WPS7969

Policy Research Working Paper 7969

Estimating Poverty Using Cell Phone Data

Evidence from Guatemala

Marco Hernandez
Lingzi Hong
Vanessa Frias-Martinez
Andrew Whitby
Enrique Frias-Martinez

WORLD BANK GROUP
Macroeconomics and Fiscal Management Global Practice Group
February 2017

Policy Research Working Paper 7969

Abstract

The dramatic expansion of mobile phone use in developing countries has given rise to a rich and largely untapped source of information about the characteristics of communities and regions. Call Detail Records (CDRs) obtained from cellular phones provide highly granular real-time data that can be used to assess socio-economic behavior including consumption, mobility, and social patterns. This paper examines the results of a CDR analysis focused on five administrative departments in the south west region of Guatemala, which used mobile phone data to predict observed poverty rates. Its findings indicate that CDR-based research methods have the potential to replicate the poverty estimates obtained from traditional forms of data collection, like household surveys or censuses, at a fraction of the cost. In particular, CDRs were more helpful in predicting urban and total poverty in Guatemala more accurately than rural poverty. Moreover, although the poverty estimates produced by CDR analysis do not perfectly match those generated by surveys and censuses, the results show that more comprehensive data could greatly enhance their predictive power. CDR analysis has especially promising applications in Guatemala and other developing countries, which suffer from high rates of poverty and inequality, and where limited fiscal and budgetary resources complicate the task of data collection and underscore the importance of precisely targeting public expenditures to achieve their maximum antipoverty impact.

This paper is a product of the Macroeconomics and Fiscal Management Global Practice Group. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://econ.worldbank.org. The authors may be contacted at marcohernandez@worldbank.org.

Estimating Poverty Using Cell Phone Data: Evidence from Guatemala

By MARCO HERNANDEZ, LINGZI HONG, VANESSA FRIAS-MARTINEZ, ANDREW WHITBY, and ENRIQUE FRIAS-MARTINEZ[1]

JEL Classification: O12, I32, B41, C80.
Keywords: Big data, cell phone call detail records, poverty measurement, Guatemala.

[1] M. Hernandez: The World Bank, Washington, D.C. (email: marcohernandez@worldbank.org); L. Hong: University of Maryland, College Park, MD. (email: lzhong@umd.edu); V. Frias-Martinez: University of Maryland, College Park, MD. (email: vfrias@umd.edu); A. Whitby: The World Bank, Washington, D.C. (email: awhitby@worldbank.org); E. Frias-Martinez: Telefonica Research, Madrid, Spain (email: enrique.friasmartinez@telefonica.com). We are grateful to the Guatemalan authorities for their support to this project, in particular the Ministry of Public Finance and the National Statistics Institute. We thank Oscar Calvo-Gonzalez, Manuela Francisco, Humberto Lopez, Pablo Saavedra, Kinnon Scott, and participants of the World Bank Group's Innovation Challenge for helpful comments. Maryam Ali-Lothrop and Sean Lothrop provided excellent research and editorial assistance. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the World Bank, its Executive Directors, or the governments they represent.

Introduction

The explosive growth of telecommunication networks in developing countries is yielding an unprecedented wealth of highly granular real-time data, and governments have only begun to leverage the enormous potential of this new information source. This paper explores analytic methodologies for using aggregated and encrypted cell phone call data to map the distribution of poverty in Guatemala. To estimate poverty rates, Guatemala, like many other developing countries, relies on conducting and analyzing household surveys and population censuses that are both expensive and administratively demanding. By contrast, analyses based on cell phone data and machine-learning technology have the potential to generate reliable and timely information on the spatial distribution of household poverty at a far lower cost than traditional household surveys or population censuses.

Poverty in Guatemala

Poverty rates in Guatemala are higher than in other comparable middle-income countries, and the distribution of poverty reflects a set of overlapping regional, rural/urban and ethnic dimensions. Contrary to the trend observed in other Latin American countries, poverty rates in Guatemala have increased in recent years. The poverty rate[2] rose from 55 percent in 2001 to 60 percent in 2014, as the number of people living below the poverty line increased by 2.8 million. Poverty rates vary dramatically by administrative department. In 2014, Guatemala's poorest department, Alta Verapaz, had a poverty rate of 83 percent and an extreme poverty rate[3] of 54 percent. Meanwhile, the poverty and extreme poverty rates in the country's wealthiest department, where the capital of Guatemala City is located, were far lower at 33 percent and 5 percent, respectively. Moreover, while urban areas are now home to a majority of the country's poor, poverty rates remain substantially higher in rural areas. In 2014, 35 percent of the rural population was living in extreme poverty, compared with 11 percent of the urban population. Poverty rates are also significantly higher among Guatemala's indigenous population. Indigenous peoples represent 42 percent of the country's total population, but in 2014 they accounted for 52 percent of the poor and 66 percent of the extremely poor in the country.[4]

Guatemala's high poverty rates and persistent income inequality are reflected in the country's weak human development indicators. Limited and unequal access to public services such as education and health care have constrained human capital formation, while a lack of basic

[2] The moderate poverty rate reflects per capita household consumption equivalent to US$4.00 per day in purchasing-power parity terms.

[3] The extreme poverty rate reflects per capita household consumption equivalent to US$2.50 per day in purchasing-power parity terms.

[4] Sanchez, Scott, and Lopez (2016).

infrastructure has increased production and transportation costs and reduced the job opportunities available to the poor. Together, these factors are contributing to a long-term decline in economic productivity. Meanwhile, Guatemala's low tax revenues limit its capacity for redistributive fiscal policy and pro-poor spending. Guatemala's high poverty rates and tight fiscal constraints underscore the critical importance of effectively targeting public spending.

The Role of Data in Poverty Reduction

Effective poverty-reduction strategies require detailed information on the current geographic distribution of poverty and the characteristics of households living below the poverty line. Censuses and household surveys can shed light on a wide range of economic and social indicators. However, these methods are expensive and time-consuming, and implementing them requires considerable institutional capacity. Moreover, adverse local conditions such as violent conflict, high crime rates, or political instability may make traditional in-person surveying impossible in certain areas. As a result, policy makers must often base critical decisions on outdated or incomplete information.

Social scientists are increasingly using so-called "big data" analysis to supplement more traditional sources of information. Satellite imagery, logs from sensors (e.g. traffic, weather), smartphone applications, and cell phone data—the subject of this paper—have already yielded important insights in numerous fields. Unlike household surveys, which are specifically designed to address certain research questions, big data are usually collected in a non-research context, often as the byproduct of a commercial activity or public service. Analyzing big data requires new research methods, many of which are still in the early stages of their development. Emerging research methodologies based on Call Detail Records (CDRs) and advanced machine-learning techniques have especially promising applications in developing countries, as they can potentially generate reliable poverty data at a far lower cost than conventional household surveys.

CDR analysis can play a vital role by filling the spatial and temporal gaps left by traditional research methods. By making inferences based on cellular network usage, CDR analysis can reliably project the evolution of poverty dynamics over a specific timeframe. Unlike censuses and household surveys, CDR analysis is quick and relatively inexpensive and can be performed by a small team of statisticians using records that are already collected by Mobile Network Operators (MNOs).

Guatemala offers a prime example of the limits of traditional data collection. The country's most recent Population and Housing Census, dates from 2002, and all national poverty data are derived from just four household surveys conducted over the past 25 years. For example, the most recent 2014 household survey (*Encuesta Nacional de Condiciones de Vida*, ENCOVI) covered

about 11,500 households, took about two years to complete, and it cost about US$2 million. By contrast, the CDR analysis performed for this report cost about US$100,000, and the majority of the costs went to the development of the computer algorithm, which is a fixed cost. Thus, if the CDR analysis is conducted again with new data, it would be significantly cheaper.

While this was the first analysis of its kind conducted in Guatemala and designed primarily to test the validity of various methodologies, a more thorough exercise would only require a small fraction of the time, funds, and human resources required relative to a traditional census or survey. Moreover, these costs would likely be even lower in later iterations of the CDR analysis, as methodological innovation and testing are replaced by the routine implementation of established techniques. While CDR analysis cannot fully supplant conventional research methods, it can greatly enhance their value by providing high-frequency updates and complementary information. Moreover, if CDR analysis can be shown to provide sufficiently accurate inferences to allow countries to slightly extend the time between traditional surveys, it could potentially generate a net savings for the national research budget.

Background

Call Detail Records (CDRs)

MNOs record and store data about their customers' phone use, primarily for billing purposes. In addition to recording cellular data consumption, MNOs collect information on each call and text message. Stored data do not generally reflect the content of a call or message. Instead, they record circumstantial details, such as the time and duration of a call, the size of a message, the identities of the parties involved and their network information. These data are referred to in the telecommunications industry as CDRs.

Figure 1: Sample Call Detail Records

Interaction	Direction	Correspondent ID	Date and Time	Call Duration	Antenna ID
Call	In	8f8ad28de134	2012-05-20 20:30:37	137	13084
Call	Out	fe01d67aeccd	2012-05-20 20:31:42	542	13084
Text	In	c8f538f1ccb2	2012-05-20 21:10:31		13087

Source: http://bandicoot.mit.edu/docs/quickstart.html

In addition to CDRs, MNOs often store certain personal details about their customers, including their name and home address, and in some cases their gender, age or other characteristics. For prepaid customers, which are very common in low- and middle-income countries, MNOs typically keep a record of credit recharges or "top ups."

Using CDRs in Social and Economic Research

Although CDRs may appear to present a narrow and technical dataset, due to the dramatic expansion of mobile phone use over the past several decades these records can provide a rich source of information about human behavior and the characteristics of communities. CDRs can be used to infer certain personal attributes about a cellular user, such as their home location, and they can be used to analyze social networks, as each call or message can be viewed as a link between MNO customers. This approach enables researchers to map social interactions, identify community nexus points and examine how information is transmitted across groups and regions.[5]

[5] These applications are described in greater detail in Blondel et al. (2015).

Box 1. The Virtual Geography of Cell Phones: Determining User Location from CDRs

Unlike satellite phones, cell phones rely on a network of towers known as "base transceiver stations," each of which operates within a limited range. This effectively divides the network's coverage area into individual "cells." As a customer moves, their connection to the network is passed from one tower to the next. When a customer places a call, sends a text, or initiates a data session, the identity of the relevant tower is recorded in the CDR. MNOs maintain a list of the coordinates of each tower, making it possible to determine a phone's general location each time it is used.

While the user's precise location cannot be identified, limited geographical areas called "Voronoi polygons" can be constructed based on the assumption that a handset always connects to the closest cell tower. In reality, factors such as different antenna strengths, capacity constraints and terrain may cause a handset to connect to a more distant antenna. Nevertheless, the Voronoi polygon remains a useful tool for approximating a cell user's location.

The network structure determines the size of each Voronoi polygon. Since both the antenna hardware and the mobile spectrum have limited capacity, MNOs tend to place more antennas in areas of higher usage to maximize throughput. Hence, polygon size tends to correlate inversely with population density—that it, densely populated places tend to have more antennae and smaller polygons. However, some places with a small residential population but high rates of commercial activity, such as business districts, shopping malls and airports, may have a large number of antennae. Polygons can range in diameter from a few hundred meters to tens of kilometers, depending on the network.

Figure 2: Sample Cell Towers around Guatemala City's Constitution Plaza (L) and the Voronoi Polygons They Create (R)

Note: The OpenCellID project collects cell tower locations based on reports from volunteer users who install a participating app. Therefore, these tower locations are approximate. No official cell tower information has been used for these plots.
Source: opencellid.org

Estimating Poverty in Guatemala using Cell Phone Data

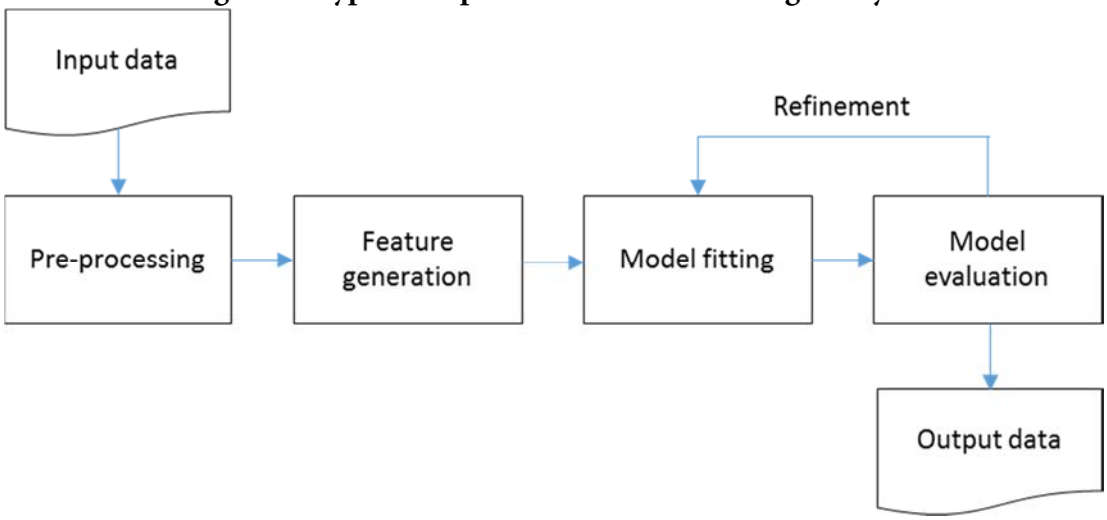
This section describes the results of a recent study of CDR-based research methods in Guatemala designed to evaluate the potential value of CDR analysis as a tool of socioeconomic research. The study's objective was to create a model using CDR data that could accurately predict the observed incidence of extreme poverty. The study focused on five municipalities (*municipios*) in the administrative departments (*departamentos*) of Quetzaltenango, Suchitepéquez, Sololá, Totonicapán and San Marcos.[6] Together, these departments represent 20 percent of the Guatemalan population.

The study addressed three questions:

(i) Can CDR data be used to reliably estimate poverty rates in Guatemala?
(ii) Are these estimates more accurate in urban areas, rural areas or at the national level?
(iii) Can poverty data derived from CDRs in 2006 be used to predict poverty rates in 2011?

To answer these questions, the study employed a machine-learning approach, which is a highly general, iterative method for uncovering the relationship between input data and output data, which in this case are cell phone records and poverty rates, respectively. Figure 3 illustrates the machine-learning methodology.

Figure 3: Typical Steps in a Machine-Learning Analysis



Source: Adapted from Hong and Frias-Martinez (2015a).

Data Sources

[6] These were the municipalities for which CDR data were available at the time. The methodology could be easily replicated to include CDR data for the rest of the country.

In order to test the validity of the CDR analysis, its findings were compared to World Bank poverty estimates, which are based on Guatemala's National Living Conditions Survey (ENCOVI) for 2006 and 2011 and the 2002 Population and Housing Census. The ENCOVI is a traditional household survey, and its sample size does not provide reliable estimates at the municipal level. However, small-area estimation techniques[7] which combine the ENCOVI data with the census data allow for poverty to be estimated at the municipal level. For the purposes of the study these estimated rates were treated as "ground truth data." In machine-learning analysis, ground truth data are obtained by direct observation, rather than by modeling or inference. In this context, however, the term refers to the poverty rates determined by standard statistical estimation methods, which provide the only existing measure of ground truth. Nonetheless, it should be borne in mind that all poverty-rate estimation methodologies are predicated on assumptions, and in this area no ground truth data can offer a perfect representation of reality.

Supervised machine-learning models, as described here, require a "training dataset," which comprises benchmark data that represent the ground truth. Since CDRs relate directly to people, they can be considered unit-record data. Then it is the detail of the ground-truth data that largely determines the resolution of the model. In the case of Guatemala, ground-truth data are aggregated This is often the only available option when the ground truth data are based on estimations using household survey data, in which no mobile phone numbers are collected during the survey. In this case, individual-level features are extracted from the CDRs, then combined to form statistical aggregates at the chosen geographic level (e.g. mean, median, maximum or quantiles by region). These aggregates are used to build an area-level predictive model calculated for each area (e.g. poverty headcount). A relatively small sample size (for example, Guatemala's 338 *municipios*) means that internal validation, such as cross-validation or hidden test data, is difficult, and some external validation may be required.

Poverty rates were calculated for each *municipio*. Aggregate rural, urban and overall poverty rates were available for 2006, but only rural poverty rates were available for 2011.[8] The study used aggregated and encrypted CDR data for August 2013, which overlaps with the ENCOVI survey period. In 2013 Guatemala had 140 cellular accounts for every 100 people.[9] The model tested two types of prediction: (i) same-survey prediction, such as predicting the 2006 urban poverty rates based on a model of the relationship between the 2006 ENCOVI data and the 2013 CDRs; and (ii) different-survey prediction, such as predicting the 2011 rural poverty rates based on a model of the relationship between the 2006 ENCOVI data and the 2013 CDRs. Same-survey

[7] See Elbers, Lanjouw and Lanjouw (2003).

[8] These data come from a 2011 census of rural areas designed to gather information for social-protection programs. No national census was conducted that year.

[9] World Development Indicators, 2016. This reflects multiple accounts per person. While this does not indicate that every person has a cell phone, it suggests that cell phone use is relatively high.

prediction is most similar to the real-world application of cellular data modeling known as "spatial infill" or "spatial extrapolation," while different-survey prediction is an example of "time extrapolation."

Pre-Processing

Data Cleaning and Enrichment

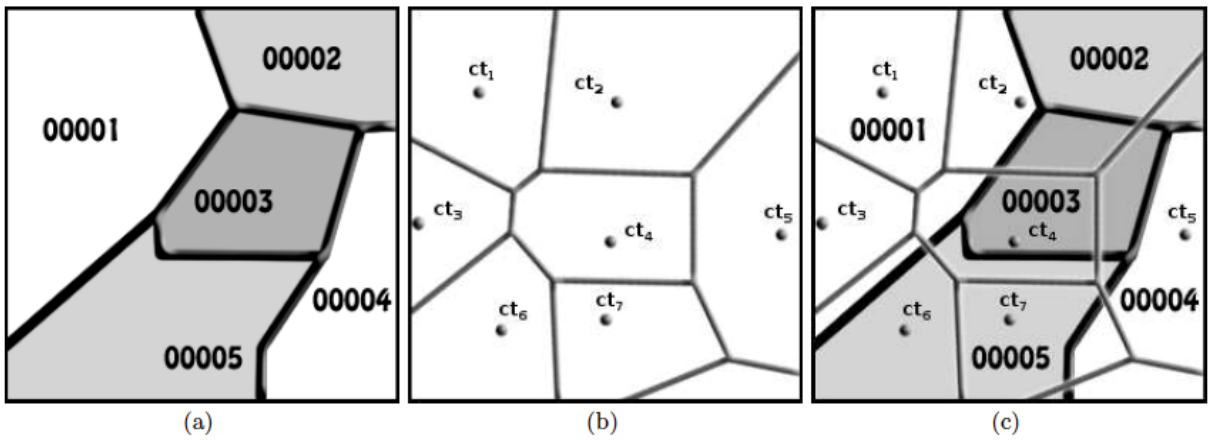
Raw CDR data are invariably noisy and require pre-processing before they can be analyzed. The primary sources of data noise are: (i) gaps or inconsistencies caused by MNO network operation decisions, including technological changes that affect CDR comparability; and (ii) the presence of business lines, text recharge numbers or other connections that do not reflect communications between individual customers. Pre-processing begins by identifying inconsistent or irrelevant CDRs and dropping them from the dataset. Each customer in the CDR dataset is then assigned a home location, which is generally inferred based on the network cell in which that customer is most active after 6pm. However, if that cell has less than 30 percent more activity than the next-most-active cell during the same period, it is assumed that the user is itinerant and no home location is assigned.

Spatial Harmonization

Ensuring that all datasets share a common spatial scale is an important step in data preparation. While the network cell is the natural spatial unit for CDR data, poverty rates are generally calculated using administrative boundaries, in this case the *municipio*. These spatial structures must be reconciled in order to analyze the data.

In the model presented below, poverty data are mapped onto the network cells. Each cell is assigned an area-weighted average of the poverty rates in the *municipios* it covers. Network cells located entirely within one *municipio* are assigned that *municipio's* value. This process, illustrated in Figure 4, ensures that all datasets use the common geography of the network cell, which then becomes the unit of analysis. More complex approaches could also be adopted, such as weighting the distribution by population density.

Figure 4: An Example of Spatial Harmonization



Source: Reproduced from Frias-Martinez et al. (2012), Figure 1.
Note: Merging (a) Geographical units from the survey data with (b) geography of the network cells defined as Voronoi diagrams. (c) The imputed poverty incidence for polygon 00003 will be "x" percent of that in ct2, "y" percent in ct4 and "z" percent in ct5. Where x+y+z=100, representing the full area of polygon 00003.

Feature Generation

Under the machine-learning approach "feature generation" is the process of collapsing a rich multidimensional dataset into a smaller number of carefully chosen dimensions, which are statistics of the underlying data. These features then form an intermediate representation of the data and are the basis for the final regression-based or classification-based model. For example, handwriting recognition is a classic machine-learning task in which an image of a handwritten digit represents a highly multidimensional dataset. In some cases, each image may be 64 by 64 pixels, for a total of 4096 dimensions. Since this may be too many for a classification model to work well, a smaller number of features are extracted using well-known, preexisting algorithms. These features usually represent the presence or absence of higher-level geometric features such as edges, curves and corners. This simplifies the final model, speeding the training process.[10]

In the current analysis, two sets of features are generated for each network polygon:

- **Customer home-oriented.** This first set is based on the records of customers who live within the polygon, as determined by their inferred home location. This is akin to taking a survey of people at their usual residence. Features are first calculated on a per-customer basis, then aggregated to the network polygon. They include measures of consumption (e.g.

[10] The "feature" paradigm in machine learning is gradually being overtaken by "deep learning" approaches, which work directly with underlying highly dimensional data. This is especially true for well-studied tasks like image recognition. For more complex tasks such as modeling social structures based on CDRs, deep learning remains an active research area.

the number of calls made or received) and measures of mobility (e.g. where, how far and how often a customer travels). Data are aggregated by taking the mean over all customers, if appropriate, or else a variety of threshold values—for example, the number of customers living in a network polygon who regularly travel beyond 80 kilometers of their home location.

- **Polygon activity-oriented.** The second set is based on activity occurring within the polygon, regardless of the home location of the customers involved. This is akin to taking a survey of people passing through an area. In this case, features based on activity include the number of customers entering the network polygon but living outside of it, how often they visit, and the volume of inbound and outbound calls that are processed. These features are calculated directly at the network-polygon level and do not require further aggregation.

Fitting the Model

Estimating poverty rates from CDR features is an example of a "supervised" machine-learning problem, or one in which input data are used to predict known outputs—in this case CDR features and poverty rates. Once the model is built, it can be applied to new input data for which the corresponding output is not known. In this case, the unknown output would be either different geographies or different points in time.

Supervised machine-learning problems are either classification-based, in which case the output variable is one of a discrete set of classes (e.g. male/female, poor/not poor, etc.), or regression based, in which the output variable is a continuous real number expressed as a decimal, ratio or percentage. Poverty rates are most naturally modeled as a continuous output variable. However, it is also possible to group the data into a small set of classes reflecting low, moderate, or high rates of poverty. Both approaches were examined in this example, which tested a variety of scenarios by mixing different training and test datasets and employing both regression and classification methods. Figure 5 illustrates the different combinations of data and methodology. [11]

[11] For more information, see: Hong and Frias-Martinez (2015).

Figure 5: Data and Methodology Combinations Tested

Cell Phone Data		Training Survey	Testing Survey		Method		
2013	×	2006 total	2006 total	×	Regression		
		2006 rural	2006 rural		Classification:		
		2006 urban	2006 urban		Equal width Equal probable K-means Feature-based	×	Baseline SVM Random forests Stochastic gradient boosting K-means Gaussian mixture Supervised Topic Models
		2011 rural	2011 rural				
		2006 rural	2011 rural				

Source: Adapted from Hong and Frias-Martinez (2015)

Box 2. Applying CDR-Based Poverty Models to Fill Data Gaps

Much of the discussion surrounding using CDRs for socioeconomic research ignores the practical conditions in which they will be applied. CDR data are almost always available in addition to conventional data, such as household surveys or censuses. There is little to be gained in showing that CDR data can be used to predict this existing survey data; instead, the fitted model must be applied to new, unseen data to answer new questions.

There are at least three different ways in which CDR data can complement conventional data sources:

1. **Spatial infill: generating small-area statistics.** Given the relatively high spatial resolution of CDRs, spatial infill likely offers the most added value as a socioeconomic research method. A particular household survey with a limited sample size may only support estimates at a relatively coarse spatial resolution, such as the *departamento* level. The high-resolution behavioral signals in CDRs can enhance the statistical strength of survey data, enabling more detailed and accurate estimates. Ideally, the CDR collection period should coincide with the period in which the survey was conducted. This approach entails the risk of ignoring the role that space and geography might play in the prediction of poverty rates.

2. **Time interpolation/extrapolation.** The potentially high frequency of CDRs can also complement conventional data sources. Household surveys are generally updated at 2-to-5-year intervals. CDRs can be used to update these estimates, providing officials with current information on policy-specific subjects. A predictive model can be constructed for a survey year using contemporaneous CDR data. This model can then be applied to more recent CDR data for which corresponding survey data are not available. There is a risk, however, that the relationship between CDR behavioral signals and the target variable, in this case the poverty rate, could change over time.

3. **Spatial extrapolation.** This is the most ambitious potential application of CDR data. In countries or regions in which no recent survey data are available, such as conflict-affected areas or countries that have experienced severe political instability, estimates can be generated by using survey data and CDRs for a similar location and then applying this model to CDR data from the target location. This approach requires significant assumptions, but it may be useful in cases where no stronger data source exists. To date, research into the practical applications of spatial extrapolation has been limited.

Evaluating the Model

One major liability of machine-learning models is a phenomenon known as "overfitting." Training data always reflect both meaningful "signal" and random "noise." Overfitting occurs when the model being fit has so many free parameters that it fits to both the signal and the noise. In such a case, the model will appear to be an excellent fit for the input data, with high R^2 and similar values, but it will perform poorly when applied to previously unobserved data.

In the case of Guatemala, a technique called "cross-validation" was used to guard against overfitting. Cross-validation splits the input data into parts, training the model on a subset of the data (75 percent of the dataset), then testing it on the remaining data (25 percent). This enables the diagnostic values from the testing set to provide a more accurate reflection of how the model would perform in a realistic out-of-sample scenario. The cross-validation subsets can be constructed multiple times in different ways, with the results averaged, to provide better estimates.

To evaluate the regression results, the accuracy of the machine-learning model is measured using R^2, root mean square error, and the correlation between real and predicted values. R^2 measures the extent to which the model explains the variability of the response data around its mean, while root mean square error indicates the difference between real values and predicted ones. The quality of the classification techniques is analyzed using two measures: accuracy and F1 score. Accuracy reflects the percentage of testing samples whose predicted class is the same as their real ones. The F1 score is a metric that combines the precision and the recall of the method, i.e., the number of samples that are correctly classified and the number of samples for which a label is given. In general, stronger methodologies have higher values for both precision and recall.

Results

All of the CDR models exhibited a significant degree of predictive value. However, the specifics of different models influenced how well they predicted poverty rates. The analysis yielded four overarching results:

Result #1: CDRs can predict poverty rates in Guatemala

Across all the models, CDR analysis consistently predicted poverty rates at the *municipio* level, though their predictive value was limited under certain experimental settings. The best models predicted total poverty levels in 2006 with an R^2 of 0.76, implying that approximately 76 percent of the variation in *municipio*-level poverty rates could be explained by 2013 cell phone data, and with F1 scores of up to 0.84 for the classification models, indicating that 84 percent of *municipios* were classified according to the correct category when three separate categories (low, medium and high poverty rates) were considered. On the other hand, the prediction rates for urban data in 2006 showed R^2 values of 0.69 for regression techniques and 0.73 for classification, indicating weaker predictive value. The lowest prediction rates were for rural data in 2011, with R^2 results of 0.46 and 0.59 for regression and classification models, respectively. Classification results were slightly better than regression results across all models due to the fact that classifying poverty rates into three classes is a simpler prediction problem than attempting to approximate real values. Experimentally, as more classes were included, predictive accuracies for classification decreased

and ultimately converged with those for regression. Therefore, the number of poverty classes selected entails a tradeoff between the accuracy and the granularity of the prediction.

Result #2: In Guatemala CDRs predict urban and total poverty more accurately than rural poverty

In both the regression and classification models, rural poverty rates were consistently more difficult to predict than either total or urban poverty rates. R^2 values fell to around 0.25 for rural poverty, implying that CDR data could explain just 25 percent of the variation in rural poverty rates. The accuracy of the classification models dropped to between 0.3 and 0.65 depending on the specification. Two hypotheses could explain this phenomenon. First, cellular penetration rates in urban areas tend to be higher, and thus CDR-based analysis provides more robust modeling signals in urban areas, where they represent the behavior of a larger share of the population. In rural areas, fewer phones and fewer calls weaken the signal that can be extracted from CDR data, and their predictive power decreases. Second, urban areas tend to have more cellular antennae per square kilometer, which results in smaller polygons. The much larger polygons in rural areas may tend to reduce the granularity of the data by aggregating behaviors, weakening the predictive power of the algorithm. Testing these hypotheses will require further research.

Result #3: Further analysis will be necessary to determine the extent to which CDR-based models based on past poverty data can be used to predict future poverty dynamics

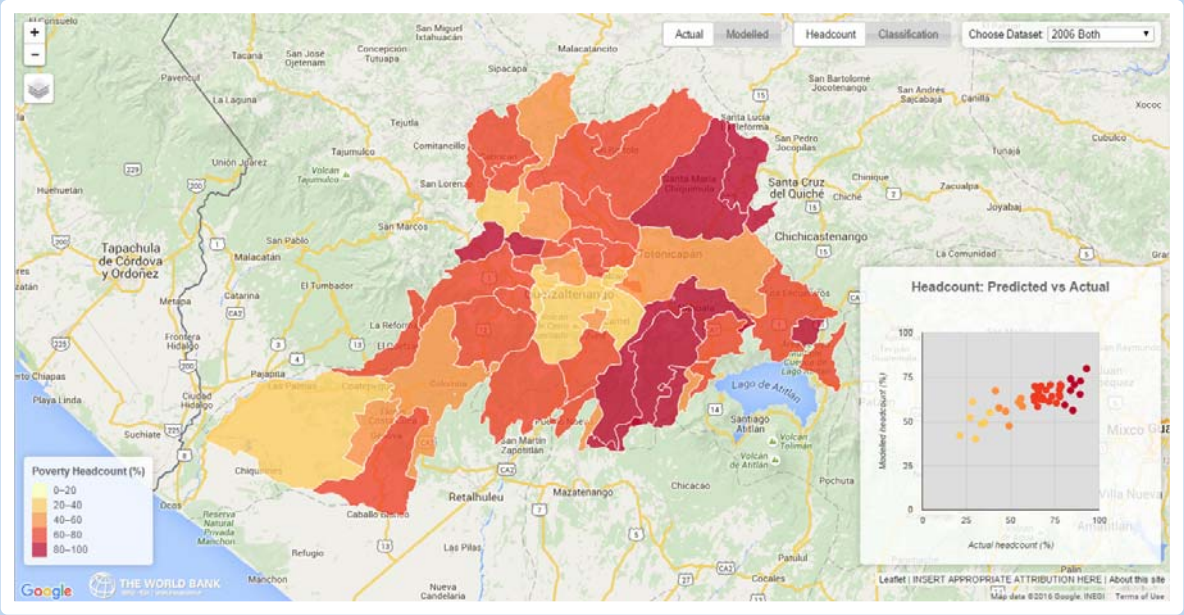
CDR-based models could potentially be used for temporal prediction or time extrapolation (see Boxes 2 and 3), but no research project has yet been able to prove that models trained on past poverty values can be used to predict future poverty levels. Such an analysis would require an extremely large and detailed dataset. For example, predicting future poverty values in Guatemala would require CDR data from 2006 and 2011, as well as corresponding poverty survey data for the same time periods. However, this analysis was based on CDR data from 2013 and rural poverty levels for 2006 and 2011, and no national or urban poverty data were provided. As a result, the predictive model was trained with the 2013 CDR data and the 2006 rural poverty data, and the 2013 CDR data were used to predict rural poverty levels in 2011. While this is the best methodological approach given the data constraints, the preliminary results showed low R^2 values at around 0.29 for regression and maximum F1 scores of 0.6.

Box 3. Using Interactive Data Visualization to Communicate Model Results

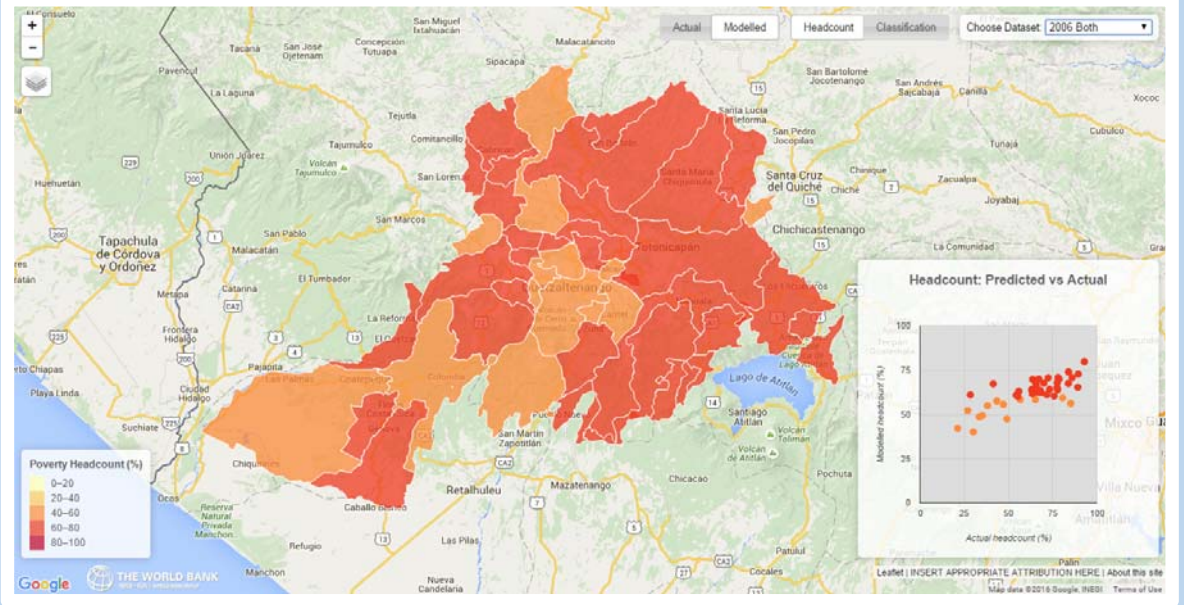
The Guatemala data were mapped on an interactive website. The map allows users to switch between regression and classification models, as well as different data periods. Shaded maps display poverty estimates at a glance, while scatterplots and "confusion matrices" provide more detailed assessments of each model's performance.

Figure 6: Actual Poverty Rates (A) and Modeled Poverty Rates (B) for Included *Municipios*

(A)



(B)



Source: Author's visualization based on data from Hong and Frias-Martinez (2015b).

Other Development Policy Applications of CDR Data

While CDR models are not yet sufficiently accurate to supplant traditional research methods, increasingly sophisticated techniques for combining CDRs with other data sources can enable cellular data to magnify the value of censuses and household surveys. Two especially promising approaches are described below.

Unit-Level Predictions Using Mobile Survey Ground-Truth Data

A recent study in Rwanda used targeted telephone surveys to collect data on personal wealth rather than poverty incidence.[12] The advantage of this approach is that telephone survey responses can be combined with CDR data at the individual level. This is typically not possible with official household surveys, which are often anonymous and do not contain a cell phone number that can be cross-referenced against CDR records.

By coordinating the collection of survey responses with CDR analysis, the study was able to develop a set of high-quality matched individual data. However, the use of telephone surveys to gather the data presents its own limitations, as it is not possible to obtain the same detailed consumption information that face-to-face household surveys usually reveal. Instead, wealth proxies must be used—in this case, questions about asset ownership. If the results are not carefully validated, this can lead to errors in the output variables. Additionally, matching the data to telephone surveys requires that the CDRs not be anonymized, which generates substantial privacy concerns.

The Rwanda study offered an encouraging proof-of-concept for the "time extrapolation" application of CDR data. It found that a CDR model for 2009 provided a more accurate poverty assessment than the outdated 2007 household survey. However, the CDR model was not uniquely suited to this task, as macroeconomic indicators such as GDP growth could be used to create a similarly accurate projection.[13] Moreover, the CDR model only partially reproduced the estimates of district-level average household wealth recorded by the household surveys. Although correlations of around 0.9 were reported at the district level—higher than in the Guatemala analysis—it appears that a few wealthy districts may have driven that result.

Combining CDR Data with Public Observable Data

[12] See Blumenstock et al. (2015).

[13] See Beegle (2016).

For the past several years the WorldPop project,[14] based at the University of Southampton, has produced gridded population maps for dozens of countries, which detail the estimated population per 100 square meters.[15] WorldPop has also produced gridded poverty estimates at a 1 kilometer resolution for a small number of countries using a similar method based on household survey data. In collaboration with the Flowminder Foundation, WorldPop is now beginning to integrate CDR-derived features into the same framework. While relatively little information has been published on this approach, it has the advantage of being able to incorporate all of the relevant data into a single format. The resulting maps should not only be highly accurate, but also as consistent as possible with the underlying datasets. One caveat to the WorldPop approach is that the inclusion of the poverty dimension is still in an early stage, and it is not clear how these maps are validated, since by design they should closely correlate with survey-based poverty estimates.[16]

Using CDR Data for Poverty Mapping

As technical approaches to using CDRs for poverty mapping continue to evolve, a number of challenges will need to be addressed in order to operationalize this methodology as a practical tool for policy analysis. Several of these challenges are described below. A range of other legal and ethical issues are considered in a separate World Bank/Data-Pop Alliance white paper.[17]

Validation and Transparency

Validation is more complicated for CDR models than it is for survey-based techniques. Survey instruments are relatively transparent. They can be inspected, and quantitative and qualitative fieldwork performed both before and after data collection can build confidence. Spatial and temporal inconsistencies can be checked, and in rare cases overlapping surveys can be cross-checked. As a result, well-established survey programs such as the Demographic and Health Surveys and Living Standards Measurement Study are often treated as close to the ground truth.
Validation is more difficult for big data research in general and for CDR models in particular. These data are not specifically collected for socioeconomic analysis, so inspection may not be useful and pre-testing is often not possible. Moreover, CDR models are usually designed to interpolate between conventional survey models in either time or space, so direct validation against a survey is not possible.

---

[14] See http://www.worldpop.org.uk/

[15] These maps are built by combining census data with physical covariates, such as climate, elevation, slope and water bodies, and human covariates, such as urban sprawl, based on satellite imagery, known settlements, roads, and points of interest under a Bayesian framework.

[16] The World Bank Innovation Lab is presently sponsoring additional CDR analysis work with Flowminder/Worldpop in Haiti, which is expected to provide deeper insight into their methods.

[17] Letouzé and Vinck (2015).

Nevertheless, rigorous validation strategies must be developed. At a minimum, in-sample validation techniques should be used.[18] The potential for out-of-sample validation remains unknown, but access to longer CDR data series, which include more than one household survey, could help address this concern.

Privacy

Data privacy is a sensitive and often controversial issue, and certain precautions must be taken before analyzing CDR data. Identifiers should be obscured before records are exported from MNO systems, so that cell phone numbers or similar fields do not remain in the output data. This process is referred to as "pseudonymization." In general, pseudonymization should ensure that the same pseudonym is applied to a given user's entire call history.

Research suggests that simple pseudonymization can be reversed by a determined individual armed with relatively easy-to-obtain auxiliary information, such as a person's home and work addresses and one or two other locations that they are known to visit at particular times.[19] Therefore, additional technical and institutional measures are necessary to restrict access to the data. This has generally entailed nondisclosure agreements between MNOs and researchers, and the research, development and MNO communities are striving to streamline this process.

In addition, any final data produced as a result of the CDR analysis should not compromise individuals' privacy. Similar precautions to those used when releasing tabulations from traditional survey data can reinforce the privacy of CDR data. These measures may include grouping data where cell sizes fall below some fixed number of people (e.g. 5 or 10), or top- and bottom-coding variables in cases where extreme values could be revealing.

Other Promising Applications of CDR Analytics

CDRs offer a rich dataset for studying populations, and numerous applications are emerging in areas beyond poverty measurement. Two applications of particular relevance to the World Bank's work in Guatemala and the Latin America region are described below.[20]

Transportation Analysis

[18] In-sample validation techniques include holdout test sets or leave-one-out cross validation.

[19] De Montjoye (2013).

[20] A more complete review is provided in Blondel (2015).

Transportation analysis is among the most promising applications for CDR models. Social and economic characteristics are only implicitly recorded in the CDR data, and thus strong assumptions and complex modeling are required to infer them. Spatial behavior, on the other hand, is explicitly recorded by the locations of cellular antennae. For this reason, purely spatial analysis tends to be simpler and more robust than other forms of CDR analysis.[21]

Because of the relative simplicity of the analysis, some MNOs and third-party consultants are beginning to offer standard transportation data products to local and national governments.

Disaster Preparedness and Response

Large population movements often occur in the wake of natural disasters, and these movements can render pre-disaster surveys and censuses obsolete. In order to provide humanitarian assistance and restore basic services in disaster-affected areas, governments and organizations require revised population data that can be collected swiftly and at a modest cost. This was the case in Haiti after the 2010 earthquake, which prompted several researchers to examine the potential of CDR data to quickly produce high-frequency tracking information on short-term population displacement. This method was found to correspond closely with other estimates of population mobility. Moreover, it was found that that detailed historical data on pre-disaster population mobility could be used to predict residents' responses to the earthquake, allowing agencies to better prepare for future disasters.[22] Similar techniques have since been applied successfully during and after other natural disasters.[23]

Conclusion

The dramatic expansion of mobile phone use in developing countries in recent years has yielded a rich and largely untapped source of information about the characteristics of communities and regions. CDR-based research methods have the potential to provide detailed and reliable estimates of poverty rates in real time and at a far lower cost than traditional surveys. These methods have especially promising applications in developing countries such as Guatemala, which suffer from high rates of poverty and inequality, and where limited fiscal and budgetary resources both complicate the task of collecting data and accentuate the importance of precisely targeting public spending.

---

[21] For example, Angelakis et al. (2013) used CDR data to examine transport patterns in Côte d'Ivoire. They found that various travel matrices, including routes and travel times, could be calculated from that data at both the national and city level.

[22] Flowminder (2016a).

[23] See for example Moumny et al. (2013) and Flowminder (2016b).

CDR analysis can complement conventional research methods by enhancing the statistical strength of survey data and by extrapolating these data across time and space. The analysis presented above was limited to encrypted and aggregated CDR data from just five administrative departments in southwestern Guatemala, and the results suggest that expanding the sample size would allow for more robust and reliable poverty estimates. Policy makers in Guatemala could obtain more comprehensive datasets by working directly with MNOs.

As analytical methodologies are further developed, CDR applications could extend well beyond poverty mapping. CDRs could enable policy makers to track patterns of crime, food insecurity, epidemic disease and other social and economic shocks in real time. Policy makers in Guatemala and other developing countries now have the ability to access a wealth of cellular data. Establishing strong partnerships with mobile operators will be the first step in harnessing the enormous potential of cellular data for socioeconomic research and policy analysis.

References

Angelakis V, Gundlegård D, Rajna B, Rydergren C, Vrotsou K, Carlsson R, Forgeat J, Hu TH, Liu EL, Moritz S, Zhao S, Zheng Y (2013) Mobility modeling for transport efficiency - analysis of travel characteristics based on mobile phone data. In: *Mobile phone data for development - analysis of mobile phone datasets for the development of Ivory Coast*. Orange D4D challenge, pp 412-422

Beegle, K., Christiaensen, L., Dabalen, A. and Gaddis I. (2016) Poverty in a Rising Africa. Washington, DC: World Bank. doi:10.1596/978-1-4648-0723-7.

Blondel, Decuyper and Krings. 2015. "A survey of results on mobile phone datasets analysis." *EJP Data Science*. http://link.springer.com/article/10.1140/epjds/s13688-015-0046-0.

Blumenstock, Cadamuro and On (2015) "Predicting poverty and wealth from mobile phone metadata," *Science* 350, 1073.

Elbers, Chris, Peter Lanjouw, and Jean Lanjouw (2003). "Micro-Level Estimation of Poverty and Inequality". *Econometrica*, Vol. 71, No. 1, pp. 355-364.

Flowminder (2016a) "Case Study: Haiti Earthquake 2010", http://www.flowminder.org/case-studies/haiti-earthquake-2010

Flowminder (2016b) "Case Study: Nepal Earthquake 2015", http://www.flowminder.org/case-studies/nepal-earthquake-2015

Frias-Martinez, Frias-Martinez and Oliver (2010) "A Gender-Centric Analysis of Calling Behavior in a Developing Economy Using Call Detail Records." https://www.aaai.org/ocs/index.php/SSS/SSS10/paper/viewFile/1094/1347

Frias-Martinez, V. and Virseda, J.(2012) "On the relationship between economic factors and cell phone usage", International Conference on Technologies and Development, ICTD.

Hong, L., Frias-Martinez, E. and Frias-Martinez, V. (2016) "Topic Models to Infer Socio-economic levels", Thirtieth International Conference on Artificial Intelligence, AAAI.

Hong, L. and Frias-Martinez V. (2015a) "Estimating Incidence Values Using Mobile Phone Data: Deliverable 2: Statistical Models". Unpublished manuscript, June 4.

Hong, L. and Frias-Martinez V. (2015b) "Prediction of Incidence Levels". Unpublished manuscript, June 4.

Letouzé E, Vinck P. (2015) "The Law, Politics and Ethics of Cell Phone Data Analytics." Data-Pop Alliance White Paper Series. Data-Pop Alliance, World Bank Group, Harvard Humanitarian Initiative, MIT Media Lab and Overseas Development Institute, April.

De Montjoye, Y.-A., Hidalgo C. A., Verleysen M., Blondel V. D. (2013) "Unique in the Crowd: The privacy bounds of human mobility". *Nature S.Rep.* 3

Moumny, Y., Frias-Martinez, V. and Frias-Martinez, E. (2013) "Characterizing Social Response to Urban Earthquakes using Cell-Phone Network Data: The 2012 Oaxaca Earthquake", Third Workshop on Pervasive Urban Applications @Pervasive'13, Zurich, Switzerland.

Sanchez, S., K. Scott, H. Lopez (2016). *Guatemala: Closing Gaps to Generate More Inclusive Growth*. Washington, D.C.: The World Bank.

Simon, P (2012) "IFC Mobile Money Scoping: Country Report: Guatemala", International Financial Corporation, World Bank Group, http://www.ifc.org/wps/wcm/connect/8b233f0043efb60d95b6bd869243d457/Guatemala+Public.pdf?MOD=AJPERES.