A Method for Estimating the Relationship Between Phone Use and Wealth

Joshua Blumenstock
U.C. Berkeley School of Information
Berkeley, CA 94720
jblumenstock@berkeley.edu

Ye Shen
U.C. Berkeley Dept. of Statistics
Berkeley, CA 94720
jackshenye@gmail.com

Nathan Eagle
The Santa Fe Institute
Santa Fe, NM 87501
nathan@mit.edu

***Abstract*—We present a novel methodology for exploring the relationship between wealth and mobile phone use in developing countries. Using data from Rwanda, we show how the methodology can be used to predict the wealth of an individual using only information from that individual's call records. The approach uses mixed methods and three distinct sources of data: anonymous call records obtained from the phone company; large household Living Standards and Measurement Surveys conducted by the government; and a short phone survey administered by the first author. The results presented in this paper are preliminary and intended primarily to encourage discussion and feedback. We therefore pay particular attention to the limitations of the current approach, and to possible improvements and directions for future work.**

***Index Terms*—ICTD, mobile phones, predicted expenditures, economic status, phone survey, Rwanda**

I. Introduction

> "Not everything that can be counted counts, and not everything that counts can be counted."
>
> *–Albert Einstein*

For many people working in developing countries, it is of critical importance to have an accurate means of assessing the economic status of individuals in a population. Economic status may not be the sole determinant of a person's well-being, but it provides a useful indication of the underlying living conditions and quality of life. Pragmatically, a better understanding of the distribution of wealth – and the distribution of poverty – helps policymakers design effective policy, helps researchers design and evaluate interventions, and helps businesses meet the needs of their target population.

However, the measurement of economic status is notoriously difficult, particularly in developing countries where a large share of economic activity takes place in the informal sector. Even in industrialized economies, standard income-based surveys can overstate the importance of short-term fluctuations in income [1]. Such problems are exacerbated in developing countries, where income data are often unreliable and incomplete. The most common alternative to income-based surveys are consumption-based surveys, which measure expenditure flows over a period of time. Consumption-based surveys provide a more stable indication of permanent income, but they are not without their own limitations. Most notably, consumption surveys are extremely time- and resource-intensive. For instance, [2] notes that a typical consumption-based survey takes many hours, and at least five times as long as its corresponding income-based counterpart.

In this discussion paper, we describe a method for predicting the economic status of an individual based on his history of mobile phone calls. Specifically, we show how three different data sets - government survey data, semi-structured phone interviews, and anonymous call detail records - can be used to construct a model relating annual expenditures to call histories. In principle, this model could then be used to predict the annual expenditures of a mobile phone user given only anonymous phone usage data.

The primary contribution of this paper is methodological, as our intent is to provide a systematic

discussion of the steps necessary to analyze the relationship between mobile phone use and wealth in the typical setting where no single dataset contains both types of information. Thus, while we demonstrate the feasibility of the method using data from Rwanda, we only superficially analyze the Rwandan results, and leave the actual *prediction* of economic status to future work.

II. Related Work

Though a vast literature debates the most appropriate metrics for measuring human welfare [3][4], and a similarly rich and nuanced body of work discusses the theoretical challenges of measuring income and consumption [2][5], we restrict our current focus to empirical work on estimating individual and household expenditures.

In particular, we review a few related methods that have been used to predict individual or household consumption using proxy indicators. This is a fairly common topic in a variety of social sciences, as researchers are often interested in measuring the economic outlook of a population but for practical reasons are unable to administer time-consuming consumption-based surveys.

In the empirical literature, methods range from taking a small number of assets as a proxy for economic status to more complex proxies that involve a composite of a large number of household assets and characteristics. As an example of the former, Muhuri use an indicator for whether a household owns at least one of five assets; as an example of the latter, [6] use principal component analysis to develop a linear index that is the weighted sum of a large number of indicators of household asset ownership. The actual method can get arbitrarily complex: [7] use a layered probit model to estimate the magnitude of other unobserved factors that help determine permanent income; [8] use stepwise regression with forward and backward selection to select proxy variables. Though each of these methods for creating a proxy for wealth has advantages and limitations, in practice the resultant metrics are often highly correlated [9][6][10]. Thus, in our analysis, we opt for a relatively simple and intuitive approach that creates a composite index using weights determined by linear regression.

We are not aware of any prior attempt to use an individual's calling history to predict his economic status. However, in recent work [11] shows that large disparities in phone usage exist between rich and poor users in Rwanda, providing suggestive evidence that call records could be used to predict wealth. There is also a growing body of research that relates mobile phone data to other individual characteristics. [12] uses Rwandan mobile phone records to try and predict the gender of the phone owner, but finds that predictive accuracy is much lower than might be expected given the social norms governing phone use. [13] similarly attempts to predict the gender of phone users in developing countries, and finds that roughly 80% accuracy could be achieved, though only when a large number of users are left without predictions.

III. Data

The later analysis relies on three different sources of data that are described in greater detail in [11]. The datasets are (i) a household-level demographic survey conducted by the Rwandan government; (ii) a phone survey of a representative sample of Rwandan mobile phone users; and (iii) a log of all phone activity by those individuals in the period from January 2005 to December 2008.

A. Rwanda Demographic and Health Survey (DHS)

We use a standard Demographic and Health Survey (DHS) conducted by the Rwandan government to explore the relationship between consumption and asset ownership. This survey was conducted in 2005 by the Rwandan government on a large, representative set of 10,272 households. The survey contains roughly five hundred questions typical of Living Standard and Measurement Surveys, with detailed modules on demographic composition and socioeconomic status [14]. Most relevant to the current analysis, roughly seventy questions were asked about asset ownership and household expenditures, which makes it possible to estimate each household's annual expenditures in a manner following [2].

B. Phone survey

In Summer 2009, the first author coordinated a phone survey of a geographically stratified group of Rwandan mobile phone users. Using a trained group of enumerators from the Kigali Institute of Science and Technology (KIST), a short, structured interview was administered to roughly 900 individuals. In addition to querying basic demographic information, the phone survey collected responses for a small subset of the DHS questions (described above) about household asset ownership and housing characteristics. In Summer 2010, a follow-up survey was conducted with the 2009 respondents. Of the original 901 respondents from 2009, 682 were contacted in 2010. An additional 1300 respondents were contacted in 2010, to bring the total number of unique individuals contacted to roughly 2,200.

C. Phone company records

Finally, for each of the users contacted in the phone survey, we obtained from the phone company an exhaustive log of all phone-based activity that occurred from the beginning of 2005 through mid-2009. Thus, for every phone call made or received by one of the survey respondents, we know the time and date of the call, as well as the proximate location (based on the cell towers through which the call was routed) of both the caller and the receiver. From these call records, we can infer a wealth of information about mobile phone usage, including the phone activation date; the total days of activity, the number of incoming minus outgoing calls, the degree of the individual (the number of unique contacts), the amount of money spent on airtime, etc. These, and other metrics contained in the CDR, are described more thoroughly in [11].

IV. Methods

The ultimate goal of this research is to develop a method for predicting the income or expenditures of an individual, using only the information contained in the call history of that individual. If there existed a large sample of users for whom we had both income information and call history information, this would present a canonical problem that could be addressed using a variety of well-established methods for prediction and classification [15]. However, in the current setting, and in most settings encountered in the real world, that ideal data set does not exist. This is because most phone companies, which have access to the call history information, do not have access to reliable information about the income or expenditures of their customers. In some cases, the phone company will have access to basic demographic or socioeconomic information, and such data can be used to generalize from a small sample to the larger population of phone users [12][16]. However, in most instances the phone company collects no economic data at the individual level, and so there is no mechanical way to associate call records with income or expenditure information.[1]

Thus, the focus of this paper is to describe a method that can be used to create a single dataset that links individual CDR to individual expenditures. This can be accomplished in three steps:

1) First, we model the relationship between assets and expenditures using data in the government-collected Demographic and Health Survey. This enables us to infer the approximate annual expenditures of a household given information about the assets owned by that household.
2) Second, we conduct a phone-survey with a subset of mobile phone users to collect information on asset ownership. Given knowledge of the assets owned by these phone users, it is then possible to predict their annual expenditures using the model developed in the previous step.
3) Finally, we obtain CDR for the individuals in the phone survey, creating a single dataset that links call histories to predicted annual expenditures. This linked dataset can then be used to model the relationship between phone use and economic status.

Each of these steps is discussed in turn in the subsections that follow.

[1]This is the norm in most developing countries, where the vast majority of cell phone contracts are prepaid and SIM cards can be bought without identification for less than a dollar.

A. Modeling the relationship between assets and expenditures

Given information on assets and housing characteristics, we seek to develop a scalar measure of economic status based on the "basket of goods" owned by the individual. We do this using data from the government Demographic and Health Survey (DHS), which contains detailed information on each household's assets, characteristics, and expenditures. Our general strategy is to create a model that maps the assets and characteristics $(X_i^1, ..., X_i^N)$ of household i to the same household's expenditures Y_i using a flexible function $f()$:

$$Y_i = f(X_i^1, ..., X_i^N) \tag{1}$$

A variety of methods exist for parameterizing $f()$ – refer to the discussion in section II for a few examples. We opt for a parsimonious approach similar to [6] and [10], which models expenditures as a weighted combination of owned assets:

$$Y_{id} = \alpha + \sum_a \beta^a X_i^a + \mu_d + \epsilon_{id} \tag{2}$$

In Equation 2, expenditures Y_{id} of household i in district d are modeled as a linear combination of the assets and characteristics X^a of i, where the weights β^a reflecting each asset's relative contribution to total expenditures. We allow for district-specific intercepts μ_d.[2] To reduce the potential bias of outliers, we remove outliers with abnormally large studentized residuals, following a standard process described in [17].[3]

B. Predicting the expenditures of phone survey respondents

After estimating Equation 2 on the DHS data, we obtain a vector of coefficients $\widehat{\beta^a}$ that can be used to predict total expenditures given knowledge of assets and housing characteristics X^a. Thus, for any individual in Rwanda, we could in principle predict that individual's annual expenditures, denoted by $\widehat{Y_{id}}$, by asking that individual a small number of questions about his household. In practice, there are some outlandish assumptions that must be made to justify this approximation, but we will defer discussion of these and other limitations to section VI. This is precisely the technique we employ to infer the annual expenditures of a random sample of mobile phone users. Through phone-based surveys (described in section III-B), we collect information on the assets and housing characteristics (i.e., the X^a) which maximize the predictive power of Equation 2.

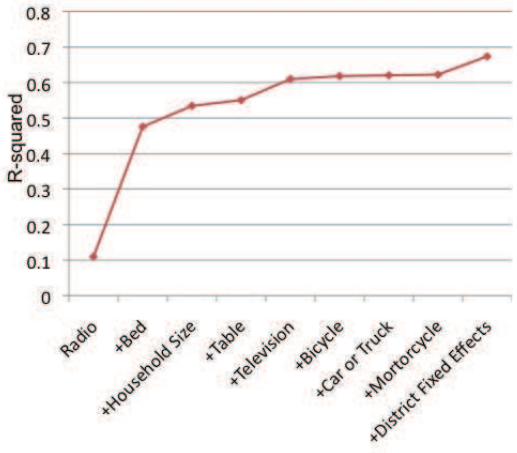

Fig. 1. Model Selection: How additional covariates affect R^2

However, there is a tradeoff that must be made in choosing the appropriate questions to ask. On the one hand, the more information gathered about the individual, the better the fit of Equation 2 will be. On the other, asking a greater number of questions takes time and money, and reduces the number of unique respondents who can be contacted given limited resources. In Figure 1, we graphically represent the added benefit of each additional question asked. We construct the figure by first running a bivariate regression of expenditures on household radio ownership among the households in the DHS government survey. We then re-estimate Equation 2 eight additional times, adding one additional

[2]Though ordinary least squares specifications are often used to model causal relationships, under no circumstances do we mean to imply that Equation 2 will recover causal effects. Our intent is rather to identify the multivariate correlations between asset ownership and expenditures.

[3]Our results change very little if we use an alternate technique for removing outliers, such as removing the top 1% or 5% of extreme values.

covariate at each iteration. On the y-axis, we plot the coefficient of determination (the R^2) from each model. As can be seen in the figure, the amount of variation explained by the model increases rapidly with the first few covariates, then shows diminishing returns after four or five X^a are included.

C. Relating call histories to predicted expenditures

Using the above technique, it is possible to obtain the *predicted expenditures* $\widehat{Y_{id}}$ for each of the individuals contacted in the phone survey. This gives us a total of roughly 2,000 individuals for whom we have a measure of predicted expenditures and detailed call history information (obtained from the mobile operator). For these individuals, it is then possible to directly evaluate the relationship between phone use and economic status:

$$\widehat{Y_{id}} = g(CDR_i) \tag{3}$$

Finding the optimal form of $g()$ is an important research topic, but is not the focus of the current paper. However, to provide some intuition on the relationship between phone use and wealth, we will later estimate a simple multivariate regression of predicted expenditures on a large number of aggregate statistics of mobile phone use, with Equation 3 parameterized in a multivariate regression.

V. Results

A. Predicting expenditures from assets and household characteristics

As can be seen in Figure 1, even a relatively simple model that accounts for only three household characteristics – the number of radios, the number of beds in the household, and the household size – explains over 50% of the variation in annual household expenditures. Adding another five covariates increases the R^2 to 0.623, and including fixed effects for each of the thirty geographic districts produces a final R^2 of 0.674. This is not to say that the ordinary least squares specification is the "correct" model. However, the high R^2 indicates that despite these shortcomings it is possible to infer a great deal about an individual's expenditures using the simple linear regression model of Equation 2.[4]

TABLE I
Regression of Expenditures on Asset Ownership

Outcome	log(Expenditures)		Expenditures	
	β^a	(S.E.)	β^a	(S.E.)
Radio	0.18	(0.02)	40090	(13007)
Television	1.14	(0.01)	2130434	(44048)
Bed	0.24	(0.04)	187061	(8266)
Table	0.13	(0.01)	57601	(9109)
Car/Truck	0.24	(0.01)	1695284	(57718)
Motorcycle	0.65	(0.04)	8229976	(197091)
Bicycle	0.22	(0.11)	138186	(20359)
HH Size	0.09	(0.02)	56168	(3198)
R^2	0.62		0.75	
RMSE	0.55		470000	
N	6900		6900	

Notes: Standard errors reported in parentheses.

Table I gives the coefficients that result from estimating Equation 2 on the DHS data. The first column uses the log of total annual household expenditures as the outcome; the second column takes the raw value. It is evident that annual expenditures are heavily correlated with asset ownership. For instance, Rwandan households with cars spend, on average and after accounting for other assets, roughly 1.7 million Rwandan Francs (USD$3,000) more per year than households without cars.

B. Predicting the expenditures of phone users

Using the estimated coefficients of Table I, we predict the expenditures of all phone survey respondents using assets and household characteristics collected over the phone. Figure 2 presents a kernel density estimation of the distribution of predicted expenditures for phone survey respondents in 2009 (blue line), and again for the same respondents in 2010 (red line). Over the one-year period, there was little change in reported asset ownership; the corresponding distributions are therefore quite similar

[4]In our preferred specification, we include district fixed effects, and allow for households to own more than one of each asset. However, the resultant R^2 does not change by much if we omit the district fixed effects, if we include dummy binary variables to indicate whether or not the household owns more than one of each asset, or if we take the logarithm of expenditures as the outcome.

and a t-test does not reject the null that hypothesis that the means are the same (p=0.254).

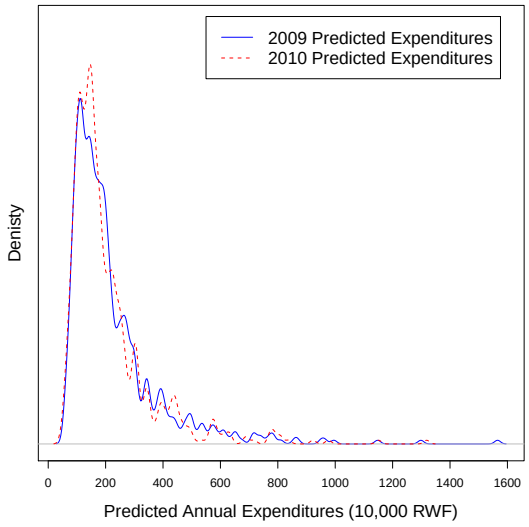

Fig. 2. Changes in predicted expenditures over time

TABLE II
REGRESSION OF PREDICTED EXPENDITURES ON PHONE USE

	Coefficient	(S. E.)
Duration (outgoing)	-0.75	(2.55)
Duration (incoming)	7.66***	(2.01)
Degree	-1038.70*	(439.98)
Int'l duration (out)	-17.60	(10.07)
Int'l duration (in)	-10.63	(7.78)
Int'l degree	10534.70***	(3883.26)
Districts called	129304.11**	(42014.00)
Districts received	-103121.07**	(36356.38)
Unique towers	2918.14	(3600.54)
Months	11916.91	(9027.69)
Avg. recharge denomination	602.87	(461.13)
Daily recharge	716.18	(794.48)
N	671	
R^2	0.21	

Outcome is predicted expenditures $\widehat{Y_{id}}$, in RWF. Standard errors in parentheses. Regression includes district fixed effects. * significant at $p < .05$; $^{**}p < .01$;$^{***}p < .001$.

C. Relating predicted expenditures to call histories

Given a measure of predicted expenditures for each phone survey respondent, we estimate Equation 3 using ordinary least squares to probe the relationship between phone use and economic status. Table II presents the results from regressing predicted expenditures on twelve different metrics of phone use. The metrics we present, defined in greater detail in [11], represent only a sliver of the hundreds of such metrics that can be computed from the call detail records. In future work, we intend to more thoroughly compare the relationship between predicted expenditures and different measures of phone use. Here, we point to a few of the most salient results.

First, the R^2 of 0.21 is indicative of a strong relationship between expenditures and phone use, but the relationship is not nearly as strong as the previously-assessed relationship between asset ownership and expenditures. This is not surprising, as we would expect that households would purchase assets in proportion to the amount of annual expenditures. With phone use, we would expect greater variance in patterns of use. Though our ex ante expectation was that richer individuals would use the phone more in aggregate individual usage patterns are evidently quite dependent on individual circumstances.

Second, there are a number of statistically significant correlations between predicted expenditures and phone use. For instance, while there is little relationship between expenditures and the amount of time spent on outgoing calls, the evidence indicates that individual households are likely to spend 8.57 additional Francs per year for every additional second spent on *incoming* calls. There are also patterns in the relationship between a person's social network and her predicted expenditures. Namely, the number of unique international contacts ("international degree") has a strong positive relationship with expenditures, while the number of unique Rwandan contacts exhibits a weakly negative association. However, we do not want to exaggerate the importance of these findings, since many of the metrics are highly collinear, and the estimated coefficients depend greatly on the other covariates included in the regression. We hope to identify the more robust relationships in future work.

Finally, we note that with this dataset linking

phone use to expenditures, it is mechanically quite simple to *predict* economic status based only on anonymous phone use data. A variety of different prediction and classification algorithms could be used in this endeavor. While prior work would indicate that this may not be a trivial task [12], we believe that for at least a subset of users it should be possible to make relatively accurate predictions [13]. We are particularly interested in examining which features, such as those presented in Table II, have the greatest impact on predictive accuracy.

VI. Limitations of our approach

The methodology described above provides a relatively straightforward method for analyzing the relationship between an individual's wealth and her use of the mobile phone. Before concluding, we discuss a few of the more problematic assumptions that we have made along the way, and their implications for future research.

Two datasets, one model: Perhaps the most troubling assumption in the above method is that the relationship between assets and expenditures identified with the function $f()$ in the 2005 DHS data will remain constant when applied to phone survey data collected in 2009 and 2010. This assumption is unjustified for at least two distinct reasons. First, the data for the two populations was collected using very different methodologies, and respondents may respond differently to questions about assets depending on whether they are asked in person or over the phone. Second, the data was collected in different years, and it is possible that the relationship between assets and expenditures would evolve over such a long interval. For instance, the strong relationship observed in 2005 between television ownership and wealth may be weaker in 2010, as electricity becomes more available and used televisions saturate the market.

While this assumption indeed limits the usefulness of our approach, we can provide suggestive evidence that a model trained on 2005 DHS data is still relevant to 2010 phone survey data. We do this by means of a "placebo test," where instead of predicting the unknown expenditures of the 2010 respondents, we instead predict a known (but hidden) item, such as television ownership. Thus, we replicate the methods described in sections IV-A and IV-B, with television ownership as the left-hand side variable Y_i. Using a probit model, we train Equation 2 on the 2005 DHS data. We then apply the learned coefficients to the 2010 phone survey data, obtaining a measure of *predicted television ownership* for all phone survey respondents. Finally, we check to see whether the predicted television ownership matches actual television ownership. In the placebo specification, our predictions are correct for 75.2% of respondents. In predicting bicycle and bed ownership, the corresponding accuracy rates are 66.3% and 97.7%. The predictions are not perfect, but clearly the function $f()$ trained on 2005 data maintains reasonable validity when applied to the 2010 data.

Functional form assumptions: At a more superficial level, we were forced to make a number of functional form assumptions when estimating equations 1 and 3 with ordinary least squares. Certainly, there is no reason to expect that expenditures would increase linearly or log-linearly in relation to a household's assets and other characteristics. Similarly, the relationship between phone use and expenditures is almost certainly rife with nonlinearities. Thus, we believe considerable improvement could be made by further investigating the parameterization of $f()$ and $g()$.

Limitations of asset-based proxies for wealth: Also problematic is the possibility, discussed in the prior literature [5], that asset-based proxies for expenditures may provide biased estimates of the expenditures of certain types of individuals. For instance, if a strong correlation is found between television ownership and assets among the aggregate population, but a small subgroup of the population has a distaste for television, this simple method would systematically underestimate the expenditures of that subgroup.

VII. Conclusion

The preceding pages have described a new methodology that can be used to analyze the relationship between mobile phone use and economic status. Using data from Rwanda, we tested

the methodology and assessed its strengths, weaknesses, and overall validity. We further presented preliminary evidence of the empirical relationship between phone use and economic status in Rwanda. Finally, we described how the method can be used to produce a dataset that can be used to predict economic status using only mobile phone records. Given the difficulty usually involved in measuring economic status in developing countries, this simple and scalable alternative offers a promising area for future research.

ACKNOWLEDGMENTS

This research was funded in part by the Institute for Money, Technology, and Financial Inclusion, the National Science Foundation, and the International Growth Centre. Paige Dunn-Rankin provided excellent research assistance.

REFERENCES

[1] M. Friedman, *Theory of the Consumption Function*. Princeton University Press, Jul. 2008.

[2] A. Deaton and S. Zaidi, *Guidelines for constructing consumption aggregates for welfare analysis*. World Bank Publications, 2002.

[3] A. Sen, *Development as freedom*. Oxford University Press, 1999.

[4] J. D. Sachs, "Investing in development: a practical plan to achieve the millenium development goals; overview," 2005.

[5] A. Deaton and J. Muellbauer, *Economics and consumer behavior*. Cambridge Univ Pr, 1980.

[6] D. Filmer and L. H. Pritchett, "Estimating wealth effects without expenditure dataor tears: an application to educational enrollments in states of india," *Demography*, vol. 38, no. 1, p. 115132, 2001.

[7] B. D. Ferguson, A. Tandon, E. Gakidou, and C. J. L. Murray, "Estimating permanent income using indicator variables," *Health systems performance assessment: Debates, methods, and empiricism. Geneva: World Health Organization*, p. 747760, 2003.

[8] S. Benin and J. Randriamamonjy, *Estimating household income to monitor and evaluate public investment programs in Sub-Saharan Africa*. Intl Food Policy Res Inst, 2008.

[9] K. A. Bollen, J. L. Glanville, and G. Stecklov, "Economic status proxies in studies of fertility in developing countries: Does the measure matter?" *Population Studies*, vol. 56, no. 1, p. 8196, 2002.

[10] M. R. Montgomery, M. Gragnolati, K. A. Burke, and E. Paredes, "Measuring living standards with proxy variables," *Demography*, vol. 37, no. 2, p. 155174, 2000.

[11] J. E. Blumenstock and N. Eagle, "Mobile divides: Gender, socioeconomic status, and mobile phone use in Rwanda," *4th International Conference on Information and Comunications Technologies and Development*, 2010.

[12] J. E. Blumenstock, D. Gillick, and N. Eagle, "Who's calling? Demographics of mobile phone use in Rwanda," *AAAI Symposium on Aritificial Intelligence and Development*, vol. Forthcoming, 2010.

[13] V. Frias-Martinez, E. Frias-Martinez, and N. Oliver, "A gender-centric analysis of calling behavior in a developing economy," *AAAI Symposium on Aritificial Intelligence and Development*, vol. Forthcoming, 2010.

[14] I. N. de la Statistique du Rwanda (INSR) and O. Macro, *Rwanda Demographic and Health Survey 2005*. Calverton, Maryland: INSR and ORC Macro, 2006.

[15] S. J. Russell and P. Norvig, *Artificial intelligence: a modern approach*. Prentice hall Englewood Cliffs, NJ, 1995.

[16] V. Frias-Martinez, J. Virseda, A. Rubio, and E. Frias-Martinez, "Towards large scale technology impact analyses: Automatic residential localization from mobile Phone-Call data," *4th International Conference on Information and Comunications Technologies and Development*, 2010.

[17] J. Fox, *Applied regression analysis, linear models, and related methods*. Sage Publications, Inc, 1997.