Monitoring Social and Economic Characteristics from Call detail records

Happy Buzaaba
Toyo University
5-28-20, Hakusan, Bunkyo-ku, Tokyo 112-8806, Japan
hbuzaaba@gmail.com

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 has led to a phenomenon known as “big data” an umbrella term for the explosion in the quantity and diversity of high frequency digital data.

These 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 examine call detail records and their potential to predict social and economic characteristics so as to complement official statistics, survey data, and information generated by Early Warning Systems, adding depth and nuances on human behaviors and experiences and doing so in real time, thereby narrowing both information and time gaps (Global Pulse, May 2012).

Keywords

Call detail records (CDR), Geographical units (GUs) Base transmitter station (BTS) National statistical Institute (NSI) Big Data Information processing society Journal (IPSJ).

1. INTRODUCTION

Timely and accurate population characteristics estimation is a significant input to social and economic research as well as policy. Poverty indices are among the key figures and indicators that are used to influence policy and thus need to be based on data that represents the population. 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. Monetary cost, time and effort of carrying them out is high that they can only be undertaken periodically. (Blumenstock, Cadamuro, & On,2015). 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 (National Institute of Statistics of Rwanda (NISR) 2014).

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. Worse still, in some developing countries, especially in Sub-Saharan Africa, Angola’s last census before 2014 was carried out in 1970 and only covered 18 districts [6] in other words no official census or survey has been undertaken in decades leading to a “dearth of reliable statistics” (Letouzé, 2014).

Needless to say, any statistics for such countries cannot be said to be accurate, a situation that the World Bank has termed as Africa's “statistical tragedy” (Letouzé, 2014).

Table1. 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

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 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 (Global Pulse, 2013).

This paper seeks to contribute as an alternative to the traditional tools by carrying out Statistical evaluation of the relationship between cell phone usage and demographic or socio economic factors as well as providing an analytical model to approximate census variables from cell phone records. The paper infers mathematical models that could be used as inexpensive soft substitutes of national census campaigns.

The widespread presence of cell phones in Rwanda is generating millions of digital footprints from cell phone usage. These records are useful in modeling the use of cell phones through variables like consumption levels, social networks or mobility patterns.

A number of researchers all over the globe have recently studied the relationship between cell phone usage patterns and their demographic or socio-economic characteristics.

In Rwanda researchers have studied relation between cell phone usage patterns from subscribers and their demographic or social economic factors [7]. The researcher computed usage patterns from a large-scale dataset of cell phone calls and also, carried out personal interviews over the phone with the subscribers, who reported their own socio-economic and demographic information.

This kind of mixed method approach limits the amount of cell phone users that can be modeled to the number of interviews that can be carried out and hence loosing the large-scale component of the analysis provided by the calls' dataset.

In this paper I attempt to address this issue by proposing an analytical approach that combines two large scale datasets of call detail records with a country wide census data from the local national statistical institute.

The combination of the two large two data sources reveals relationship between cell phone usage and census data without carrying personal interviews.

Later in the research will provide an analytical model to formalize the relationship between cell phone usage and demographic or Social and economic variables.

This model would be used to approximate the unknown census variables of a geographic region based only on the cell phone usage records.

This would serve as a useful alternative to the expensive and time consuming computation of census in developing and low resource countries like Rwanda.

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 social economic indicators.

.

3. RESEARCH QUESTION

How can mobile phone data be used to monitor social economic data?

Can it be relied on to accurately approximate census data?

4. RELATED WORK

Studies analyzing the relationship between socio-economic factors and cell phone usage have been done using large scale datasets.

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.

More qualitative studies were carried out by Kwon et al. [12] He conducted a study to understand the impact of demographics and socio-economic factors on the technology acceptance of mobile phones. For that purpose, they circulated a four page survey with 33 questions 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.

Blumenstock et al [7]. Analyzed the impact that factors like gender or socio-economic status have on cell phone use 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.

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 250million cell phone users whose geographical location within a region in the UK was known, and (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 first step towards understanding the impact of socio-economic parameters on mobile use at a regional level, in this paper I go ahead to elaborate on the analysis that can draw correlations between human factors and cell phone usage at even smaller scales like districts in Rwanda.

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 [1]. To understand the relationship between cell phone use and specific human factors. The main findings revealed that there exist moderate and strong correlations.

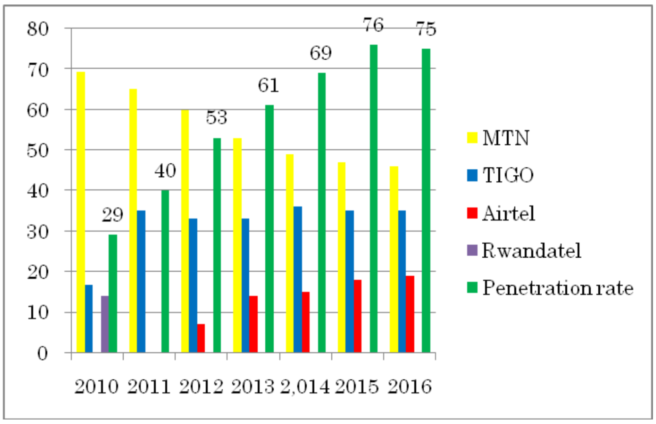
5. CASE STUDY

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).

In Rwanda mobile penetration rates defined as 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 (http://www.rura.rw/).

To understand this number, two things must be noted: Many Rwandans use more than one SIM card simultaneously in order to benefit from lower on-network calling rates, implying that the penetration rates defined above over estimates the fraction that actually uses cell phones in Rwanda.

Figure 1. Mobile penetration by telecom operator in Rwanda.



Source: http://www.rura.rw/

This is referred to as multi-SIMming. On the other hand, the proportion of children below the age of 15 in Rwanda is 42.1% (UN 2014). Assuming that children under 15 do not have their own SIM card, then adult adjusted penetration rate can be inferred to reach 132% nationwide. I could not find the average number of SIM cards per mobile user for Rwanda, but this figure has been estimated at 1.96 for Africa (GSMA (2012)). In the absence of multi-SIMming figures specific to Rwanda, the best estimate of mult-SIMming adjusted to adult mobile phone ownership is 132%/1.96, or 67%. [8]Source: (http://www.rura.rw/)

This multi-SIMming adjusted adult mobile ownership of above 50% opens the door for remarkable data gathering. Cell phone metadata such as call detail records (CDR), which include call location, length, call recipient and emitter, are usually stored by mobile phone operators for billing and business intelligence purposes. They can be used to locate individuals and populations, identify their interactions and social network, their mobile phone credit patterns, etc. The possible applications of such data are countless. Just to name a few: migrations flows can be identified, commuting times and traffic can be measured, communities can be identified, and income can be inferred.

6. METHODOLOGY

To understand the relationship between cell phone usage and census variables, analysis of variables is carried out.

The two datasets; call detail records from telecom operator in Rwanda and demographic or social and economic dataset from National statistical institute of Rwanda form dependent and independent variables respectively.

Table2. Variables

Cell phone	**Census**
Consumption	Education
Social network	Demographic
Mobility	Property Ownership

6.1. DATA DESCRIPTION

Call Detail Records

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.

The location of the subscriber within the coverage area is not known. From all the information contained in a 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 one year period (May2008-May2009) Call detail records (CDRs) dataset of 1.5 million subscribers from a mobile operator in Rwanda available through the inter-university Consortium for Political and social Research (ICPSR) www.icpsr.umich.edu, was collected.

Cell phone networks are built using a set of base transceiver stations (BTS) that are responsible for communicating cell phone devices within the network. 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 diagrams [14].

To be able to model cell phone usage from the call detail records, three sets of variables were computed per subscriber: ie; consumption ($\vec{B}$), social network ($\vec{S}$) and mobility ($\vec{M}$).

The consumption variables represent the general cell phone use statistics of a person, measuring, among others, the number of input or output calls, the duration of the calls. The social network variables compute measurements relative to the social network that subscribers build when communicating with others. They approximate the number of people a person typically calls to or receives calls from (i.e, input and output degree of the social network), the social distance between contacts (diameter of the social network) or the strength of the communication ties (strong/frequent contacts versus weak ones) and the mobility variables characterize the geographic areas where a person typically spends most of his/her work and leisure time as well as the spatio-temporal mobility patterns of individuals with the granularity of the area of coverage of a BTS.

Table 3: Description of call detail records (CDR) dataset with independent variables.

CDR	**Variables**	**Description of variables**
Caller number Receive number Date of call Time of call Duration of call BTS	Consumption	general phone usage
	Social net work	Number of people one calls or receives calls from, & duration
	Mobility	Geographical location and mobility patterns.

Census Data

In order to gather country wide demographic, socio and economic information for Rwanda, census data collected by the National Statistical Institute is used. The NSI carries out individual and household surveys at a national level every ten years. These surveys employ a large number of census takers that are responsible for interviewing every household head within their assigned geographical area.

Paper survey forms are still very common in census data collection which makes the collection process even more expensive and time consuming. During the census process, trained staffs are involved and the area is divided into geographical units (GUs).

For the analysis of this paper three groups of dependent variables ($\vec{C}$) from the census data are considered 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, property ownership is used as a proxy of the purchasing power of a person, measuring parameters like the existence of electricity, water, mobile phone or TV in the household.

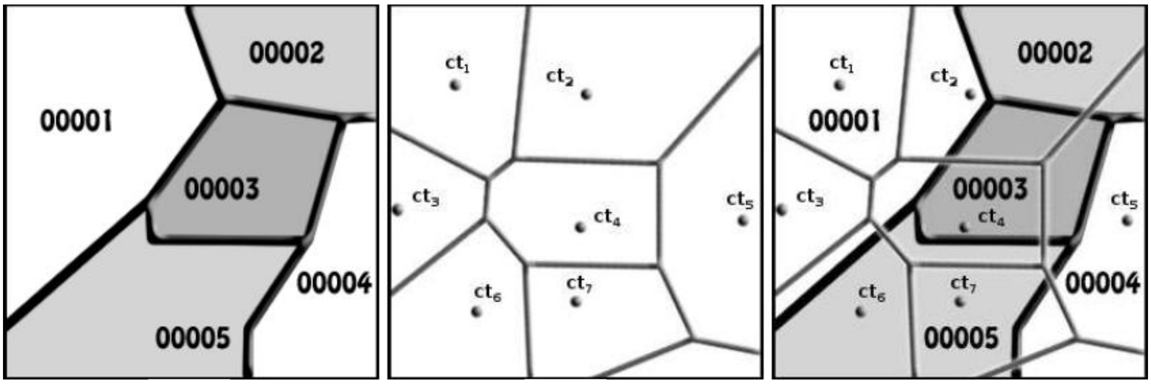
Table 4: Represents the list of census variables per household defined by the Rwanda national institute of statistics. [10]

Variable Type	Description
Education	% of Population with primary Education
	% of Female population with primary Education
	% of Male population with primary Education
	% of Population with Secondary Education
	% of Female population with Secondary Education
	% of Male population with Secondary Education
	% of illiterate population
	% of Female illiterate population
	% of Male illiterate population
Demographics	% of Female population
	% of Male population
	% of Young population (< 15)
	% of Middle age population (15-59)
	% of Senior age population (60+)
Property Ownership	% of Houses with cemented floor
	% of Houses with metal sheets roof
	% of Houses with oven fired bricks
	% of Houses with Electricity
	% of Houses with Water
	% of Houses with TV
	% of Houses with Mobile phone

Combining the two data sets

Figure 2. Shows an example of a census map and cell phone usage map for the area under study. Figure 2 (a) represents the (GUs) that the area is divided into to carry out the census surveys. Each (GU) will is associated to a set of variables (Education, demographic and property) that represent the average value of the population. Figure 2 (b) shows the BTS coverage map of the same area. Each BTS will be associated to a set of variables (consumption, social and mobility variables) averaged across all cell phone users whose residential location lies within the BTS coverage area.

Figure 2 (c) represents the merging between the two maps and as shown above GUs and BTS coverage areas may not necessarily match, so carrying out this step residential detection (scanline) algorithm is used which associates to each BTS area a set of Geographical units whose area are partially or totally included in the geographical area enclosed by each Voronoi polygon [15]. With the Scan line algorithm, the BTSs is represented as $BTSi = s * GUa + v * GUb + \ldots\ldots + w * GUd$ where ***s, v,….w*** represents the fraction of geographical units ***GUa, GUb, ... GUd*** that cover ***BTSi.*** This process is repeated for each census variable and BTS and the final result associates to each BTS a set of cell phone usage and census variable representing average values over the population that lives under the BTS coverage area.



(a) (b) (c)

Figure 2 Merging cell phone use maps with census maps: (a) Geographical units (GUs) with census data for the area under consideration; (b) Distribution of BTSs in the same area (Voronoi diagrams); (c) Overlapping of the Voronoi Diagrams with the geographical units GUs.

6.2. STATAISTICAL ANALYSIS OF VARIABLES

To understand whether there exists statistically significant difference between cell phone usage variables $(\vec{B},\vec{S},\vec{M})$ and census variables $(\vec{C})$, across different social groups, ANOVA tests are carried out.

The range of each of the census variable $(\vec{C})$ specified is divided into four quartiles *qi*(*q1...q4*) that are used to represent the social groups in the statistical test where each *qi* represents a social group in the population associated to low, medium-low, medium-high or high values for a specific census variable.

The quartiles for each census variable are computed by dividing the range between the minimum and the maximum percentage for that variable into four different subsets eg.

$$q1 = (\min, \min + \frac{\max - \min}{4}), \ldots\ldots\ldots.q_4 \qquad 1$$

Computation is then carried out to tell if there exist statistically significant differences between cell phone usage variable across four social groups.

To model the relationship between a single dependent variable Y (Census) and independent variables X1,….,Xp, (CDRs') **,** multivariate linear regression with ordinary least squares is applied; where p = 1 is called simple regression and p > 1 is multivariate regression.

$$Y = \beta o + \beta 1X1 + \beta 2X2 + \ldots \beta pXp + \varepsilon \qquad (1)$$

βi, i = 0,1,2,3 are unknown parameters, βo is the intercept term and ε is the error in representation.

It is important to note that the spatial component for this study is the district in Rwanda which is the second largest administrative level after the province. The district was selected because the available census data in Rwanda are district averages and therefore this test only reveals that there exist significant differences in the mean between the distributions of one or more groups.

This research is twofold, first part is about statistical evaluation of the relationship between cell phone usage and demographic or socio economic factors where ANOVA tests and Pearson's correlations are carried out to find out variables with statistical significance across the social groups. For this paper only one cell phone variable Consumption is considered using anonymized call detail records

CONSUMPTION VARIABLE

Figure3. Pairs graph showing relation between consumption and census variables

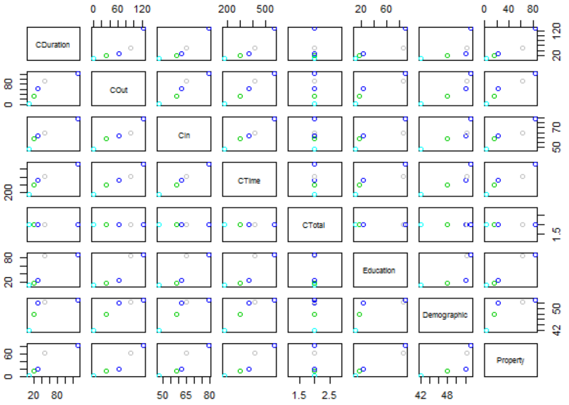


Table 3: Census variables and consumption variables. Numerical values represent ANOVA p-values; (*) represents moderate correlation, (**) strong correlation, (***) perfect correlation; and (+) and (-) represent a positive or a negative correlation between each two distributions.

	Calls				
Census Variable	CDuration	COut	CIn	CTime	CTotal
Intercept	0.326	0.186	0.1082	0.0916 .	8.65e-16 ***
Education	0.15	0.668	0.0460 *	0.0404 *	0.863
Demographic	0.335	0.172	0.0833 .	0.0455 *	0.319
Property ownership	0.101	0.245	0.0332 *	0.0268 *	0.984
Residual standard error	9.013 on 1DF	6.619 on 1DF	0.4912 on 1DF	4.184 on 1DF	3.381e-16 on 1DF
Multiple R-squared	0.992, Adjusted R-sq: 0.9678	0.9953, Adjusted R-sq: 0.9812	0.9995, Adjusted R-sq: 0.9981	0.9998, Adjusted R-sq:0.9992	0.3014, Adjusted R-sq:-1.795
F-statistic	41.08 on 3 and 1 DF	70.68 on 3 and 1 DF	1594 on 3 and 1 DF	1594 on 3 and 1 DF	0.1438 on 3 and 1 DF
P-value	p-value: 0.1141	p-value: 0.08716	p-value: 0.01841	p-value: 0.01841	p-value: 0.9222

The findings show that there exist statistically significant differences between time of the call and education level, the incoming calls were also significantly different depending on the education level, we also see significant difference between time of the call and demographics. The incoming calls and time of call also show statistical significance across property ownership. However, no significant difference is observed for the duration of the call and output calls. The correlation results show moderate positive correlations for incoming calls and education level as well as property ownership and the same is seen between time of the call, education, demographics and property ownership. All variables were computed as weekly averages for subscribers living in the same area.

Going forward the same analysis will be done for two more cell phone variables ie: social network and mobility to find variables that reveal statistical significance with respect to socio and economic information.

For social network, reciprocity of the communication and the physical distance between contacts will be considered. Reciprocity (R), will refer to the number of reciprocal voice calls between a person and her contacts. Three possible values will be evaluated: at least one, at least two and at least five reciprocal communications.

The physical distance which refers to the average distance between a person's residential area and the residential area of her cell phone contacts will also be evaluated to determine statistical significance.

For mobility variables, number of different BTSs visited by a person, distance travelled by a person will be computed as the distance between each pair of consecutively visited BTSs, the radius of gyration will be computed as the weighted average of the number of visits to each BTSs for a given person and the diameter will also be obtained as the maximum distance between the BTSs typically visited by a person.

6.3. CONCLUSION AND FUTURE WORK

The wide spread of mobile phone technology in developing countries provides large opportunity for data collection, briefly we highlighted different ways this data can be of advantage to organizations, policy makers and researchers. We go ahead to show that there exist statistical significance between mobile phone consumption variable and census variables.

Going forward will find out if there exist statistical significance between the social network and mobility variable and those variables that show statistical significance will be used to develop a predictive model that will allow the prediction of a variety of socio and economic variables from call detail records. This predictive model would be used as an alternative to the already existing tools that are expensive and time consuming especially for developing countries.

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. Changess 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.