Estimating Economic Characteristics with Phone Data

By JOSHUA E BLUMENSTOCK*

Historically, economists have relied heavily on survey-based data collection to measure social and economic well-being. More recently, the proliferation of large-scale digital data has enabled new approaches to measurement. The use of satellite imagery is now commonplace in economics research (Donaldson and Storeygard, 2016), and related work indicates that regional patterns of phone and internet use correlate with regional measures of wealth and unemployment (Eagle, Macy and Claxton, 2010; Llorente et al., 2015). The general focus of such analysis has been to identify a functional mapping between a regionally-aggregated measure of economic activity (such as the average wealth of a village) and a regionally-aggregated source of passively-collected digital data (such as aerial photographs of the village, or traffic passing through nearby cell phone towers).

Here, we address the question of whether the "digital footprints" of an *individual* can be used to infer his or her socio-economic characteristics. This builds on recent work showing that it is possible to predict the wealth of an individual from his or her mobile phone records (Blumenstock, 2014), and that these phone-based predictions can be aggregated into accurate national statistics (Blumenstock, Cadamuro and On, 2015). We focus on assessing the generalizability of this approach, and show that the same basic recipe works well in two very different economic contexts. Specifically, a simplified version of the original method, which was developed on a sample of 856 respondents to a phone survey conducted in Rwanda in 2009, can similarly be used to estimate the wealth of 1,234 respondents to a face-to-face survey conducted in Afghanistan in 2015. However, we find that such models are relatively brittle, and that a model trained in one country cannot be used to estimate characteristics in another. These results suggest several promising applications and directions for future work.

* University of California, Berkeley, School of Information, 102 South Hall, Berkeley, CA 94720-4600, jblumenstock@berkeley.edu

I. Supervised learning

Broadly, our goal is to infer the characteristics of an individual from the "digital footprints" that she leaves behind through the use of digital devices such as phones, social media, and other technology. Building on the example in Blumenstock, Cadamuro and On (2015), we start with the specific task of estimating the wealth Y_i of individual i from an administrative source of data X_i that captures i's history of mobile phone use. We assume we have access to a *training sample* for whom both Y_i and X_i are observed – the details of these training samples are described in Section II below. The estimation then proceeds in two steps.

We first transform i's raw digital device data into a vector of K metrics $X_i = \langle x_{i1}, ..., x_{iK} \rangle$ that quantify different dimensions of mobile phone use, such as the total duration of i's phone calls, the number of unique cell towers used by i, and so forth. Many approaches to this "feature engineering" step are possible. Blumenstock, Cadamuro and On (2015) develop a recursive, combinatoric algorithm to perform this transformation, which produces an expressive vector quantifying phone use in several thousand dimensions. Here, we take a shortcut and rely instead on a Python library designed specifically for the purpose of converting mobile phone data into structured vectors,[1] which produces a vector X_i of roughly 350 such metrics. We show later that this approach is considerably less expressive than the original method, and the predictive performance of the downstream model is degraded as a result. However, this shortcut greatly simplifies the exposition, and hopefully facilitates future replication and extension.

The second step is to fit a model $Y_{it} = f(X_{it})$ that captures the relationship between the target characteristic and the vector of phone use metrics. Of key concern is ensuring that the model

[1] See http://bandicoot.mit.edu/

$f()$ is both flexible (to express the relationship between phone use and economic characteristics) and parsimonious (since in many practical settings the number of metrics, K, will approach or exceed the number of individuals in the training sample, N). In what follows, we fit $f()$ using a "gradient boosting" algorithm, a flexible supervised machine learning model. This algorithm is closely related to the more common random forest algorithm, but can be more easily parallelized for computation, and contains several tweaks that lead to modest improvements in a variety of predictive tasks (Chen and Guestrin, 2016).[2]

The gradient boosting algorithm contains a set of hyperparameters Θ that jointly determine model representation and optimization. In particular, a number of these hyperparameters—such as the the maximum depth of the decision trees, and the L_1 and L_2 regularization penalties—impact the degree of regularization imposed during model fitting. To select the optimal set of hyperparameters $\Theta^\star$, we perform grid search across a very large range of possible combinations of hyperparameters, using three repeats of 10-fold cross-validation. Thus, for each combination of hyperparameters, we estimate the root mean squared error of predictions in the 30 different held-out folds, and select the parameter set that minimizes the average error across these held out folds.

II. Training data

We replicate all experiments using two independently collected datasets. The first dataset covers a sample of 856 mobile phone subscribers in Rwanda. Full details on the sampling frame and methodology are provided by Blumenstock and Eagle (2012). In summary, a 20-minute phone survey was conducted with a geographically stratified sample of subscribers in July of 2009, with undergraduate enumerators from the Kigali Institute of Science and Technology. All respondents were active on the nation's near-monopoly mobile phone network, which at the time covered approximately 10% of the total Rwandan population. Each individual's responses to the phone survey were then merged with a large database of mobile phone records describing all transactions made by each subscriber since 2005.

The second dataset was collected in Afghanistan in 2015-2016. Working with a local Afghan survey firm, we conducted several rounds of face-to-face and phone-based interviews with 1,234 Afghan citizens. Unlike Rwanda, where respondents were sampled from all districts in the country, the Afghan survey focused on male heads of household in just two provinces, Kabul and Parwan. Only individuals with active accounts on the Roshan mobile phone network were eligible to participants. At the time of the survey, mobile phone penetration in Afghanistan was roughly 70%, of whom 30% were Roshan subscribers. The Afghan sample is thus considerably more homogeneous than the Rwandan sample. As in Rwanda, each respondent's survey responses were matched to his mobile phone transaction records, which we obtained directly from the operator, for the period starting in January 2014.

In both countries, informed consent was received from subjects prior to data collection, and both research protocols were reviewed and approved by our institutional human subjects review board. The economic characteristic that we focus on predicting below is the wealth of the subscriber. We measure wealth as the first principal component of a set of responses related to asset ownership and household characteristics. In both countries, we use the same set of responses as input to the principal component analysis, but allow for the basis vectors to differ between countries.

The mobile phone data used as the basis for the construction of the X_i vector (see Section I) are the Call Detail Records (CDR) collected by the mobile phone operators. These CDR capture basic metadata on all transactions mediated by the mobile phone network, including phone calls, text messages, airtime purchases, and mobile money use. In total, we observe tens of thousands of transactions, each of which contains several fields, including: the identity of the caller and receiver, the date and time of the event, the duration and cost of the call, and the location of the cell phone tower nearest to both parties at the time the event is initiated. In all the experiments that follow, we use the two months of mobile phone activity immediately prior to

[2] In results available upon request, we find that the choice of the learning algorithm does not qualitatively affect the main results.

the date of the survey to construct X_i.

III. Prediction experiments

Our first set of results demonstrate that a simplified version of the approach developed in Blumenstock, Cadamuro and On (2015) can be used to estimate the poverty of mobile phone owners in Rwanda and Afghanistan. The simplified version deviates from the original in the following ways: (i) we use a public library to extract features of mobile phone use, instead of the more computationally intensive deterministic finite automata; (ii) only two months of phone activity are used, rather than two years; (iii) a non-linear gradient boosting algorithm is used for supervised learning, instead of an elastic net regression; (iv) to standardize across countries, a slightly different set of asset measures was used to form the wealth index.

The results for models trained and tested in Rwanda and Afghanistan are shown in Figure 1. The main left panel plots, for each of the 856 phone survey respondents in Rwanda, the actual wealth index (x-axis, as inferred from phone survey questions) and the predicted wealth index (y-axis, as predicted from the supervised learning algorithm described above). The average cross-validated R^2 is 0.33, which is comparable to the performance of the random forest model originally reported by Blumenstock, Cadamuro and On (2015). In results not shown, we find that the primary source of this discrepancy is the simplified feature engineering process; when the original finite automata is used to generate features, performance improves to the original benchmark of 0.40.

[FIGURE ?? GOES HERE]

The entire model fitting and cross-validation process is repeated using the Afghan dataset, with results presented in the main right panel of Figure 1. We observe comparable predictive performance, despite the vastly different circumstances used to collect the data and construct the sample frames.

Critically, however, we find that a model that is trained using data from one country cannot be used to infer the characteristics of individuals in another. This can be seen in the two inset figures in Figure 1. The left inset (labeled "Afghanistan test") is constructed by applying the model trained on Rwandan survey respondents to predict the wealth of Afghan survey respondents. The right inset ("Rwanda test") uses the Afghan model to predict the wealth of Rwandans. While both models do better than random guessing ($R^2 = 0.05$ and 0.07 for the left and right insets, respectively), the estimates are quite inaccurate. This finding reflects recent results by Khan and Blumenstock (2016), who find that models trained to predict mobile product adoption in one country cannot be directly applied to another country, absent model retraining. The implications of these results are discussed below.

IV. The potential for mobile phone data

By 2020, roughly three quarters of the world's population—5.7 billion people—will own a mobile phone. Even in sub-Saharan Africa, the least connected region, mobile cellular penetration is expected to soon surpass 50% (GSMA, 2017). The fact that the data generated by the everyday use of this platform can be used to estimate the economic characteristics of individual subscribers can enable many novel applications, and creates exciting opportunities for future work.

Perhaps the most immediate potential application is in basic measurement. For instance, Blumenstock, Cadamuro and On (2015) find that predictions of wealth based on mobile phone data can be used to generate district-level estimates of the distribution of wealth that are roughly as accurate as a 5-year old nationally-representative household survey. Phone-based estimates should never replace more robust data collection, but in resource-constrained environments this provides an option for quantitative measurement at a fraction of the expense of traditional methods.[3] Closely related, such estimates might provide scalable methods for targeting: many of the largest development interventions currently use wealth proxies to determine program eligibility; phone-based indices could reasonably be used as a supplement to or substitute for proxies that are more costly to collect.

Accurate indices of individual and household welfare can also lead to new paradigms for pro-

[3] A related literature indicates that satellite imagery can provide similarly accurate estimates of sub-regional wealth (Jean et al., 2016; Blumenstock, 2016), though it is not yet known whether such data can generalize to other measures of human development (Head et al., 2017).

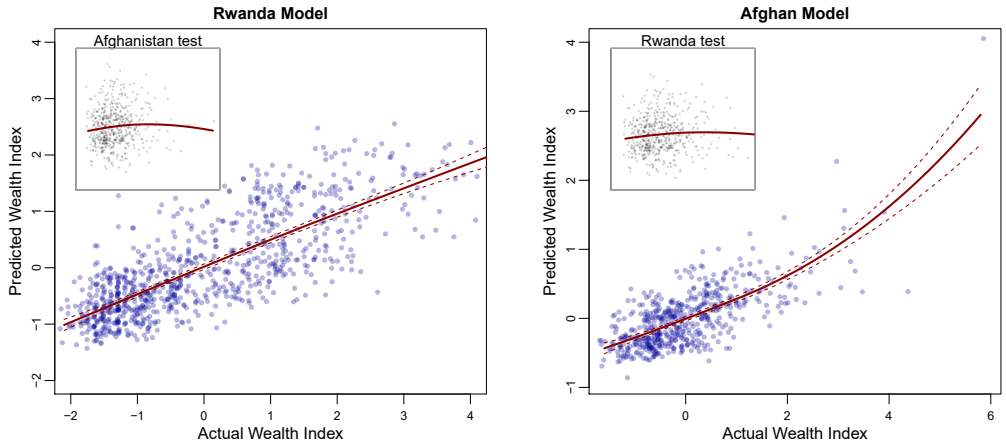

FIGURE 1. MODEL PREDICTIONS AND PERFORMANCE.

Note: Left figure compares the predicted wealth of Rwandan survey respondents (as inferred from their patterns of mobile phone use) to the actual wealth of those respondents (as reported in a phone survey). Each dot represents one of 856 survey respondents; the red line represents the local quadratic regression line of best fit. Model is trained and cross-validated using the Rwandan survey sample. Left inset figure uses the model trained on the Rwandan sample to predict the wealth of 1234 Afghans for whom mobile phone data was collected. Right figure is analagous: the main figure compares predicted to actual wealth for the Afghan survey sample, using a model trained on that sample; the inset shows the predictions of the Afghan model applied to the Rwandan sample.

gram monitoring and impact evaluation. If the dynamic well-being of an individual can be measured repeatedly over time, this facilitates inference about the causes of those changes. However, these dynamic extensions require two fundmental assumptions that have not yet been tested in the research literature. First, digital footprint data must contain sufficient signal to infer changes in welfare over time. This is not a foregone conclusion; for instance, it is quite possible that phone usage reflects a measure of permanent income, but cannot be used to recover measures of vulnerability or detect idiosyncratic shocks. Second, models trained using data from one period must be able to generalize to another. Here too there is reason for skepticism; several well-documented examples exist of machine learning algorithms whose performance quickly degrades over time (cf. Lazer et al., 2014). Indeed, the results in this paper indicate that naive models are brittle across geographic contexts; if they are similarly brittle across temporal contexts, dynamic inference may prove challenging.

There also exist plenty of private sector applications for phone-based estimates of economic characteristics. In industrialized nations, related techniques are frequently used for consumer profiling, targeting, and market segmentation; recent work indicates that in developing economies, phone data can be similarly analyzed to accurately predict product adoption (Khan and Blumenstock, 2016, 2017). Methods almost identical to those described in this paper are also now being used to develop credit scores: rather than using phone data to predict wealth, these "digital credit" applications use phone data to predict loan repayment, training the machine learning algorithm on a sample of loan applicants for whom repayment is observed (Francis, Blumenstock and Robinson, 2017; Bjorkegren and Grissen, 2015).

Yet the results in this paper also sound a note of caution. The fact that a model trained in one country performs so poorly when applied "off the shelf" in another suggests that considerable work is needed before these algorithms can be applied at scale. There may be empirical techniques to enable such cross-context generalization, for instance by "over-regularizing" the model, by manipulating the weights assigned to training instances, or through more thoughtful application of active and semi-supervised learning. However, some limitations may be fundamental—in particular, differences between how rich and poor people use phones in one country may not be relevant to another. For instance, in many countries "missed calls" are quite common (where person A calls person B

but hangs up before B answers, as a signal that B should call A) and often indicate relative wealth (i.e., that B is wealthier than A); in other countries, no such norm exists. More generally, little is know about the extent to which complex, non-parametric algorithms can generalize from one geographic or temporal context to another. Thus, while the mass adoption of mobile phones is opening up new frontiers for quantitative research in developing countries, many basic questions must be addressed before the value of these data is known or realized.

REFERENCES

Bjorkegren, Daniel, and Darrell Grissen. 2015. "Behavior Revealed in Mobile Phone Usage Predicts Loan Repayment." *Available at SSRN 2611775*.

Blumenstock, Joshua Evan. 2014. "Calling for Better Measurement: Estimating an Individuals Wealth and Well-Being from Mobile Phone Transaction Records." New York, NY.

Blumenstock, Joshua Evan. 2016. "Fighting poverty with data." *Science*, 353(6301): 753–754.

Blumenstock, Joshua Evan, and Nathan Eagle. 2012. "Divided We Call: Disparities in Access and Use of Mobile Phones in Rwanda." *Information Technology and International Development*, 8(2): 1–16.

Blumenstock, Joshua Evan, Gabriel Cadamuro, and Robert On. 2015. "Predicting poverty and wealth from mobile phone metadata." *Science*, 350(6264): 1073–1076.

Chen, Tianqi, and Carlos Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." *KDD '16*, 785–794. New York, NY, USA:ACM.

Donaldson, Dave, and Adam Storeygard. 2016. "The View from Above: Applications of Satellite Data in Economics." *Journal of Economic Perspectives*, 30(4): 171–198.

Eagle, Nathan, Michael Macy, and Rob Claxton. 2010. "Network Diversity and Economic Development." *Science*, 328(5981): 1029–1031.

Francis, Eilin, Joshua E Blumenstock, and Jonathan Robinson. 2017. "Digital Credit: A Snapshot of the Current Landscape and Open Research Questions." *BREAD Working Paper No. 516*.

GSMA. 2017. "The Mobile Economy 2017: Measuring the future of mobile." GSMA Intelligence.

Head, Andrew, Melanie Manguin, Nate Tran, and Joshua E Blumenstock. 2017. "Can Human Development be Measured with Satellite Imagery?" *ICTD '17*. Lahore, Pakistan:ACM.

Jean, Neal, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell, and Stefano Ermon. 2016. "Combining satellite imagery and machine learning to predict poverty." *Science*, 353(6301): 790–794.

Khan, Muhammad Raza, and Joshua Evan Blumenstock. 2016. "Predictors without Borders: Behavioral Modeling of Product Adoption in Three Developing Countries." *Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '16)*.

Khan, Muhammad Raza, and Joshua Evan Blumenstock. 2017. "Determinants of Mobile Money Adoption in Pakistan." *The Thirty-first Annual Conference on Neural Information Processing Systems (NIPS 17), Workshop on Machine Learning for the Developing World*.

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. "The Parable of Google Flu: Traps in Big Data Analysis." *Science*, 343(14 March).

Llorente, Alejandro, Manuel Garcia-Herranz, Manuel Cebrian, and Esteban Moro. 2015. "Social Media Fingerprints of Unemployment." *PLOS ONE*, 10(5): e0128692.