Calling for Better Measurement:

Estimating an Individual's Wealth and Well-Being from Mobile Phone Transaction Records

Joshua E. Blumenstock
University of Washington
Seattle, WA
joshblum@uw.edu

ABSTRACT

We provide evidence that mobile phone records can be used to predict the socioeconomic status and other welfare indicators of *individual* mobile phone subscribers. Combing several terabytes of anonymized transactional mobile phone records with data collected through 2,200 phone-based interviews, we test the extent to which it is possible to predict an individual's responses to survey questions based on phone records alone. We observe significant correlations between asset ownership and a rich set of measures derived from the phone data that capture phone use, social network structure, and mobility.

Simple classification methods are able to predict, with varying degrees of accuracy, whether the respondent owns assets such as radios and televisions, as well as fixed household characteristics such as access to plumbing and electricity. More modest results are obtained when attempting to predict a broader set of development indicators such as an individual's response to the question, "Have you had to pay unexpected medical bills in the past 12 months?" While these methods offer a powerful opportunity for policymakers and researchers working in developing countries, we argue that considerable calibration and refinement is needed before such methods can be deployed.

Keywords

Mobile phones, development, big data, call detail records, wealth, mobility, social networks, regression

1. INTRODUCTION AND MOTIVATION

Reliable, quantitative data is a critical input to development policy, social science research, and to the decision-making process of firms and organizations interested in promoting social good. However, the basic measurement of key development outcomes – such as poverty, physical security, and happiness – is notoriously difficult in developing countries, where a lack of physical infrastructure and resources is often compounded by market failures and fragile institutional capacity [10].

Such problems are exacerbated in fragile and conflict-affected regions, where concerns over corruption and the physical security of enumerators and respondents make the regular collection of representative household survey data all but impossible. For example, Angola's last census was in 1970, and covered just 18 districts [11]. As a result, researchers and policymakers typically rely on data from large-scale national surveys (which occur infrequently), or specialized panel survey modules (which are typically administered to small, local populations). Neither traditional source captures fine-grained variation in development outcomes over both space and time.

In this paper, we describe preliminary results from efforts to develop models for predicting an *individual's* socioeconomic status and related development outcomes based upon anonymous, high-frequency data passively registered through use of mobile phone networks. A key innovation of this approach is our ability to link individual survey responses collected in phone interviews with incredibly rich social network and communication data obtained from mobile phone operators. Such an approach can be used to model the relationship between passively collected metrics of mobile phone use and explicitly queried socioeconomic phenomena. For instance, it will be possible to tell whether an individual's communication history can be used to predict whether that individual agrees with a survey-based statement such as, "*I believe the current economic situation will improve in the coming year*," or "*I feel connected to other members of my local community.*"

Here, we focus on results from the analysis of data collected in Rwanda in 2009 and 2010. This work extends a previous workshop paper that used a simple regression model to illustrate the strong relationships between simple metrics of phone use and a composite indicator of socio-economic status [1]. To our knowledge, no other prior work has investigated the relationship between individual communication histories and individual development outcomes. However, a series of recent studies have shown that geographically-aggregated communication records are strong predictors of regional census data [5, 8, 9]. A closely related set of work uses individual phone records to model gender and related (fixed) demographic characteristics [4, 7]. These approaches are strongly complementary, and we expect that over the next several years these methods will significantly advance our ability to measure, model, understand, and improve the lives of historically marginalized populations.

KDD-Data Science for Social Good 2014 New York, NY USA

Sample Y_{it} Indicator ("development outcome")	Sample X_{it} Indicator ("feature")
Household owns a motor vehicle	Average number of outgoing phone calls per day
Amount of land owned by individual	Number of unique contacts in social network
Recent illness or other negative economic shock	Number of geographic regions visited in past month
Total expenditures in last month	Total expenditures on mobile phone-based communication
Value of recent agricultural harvest	Eigenvector centrality of respondent
Financial outlook on 7-point Likert scale	Percentage of closed triangles in social network

Table 1: Sample Y_{it} development status indicators to be modeled as a function of sample X_{it} features

2. METHOD OVERVIEW

Our practical goal is to predict different aspects of an individual's wealth and well-being using the high-frequency spatiotemporal transaction logs generated by that same individual's use of a mobile phone. To provide a few concrete examples, Column 1 of Table 1 lists several such livelihood indicators; Column 2 lists examples of the sort of features that can be extracted from communication logs. These lists are far from exhaustive, and simply represent the sort of outcomes that would be useful to policymakers, and the sort of metrics that can be derived from transaction logs.

In principle, these data could come from different sources. For instance, a smartphone application could collect information on an individual's use of the phone and separately provide an interface for that individual to input demographic and economic information. In practice, for the purposes of this study we combine data collected in phone-based surveys in Rwanda with data obtained from the country's primary mobile phone network operator.

Formally, we assume that for each of N individuals indexed by i, we are able to collect a vector of socioeconomic and development outcomes Y_{it} for each of T periods indexed by t. We similarly assume that we can extract individual i's mobile phone history, and that this rich transaction record can be appropriately aggregated into a vector of metrics $X_{it} = \langle x_{i1t}, ..., x_{iKt} \rangle$ such as those contained in Table 1, column 2. For each Y_{it}, we can then build a model $Y_{it} = f(X_{it})$ that captures the relationship between the reported development indicator and the observed patterns of call activity.

A primary technical concern is the choice of an appropriate functional form for $f()$. Here, we discuss results that employ a simple regression benchmark estimator,

$$Y_{it} = \alpha + \sum_{ik}^{K} \beta_k x_{ikt} + \epsilon_i \tag{1}$$

where we assume we can derive K features indexed by k. We will also present results from a more flexible regression specification that includes regional fixed effects μ_d to account for unobserved heterogeneity at the regional level,

$$Y_{id} = \alpha + \sum_{\delta} \sum_{ik}^{K} \beta_{k\delta} x_{ikt}^{\delta} + \mu_d + \epsilon_{id} \tag{2}$$

Model (2) further allows for polynomial functions of each x_{it} of arbitrary degree δ; it is simple to see how additional flexibility could be added by allowing for interactions between different x_{it} and x_{jt} to allow for joint conditional relationships (e.g. if wealth is unusually high for individuals with both high outgoing calls and high degree, but not necessarily high for individuals with high outgoing calls *or* high degree).

More sophisticated supervised learning models (kernel-based methods, regression trees, ensemble approaches, etc.) would almost certainly increase the model's predictive power, but possibly at the expense of reduced interpretability, and increased difficulty of implementation by other researchers and policymakers. This is indeed an area of active ongoing work, and one that would benefit greatly from additional feedback. In section 5, we briefly discuss possible extensions, other approaches to modeling the data, and the strengths and weaknesses of the current approach.

3. CASE STUDY: RWANDA

We evaluate this method using survey data and mobile phone records from Rwanda. The surveys we use were collected via phone interviews with a geographically stratified group of mobile phone users in 2009 and 2010. 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. A focus of the survey was to collect basic information on household characteristics, assets, and expenditures, which previous research has shown to be highly correlated with wealth and socioeconomic status. In total, we contacted roughly 2,200 individuals through these surveys.

For each of the 2,200 phone survey respondents, we obtained from the phone company an exhaustive log of all phone-based activity that occurred from the beginning of 2005 through mid-2009.[1] Thus, for every phone call and text message in which each respondent was invovled, we know the time and date of the call, as well as the approximate geographic location of both the caller and the receiver (based on the cell towers through which the call was routed). From these call detail records (CDR), we derive hundreds of features of mobile phone usage such as those described in column 2 of Table 1, including: account activation date; total days of activity, number of incoming minus outgoing calls, degree of the individual (the number of unique contacts), amount of money spent on airtime, etc. These, and other metrics contained in the CDR, are described more thoroughly in [3]. Additional details fo the phone survey are discussed in [2].

[1] In accordance with our approved human subjects protocol, informed consent was obtained from each respondent at the beginning of the phone survey. Respondents were compensated roughly USD$1.50 for their participation in the study. All personally identifying data was removed after merging with phone records and prior to analysis.

	Accuracy	Recall	Precision	F	AUC	% Answered Yes
Panel A: Assets and Housing						
Owns a radio	0.976	1.000	0.976	0.988	0.899	0.973
Owns a bicycle	0.676	0.552	0.678	0.609	0.722	0.456
Household has electricity	0.819	0.533	0.761	0.627	0.828	0.285
Owns a television	0.855	0.497	0.738	0.594	0.814	0.214
Has indoor plumbing	0.887	0.250	0.842	0.386	0.843	0.142
Owns a motorcycle/scooter	0.899	0.011	1.000	0.022	0.772	0.102
Owns a car/truck	0.945	0.213	0.867	0.342	0.849	0.068
Owns a refrigerator	0.954	0.180	1.000	0.305	0.878	0.055
Has landline telephone	0.992	0.125	1.000	0.222	0.562	0.009
Panel B: Social Welfare Indicators						
Hospital bills in last 12 months	0.633	0.890	0.633	0.740	0.653	0.587
Very ill in last 12 months	0.686	0.188	0.550	0.280	0.671	0.325
Death in family in last 12 months	0.665	0.183	0.632	0.284	0.619	0.363
Flood or drought in last 12 months	0.788	0.086	0.607	0.151	0.706	0.219
Fired in last 12 months	0.901	0.022	1.000	0.043	0.731	0.101

Table 2: Model performance at predicting responses from survey respondents based on call records data

4. PRELIMINARY RESULTS

In ongoing work, we are conducting additional phone surveys to collect a broader range of development outcomes such as those listed in column 1 of Table 1. Here, we focus on the simplified task of predicting responses to relatively well-defined questions with concrete answers that were collected in the short interviews conducted in 2009 and 2010. Section 4.1 describes results from predicting asset ownership and housing characteristics; section 4.2 describes initial results at predicting more general measures of social welfare; and section 4.3 describes results from predicting a composite index of respondent wealth.

4.1 Predicting asset ownership and housing characteristics

In Panel A of Table 2, we present the results from the use of a logistic regression to predict binary responses to survey questions about fixed assets and housing characteristics such as, *"Does your household own one or more radios?"* or *"Does your household have electricity?"* We fit a version of model (2) with regional fixed effects and roughly twenty aggregated measures of phone activity such as those in column 2 of Table 1, including measures of phone use, SMS use, geographic mobility, and social network structure. The model is fit using 10-fold cross-validation on a sample of roughly 900 respondents who answered all survey questions, where the binary classification threshold is determined to maximize accuracy and the other performance metrics are reported at that threshold. Figure 1 shows the ROC curves for three representative questions asked in the survey.

In general, this rudimentary approach to modeling the relationship between phone use and asset ownership shows signs of modest success. For most of the outcome variables we seek to model, we can achieve relatively high accuracy, but these rates are only marginally higher than the naive baseline of predicting the majority class. For instance, the model accuracy of 85% in predicting television ownership is only an 8 percent (6 percentage points) increase over a model that predicts all respondents do not own televisions.

4.2 Predicting welfare indicators

Panel B of Table 2 presents similar results from our attempts to predict more subjective responses to broader development questions such as *“Has your household had to pay significant hospital bills in the past 12 months?”* and *“Have you lost your job in the last 12 months”*. Here, performance is lower than with the asset ownership questions; we find that our models are only able to offer marginal improvements over naive baseline predictions.

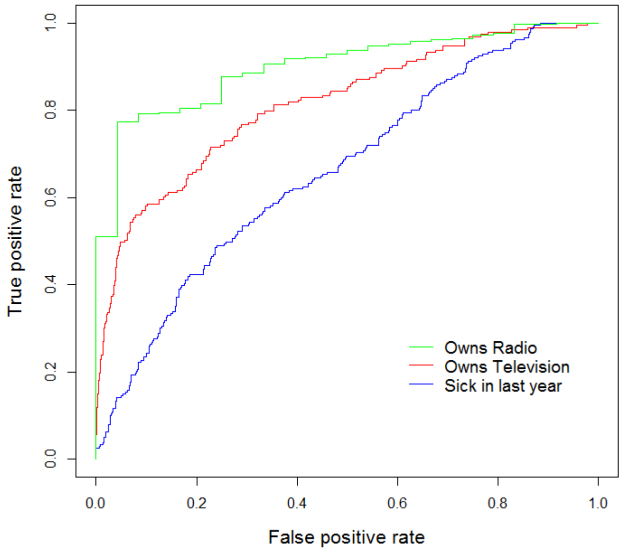

Figure 1: ROC curve for three survey outcomes

4.3 Predicting composite socioeconomic status

Finally, we test the ability of this approach to predict composite index of socioeconomic status. To create this aggregate metric from the survey responses, which we denote by $\widehat{Y_{id}}$, we take the first principal component of the 9 asset and housing characteristics listed in Panel A of Table 2.

The first principal component of wealth explains 27.24% of the variance of the 9 asset categories. Similar results obtain when creating a composite based on the first principal component of a much larger number of assets and housing characteristcs.[2]

In Table 3, we present the results from fitting an ordinary least squares regression of this first principal wealth component on a representative sample of mobile phone use metrics. While the explanatory power of this regression is rather limited ($R^2 = 0.29$), there are strong relationships between the wealth composite and several of the measures of phone use and network structure. Note that the sign and magnitude of each of the regression coefficients is highly dependent on the set of regressors included; because of the natural dependencies in the phone data, inclusion or exclusion of additional features substantively changes the estimated coefficients (though such tinkering has relatively little effect on the fit of the model).

Table 3: Regression of first principal component of assets on selected measures of phone use

	$Coefficient$	$(S.E.)$
Active days	-0.04	(0.03)
Calls per day	2.49	(2.28)
Outgoing calls	0.01	(0.01)
Incoming calls	$-0.01^{\dagger}$	(0.01)
Degree	0.08^{**}	(0.03)
Int'l outgoing calls	-0.59^{*}	(0.26)
Int'l incoming calls	-1.09	(0.72)
Int'l degree	0.38^{*}	(0.17)
Towers visited	-0.03	(0.21)
Avg. recharge denomination	0.01	(0.03)
Daily recharge	-0.27^{***}	(0.06)
Clustering	-505.80^{***}	(148.09)
Betweenness	137.06^{*}	(56.66)
N	897	
R^2	0.29	

Results show regression of first principal component of the wealth ($\widehat{Y_{id}}$), scaled by 100 to simplify presentation. Standard errors in parentheses. Regression includes district fixed effects but coefficients are omitted from table for clarity. † significant at $p < .10$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

To further illustrate the strong correlations between phone use and wealth, we perform a second principal component analysis on a large set of different metrics of mobile phone activity. In this case, the first principal component of 97 metrics of phone use explains 34.63% of the variance of the full dataset. In Figure 2, we plot for each of the survey respondents the first principal component of wealth (y-axis) against the first principal component derived from the phone use data (x-axis). The strong positive relationship between these two components is illustrated by the Nadaraya-Watson kernel regression shown in red.

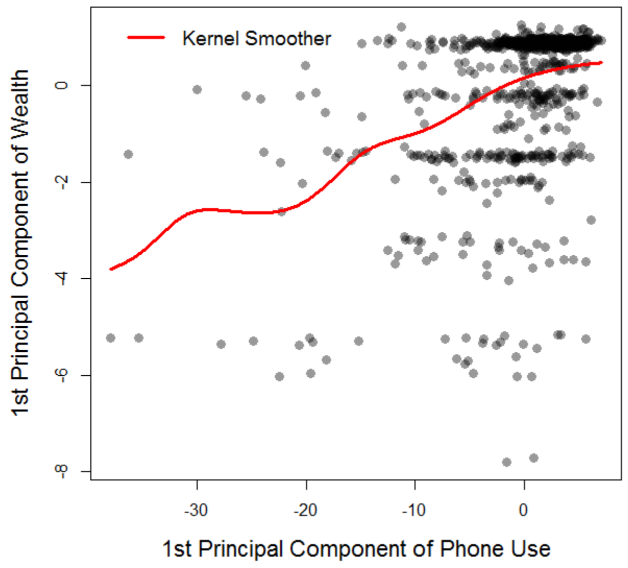

Figure 2: Relationship between 1st principal components of wealth and phone use

5. DISCUSSION AND CONCLUSION

We have presented preliminary evidence that it is possible to predict a variety of indicators of individual socioeconomic status and welfare using mobile phone call records. If these results can be further calibrated and improved upon, this technique could provide policymakers and researchers with a novel quantitative perspective on populations for whom good data has historically been hard to find. Compared to traditional methods for collecting individual and household data, the use of call records represents a considerably cheaper alternative, with dramatically higher spatial and temporal precision. In principle, such fine-grained development indicators could be applied in a variety of settings, from program monitoring and evaluation to social welfare targeting and analysis.

While provocative, we do not want to overstate the accuracy of the methods tested thus far, or imply that such techniques will ever supplant alternative modes of data collection. The predictions presented in this paper are relatively inaccurate, and the methods, models, and data leave considerable room for improvement. In ongoing work, we are working to develop improved statistical and computational models, and are collecting a large amount of survey data that will allow for better calibration and testing.

6. ACKNOWLEDGMENTS

The authors would like to thank Nathan Eagle, Daniel Gillick, and Ye (Jack) Shen for assistance procuring and analyzing the original datasets collected for this paper. Michael Callen, Tarek Ghani, and Jacob Shapiro provided invaluable input in conceptualizing and framing this work.

[2] In earlier work, we have taken a different approach that develops a composite index of "predicted expenditures" using publicly available household survey (DHS) data to approximate the estimated annualized household expenditures of survey respondent [1]. See [6] for a related approach to developing a composite wealth index from survey data.

7. REFERENCES

[1] J. Blumenstock, Y. Shen, and N. Eagle. A method for estimating the relationship between phone use and wealth. *QualMeetsQuant Workshop at the 4th International IEEE/ACM Conference on Information and Communication Technologies and Development*, 2010.

[2] J. E. Blumenstock and N. Eagle. Mobile divides: Gender, socioeconomic status, and mobile phone use in Rwanda. *4th International IEEE/ACM Conference on Information and Communications Technologies and Development*, Dec. 2010.

[3] J. E. Blumenstock and N. Eagle. Divided we call: Disparities in access and use of mobile phones in Rwanda. *Information Technology and International Development*, 8(2):1–16, 2012.

[4] 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*, 18:116–117, 2010.

[5] N. Eagle, M. Macy, and R. Claxton. Network diversity and economic development. *Science*, 328(5981):1029–1031, May 2010.

[6] D. Filmer and L. H. Pritchett. Estimating wealth effects without expenditure data - or tears: an application to educational enrollments in states of india. *Demography*, 38(1):115 – 132, 2001.

[7] 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*, Forthcoming, 2010.

[8] V. Frias-Martinez and J. Virseda. On the relationship between socio-economic factors and cell phone usage. In *Proceedings of the Fifth International Conference on Information and Communication Technologies and Development*, ICTD '12, pages 76–84, New York, NY, USA, 2012. ACM.

[9] T. Gutierrez, G. Krings, and V. D. Blondel. Evaluating socio-economic state of a country analyzing airtime credit and mobile phone datasets. *arXiv preprint arXiv:1309.4496*, 2013.

[10] M. Jerven. *Poor numbers: how we are misled by African development statistics and what to do about it.* Cornell University Press, 2013.

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