Supplemental Appendix: Predicting Well-Being with Mobile Phone Data: Evidence from Four Countries

By EMILY AIKEN, JOSHUA E. BLUMENSTOCK, SVETA MILUSHEVA, AND M. MERRITT SMITH

I. Methods Appendix

A. *Survey Datasets*

We use data from four household surveys. For Afghanistan, we use data collected as part of the Targeting the Ultra Poor impact evaluation described in Bedoya et al. (2019). For Côte d'Ivoire, we use the SWIFT phone survey conducted by the World Bank in 2023. For Malawi, we use proprietary survey data collected by the Global Opportunity Lab in 2024. For Togo, we use the 2018-19 LSMS survey. Survey characteristics are summarized in Table 1.

B. *Outcome Variables*

Consumption Expenditures: We measure total household consumption as the sum of expenditures across four categories: food, housing, durable goods, and non-durable non-food goods. All consumption measures are expressed on a per capita, per day basis by dividing total household expenditure by household size (as reported in the survey) and the number of days in the recall period. We also construct two subsidiary consumption measures: food consumption (food expenditures only) and non-food consumption (the sum of housing, durable goods, and non-durable non-food expenditures). All consumption measures are logged to account for right-skewed distributions.

Asset Index: We construct a household asset index from survey questions about asset ownership. The specific assets included vary by country based on each survey's questionnaire, but typically include items such as televisions, refrigerators, motorcycles, and so on. Asset variables are measured either as binary indicators (e.g. "Do you own a TV?") or as counts (e.g., "How many TVs do you own"). We normalize all count variables to have mean zero and unit variance, then apply Principal Components Analysis (PCA) to the full set of asset variables (both binary and normalized counts). We use the first principal component as our asset index measure. For Afghanistan, we use a pre-computed wealth score provided in the Bedoya et al. (2019) dataset.

Multidimensional Poverty Measure (MPM): We construct the MPM following the methodology described in the World Bank's Poverty, Prosperity, and Planet Report (World Bank, 2024). This measure is guided by other global multidimensional measures, particularly the Multidimensional Poverty Index (MPI). The MPM aggregates deprivations across three dimensions of wellbeing: monetary poverty, education and basic infrastructure services. Each dimension is composed of several indicators, with households considered deprived if they fall below defined thresholds for each indicator. The overall MPM score represents the weighted sum of deprivations experienced by the household. Specific indicators and weights vary by country based on data availability in each survey, but follow the standard framework outlined in the World Bank methodology. The limited set of indicators required for the MPM (compared to other multidimensional poverty measures like the MPI) are more universal across the survey datasets, allowing the construction of more comparable indices.

*Food security index:*We construct a food security index using the same PCA methodology as the wealth index. Food security variables (which vary by survey) typically include indicators of dietary diversity, food consumption frequency, and experiences of food insecurity. We normalize all variables and apply PCA, using the first principal component as the food security index. For Afghanistan, we use use a pre-computed food security index provided in the Bedoya et al. (2019) dataset.

Mental Health: We measure mental health using validated screening instruments that vary by country. For Malawi, we use the Patient Health Questionnaire-9 (PHQ-9), a 9-item instrument that assesses depressive symptoms over the previous two weeks, with scores ranging from 0 to 27 (Kroenke, Spitzer and Williams, 2001). For Afghanistan, we use a pre-computed z-score based on the Center for Epidemiologic Studies Depression Scale (CES-D) provided in the Bedoya et al. (2019) dataset. Higher scores on both instruments indicate greater severity of depressive symptoms.

Monthly income: We measure household monthly income using country-specific methods based on available survey data. For Malawi, we use direct responses to a question asking respondents to report their total household income in the previous month. For Togo, we construct monthly income by aggregating reported earnings across multiple sources, including salaries, wages, and agricultural sales.

C. Mobile Phone Datasets

Mobile phone data were provided by mobile network operators in each country and cover the periods specified in Table 1. The datasets include several types of behavioral records for matched survey respondents:

Call Detail Records (CDR): Records of voice calls, including timestamps, call duration, directional information (incoming vs. outgoing calls), and the cell tower used by the caller and receiver. CDR data are available for all four countries.

SMS records: Records of text messages sent and received, including timestamps and directional information. SMS data are available for all four countries.

Mobile money transactions: Records of mobile money transfers, including transaction amounts, timestamps, and transaction counts. Mobile money data are available for Malawi and Togo.

Airtime recharge events: Records of when users purchased airtime credit, including recharge amounts and timestamps. Recharge data are available for Afghanistan and Malawi.

Mobile data usage: Records of data consumption, including data volume and timestamps. Mobile data usage records are available for Afghanistan and Malawi.

D. Feature Engineering

From the raw mobile phone data types, we extract behavioral features using a custom featurization pipeline described in Aiken et al. (2022). Features are organized into several categories based on the underlying data source and behavioral construct:

Communication and interaction patterns: We calculate aggregate metrics capturing overall communication behavior across calls and texts combined, including the number of unique contacts, entropy of contacts measuring diversity of communication partners, balance of contacts measuring symmetry between incoming and outgoing interactions, interactions per contact, the proportion of interactions initiated by the user, inter-event time measuring gaps between consecutive interactions, and indicators for international communication. We also compute the concentration of interactions using Pareto distributions. Finally, we calculate the number of active days, defined as days with any recorded activity, calculated both overall and separately for weekdays and weekends.

Call-specific features: Total call duration and international call indicators, calculated separately from text-based interactions.

*Text-specific feature*s: Response delay measuring time between receiving and responding to messages, response rate, and international text indicators.

Mobility patterns: Features derived from cell tower locations, including the number of unique antennas accessed, the frequency of visits to the most common antennas, entropy of antenna usage measuring spatial diversity, the percentage of time spent at the identified home location, and radius of gyration measuring the spatial extent of movements.

Mobile money activity: Transaction counts and total monetary values for mobile money transfers.

Airtime recharge behavior: Recharge event counts and total recharge amounts.

Mobile data consumption: Total data usage volume and breakdowns thereof.

Location-based features: Administrative region indicators derived from the most frequently used cell towers.

For most feature categories, we calculate variants aggregated at different temporal scales (total across observation period, average per active day, average per week) and time windows (all hours, daytime only, nighttime/nocturnal only, weekdays only, weekends only), as well as directional breakdowns (incoming vs. outgoing) where applicable. Temporal aggregations allow us to capture both overall activity levels and intensity of usage conditional on being active.

E. Machine Learning Training and Evaluation

We compare three regression models for each outcome-country combination: Lasso (L1-regularized linear regression), Histogram-Based Gradient Boosting Regression, and Random Forest. We train separate models for each outcome variable and country. For hyperparameter optimization, we employ Bayesian optimization using Weights & Biases, conducting 20 optimization runs for each model-outcome-country combination (Biewald, 2020). The hyperparameter search spaces are detailed in Table 1. We use 5-fold cross-validation during hyperparameter tuning.

We partition each dataset into training (80%) and holdout (20%) sets. The training set is further split during cross-validation for hyperparameter tuning. The holdout set is never used during training or hyperparameter selection and serves solely for final model evaluation after all model selection is complete.

We apply several preprocessing steps to handle missing values and prepare features for modeling, with careful attention to preventing data leakage between training and test sets.

Missing value imputation: For numeric features, we impute missing values using the median if the feature contains zeros, and the mean otherwise. These statistics are computed on the training data and applied to test and holdout sets. For non-numeric features (primarily geography indicators), we impute missing values using the mode from the training data.

Binary feature encoding: Numeric features that contain only 0 and 1 values are decomposed into two binary indicators: one for the value being 1, and one for the value being missing.

Categorical encoding: Non-numeric features, including geography indicators, are one-hot encoded using dummy variables (with the first category dropped to avoid multicollinearity). For test and holdout sets, we ensure the same dummy variables exist as in the training set, adding zero columns for categories present in training but absent in test data, and removing categories present in test but not in training. All preprocessing statistics (medians, means, modes, and dummy variable structures) are learned exclusively from the training data and applied consistently to test and holdout sets to prevent data leakage.

We evaluate model performance using four metrics: coefficient of determination (R^2), root mean squared error (RMSE), Pearson correlation coefficient, and Spearman rank correlation. For model selection during hyperparameter tuning, we select the model configuration that minimizes RMSE on the test set. All results reported in the main text reflect Pearson correlations computed on the holdout set using the model with the best test set RMSE.

For the sample size evaluation (Figure 2, Panel A), we repeat the full modeling procedure described above at training sample sizes ranging from 10 to 5,000 users. Specifically, we test sample sizes of 10, 25, 50, 75, 100, 125, 150, 175, 200, 225, 250, 275, 300, 325, 350, 375, 400, 425, 450, 475, 500, 750, 1000, 1250, 1500, 1750, 2000, 2250, 2500, 2750, 3000, 3250, 3500, 3750, 4000, 4250, 4500, 4750, and 5000. For each sample size, we draw 5 bootstrap resamples from the full training set and train models on each resample using the same hyperparameter optimization and cross-validation procedures. We compute performance metrics on the holdout set for each bootstrap iteration and report the mean and 95% confidence intervals across the 5 resamples.

II. Appendix Tables

Table 1—: Hyperparameter search spaces for Bayesian optimization

Model	**Hyperparameter**	**Search Space**
Histogram Gradient Boosting	max_iter	5000 (fixed)
	learning_rate	Log-uniform [0.00001, 0.1]
	max_depth	Uniform integer [3, 15]
	l2_regularization	Log-uniform [0.0001, 0.01]
	min_samples_leaf	Uniform integer [1, 50]
Lasso	max_iter	1000 (fixed)
	tol	0.001 (fixed)
	alpha	Log-uniform [0.00001, 0.01]
Random Forest	n_estimators	Uniform integer [250, 1000], step 50
	max_depth	Uniform integer [3, 20]
	min_samples_split	Uniform integer [2, 20]
	min_samples_leaf	Uniform integer [1, 10]

III. Appendix Figures

Figure 1. : Extensive Margin of Mobile Phone Usage

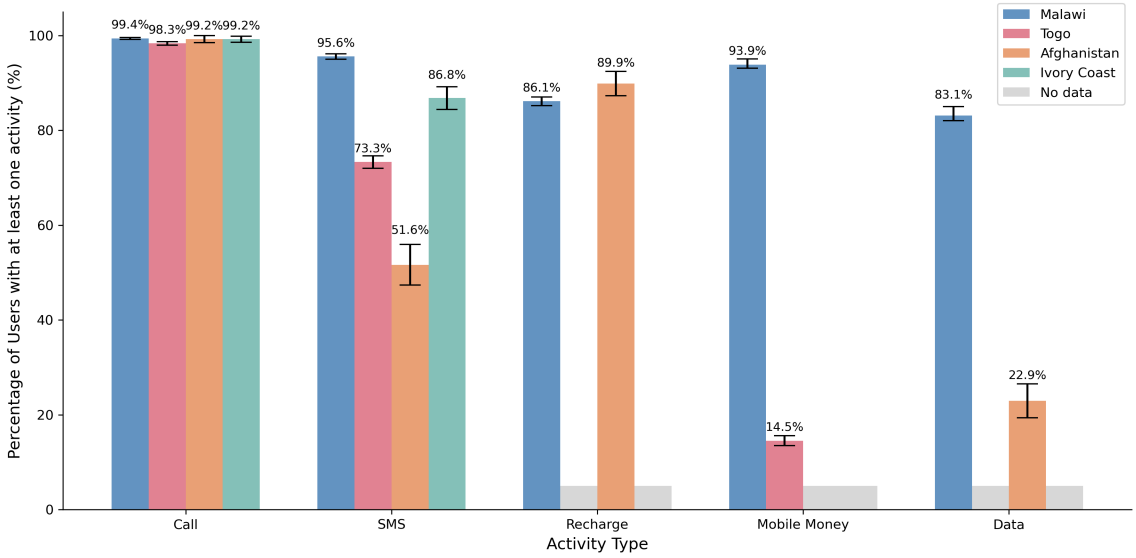
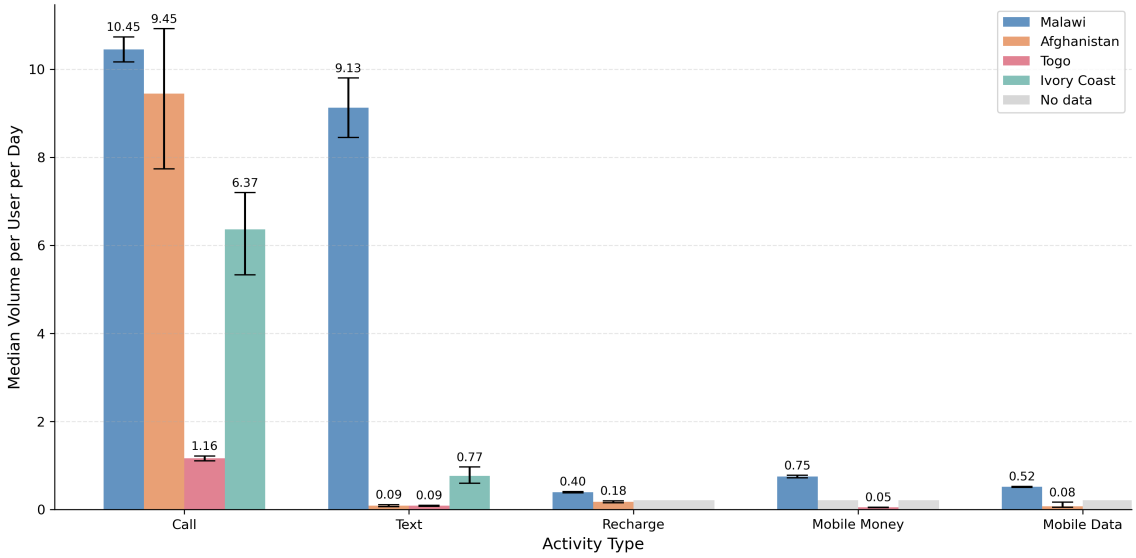


Figure 2. : Intensive Margin of Mobile Phone Usage



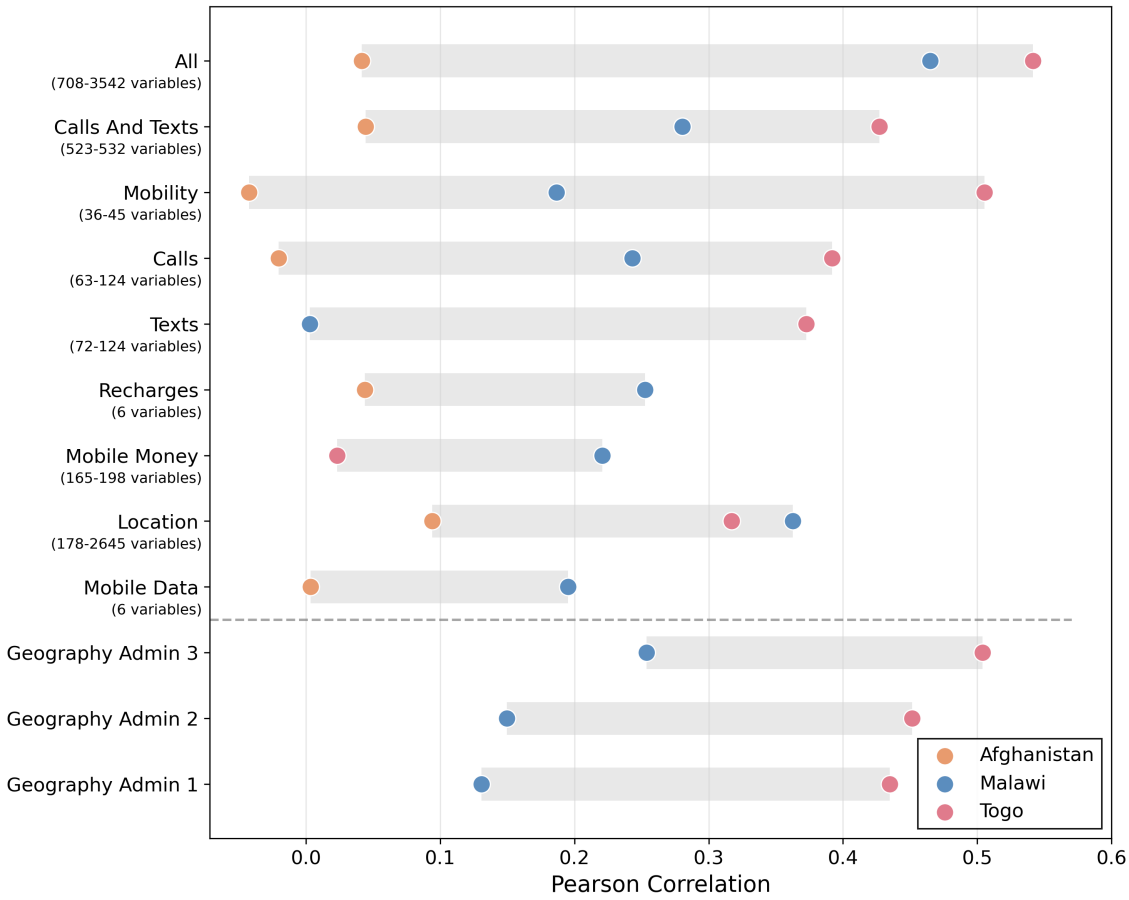


Figure 3. : Pearson correlations when using different mobile phone transaction types to predict total consumption per capita per day. Where data types are not available, no results are shown. We do not have consumption information for Côte d'Ivoire and therefore do not report it here.

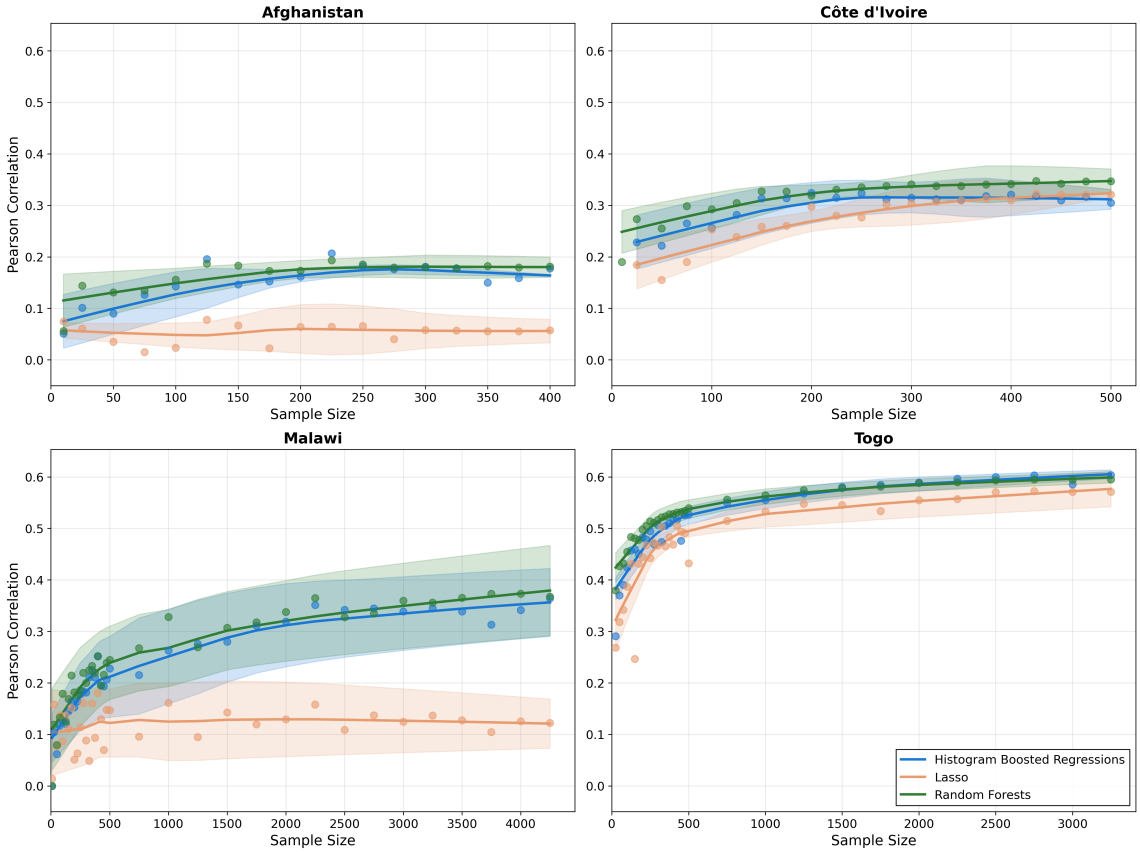


Figure 4. : Pearson correlations as sample size increases, split out by the type of model used. Points are the mean holdout Pearson ρ for the given sample size across bootstraps, and the shaded region is the 95% confidence interval derived from 5 bootstrapped runs of the same sample size.

REFERENCES

Aiken, Emily, Suzanne Bellue, Dean Karlan, Chris Udry, and Joshua Blumenstock. 2022. "Machine learning and phone data can improve targeting of humanitarian aid." *Nature*, 603(7903): 864–870.

Bedoya, Guadalupe, Aidan Coville, Johannes Haushofer, Mohammad Isaqzadeh, and Jeremy P Shapiro. 2019. "No household left behind: Afghanistan targeting the ultra poor impact evaluation." National Bureau of Economic Research.

Biewald, Lukas. 2020. "Experiment Tracking with Weights and Biases." Software available from wandb.com.

Kroenke, Kurt, Robert L Spitzer, and Janet BW Williams. 2001. "The PHQ-9: validity of a brief depression severity measure." *Journal of general internal medicine*, 16(9): 606–613.

World Bank. 2024. "Poverty, Prosperity, and Planet Report 2024." World Bank Report, Washington, DC.