	60.0	93.7
Occupation	Agriculture	87	68.2	94.3
	Non-agriculture	81	96.9	69.8
	Non-labor force	86	72.9	86.2

Note: XGBoost is the best performing ML model in all classification problems, except "Owning land," for which SVM with rbf Kernel performs best.

TABLE 5. PREDICTIONS FOR THE FULL USER BASE

INDICATOR	CATEGORY	PREDICTED SHARE OF FULL USER BASE (%)	PREDICTED SHARE OF THOSE PREDICTED AS WOMEN (%)
Sex	Female	31	100
	Male	69	0
Decision	Alone/spouse	99	50
	None	1	50
House	Owning	70	50
	Not owning	30	50
Land	Owning	54	50
	Not owning	46	50
Occupation	Agriculture	43	21
	Non-agriculture	3	6
	Non-labor force	54	73

ANNEX

TELEPHONE SURVEY

Study ID		Reference. No.	
Interviewer No.		Interview Length	
Contact Number			

Good morning/afternoon/evening. My name is .....I am an interviewer from ..., which is a market research company. We regularly conduct market surveys for different products and are currently conducting one such survey which includes questions regarding your use of mobile phone as well as questions regarding your access to resources. The survey will be used for research purposes only. There is minimal anticipated risk associated with participation in the study. All your answers will be kept strictly confidential. The data will not be shared outside the research team. This interview is completely voluntary; you may refuse to answer any specific questions; and you may stop the interview at any time. The survey will take approximately 5 minutes. Your participation will be highly appreciated.

I am happy to share my supervisor's contact details and he/she can provide you with more information. Do you agree to participate in the survey?

Consent given:
Yes -- 1
No -- 2

Q1. (Instruction: If gender has not been inferred till now from the name in Q1 and the voice of the respondents, please ask the following question)
How should I greet you?

Madam / Miss	1	
Sr	2	

Q2. What year were you born?

Q3. Can you please tell me your marital status?

Single	1	
Married monogamous	2	
Married polygamous	3	
Living together	4	
Divorced	5	
Widowed	6	
(Refused)	7	

Q4. Are you the main user of this phone number/SIM? By main user, I mean do you use this SIM more than anybody else does? **[Instruction: if the respondent does not understand the word SIM card, then please read the definition]** SIM cards are inserted into a mobile phone handset or other device to enable customers to access a mobile operator's network. The SIM card may have been installed in the handset by the supplier of the phone and it can come in different sizes. Every SIM has a unique telephone number.

YES	1	Continue
NO	2	Please ask for the main user of the SIM/ number. **[Instruction: If not available, then ask when s/he will be available and schedule an interview accordingly. Please make at least 3 attempts]**
(DK)	3	
(Refused)	4	

Q5. Can you please tell me your main occupation (the one you get most of your income from) in the last 12 months?

Subsistence farmer (Production intended mainly for consumption by your household)?	1	
Commercial farmer – producing mainly for sale	2	
Agricultural wage worker	3	
Non agriculture self-employed	4	
Non agriculture wage-workers with contract	5	
Non agriculture wage-workers without a contract	6	
Retired or unemployed and not looking for work	7	
Student	8	
Housewife	9	
Unemployed	10	
(Refused)	11	

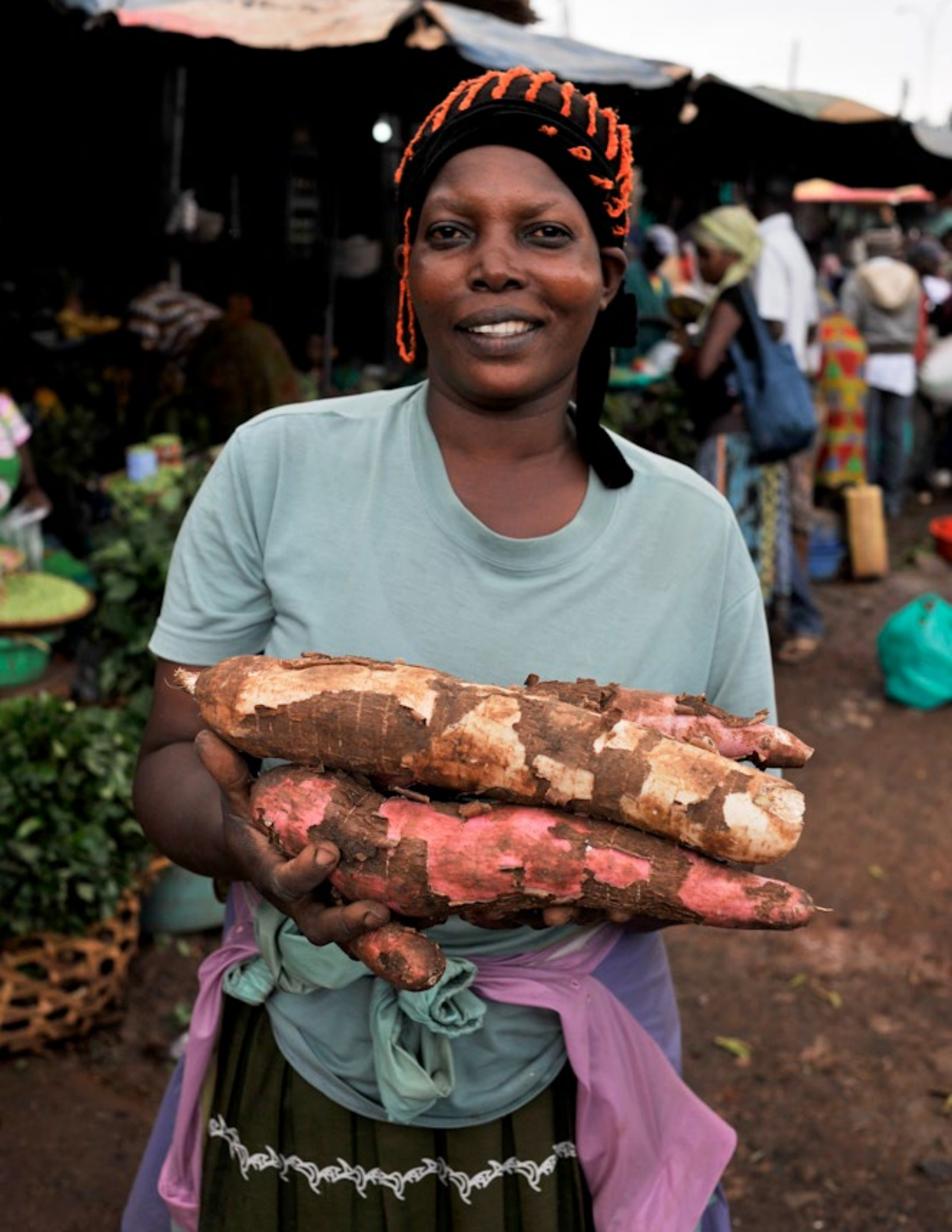
Q6. Thinking about the last 12 months, who usually makes decisions about how to use total HOUSEHOLD income?

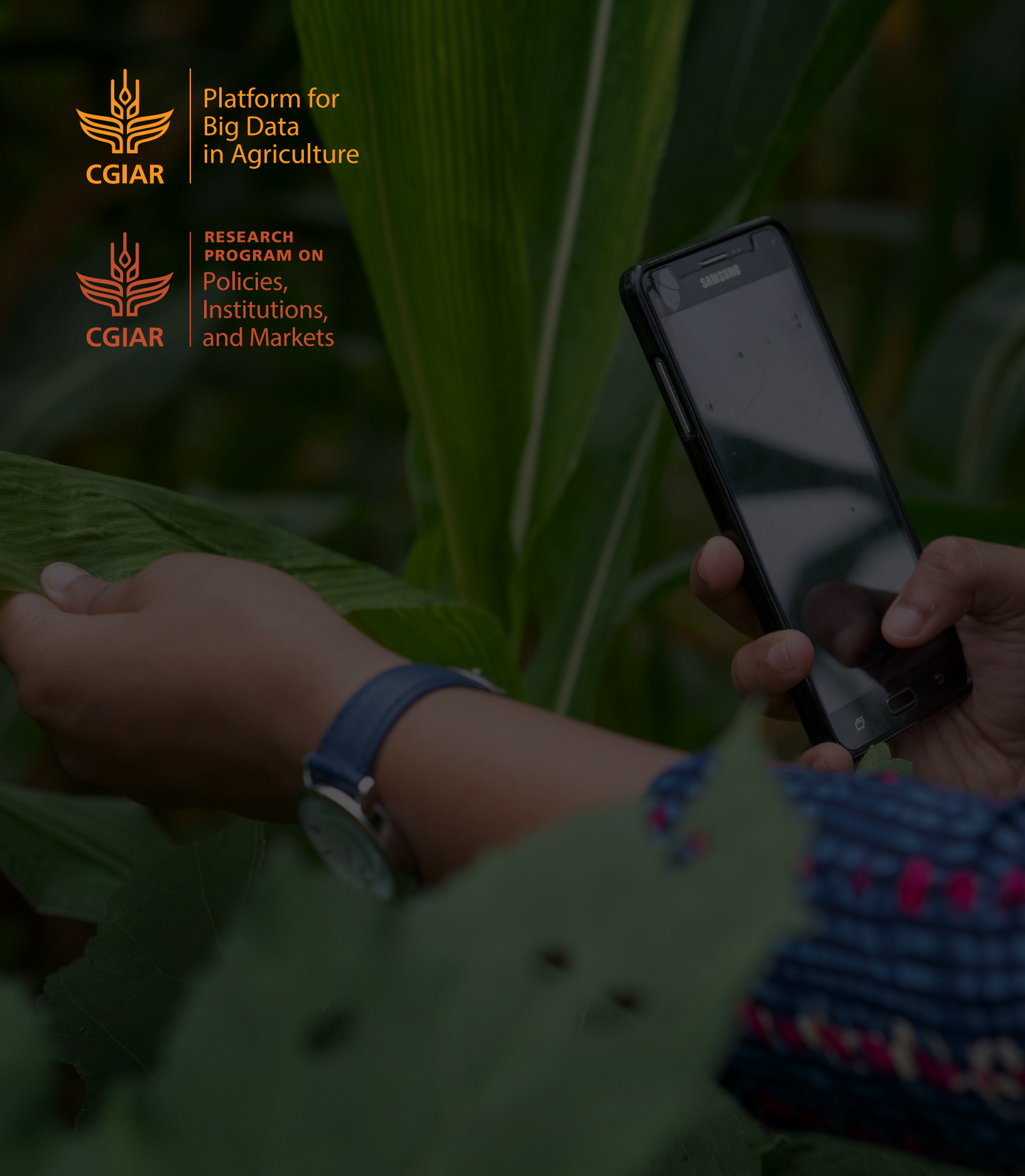
Respondent	1	
Spouse/partner	2	
Respondent and spouse/partner jointly	3	
Someone else	4	
(Refused)	5	

Q7. Do you own your own house or agricultural land, either solely or jointly with someone else?

Land	1	
House	2	
Both	3	
Neither	4	

Thank you for your time.





CGIAR

Platform for Big Data in Agriculture

CGIAR

RESEARCH PROGRAM ON Policies, Institutions, and Markets

Dalberg

Alliance

Bioversity International

CIAT

International Center for Tropical Agriculture

Since 1967 Science to cultivate change

KIT Royal Tropical Institute

IFPRI

INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE