Constructing spatiotemporal poverty indices from big data

Article *in* Journal of Business Research · August 2016

DOI: 10.1016/j.jbusres.2016.08.005

2 authors, including:

Christopher Wambugu Njuguna
Carnegie Mellon University, Rwanda
1 PUBLICATION **27** CITATIONS

Constructing spatiotemporal poverty indices from big data

Christopher Njuguna[a], Patrick McSharry[a,b]

[a] ICT Center of Excellence, Carnegie Mellon University, Kigali, Rwanda.
[b] Smith School of Enterprise and the Environment, University of Oxford, UK.
Email: chris.njuguna@gmail.com, patrick@mcsharry.net

Abstract

Big data offers the potential to calculate timely estimates of the socioeconomic development of a region. Mobile telephone activity provides an enormous wealth of information that can be utilized alongside household surveys. Estimates of poverty and wealth rely on the calculation of features from call detail records (CDRs), however, mobile network operators are reluctant to provide access to CDRs due to commercial and privacy concerns. As a compromise, we show that a sparse CDR dataset combined with other publicly available datasets based on satellite imagery can yield competitive results. In particular, we build a model using two CDR-based features, mobile ownership per capita and call volume per phone, combined with normalized satellite nightlight data and population density, to estimate the multi-dimensional poverty index (MPI) at the sector level in Rwanda. Our model accurately estimates the MPI for sectors in Rwanda that contain mobile phone cell towers (cross-validated correlation of 0.88).

Keywords

Call Detail Record (CDR); Poverty index; Machine learning; Big data; Socioeconomic level; Rwanda

1. Introduction

According to the United Nations (UN), by 2015, about 12% of the world's population or approximately 800 million people were considered extremely poor [1]. While this constituted a significant decline in extreme poverty from a global estimate of 36% (approximately 1.9 billion) in 1990 and an over-achievement of the millennium development goal (MDG) target to halve the global population that is extremely poor by 2015, the number of global poor still remains a major challenge. The UN General Assembly in adopting the agenda for sustainable development stated that ending poverty in all its forms and dimensions, including extreme poverty, is "the greatest global challenge and an indispensable requirement for sustainable development" [2]. It then follows that the prime sustainable development goal, SDG1, pledges to "end poverty in all its forms everywhere" [2] and sets its first target "to eliminate extreme poverty for all people everywhere by 2030" [2]. To eliminate poverty it must first be defined and measured and there are a number of poverty measures used in the world today falling into 2 main categories: monetary poverty indices and non-monetary poverty indices [3]. The monetary indices are based on monetary income measures in local or global currency and include the different national poverty lines and the World Bank $2 a day and $1.25 a day poverty indices. On the other hand, non-monetary indices rely on proxies for wealth and assign poverty level by how much one is deprived of the listed proxies. One such non-monetary index is the Multi-Dimensional Poverty Index (MPI) developed at the Oxford University [4]. The UN currently measures extreme poverty using the $1.25 a day poverty index [2].

Poverty indices are among the key figures and indicators that are used to influence policy and, thus, need to be

based on data that are representative of the population. How is this data collected? Traditionally, the census and survey are two tools that have been commonly used to obtain data but, while they have been proven over the years to be accurate and reliable, the great monetary cost, time and effort of carrying them out means that they can only be undertaken periodically - censuses are typically implemented in ten-year cycles and surveys in a smaller multiple of years, typically three to five years [5]. For example, in Rwanda, the National Institute of Statistics of Rwanda (NISR) undertakes the Rwanda Population and Housing Census (RPHC) every ten years while the Rwanda Economic and Living Conditions survey (EICV) is carried out every three years. Meanwhile, the Rwanda Demographic and Health (RDHS) survey usually undertaken every five years is now carried out every three years [6]. Thus, the census approach to data collection presents a one-off detailed snapshot of the socioeconomic status of the country and with these snapshots, decision-makers can only assess the impact of their policies after three to ten years. Furthermore, decisions made using census data are essentially made using old data since many factors may have changed by the time they are analyzed and disseminated. Worse still, in some developing countries, especially in Sub-Saharan Africa, there has not been an official census or survey in decades leading to a “dearth of reliable statistics” [7]. Needless to say, any statistics for such countries cannot be said to be accurate – a situation that the World Bank has termed Africa’s “statistical tragedy” [7]. This lack of data and, where data do exist, their low update frequency, prompted a global search for alternatives to these traditional tools.

The introduction of digitized records, the growing use of mobile phones and other digital devices and the proliferation of remote sensing devices has led to an explosion in the amounts and variety of data available. This explosion of data has led to a phenomenon known as “big data” which is showing promise as an answer to the search for an alternative to the traditional data sourcing tools [8]. Big data, which due to its novelty still does not have a fixed definition, has been described in various ways. One common description of big data is that it comprises data that have high volume, velocity and variety [7]. Another definition geared towards the use of big data in development describes big data as data that have all or some of the following characteristics: they are digitally generated, passively produced (i.e. generated as a byproduct of everyday life), automatically collected and geographically or temporally trackable with the ability to be continually analyzed [9]. This definition de-emphasizes the volume aspect of data arguing that the kind of data and their source are more relevant than their size. Thus, some regard the name “big data” as something of a misnomer [7].

Big data is a combination of data and the process of analyzing and converting them into actionable insights. In the business world, this has brought about the idea of business intelligence and analytics (BI&A) – the ecosystem that converts data into insights about a business and its environment and informs decisions in a timely manner. Big data and big data analytics, then, have emerged as the third stage of a BI&A evolution over the years. To begin with, the advent of the digital revolution saw organizations shift from manual to digitized records introducing the age of business intelligence (BI&A 1.0). The rise and evolution of business intelligence into business analytics (BI&A 2.0) and finally to big data analytics (BI&A 3.0) has been prompted by the kind and size of data and the tools used to manage and derive value from them [10].

More precisely, business intelligence (BI&A 1.0) appeared in the 1990s and involves the statistical analysis of structured records stored in relational database management systems (RDBMSs). It has been fully adopted into commercial systems today with features such as online analytical processing, reports, statistical analysis, data mining and prediction as examples. Business analytics appeared in the 2000s with the advent of the internet and the web [10] and builds on BI&A 1.0 by adding IP information and personal interaction information stored in server and transaction logs. It also includes unstructured data such as pictures. This data allows for a deeper understanding of customer needs and business opportunities, is significantly larger in volume than BI&A 1.0 and is processed using text and web analytics. Some of the features of BI&A 2.0 have been adopted into commercial systems and include web mining and social-network analysis. Finally, in the third phase of the evolution in the 2010s, big data appeared and it represents an exponential growth in the volume of data. As mentioned earlier, this is brought about by the explosion in the number of phones and sensor-based internet-enabled devices. To illustrate this exponential growth, in 2012 it was estimated that 90% of all the data in the world had been generated within the period 2010-2012 - slightly over two years [7]. The storage, analysis and use of this data is still in the research phase and will not likely be adopted by commercial systems in the near future [10].

Big data has the advantage that it provides a “highly mobile, location-aware, person-centered and context-relevant” [10] dataset and its benefits are starting to be appreciated with the impetus for using big data in improving official statistics gaining momentum recently. The United Nations has recognized the positive impact of big data in projects and called for the increased use of big data to inform decision-making and to assist in working towards the achievement of the newly established global goals for sustainable development terming it the “data revolution for development”. In line with this new direction, the UN Global Pulse was created and it describes itself as a “flagship innovation initiative of the United Nations Secretary-General on big data” which is based on the recognition that “digital data offers the opportunity to gain a better understanding of changes in human well-being and to get real-time feedback on how well policy responses are working” [8]. This activity within the UN demonstrates that big data is not only of academic interest but is already being considered for policymaking.

Yet, with all the promise that big data heralds it also has its limitations. To start with, one of the characteristics of big data is that it is not usually collected with the intent to analyze it [7]. This means that the sample of the data collected may not be representative of the population due to various biases. For example, due to the fact that big data is collected from digital platforms, certain sections of the population such as the young, and those with more disposable income may be higher represented than other demographics simply because they are more likely to use those platforms, thus, leading to a selection bias [9]. Second, the power of big data is in the ability to store and process large amounts of data. However, due to cost and capacity, this capability is less likely to be found in the communities who could actually benefit most from this analysis – the poor countries. Big data could, then, actually end up benefiting wealthier nations rather than the developing world where the benefits are needed most and in the process widening the digital divide rather than narrowing it [7]. A third limitation is in the disproportionate power given to certain actors in the big data ecosystem. For example, there is a greater responsibility given to the data analyst who in her/his analysis can cause certain salient aspects of the data to, say, disappear. A case in point would be where a gender variable is left out of a dataset or analysis and, hence, inequalities in gender representation are not surfaced. The data analyst, then, must have good intentions coupled with a good grasp of the particular domain the data represents in order to offer a fair analysis, since these affect the data they choose to collect, the methods they use to analyze them and how these resutls are presented [11]. Finally, big data has limitations in access - most big data is not open and easily accessible [9]. This is true with both private data owned by companies and public data held by governments. Due to privacy and commercial concerns private companies want to mine their own data in the quest to find value from it and are afraid others might glean information that they have missed. Meanwhile, public institutions may not see “opening” data as being in their interest such as where performance- or corruption-related issues may be revealed. One of the ways to overcome these problems is for organizations to provide aggregated data which would prevent sensitive data from being revealed. However, it is important to note that merging datasets, even aggregated ones with other datasets may still pose a privacy risk as linking and merging datasets may have the effect of revealing individual data [11]. Another way to overcome closed data is to involve different international actors with superior access to data and analytical capacity. Such actors can show the potential value of data by showcasing benefits achieved in their national contexts and hopefully [11]. As big data grow in influence and as more uses are found for them, it is foreseen that addressing the challenges of big data will continue to be a crucial subject area for researchers and practitioners ahead of mainstream acceptance.

Big data can be classified into three categories based on their origin: data explicitly provided by users (e.g. surveys and social media posts), observed data (e.g. web logs and phone records) and data obtained from inference and as output from algorithms (e.g. an individual’s social network). Researchers can combine these data with other data sources to yield high-resolution datasets which can be used in analysis and prediction when compared to a ground truth variable such as census data [11]. One form of big data that is gaining attention from researchers is the phone record collected by mobile companies for billing purposes. Mobile phone records otherwise called call detail records (CDRs) are a form of observed big data and provide a cost-effective and much less labor-intensive process that has the potential of providing a continuous and timely measure of the pulse of a region. This kind of data is continuously available at a daily frequency and high spatial resolution associated with individual cell towers. In addition, such CDR data could be available to government at a reduced

cost since it is passively collected by mobile network operators each time a call is made or an SMS is sent or received and can therefore be viewed as a social good.

Table 1. Censuses and surveys in Rwanda as of Dec 2015

Name	Frequency (Years)	Last Year	Households sampled
Population and Housing Census (RPHC4)	10	2012	Whole population
Housing and Living Conditions Survey (EICV4)	3	2013/14	14,419
Demographic and Health Survey (DHS5)	3	2014/15	12,793

As mentioned earlier, due to privacy concerns, access to big data datasets is a challenge. The situation is no different in the case of CDRs - mobile operators are reluctant to release such records to the public domain due to concerns that the datasets may contain data that is commercially sensitive and may give competitors an unfair advantage. There is also a growing fear of privacy risk, whereby the identity of individual subscribers could be uncovered using sophisticated data analysis techniques even when the data is anonymized [12]. As a compromise, the use of sparse CDRs containing limited fields in combination with other publicly available datasets may provide a reasonable balance between risk and opportunity and pave the way for a workable solution. We find that the use of such a pared down CDR dataset in conjunction with such datasets as satellite night lights can result in similar or better estimates of socioeconomic indicators than those derived using richer CDR datasets alone for example.

Academic challenge

In the course of this research some of the big data challenges stated above were addressed and we hope that in sharing how we overcame them we can contribute to the potential of big data and the immediate necessity of establishing open data policies.

First, we affirm the ability to determine the poverty levels of regions using big data in the form of CDRs as the main dataset thus vindicating the growing use of CDRs as an important addition and in some cases an alternative to census and survey methods. The extent to which the results are accurate will help prove the effectiveness of big data methods in deriving useful metrics and establish these methods for other areas in the policy-making framework. Research in this area has the potential to develop new quantitative methods which can be applied in different countries. To this end, an important contribution of this paper is the cross-validated evaluation of performance against an important poverty metric that relies on survey methods. It is hoped that the accuracy of our approach will promote the general acceptance of big data for policy-making alongside national and official statistics.

Second, having faced the difficulty of accessing data and having successfully worked with a pared down CDR dataset, we hope that we can boost the theory that combining big data datasets with other sources of data from satellites can yield a merged dataset that is sufficient for deriving accurate estimates. We hope that the success of our approach strengthens the open data debate by proposing a compromise between the different stakeholders. Private and public organizations that are opposed to “opening” their data can opt for the release of a subset of anonymized and aggregated data that mitigates commercial and privacy risks. Researchers can then augment this data with other sources of information to yield competitive results with applications that benefit society.

1.1. Objectives

The objectives of this study are:

- To build an accurate poverty index with fine spatial and temporal resolutions
- To affirm the use of big data in development related uses, in this case, predicting poverty indices
- To analyze the usefulness of pared down big data datasets combined with other data sources in building predictive models as a possible contribution to the open data debate

2. Background

2.1. Literature Review

CDRs have already been used in various big data projects to determine human mobility, assist road traffic and infrastructure capacity planning and to determine socioeconomic levels of regions. They are widely recognized as a means of understanding human behavior and facilitating improved policymaking.

In 2009, Blumenstock and Eagle [13] carried out a study in Rwanda to determine the socioeconomic status of individuals based on their CDRs. Their study used interviews of subscribers and analysis of personal phone records to show significant differences in phone records of individuals with different socioeconomic status, specifically that higher socioeconomic status was accompanied by greater mobility and higher call volumes. In particular, they selected 75 questions from the DHS3 related to socioeconomic status to which 856 people responded. They also obtained informed consent from the 856 users to obtain their CDRs from the network operator for the period of a year (May 2008 – May 2009), a total of 1B records. Using the 2 datasets, the researchers then built models that could predict asset ownership of their subscribers. This model was then used to estimate the asset ownership of 15M other subscribers who had not taken part in the surveys. Their results showed they were able to predict asset ownership with a correlation coefficient of 0.917 to the 2007 DHS3 asset ownership data.

Blumenstock et al. (2015) [5] extended their 2009 research, this time using the 2 datasets, CDR data and questionnaire responses, to predict the wealth of the 856 respondents. They first auto-generated thousands of features that quantify phone usage factors such as total volume, intensity, timing, direction and migration patterns then used elastic nets to eliminate irrelevant factors. Using the DHS composite wealth index at the district level as ground truth, they constructed a composite wealth index using the first principle component of the survey responses. They were able to estimate their socioeconomic status with a cross-validated correlation coefficient $r = 0.68$. Farther, the researchers constructed their model to be able to estimate socioeconomic status at the cell and smaller spatial resolution. Using the centroid of the towers used by users weighted by the number of calls made at each tower, they were able to distribute users' data which was aggregated at the tower level. This means that the model could be used to estimate the socioeconomic status for any arbitrarily sized geographical region.

Smith-Clarke et al. [14] determined the socioeconomic levels in regions in two developing countries – Cote d'Ivoire and an anonymous region named Region B. This study used the flow of calls and SMSs between regions to determine the socioeconomic status of a particular region. They found, among other things, that regions which had higher call volumes from a higher number of other regions were more likely to have a higher socioeconomic status while "introverted" regions – regions with fewer and smaller volumes of calls with other regions - were more likely to have a lower socioeconomic status.

Soto et al. [15] were able to determine the socioeconomic level of a region with a classification accuracy of up to 80% based on CDRs. Having classified regions into three socioeconomic levels - low, medium and high - they were able to build a model using features derived from CDRs that could accurately determine the socioeconomic level of a region. However, their model used 279 features extracted from CDRs and various datasets which were reduced to 38 and 17 respectively for the two best models. The study analyzed results obtained from support vector machines (SVMs), random forests and regression to model their data. Results from the three different models were similar to the SVM model, with the latter providing the best result.

On the other hand, satellite nightlight data have been used to determine economic activity at national and sub-national levels. Here we sample two studies that use satellite nightlights.

In 2009, Elvidge et al's. [16] seminal research built a poverty index based on nightlight data and Landscan population data. In their study, they were able to estimate the poverty levels at the global level and at the national level in 233 countries. In particular, they were able to estimate a global total of 2.2 billion people below the poverty line against World Bank estimates of 2.6 billion – representing an error of 15%.

Mellander et al [17] studied night-time lights to determine if they could be used as a proxy for a number of variables among them economic growth. Using a "fine-grained geo-coded residential and industrial full-sample micro-dataset for Sweden", they find a strong correlation between the logs of both nightlights (saturated light and light density) and economic activity. They use graphically weighted regression (GWR) and compare it with ordinary least squares regression as a baseline. Using the Aikike Information Criteria (AIC) they find that the GWR yields better results than OLS regression and that the nightlight-based features, especially light density, are a good proxy for economic activity.

An overview of some of these studies and the sources of data are provided in Table 2. In this study, we demonstrate that just a few carefully chosen CDR features combined with other publicly available datasets are sufficient to accurately determine the poverty level of a sub-national spatial region. This is important because in many situations researchers only have access to limited datasets due to the commercial and privacy concerns of mobile network operators. The successful use of this restricted and anonymized dataset to determine the poverty levels of a region could offer considerable opportunities to enhance the work of policymakers, donors and non-governmental organizations (NGOs).

Table 2. Chronological sample of related studies and a description of the sources of data.

Author	Data source	Data Year	Country / Region / Individual	Population Sample size	Time Period	# Records	Poverty measure	Model	Correlation
Blumenstock & Eagle (2009)	CDR & Survey	2009	Individual	1.5M & 856	9 months	1B+ & 64K	-	-	-
C. Elvidge et al (2009)	Nightlight	2009	Global & national	-	12 months	-	WDI[a] $2 a day international poverty line (2006)	Regression	0.85
Soto et al	CDR	2010	Main city in Latin American country	500K	6 months	Unknown	Socioeconomic Levels (A, B, C) from NSI[b]	Support Vector Machine	0.80
Smith-Clarke et al	CDR	2011/ 12& 2012	Cote d'Ivoire & Anonymous Region B	5M & 928K	20 weeks & 6 weeks	471M & 40M	IMF poverty rate estimates (2008)	OLS[c] regression	-
Blumenstock et al (2015)	CDR & survey	2009	Rwanda	1.5M & 856	9 months	1B+ & 64K	DHSIV Composite Wealth Index	Linear regression	0.68
This study	CDR, RPHC4, Landscan & nightlight	2014/ 15	Rwanda	4.8M	9 months	3.8B	RPHC4 MPI	Linear regression	0.88

[a]. World Development Index.
[b]. National Statistics Institute.
[c]. Ordinary Least Squares regression.

3. Methodology

Unlike other studies that use automated methods (e.g. [18][5][14]), we employ manual feature engineering, analysis and selection of the features in the final model. An overview of the procedure follows.

After obtaining the data, we manually generated a small number ($n < 20$) of intuitive features that we expected to be good proxies for economic activity or socioeconomic status. The data were aggregated at the sector level to match the spatial resolution of our ground truth – the MPI. In addition, in line with our objective to build a self-updating index, we left out features that had low update frequencies (e.g. population from census data and number of towers per sector). The initial list of engineered features included call volume per day/week/month, number of phones and number of towers used per day/week/month. We then checked the correlation of these features to the MPI. Any features that had low correlation to the MPI ($r < 0.4$) were eliminated at this point.

Once we had selected the features with promising correlations to the MPI, we proceeded to analyze the distributions of their data. We noted that the actual MPI is approximately normally distributed while all our predictors, i.e. nightlights, call volume, mobile ownership and population were highly skewed. This, we hypothesized, was due to very bright lights around major towns compared to the rest of the country for nightlights, high call volumes and mobile ownership in a small number of sectors with high populations and finally high populations in large sectors. We normalized these variables to reduce the effects of population and geographical area.

To select the features that we would use to build our final model, we used LASSO (Least Absolute Shrinkage and Selection Operator) to eliminate any features that were not significant to the model. Ridge regression on the other hand was used to eliminate any features that exhibited collinearity. Elastic nets ensured our models were sparse. To obtain significance values not provided by LASSO, Ridge regression and elastic nets, and as a farther confirmation of the features selected for our model, we used backwards stepwise regression.

In the next sections we describe the different datasets used, how they were processed and transformed to form the features used to build the final model.

3.1. Data

3.1.1. Sectors

The spatial component for this study is the sector in Rwanda which is the third administrative level after the province and district. There are 416 sectors, 30 districts and 5 provinces in Rwanda. For this study we utilized sector maps from the Rwanda Land Use Planning Portal [19]. Maps were projected to the WGS84 coordinate reference system (CRS) to standardize their positions and facilitate accurate mapping operations. Sectors in Rwanda have areas ranging from 1 km^2 to 650 km^2 and an average of 58 km^2.

3.1.2. The Multi-dimensional Poverty Index (MPI)

The multi-dimensional poverty index (MPI) is an index that was developed at Oxford University's Oxford Department of International Development under the Oxford Poverty and Human Development Initiative (OPHI) [4]. The MPI measures three dimensions like the Human Development Index: health, education, and standard of living. The MPI is calculated using "several factors that constitute poor people's experience of deprivation – such as poor health, lack of education, inadequate living standard, lack of income (as one of several factors considered), disempowerment, poor quality of work and threat from violence" [4]. The MPI has been measured globally by OPHI and the United Nations Development Program (UNDP). The MPI is one of the measures of poverty used in Rwanda alongside monetary poverty. The MPI data used in this study were retrieved from calculations made by the National Institute of Statistics of Rwanda using data from the 2012 Rwanda Population and Housing Census (RPHC4). Thus, the MPI is based on data collected from the whole population of Rwanda and is representative of the levels of poverty per sector in the country.

Advantages

- The MPI is calculated at a relatively high spatial resolution (416 sectors): Other poverty measures are at lower resolutions e.g. district
- The MPI is employed in multiple countries hence may allow for international comparisons
- The MPI is representative of the population: the sample is the whole population of Rwanda

Disadvantages

- The MPI is calculated at a low temporal resolution: It is based on the RPHC census data which is updated once a decade

3.1.3. Call Detail Records (CDRs)

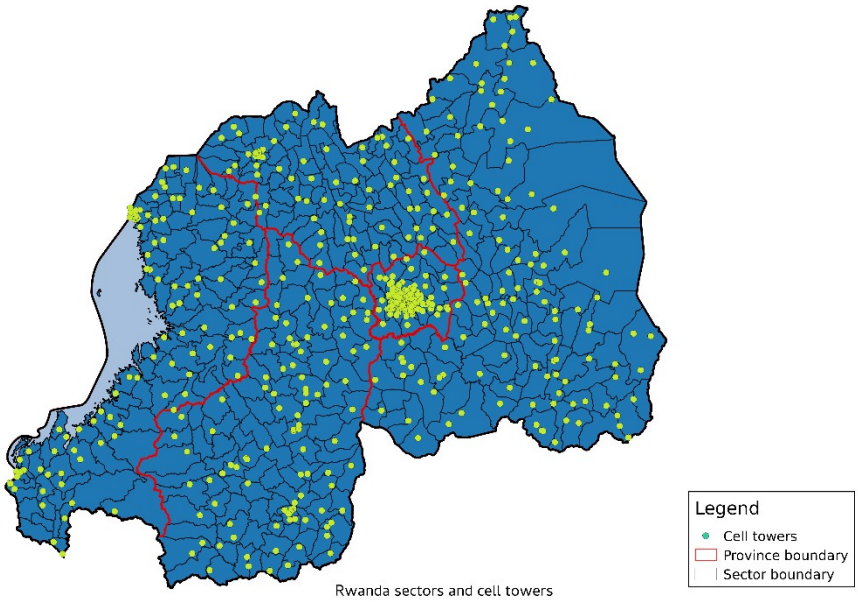
We use mobile CDRs from a leading mobile network operator (MNO), with about 49% market share (as of the last month in the dataset) in Rwanda, representing calls and SMSs between 1st July 2014 and 31st March 2015. The choice of Rwanda is first and foremost because the CMU campus is based in Kigali, Rwanda. Also, a partnership between the university and the Rwanda Utilities and Regulatory Authority meant that the data, especially CDRs, was more accessible for this research (see acknowledgements). In this dataset, calls are not distinguished from SMSs, do not contain call duration, and do not indicate the receiving party. The data is anonymized by providing a unique ID that can be used to identify events attributed to an individual user but cannot be mapped back to any personal information about them. The scant features of this data posed a substantial research challenge, namely to find out how much value can be derived from such a limited dataset. On the other hand, there are fewer privacy concerns with making use of this data. The call records contained the following fields: timestamp, anonymized subscriber ID, start cell and end cell. Summary statistics of the CDRs are provided in Table 3.

The CDRs are mapped to 580 towers located across the country as shown in Figure 1. Each tower hosts a number of cells that subscribers actually connect to. The location of a subscriber during a call or when sending an SMS is estimated to be the geographical location of the tower that hosts the cell that the user is associated with at the time of the event. It is possible for the subscriber to be connected to a different cell by the end of a call and this is recorded by having a different value in the End Cell field than in the Start Cell field.

Table 3. Statistics from the Rwanda CDR Dataset: July 2014 – March 2015.

Number of calls/SMSs	Number of subscribers	Number of cells	Number of towers
3.8B	4.8M	3075	580

In this study we focus on two variables that are extracted from the CDRs: mobile ownership and call volume. To store and manipulate this data we use the Hadoop ecosystem, in particular Apache Hive to store the records in tables, Cloudera Impala for ad-hoc queries, and Apache Pig for more complex operations all built atop the YARN resource manager and the HDFS filesystem.



Rwanda sectors and cell towers

Figure 1. Rwanda sector map showing cell tower locations.

Advantages
- CDRs are spatially and temporally disaggregated allowing for rich analysis and feature engineering
- CDRs are relatively cheap to produce since they already exist

Disadvantages
- CDRs require large storage and processing resources which may incur costs and also require specialized skills
- CDRs are difficult to obtain due to commercial sensitivities and privacy concerns
- CDRs are biased by the characteristics of the "owning" mobile operator as well as the demographics of people likely to use mobile services

3.1.4. Nightlight satellite data

Nightlight data are records of artificial light on the earth's surface collected during the nighttime by the visible infrared imaging radiometry suite (VIIRS), a remote sensing instrument on the Suomi National Polar-orbiting Partnership (S-NPP) satellite. The data is distributed by the National Oceanic and Atmospheric Administration (NOAA) in the form of monthly composited cloud-free maps to maximize the amount of light received from the earth's surface – the satellite cannot "see" through clouds. The nightlight maps have a resolution of 15 arc-seconds which corresponds to about 450m x 450m at the equator and have a range of 3×10^{-9} $W.cm^{-2}.sr^{-1}$ to 0.02 $W.cm^{-2}.sr^{-1}$ [20]. The data used is a monthly composite of daily nightlight intensities for the month of December 2014.

Advantages
- Nightlights are freely available
- Nightlights have a fine spatial and temporal resolution

Disadvantages
- Nightlights may be distorted by other sources of light e.g. gas flares and wild fires

3.1.5. Landscan population data

The Landscan population database produced by US Department of Energy, Oak Ridge National Library since 1998/99 is a spatially disaggregated global population count dataset. Unlike census data which represents residential populations, Landscan represents a 24-hour population count which results in populations for such places as public areas [21]. We use this dataset to show that the poverty levels can be calculated using wholly independent and unofficial sources. While the producers do not recommend a one-to-one comparison of the Landscan population data to population census data, we found a correlation of 0.87 between the 2012 Landscan and the 2012 census (RPHC4) population counts for Rwanda at the sector level.

Advantages
- Accurate and regularly updated
- Distributed population so that counts are not only in residential areas

Disadvantages
- Not free: requires purchase

3.2. Feature Engineering

To determine the socioeconomic status of the sectors in Rwanda, we calculated a small number of meaningful features from the CDRs combined with the nightlights and Landscan population datasets. This was done by

calculating a number of features that were thought to describe economic activity or socioeconomic status and then performing feature selection against our dependent variable multi-dimensional poverty index (MPI). The selection of features was unanimous across a variety of feature selection methods such as backwards stepwise regression, LASSO, elastic nets and ridge regression, thereby indicating the robustness of the approach. We now provide a detailed description of the construction of each of the selected features.

3.2.1. Nightlight per capita

While nightlights are used as a proxy for economic activity, the amount of activity that can be attributed per capita yields a socioeconomic profile for the individual and hence of the area. Higher intensities with lower populations would yield the highest socioeconomic status of a region while low nightlight intensities of areas with high populations would imply a lower socioeconomic status. We use the zonal statistics function in QGIS to intersect the raster nightlight images with sector polygons to yield the number of pixels, sum of intensity and mean intensity per sector. We divide the total intensity in each sector by the population of the sector to yield a light intensity per capita.

3.2.2. Mobile ownership per capita

Mobile phone ownership is a good indicator of socioeconomic status. In the 2012 Rwanda Population and Housing Census (RPHC4) 54% of households were reported to own a mobile phone. This corresponded to 84% of urban and 48% of rural households. In addition, 85% of households in the Kigali Province owned a mobile phone compared to a low of 46% in the Southern Province [6]. CDR data are not fully representative of the population since the poorest and most vulnerable section of the society may not be represented.

Mobile ownership is derived from CDRs by determining unique mobile identities and the towers at which they are most active. This is done by aggregating all users' calls by the towers they are associated with. The tower where they make the most calls is the considered their "home" tower. The number of subscribers per tower are then aggregated to the sector in which the towers are located and this number then becomes the mobile ownership for the sector.

Note that only 295 of the 416 sectors in Rwanda have towers within their boundaries. While it is a fact that the sectors without towers are served by towers in neighboring sectors, it is difficult to determine what ratio of calls to assign to each of the sectors. As such, sectors without towers are not included in this model. In the discussion at the end of this paper we consider ways that could be used to distribute the call volume and mobile ownership so as to allow the estimation of poverty for all sectors in the country and to allow for this model to be adapted to finer spatial resolutions.

3.2.3. Average daily call volume per phone

Call volume per phone is one of the metrics that we used as a proxy for socioeconomic status. It has been previously shown that the higher the socioeconomic level, the more calls a subscriber would make on a daily basis [18]. Call volume per phone is calculated as the total number of events (calls and SMSs) made from the mobiles owned in a particular sector for the entire nine month period for which we have CDRs and averaged to provide a daily call volume. We assign the calls to the subscribers "home" sector which is identified as described in the mobile ownership per capita section below. We proceed to count the total number of calls and SMSs made during the nine months and divide this number by the number of days in the dataset to yield the average daily call volume per phone.

3.2.4. Population density

Global trends show that urbanization is increasing and this provides an indicator of industrialization. It is assumed that the gathering of people in an area implies a source of livelihood. The greater the population density, the higher the probability of economic activity.

To calculate the population density we use the sector population data from the Landscan database and divide it by the sector area as calculated from the sector map in QGIS using the $area function in the polygon field calculator. This yields the geographic area of each sector in the country and is validated against the sector areas as listed in the Rwanda Integrated Living Conditions (EICV4) survey. We find less than a 4% difference in the total area of the sectors in the country.

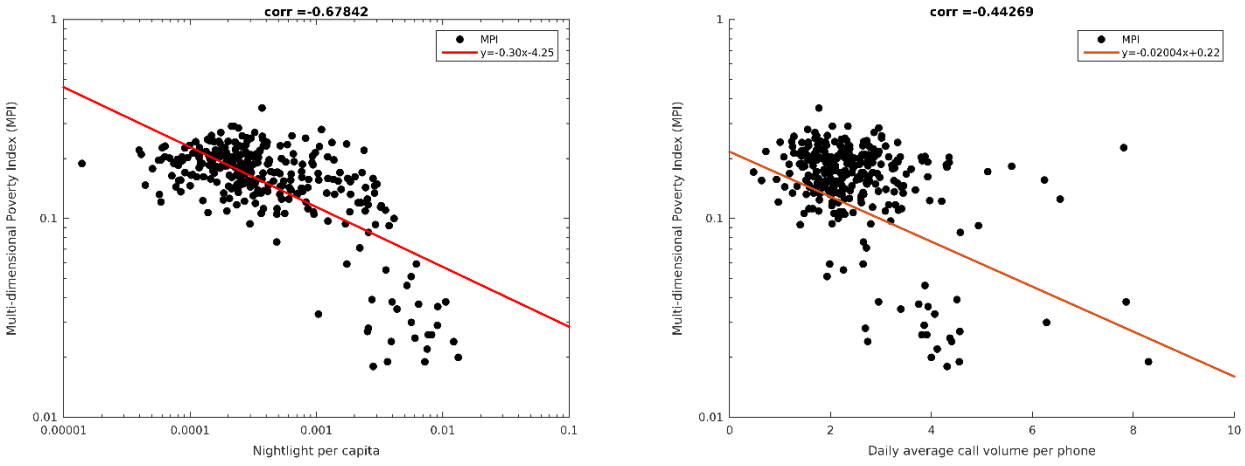
4. Results

The predictive features are carefully constructed so as to provide proxies for economic activity in each sector. For this reason, we have to ensure that each feature is normalized such that neither the population nor area of the sector influence the measured values. In addition, we need to transform each feature so that it is approximately normally distributed to facilitate the construction of a quantitative model.

Analysis of the scatter plots of the four features against the MPI suggested possibly non-linear relationships between nightlights per capita and population density and MPI. To attempt to fit a linear model to this data, we transformed the data by taking their logarithms. To assess the validity of this approach, we used Least Absolute Shrinkage and Selection Operator (LASSO) feature selection to decide whether the log or non-log versions of the variables should be employed. The log versions of the call volume per phone and the number of phones per capita were dropped prompting the use of the non-log versions of these particular variables. Table 4 provides summary statistics for the four predictive features and the MPI. The scatter plots showing the transformed relationships are provided below in Figure 2.

Table 4. Summary statistics of the features and MPI.

Feature	Average	Median	Minimum	Maximum
Nightlights per capita	0.00008	0.00028	0.00001	0.01328
Daily call volume per phone	2.54	2.30	0.47	8.30
Mobile ownership per capita	0.523	0.413	0.007	2.601
Population density	1032	551	45	27333
MPI	0.1703	0.1730	0.0180	0.3590



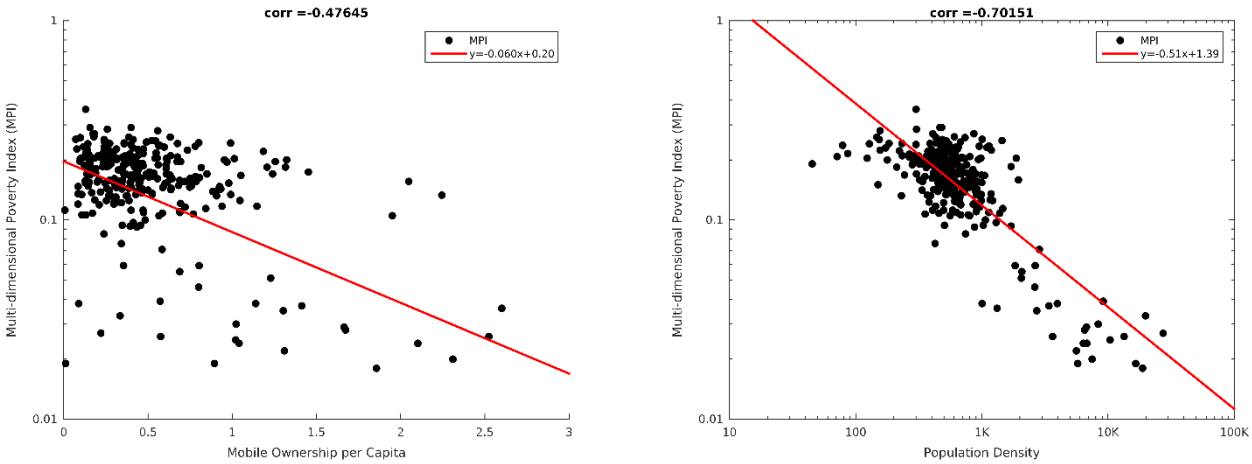


Figure 2. Scatter plots of the MPI versus the four predictive features showing their correlations.

Using LASSO, we fit a model between the four features, nightlights per capita, call volume per phone, mobile ownership per capita and population density, and the dependent variable MPI yielding a linear model with the following structure:

$$\log(MPI) = \beta_0 + \beta_1 x_1 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 . \quad (1)$$

The parameter estimates are given in Table 5 and all were statistically significant ($p < 0.05$). A comparison of the predictability of the individual features and the final model is provided in Table 6. The cross-validated correlation coefficient of our predicted log (MPI) versus the actual log (MPI) is 0.88. This corresponds to an R-Squared coefficient of 0.76 and means that the selected model explains 76% of the variance of the MPI across the 295 sectors in Rwanda. A scatter plot shows the predicted MPI versus the actual MPI in Figure 3. With this method we can then rank the sectors in deciles in terms of their poverty level to determine the sectors with the highest and lowest poverty levels in Rwanda. This information can then be used to construct a visual graphic in the form of a map as shown in Figure 4.

The maps in Figure 4 show the decile ranking of sectors based on their MPIs, the first being the MPI obtained from the NISR and the second the estimated MPI from our model. The estimated MPI map is created using data imputed using the nearest neighbor technique for the 121 sectors that do not have cell towers in order to aid visual analysis – the map with only 295 estimated MPI values is much harder to visually compare against the 426 sectors in the actual MPI map due to the gaps where no value is estimated. As such, we expect a lower correlation to the MPI map but a display that makes it easy to visually analyze the model's results. The correlation in the maps is most visible where there are clusters of sectors that fall in the same or neighboring deciles. This can be seen in the around Kigali (in the center) where there is low MPI, and in clusters to the south, south west and north west where the MPI is high.

Table 5. Estimated parameter values of the multivariate linear model consisting of four predictive features.

i	x_i	β_i
0	Intercept	-0.5883
1	Log (nightlights per capita)	-0.1229
2	Call volume per phone	-0.0001
3	Mobile ownership per capita	-0.1693
4	Log (Population density)	-0.3276

Table 6. Correlations of the four predictive features and the final model prediction with the MPI.

Feature	Correlation	t-stat	p-value
Log (nightlights per capita)	-0.6784	-9.3723	2.09×10^{-18} ***
Call volume per phone	-0.4427	-3.5340	4.76×10^{-4} ***
Mobile ownership per capita	-0.4765	-6.0115	5.51×10^{-9} ***
Log (Population density)	-0.7015	-15.552	4.42×10^{-40} ***
Model	0.8775		

*significant at p = 0.05.
**significant at p = 0.01.
***significant at p = 0.001.

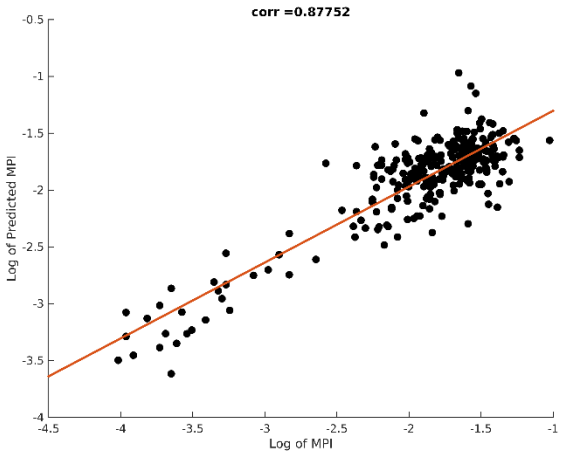


Figure 3. Scatter plot of the predicted MPI versus the actual MPI.

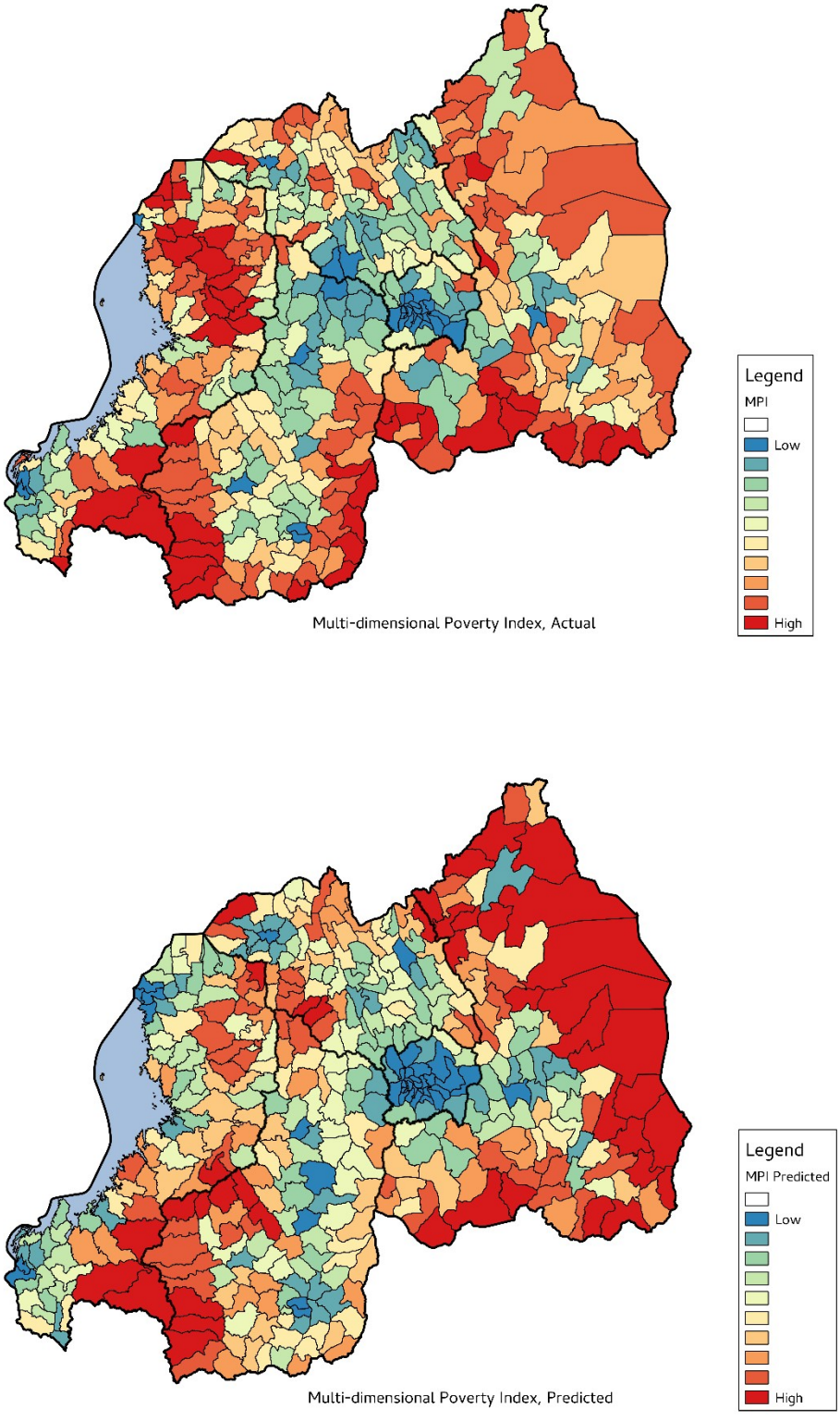


Figure 4. Maps showing the actual (top) and predicted (bottom) MPI in deciles.

5. Discussion

5.1 Limitations

This study utilizes CDRs obtained from a leading MNO, however, it is worth noting that the provider has a greater market share in the urban areas, especially in Kigali, than in the rural areas and may not be representative of all segments of Rwandan society. As such, we find that the correlations of CDR-based features, call volume per phone and mobile ownership per capita, to MPI, while significant cannot be said to be strong. Stronger correlations may exist between these two CDR-based features and the MPI in Rwanda if CDRs from all MNOs were obtained. A complete set of CDRs would eliminate biases introduced by the positioning and marketing campaigns of MNOs and provide statistics that are more representative of the various socioeconomic levels in the country.

Second, only 295 of the 416 sectors contain mobile phone cell towers. This means that 121 sectors do not have a direct measure of CDR statistics. Various methods are used to distribute calls evenly across a geographical region one of which is the employment of voronoi partitions. Voronoi partitions were used but yielded low correlations for reasons that were not immediately clear. Therefore, this study proceeded to assign all calls to the sector in which the cell towers were found. We realize that this means that we do not have estimates for 121 sectors in the country and that there is a definite bias introduced for the sectors where data is available and estimates are calculated.

Finally, there is a lag between the predictor data and the ground truth data i.e. the MPI. While the nightlights, Landscan population and CDRs are all gathered from periods around the year 2014, the MPI is calculated using data from the 2012 Rwanda population census. This lag is likely to have negatively affected our results and so we expect that a future study with all datasets picked from the same time period should yield even better results. Since in Rwanda the next population census will be in 2022, future studies may choose to use the DHS composite wealth index as an alternative ground truth even though it has a lower resolution being calculated at the district level. Another alternative, is to carry out the study in another country or region.

5.2 Practical Implications

The study demonstrates that alternative sources of big data, from satellites and mobile network operators, offer a means of accurately predicting multidimensional poverty. If mobile CDRs were made available via an API, these predictions could be generated in real-time, thereby reducing costs and effort in collecting data. In contrast to census data, which is updated every ten years, this alternative approach can be updated on a monthly or quarterly basis. This means that stakeholders and policymakers can obtain up-to-date information that will help them monitor and evaluate programs on a regular basis. Frequent feedback offers the basis of an early warning system to alert policymakers so that corrective and timely action can be taken before the situation deteriorates. The ability to disaggregate poverty information by geographical location can help target resources and interventions.

A dashboard could be created to allow policymakers to drill down and visualize the evolution of poverty levels on a monthly basis at a neighbourhood scale, allowing poverty trends to be shown for each particular neighbourhood. The dashboard could include a dynamic map that enables policymakers to focus on specific geographical locations. With sufficient historical information it will be possible to empirically determine the impact of different development programs active in these areas.

The open data movement advocates for the publication of data by private companies and governments for the use of researchers and innovators as a way to unlock the potential held in numerous disparate datasets. Although companies and governments acknowledge the potential of data analytics, many feel they have much more to lose than gain by sharing data. From commercial sensitivities to privacy concerns, there are good reasons against the publication of data held by these entities as far as the perceived risks are concerned. Having faced this challenge in this particular study, even with regulatory support from government, it is recognized that only

once the rewards of using CDRs are quantified will these entities become more willing to share or release their valuable data.

The ability to combine disparate datasets offers great potential but more successful examples are required to break the stalemate that exists between data gatekeepers and researchers. Our approach has been to propose a compromise whereby organizations and researchers can negotiate on the minimum dataset that reduces privacy and commercial concerns. Researchers can then demonstrate the social value of these datasets by linking and merging them with other sources of data. It is hoped that once data holders see the impact of their data with hopefully minimal impact on their competitiveness, they will be willing to discuss releasing richer datasets.

5.3 Future work

This study utilizes CDRs that are located only in the sectors where cell towers exist. Other studies have used various techniques such as Voronoi partitions and user centroids weighted by call volume to distribute the tower-based statistics e.g. call volume and mobile ownership to finer geographical resolutions in the country via an intersection and merge operation. It is envisaged that, with some analysis in this direction, suitable features that are representative of economic and socioeconomic status for arbitrary geographical areas can be estimated.

Finally, future work should result in the building of a self-updating system with pipelines from all data sources and all processing automated where possible and semi-automated where manual interventions are required (e.g. cleaning of CDR datasets). An example frontend would be a dynamic web-based country map showing the sectors ranked and color coded based on the estimated MPI that could have a playback function to show how the ranking has changed over time.

6. Business Research Implications

The outcomes of the research presented here suggest potential for a number of business applications in microfinance. At present many people in developing countries live in a rural setting and are unbanked with little opportunity to borrow money from financial institutions. Any loans that are available are at prohibitive interest rates. The ability to use CDRs to assess socioeconomic status could be combined with other features to determine credit risk scores. Although the amounts being borrowed might be small, the potential for a large number of transactions across Africa is considerable. Data science could provide a means of establishing risk scores and underpin a mechanism for offering affordable loans and this would improve financial inclusion. Furthermore, information gathered from a variety of sources helps financial institutions to “know your client” even before meeting them in person. At this point there is ongoing between existing brick and mortar financial institutions such as banks and microfinance institutions and the new players in the form of mobile network operators. It is likely that the winners will be those that figure out how best to harness the important information that can be extracted from big data.

7. Conclusion

With the census and surveys being resource-intensive and being carried out in cycles spanning multiple years, big data has been shown to be an effective means of obtaining socioeconomic estimates of regions which can be used to continuously monitor the effects of policy and help shape future decisions. We have demonstrated that the use of independent, publicly available and relatively cheap data sources such as call detail records, nightlights and the Landscan population dataset can be used to accurately determine the multidimensional poverty levels of the sectors in Rwanda. We fit a multivariate linear model that describes the relationship between four features, nightlight intensity per capita, call volume per phone, mobile ownership per capita and population density to the multi-dimensional poverty index derived from the Rwanda Integrated Living

Conditions survey. We estimate MPIs for 295 of the 416 sectors in Rwanda that have mobile phone towers within their boundaries. The key findings of the research are the following:

- Demonstrating the predictability of poverty levels based on proxies for disposable income such as nightlight and mobile telephone ownership and usage.
- Showing that a sparse subset of call details records, mitigating commercial and privacy risks, can be used to accurately quantify poverty.
- Constructing of a model based on big data sources that allows for accurate prediction of the MPI with a cross-validated correlation coefficient of r=0.88.
- Establishing the potential of big data for informing policymakers and monitoring poverty in real-time by geographical region.
- Providing the basis of an approach for government and donors to target resources and monitor the impact of different interventions.
- Paving the way for business applications to assess the potential of providing banking services for poorest of the poor, improve financial inclusion and facilitate credit risk ratings.

Importantly, this study yields competitive results from a sparse CDR dataset containing only four fields without requiring the use of other highly sensitive fields that mobile network operators are, in most cases, unwilling to provide due to commercial sensitivity and privacy risk. We show that through the combination of sparse CDRs with other publicly available datasets based on satellite imagery, we can produce effective poverty estimates that can be used to help decision-making. We believe that this may provide a compromise between mobile network operators and policymakers in the quest to unlock the wealth of data that CDRs represent and offer what promises to be a considerable social good.

Finally, we show that big data is a valuable resource for estimating socioeconomic indicators and is reaching maturity where it can and should be accepted as a dependable and accurate source of information. Timeliness and high temporal resolution are two enormous advantages of big data that should be emphasized. The CDR and satellite imagery required for predicting multidimensional poverty could be updated every month and accessed immediately, which therefore offers a means of tracking poverty at the level of individual sectors over a relatively fine temporal scale. It can be argued that the minimal loss of accuracy in predicting MPI is more than made up for by the near real-time aspect of this policy tool based on big data.

Acknowledgements

We would like to thank Patrick Nyirishema the Director General of Rwanda Utilities Regulatory Authority (RURA) for partnering with the Carnegie Mellon University on this project. We would also like to thank Francis Ngabo, the then Director of the Frequency Monitoring Division of RURA, Georges Kwizera, Protais Kanyankore and the staff of RURA in general.

We are grateful to Rajiv Ranjan, Governance Unit of the United Nations Development Program (UNDP) and ICT Advisor to the National Institute of Statistics of Rwanda (NISR) and Tom Bundervoet, Senior Economist with the World Bank for their invaluable help and support.

We acknowledge Chuma Vuningoma and the staff of MTN, Rwanda for their support and assistance with CDR acquisition.

References

[1] United Nations, “The Millennium Development Goals Report 2015,” *Dep. Econ. United Nations. Dep. Public Inf.*, p. 75, 2015.

[2] “Transforming our world: the 2030 Agenda for Sustainable Development,” 2015. [Online]. Available: http://www.un.org/ga/search/view_doc.asp?symbol=A/RES/70/1&Lang=E.

[3] F. Bourguignon and S. Chakravarty, “The measurement of multidimensional poverty,” *J. Econ. Inequal.*, vol. 1, no. 1, pp. 25–49, 2003.

[4] Oxford Poverty and Human Development Initiative, “Rwanda Country Briefing,” *Multidimens. Poverty Index Data Bank. OPHI, Univ. Oxford*, 2013.

[5] J. E. Blumenstock, G. Cadamuro, and R. On, “Predicting Poverty and Wealth from Mobile Phone Metadata,” *Science (80-.).*, vol. 350, no. 6264, 2015.

[6] National Institute of Statistics of Rwanda (NISR) and M. of F. and E. P. (MINECOFIN), “Fourth Population and Housing Census, Rwanda, 2012. Thematic Report: Characteristics of households and housing,” pp. 9–19, 2014.

[7] E. Letouzé, “What is big data, and could it transform development policy?,” no. May, 2014.

[8] Global Pulse, “Mobile Phone Network Data,” *United Nations Glob. Pulse*, no. October, pp. 1–12, 2013.

[9] M. Hilbert, “Big Data for Development,” *Dx.Doi.Org*, no. May, pp. 1–41, 2013.

[10] H. Chen, R. H. L. Chiang, and V. C. Storey, “Business Intelligence and Analytics: From Big Data to Big Impact,” *Manag. Inf. Syst. Q.*, vol. 36, no. 4, pp. 1165–1188, 2012.

[11] L. Taylor, J. Cowls, R. Schroeder, and E. T. Meyer, “Big Data and Positive Change in the Developing World,” *Policy & Internet*, vol. 6, no. 4, pp. 418–444, 2014.

[12] E. Letouzé and P. Vinck, “The Law, Politics and Ethics of Cell Phone Data Analytics,” 2015.

[13] J. Blumenstock and N. Eagle, “Mobile Divides: Gender, Socioeconomic Status, and Mobile Phone Use in Rwanda,” *Proc. 4th ACM/IEEE Int. …*, pp. 6:1–6:10, 2010.

[14] C. Smith-Clarke, A. Mashhadi, and L. Capra, “Poverty on the cheap,” in *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI ’14*, 2014, pp. 511–520.

[15] V. Soto, V. Frias-Martinez, J. Virseda, and E. Frias-Martinez, “Prediction of socioeconomic levels using cell phone records,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6787 LNCS, no. 1, pp. 377–388, 2011.

[16] C. D. Elvidge, P. C. Sutton, T. Ghosh, B. T. Tuttle, K. E. Baugh, B. Bhaduri, and E. Bright, “A global poverty map derived from satellite data,” *Comput. Geosci.*, vol. 35, no. 8, pp. 1652–1660, 2009.

[17] C. Mellander, J. Lobo, K. Stolarick, and Z. Matheson, “Night-time light data: A good proxy measure for economic activity?,” *PLoS One*, vol. 10, no. 10, pp. 1–18, 2015.

[18] J. E. Blumenstock and N. Eagle, “Divided We Call : Disparities in Access and Use of Mobile Phones in Rwanda,” *Inf. Technol. Int. Dev.*, vol. 8, no. 2, pp. 1–16, 2012.

[19] “Rwanda Land Use Planning Portal.” [Online]. Available: http://www.rwandalanduse.rnra.rw. [Accessed: 08-Dec-2015].

[20] K. Shi, B. Yu, Y. Huang, Y. Hu, B. Yin, Z. Chen, L. Chen, and J. Wu, “Evaluating the ability of NPP-VIIRS nighttime light data to estimate the gross domestic product and the electric power consumption of China at multiple scales: A comparison with DMSP-OLS data,” *Remote Sens.*, vol. 6, no. 2, pp. 1705–1724, 2014.

[21] A. N. Rose and E. Bright, “The LandScan Global Population Distribution Project: Current State of the Art and Prospective Innovation,” 2014.