Monitoring Socio-Economic Characteristics from Call detail records

HAPPY BUZAABA[†1] CHIEKO NAKABASAMI[†2]

Abstract
The growing use of mobile phones and other digital devices all over the world has led to an explosion in the amounts and variety of data available. This data hold the potential as yet largely untapped to allow decision makers to track development progress, improve social protection, and understand where existing policies and programs require adjustment.
In this paper we present a large scale study to analyze the relationship between socio-economic factors and how people use phones in Rwanda. Our approach combines large scale call detail records dataset with country wide household survey data to reveal findings at the national level. The results show correlation between cell phone usage and socio-economic characteristics with $R^2 \approx 0.505$.

Keywords
Call detail records (CDR), Base transmitter station (BTS), National Institute of Statistics (NIS), Demographic and household survey (DHS).

1. Introduction

Socio-economic indicators are among the key figures that are used to influence policy and thus need to be based on accurate data that represents the population.
Traditionally, the census and survey are two tools that have been commonly used to obtain socio-economic information. While they have been proven over the years to be accurate and reliable, monetary cost, time and effort of carrying them out is high that they can only be undertaken periodically. (Blumenstock, Cadamuro, & On,2015)

Table 1. Countries with outdated census before 2014

Country	Year of last census	Years Since
Somalia	1986	28
Congo, Dem, Rep.	1984	30
Eritrea	1984	30
Afghanistan(2011 socio-economic and demographic survey)	1979	35
Lebanon	1943	71
Angola before	1970	44

Source: United Nations Department of Economic and social Affairs, Population Division, based on analysis in July 2014 of the Implementation of the World Population and Housing Census program since 1948.

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) and the Rwanda Demographic and Health Survey (RDHS) are carried out every three years as shown in Table 2.

Table2. Census and surveys in Rwanda as of Dec.2015

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

Source: National Institute of Statistics of Rwanda 2014.

In countries with a limited budget, decisions made using census data are essentially made using old data since many factors may have changed by the time they are analyzed and dispersed.
This lack of data and where data do exist their low update frequency prompted a global search for alternatives to these traditional tools.
The growing use of mobile phones and other digital devices has led to an explosion in the amounts and variety of data available. This explosion of data is showing promise as an answer to the search for an alternative to the traditional data sourcing tools (Global Pulse, 2013).
The mobile phone adoption in Rwanda is generating millions of digital footprints from cell phone usage which would be modeled into variables like consumption levels, social networks and mobility patterns.
This paper statistically evaluates the relationship between cell phone usage and socio-economic factors by combining two datasets, one of cell phone records from the second largest telecommunication operator in Rwanda and a demographic and household survey dataset gathered by the Rwanda national

†1 Toyo University.
†2 Toyo University

statistical institute. The call detail records dataset from the telecommunication operator in Rwanda is used to generate cell phone usage patterns for mobile phone users by geographical location and the demographic and household survey dataset is used to generate social, economic and demographic variables by geographic area in Rwanda. We combine the two datasets and infer mathematical model that formalizes the relationship between cell phone usage and household survey data at a large scale without conducting personal interviews.

Such models could be used to approximate census variables of individuals or geographical regions basing on call detail records. This could be used to complement the computation of expensive and time consuming census maps specifically for developing countries with a limited budget.

2. OBJECTIVE

This paper proposes an analytical approach of approximating census variables from behavioral patterns collected through cell phone records with the objective of examining mobile phone call detail records and their potential to accurately approximate socio-economic indicators.

In this paper we will be answering the following questions:

How can call detail records be used to monitor social economic data?

Can they be relied on to accurately approximate census data?

3. RELATED WORK

Several studies analyzing the relationship between socio-economic factors and cell phone usage have previously been conducted using large scale datasets.

An analysis proposing a model closely related to this work of combining cell phone data set and the census data set collected at the national level, was carried out across large and middle sized cities in Latin America by Vanessa Frias-Martinez together with Jesus Virsesa of Telefonica Research in 2012 to understand the relationship between cell phone use and specific human factors. The main findings revealed that there exist moderate and strong correlations [1].

An analysis to find out the impact that factors like gender or socio-economic status have on cell phone use was conducted by Blumenstock et al in Rwanda. He combined two datasets, one containing call detail records from a telco company in Rwanda and the other one containing socio economic variables computed from personal interviews with the company's subscribers. Their main findings revealed gender based differences in the use of cell phones and large statistically significant differences across socio-economic levels with higher levels showing larger social networks and larger number of calls among other factors. This approach succeeds to reveal findings at an individual level [7].

Donner et al. presented a survey of 277 micro entrepreneurs and mobile phone users in Kigali, Rwanda, to understand the types of relationships with family, friends and clients, and its evolution over time [11]. Among other findings, the author discovered an inverse correlation between the age of the user and the probability of adding new contacts to its mobile based social network. He also mentioned that users with higher educational levels were also more prone to add new contacts to their social net works.

A study to understand the impact of demographics and socio-economic factors on the technology acceptance of mobile phones was conducted by Kwon et al. During the study, a four page survey with 33 questions was circulated to 500 cell phone subscribers and found that older subscribers felt more pressure to accept the use of mobile phones than their younger counterpart. In fact, cell phones were generally given as presents by family members for security purposes [12].

Eagle et al. studied the correlation between communication diversity and its index of deprivation in the UK [13]. The communication diversity was derived from the number of different contacts that users of a UK cell phone network had with other users. He combined two datasets:

I. A behavioral dataset with over 250 million cell phone users whose geographical location within a region in the UK was known.
II. A dataset with socio-economic metrics for each region in the UK as compiled by the UK Civil Service. He found that regions with higher communication diversity were correlated with lower deprivation indexes. These results represent an important step towards understanding the impact of socio-economic parameters on mobile use at a regional level.

3.1. STUDY MOTIVATION

Mobile phone adoption in developing economies has occurred faster than any other communication technology. As they have become increasingly affordable, penetration rates have soared in Africa. This has undoubtedly enhanced the social and economic integration of the region: migrant workers stay in contact with their families more easily than before (Jansen, 2007) and mobile money services now facilitate business transactions, remittances and even micro-credit. In the past 3 to 6 years, these high levels of penetration and usage have become very significant even in the most rural areas of the least developed countries (GSMA 2014).

The mobile phone penetration rate in Rwanda has been the main driver of this research. This rate defined as the number of active mobile phone numbers divided by the population have soared in the past 7 years to reach almost 76% of the total population in 2016 [8]. This penetration rate opens door for remarkable data gathering.

4. METHODOLOGY

4.1. DATA DESCRIPTION

4.1.1. Call Detail Records

Cell phone networks are built using a set of base transceiver stations (BTS) that are responsible for communicating cell phone devices within the network. Call Detail Records (CDRs) are generated whenever a cell phone connected to the network makes or receives a phone call or uses a service (e.g., SMS, MMS). In the process, the BTS details are logged, which gives an indication of the geographical position of the user at the time

of the call. For this analysis the maximum geo-location granularity that we could achieve was that of the area of coverage of a BTS. Each BTS or cellular tower is identified by the latitude and longitude of its geographical location. The area of coverage of a BTS was approximated using spatial voronoi tessellation [14]. The location of the subscriber within the coverage area is not known. From all the information contained in CDR, our study only considers the encrypted originating number, the encrypted destination number, the time and date of the call, the duration of the call and the BTS that the cell phone was connected to when the call was placed. A 6 months period (August 2016 - February 2017) Call detail records (CDRs) dataset of 3.2 million subscribers provided by the second largest mobile telecommunication operator in Rwanda was used. From the call detail dataset, three set of variables were computed: (1). Consumption variable characterizing the cell phone usage measuring the total number of calls both input and output and the duration of the call or the expenses. (2). Social network variable estimates the social network a person builds when communicating with others. This approximates parameters like the number of people one receives calls from; we consider those that receive at least one, at least two and at least five calls in a week. (3). Mobility variable which characterizes the geographic area where a person spends most of his or her time. The factors considered are the number of sites visited (movement patterns) and radius of gyration or number times a particular site is used.

Table 3: Description of independent variables from (CDR)

Variable Type	Description of variables
Consumption	Total number of calls Average call Duration
Social net work	Reciprocity of communication At least one (R1) At least two (R2) At least five (R5)
Mobility	Total number of sites visited Radius of gyration

4.1.2. Demographic and household survey data

In order to gather country wide demographic, socio and economic information for Rwanda, demographic and household survey (DHS) data collected in 2014 by the National Institute of Statistics is used. During the survey, a total number of 12,792 households were interviewed around the country. The country was divided into 492 clusters and 26 households were randomly selected from each of the cluster for interview. Paper survey forms are applied during the interview which makes the collection process even more expensive and time consuming.

For analysis of this paper three groups of variables are used from the demographic and household data ie: education to measure citizen's education level whether they are illiterate or have finished up to certain educational level, demographic to measure gender and age variables and property ownership is used as a proxy of the purchasing power of a person. Property ownership measures parameters like the existence of electricity, Refrigerator, computer or TV, in the household.

Table 4: Description of dependent variables from (DHS)

Variable Type	Description
Education	% of Illiterate Population % of population with Primary Education % of population with Secondary Education % of population with University Education
Demographics	% of Female population % of Male population % of population under (Age < 35) % of Middle aged population (35 - 65) % of Senior aged population (65+)
Property Ownership	% of households with Electricity % of households with Refrigerator % of households with Computer % of households with a Bicycle % of households with a Moto % of households with a car

4.2. Combining the two data sets

To find the relationship between cell phone usage variables and socio-economic variables we combined the two datasets by associating clusters to the BTS residential locations they belong to using their GPS locations.

Figure1. Shows the BTS coverage map of Rwanda, the red dot is the position of the BTS and the polygon is the estimated area of coverage of the BTS. We estimate the area of coverage of the BTS using Voronoi tessellation [14].

Figure 2. Shows the mapping of clusters to the BTS residential locations they belong to. The blue dot is the cluster location in the polygon area of coverage of the BTS.

In order to associate clusters to the BTS residential area they belong to we apply the haversine formula to calculate the difference between latitude and longitude [18].

When a BTS is set up, it is estimated to cover up to 2 km^2 in urban areas and up to 10 km^2 in rural areas. During the survey, a cluster was estimated to cover up to 2 km^2 and 10 km^2 in urban and rural areas respectively. In the process of associating clusters to their BTS locations, if the absolute difference between the BTS and cluster coordinates was less than 0.2 in urban areas, the cluster was associated to that BTS location. For rural areas if the absolute difference between the BTS and cluster coordinates was less than 0.5, the cluster was associated to that BTS.

4.3. STATISTICAL ANALYSIS OF VARIABLES

To understand whether there exists statistically significant difference between cell phone usage variables and socio-economic variables in their residential locations, multiple linear regression is applied to carry out ANOVA tests

It is important to note that the relationship such analysis reveal is only valid in principle for the users in our sample.

In order to extend our findings to mobile users beyond our

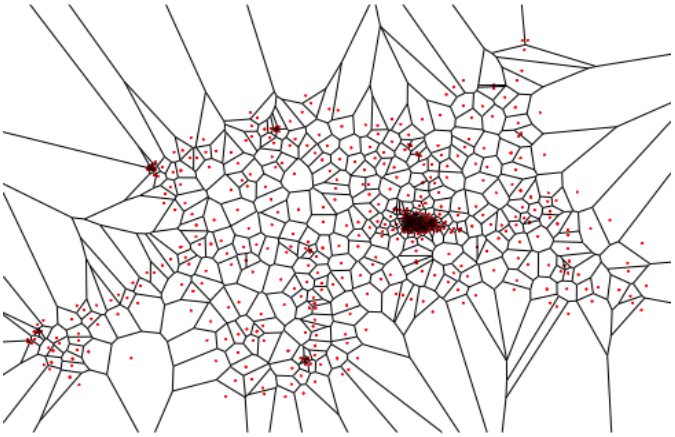

Figure1. Map of Rwanda showing area of coverage of BTS as approximated by voronoi tessellation.

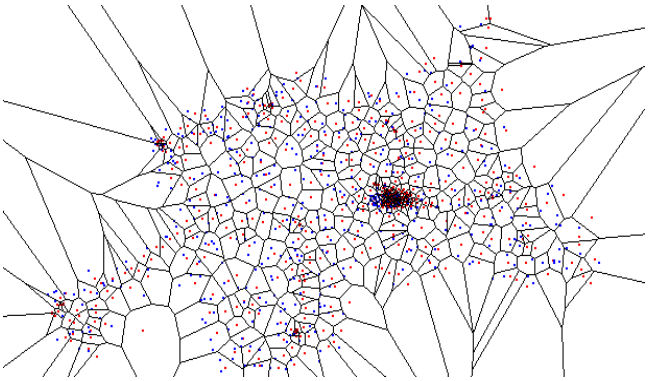

Figure2. Map of Rwanda showing clusters in blue associated to their respective BTS locations

sample, it is necessary to guarantee that they represent a distribution of citizens similar to the general distribution of the country under study.

To apply multiple linear regression we formulate the parameters:

Assuming Y*i* is an ideal function that predicts social and economic status based on features X*i*, then Equation 1;

$$Yi = \beta o + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots\ldots + \beta_7 X_{i7} + \varepsilon$$

Where;

Y_i is the actual Social and economic status from the dependent variables,

X_{i1}– Total calls per person,

X_{i2} – Average call duration per son,

X_{i3} – R(5) at least 5 calls a week,

X_{i4} – R(2) at least 2 calls a week,

X_{i5} – R(1) at least 1 call a week,

X_{i6} – total number of sites a person uses,

X_{i7} – Radius of gyration or average number of visits by a person to a particular site. $\beta_0, \beta_1 \ldots \beta_7$ are coefficient estimates and ε is the irreducible error.

Therefore the matrix representation is:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{17} \\ x_{12} & x_{22} & \cdots & x_{27} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{n7} \end{pmatrix} * \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_7 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} + \beta$$

To estimate Y_i, we fit a model function $\hat{f}(x)$;

$$\hat{f}(x) = \hat{\beta}_o + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \ldots + \hat{\beta}_7 X_{i7}$$

The hat "λ" is used to refer to an estimate of a given variable or function. We use R Software to determine the quality of the fitted model and tell whether there exists statistical significance as well as the correlation.

We report statistical results for each group of variables characterizing cell phone usage (Consumption, Social net work and mobility) and socio-economic variables.

Only those variables that show significant statistical results are reported because the number of variable pairs evaluated is large. For (ANOVA) tests we report results for the p-value < 0.05 and Pearson's correlation with $R^2 > 0.4$.

The (*), (**) and (***) represents moderate, strong, and very strong significance respectively the (+) and (-) refer to a positive or negative correlation between two distributions.

a. Consumption Variables

Consumption variables that showed statistical significance are total number of calls made or received and the average call duration. All variables refer to weekly averages computed for mobile phone users living within the same BTS coverage area.

Table 5 shows statistical results for ANOVA tests and Pearson's correlation. The findings show that there is statistical significance between the total calls made and the census variables with the exception of households reported to own a computer. Call duration also show statistical significance with all the socio economic variables apart from Call duration.

In terms of correlation results, the demographic variable show interesting results, the percentage of (age <= 35 and age >= 65) show a positive correlation with the total number of calls made or received. This means that in an area with a big population under these age ranges, the more the total number of calls made or received. There is however a negative correlation with Call duration which means that they make more calls that last for a short time.

The percentage of age (35 < 65) seems to show a positive correlation with call duration meaning that the higher the population with this age range in the region, the large the call duration in that region. There is however a negative correlation with the total number of calls made, this means that they do make fewer calls but the calls last for a long time which is associated to the expense i.e. more money is spent.

There is also a positive correlation between primary school completion and duration of call. This implies that the more the population with a primary school completion the more the call

duration on contrary to the illiterate percentage which show a negative correlation with duration of call. Households which report computer, refrigerator, moto and car ownership show positive correlation with call duration. In other words they tend to make calls that last for a significant amount of time. Whereas households with bicycle tend to make more calls that last for a short time.

Table5. Consumption variable

Census Variable	N_Total	Call Duration
Age <=35	< 2e-16***+	3.86e-10***-
Age 35 < 65	< 2e-16***-	2.67e-16***+
Age >= 65	<2e-16***+	
Illiterate	< 2e-16***+	9.1e-12***-
Primary	< 2e-16***-	0.000112***+
Electricity	< 2e-16***+	1.4e-06***-
Computer		3.29e-06***+
Refrigerator	< 2e-16***-	0.000819***+
Bicycle	< 2e-16***+	0.000186***-
Moto	7.04e-05***+	0.008184**+
Car	< 2e-16***-	4.21e-11***+

b. Social network variables

In Table 6, reciprocity of communication is the social network variable that showed statistical significance with the socio-economic variables. Reciprocity (R) refers to the number of reciprocal communication between a person and her contacts, in other words out of the people one communicates with, how many calls him/her back in a week. We evaluate three possible values, at least one (R1), at least two (R2) and at least five (R5). All the socio-economic variables showed statistical significance for (R1) and (R2), we however notice no significance between primary school completion, owning a computer and (R5).

Households that reported to have electricity show a positive correlation across (R1, R2, and R5) this means that in an area with a big percentage of households with electricity, a high number of reciprocal communications is observed than others which is an interesting finding since in Rwanda the total electricity connectivity is at 25% [18].

c. Mobility Variables

Table 7, show mobility variables that show statistical significance include the number of sites visited by a person and the radius of gyration referred to as the number of times a person uses a particular site. In Table 7, Sites visited show statistical significance with all socio-economic variables, there is however no statistical significance with the age >= 65. The radius of gyration show statistical significance with all socio-economic variables with the exception of households that reported to own Moto. In terms of correlation coefficient we observe a positive correlation between the number of sites visited and age < 35 which means that the higher the percentage of population in an area under this age rage, the more the sites visited. In other words young people tend to visit more BTSs contrary to old people who show a positive correlation with radius of gyration. We also see another interesting finding, households with electricity, car and computer ownership show a positive correlation with sites visited yet those who reported to own a bicycle show a positive correlation with the radius of gyration.

Table 6. Social network variable

Census variables	R(1)	R(2)	R(5)
Age 35 < 65	<2e-16***-	<2e-16***-	<2e-16***-
Age >= 65	<2e-16***-	<2e-16***-	<2e-16***-
Illiterate	<2e-16***-	<2e-16***+	<2e-16***+
Primary	6.86e-14***+	0.00785**-	
Secondary	7.92e-06***-	<2e-16***-	<2e-16***-
Electricity	<2e-16***+	<2e-16***+	<2e-16***+
Moto	<2e-16***-	0.0112*+	0.0033**-
Car	<2e-16***-	<2e-16***-	<2e-16***-
Computer	<2e-16***-	<2e-16***+	

Table 7. Mobility variable

Census variables	Sites visit	Radius of Gyration
Age <= 35	<2e-16***+	3.42e-10***-
Age 35 < 65	<2e-16***-	<2e-16***+
Age >= 65		<2e-16***+
Illiterate	<2e-16***+	<2e-16***-
Secondary	<2e-16***-	<2e-16***+
Electricity	< 2e-16***+	6.93e-05***-
Bicycle	<2e-16***-	<2e-16***+
Moto	0.000123***-	
Car	0.000399***+	< 2e-16***-
Computer	<2e-16***+	<2e-16***-

5. PREDICTIVE MODEL

This section explores whether cell phone usage variables can be used to predict socio-economic status. This could be used as a cheap alternative to the existing expensive census and survey tools.

In this part we use the cell phone consumption composite variable as a measure of expenditure and property ownership composite variable as a measure of purchasing power parity to show that mobile phone subscribers' wealth can be predicted from his or her historical patterns of phone usage with a correlation coefficient of $R^2 \approx 0.505$.

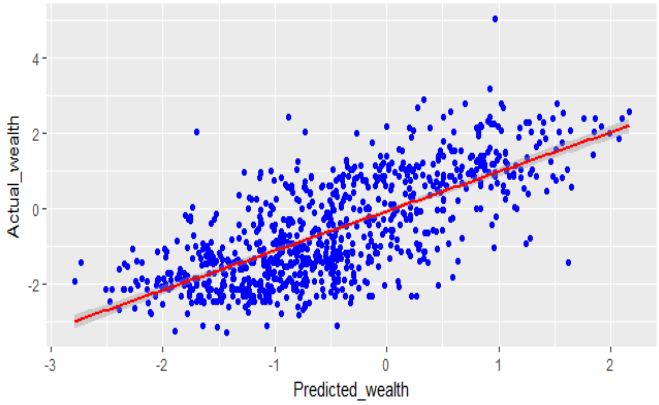

Figure 1. Relation between actual wealth as reported by DHS survey and predicted wealth as inferred from CDR for each of the BTS location with $R^2 \approx 0.505$.

5.1. DISCUSSION

The study uses CDRs from one telecom operator in Rwanda to reveal the relationship between cell phone usage and socio-economic factors. It is worth noting that the provider has a large market share in rural areas than urban and may not be a representative of all segments. We therefore find that while the correlation is significant, there could be a stronger correlation if data from all mobile operators is used.

Secondary, only 12,792 households are covered in the DHS data used, meaning that a large portion of the population is not represented by the results. More significant results would be generated if census data for the entire population is used.

There is also a lag between the predictor data and the ground truth data. i.e. while the CDR data is gathered in 2016/17, the relation is calculated using data from 2014 Rwanda demographic and household survey.

5.2. APPLICATIONS

The results show potential of revealing behavioral patterns which could be applied to estimate the socio-economic variables that are critical inputs to public policy, ie infrastructure investment, aid distribution and political redistricting cheaply and in real time that would outperform the 3 years survey and the 10 year census in Rwanda.

The results also show a capacity to identify wealth neighborhoods, it could be used for tax collection initiatives or locate systematic tax evasions, though it might not be reliable to identify tax potential, it could be useful at neighborhood level.

Knowing the geographical location of poor and rich populations helps in making sure that anti-poverty spending reaches those it was designed to reach without leakages to non poor.

A combination of poverty geographical information with other data sets, such as those relating to climate or infrastructure, may help us to better understand some of the drivers of poverty.

Information regarding communicable diseases can be selectively transmitted to phone numbers geographically identified (i.e in terms of sites they use or visit) to be at risk.

6. CONCLUSION

We have presented a study to understand the relationship between various census variables and cell phone usage in Rwanda combining cell phone records dataset with demographic and household survey to reveal findings without the need to carry out expensive personal interviews or questionnaires.

The main findings reveal that there exist moderate and strong correlations between the specific census variables and the total number of calls, call duration of a person, the reciprocity of her communications and sites visited or geographical area where a person typically moves to or spends most of their time.

We also provided a predictive model that allows predicting the wealth of a geographical location exclusively from cell phone records.

Such predictive models can be used to cheaply approximate expensive census maps especially for emerging economies. Future work will focus on computing similar analysis with census data covering the entire population and call detail records from all the telecommunication operators in the country to test the significance.

References

[1] Vanessa Frias-Martinez of Telefonica and Jesus Virsesa On relationship between social economic factors and cell phone usage.

[2] Trevor Hastie, Robert Tibershirani and Jerome Friedman, The Elements of Statistical Learning, Data Mining, Inference, and Prediction. (Springer Series in statistics)

[3] National Institute of Statistics of Rwanda (NISR) [Rwanda], Ministry of Health (MOH) [Rwanda], ICF International, "Rwanda Demographic and Health Survey2010," DHS Final Reports (publication ID FR259, NISR, MOH, and ICF International, Calverton, MD, 2012).

[4] H. Zou, T. Hastie, J. R. Stat. Soc. Ser. B 67, 301–320 (2005).

[5] L. Breiman, J. Friedman, C. J. Stone, R. A. Olshen, Classification and Regression Trees (Chapman and Hall/CRC Press, New York, ed. 1, 1984).

[6]. A. J. Tatem and S. Riley. E_ect of poor census data on population maps. Science, 318(5847):43, Oct. 2007.

[7] J. Blumenstock and N. Eagle. Mobile divides: Gender, socioeconomic status, and mobile phone use in Rwanda. In Proceedings of the 4th International Conference on Information and Communication Technologies and Development, 2010.

[8] "Rwanda utilities regulatory authority." http://www.rura.rw/fileadmin/docs/Monthly_telecom_subscribers_of_December__2015_Chgd.pdf on 12/10/2016

[9] "GSMA-Intelligence." https://gsmaintelligence.com/analysis/2012/11/two-thirds-of-africans-yet-to-join-the-mobile-revolution/357/

[10] "National Institute of Statistic of Rwanda, Integrated Survey on Life Conditions." http://www.statistics.gov.rw/publications/article/rwanda%E2%80%99s-mobile-phone-penetration-rised-over-past-five-years on 12/10/2016

[11]. Donner. The use of mobile phones by micro Entrepreneurs in Kigali Rwanda. Changes to social and business network. Information Technologies and International Development, 3(2), 2007.

[12]. Kwon and L. Chidambaram. A test of the technology acceptance model: The case of cellular telephone adoption. In Proceedings of the 33rd Hawaii International Conference on System Sciences, 2000.

[13]. Eagle. Behavioral inference across cultures: Using telephones as a cultural lens. IEEE Intelligent Systems, 23:4:62–64, 2008.

[14]. G. Voronoi. Nouvelles applications des param`etres continus `a la th´eorie des formes quadratiques. Journal fur die Reine und Angewandte Mathematik, 133:97–178, 1907.

[15] J. M. Lane, L. C. Carpenter, T. Whitted, and J. F. Blinn. Scan line methods for displaying parametrically defined surfaces. *Communications* ACM, 23(1):23–34, 1980.

[16]. M. Gonzalez, C. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. Nature, pages 453:479–482, 2008.

[17]. A. Rubio, V. Frias-Martinez, E. Frias-Martinez, and N. Oliver. Human mobility in advanced and developing economies: A comparative study. In AAAI Spring Symposium on Artificial Intelligence for Development (AI-D), 2010.

[18]. http://www.rdb.rw/rdb/energy.html