Mapping Socioeconomic Indicators Using Social Media Advertising Data

Masoomali Fatehkia[a], Isabelle Tingzon[b], Ardie Orden[b], Stephanie Sy[b], Vedran Sekara[c], Manuel Garcia-Herranz[c], and Ingmar Weber[a*]

[a]Qatar Computing Research Institute, HBKU, Doha, Qatar
[b]Thinking Machines, Manila, Philippines
[c]UNICEF Innovation, New York, USA

1 Supplementary Information

1.1 Summary statistics of the distribution of Wealth Index

Tables S1 and S2 show summary statistics of the distribution of the DHS Wealth Index for different subsets of clusters in the Philippines and India respectively. The tables show the distribution of Wealth Index for all clusters, the clusters that were used in the analysis and clusters that were excluded due to missing geo-location information or lack of Facebook users.

In both countries the clusters that were used in the analysis had on average slightly higher Wealth Index than all the clusters combined. In terms of the spread of the distribution, the standard deviations were roughly similar for the clusters used in the analysis and for all clusters combined. This suggest that clusters from throughout the Wealth Index distribution were included in the data analysis for both countries. However, it should be noted that the clusters that had to be excluded due to missing geo-locations or lack of FB users were overall from the poorer end of the Wealth Index distribution with lower median/mean Wealth Index than observed in the overall group of clusters.

Table S1: Summary statistics of the distribution of the DHS Wealth Index for different subsets of clusters. Data is for the Philippines. Std is the standard deviation.

Subset of clusters	N	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	std
All DHS clusters	1,249	-183,335	-48,272	495	4,130	58,356	220,884	71,532
Geo-located clusters	1,213	-183,335	-46,820	1,196	4,892	58,356	220,884	70,973
Clusters used in the analysis	1,205	-183,335	-46,213	1,750	5,599	59,145	220,884	70,626
Clusters that were not included in the analysis								
Excluded (all)	44	-177,150	-104,186	-37,524	-36,105	36,836	150,971	84,425
Excluded due to missing geo-location	36	-177,150	-83,614	-22,532	-21,559	46,319	150,971	85,573
Excluded due to no (< 100) FB users	8	-150,668	-121,767	-104,608	-101,564	-85,914	-41,858	34,409

[*] corresponding author; E-mail: iweber@hbku.edu.qa

Table S2: Summary statistics of the distribution of the DHS Wealth Index for different subsets of clusters. Data is for India. Std is the standard deviation.

Subset of clusters	N	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	std
All DHS clusters	28,524	-179,271	-62,363	-7,417	783	62,271	223,809	79,299
Geo-located clusters	28,393	-179,271	-62,338	-7,198	914	62,581	223,809	79,352
Clusters used in the analysis	28,043	-178,461	-62,015	-6,802	1,346	63,397	223,809	79,390
Clusters that were not included in the analysis								
Excluded (all)	481	-179,271	-81,201	-37,089	-32,035	15,194	173,493	66,027
Excluded due to missing geo-location	131	-165,579	-68,767	-37,032	-27,510	12,186	140,853	60,739
Excluded due to no (< 100) FB users	350	-179,271	-85,009	-37,363	-33,729	17,450	173,493	67,906

1.2 The choice of radius of data collection

The DHS survey locations are called clusters with each cluster consisting of a group of surveyed households (median of 23 households surveyed per cluster in the Philippines and 21 in India). The DHS clusters correspond to Primary Sampling Units (PSU) from the respective country's census: in the Philippines DHS data used here, these correspond to "a barangay, a portion of a large barangay, or two or more adjacent small barangays"[1](1) and in India they correspond to "villages in rural areas and Census Enumeration Blocks (CEBs) in urban areas"(2). Based on geographic shape file for the barangays in the Philippines[2], we estimate a barangay to have an average area of 7.06 km^2 (median: 2.86 km^2, IQR: 5.26 km^2); we could not find corresponding shape files for villages and census enumeration blocks in India.

The DHS reports cluster locations in the form of latitude and longitude coordinates of the cluster centroid. In order to preserve respondent confidentiality, the reported location coordinates are perturbations of the actual coordinates. The true location coordinates are perturbed up to 2km for urban clusters and up to 5km for rural clusters with a further 1% of rural clusters displaced up to 10km.

In order to collect estimates of Facebook users for each cluster location, we used the reported latitude and longitude coordinates in combination with a given radius around that location for which to collect data. A radius of 1km is the smallest radius around a given coordinate for which data can be collected from the Facebook marketing API[3]. However, due to the perturbation of the reported DHS cluster locations this would then not necessarily contain the actual location, while also likely suffering from data sparsity given the small radius. Previous studies that used satellite imagery to predict the DHS Wealth Index, chose radii of data collection ranging from 2km/5km for Urban/Rural areas (3) up to 10km (4) to ensure that the Satellite imagery from the actual location was included in the data. We follow this approach while also trying to collect data for a large enough radius so as to minimizes data sparsity (few or no Facebook users).

As an initial analysis, we collected estimates of Facebook users (aged 18+) at a radius of 2km for urban clusters and 5km for rural clusters (corresponding to the amount of perturbation of cluster locations) as well as for a larger radius of 5km for urban and 10km for rural clusters. Table S3 reports the levels of sparsity observed for different choices of radius of data collection. In the Philippines, the choice of 2km radius for urban and 5km radius for rural locations was made. In India, where Facebook penetration is lower than in the Philippines, a larger radius of data collection was needed to achieve the same level of sparsity (% of clusters with <= 1000 FB users) as in the Philippines; hence, in India data were collected at a radius of 5km for urban and 10km for rural clusters.

[1] A barangay is the smallest administrative division in the Philippines.

[2] `http://philgis.org/country-barangay/country-barangays-file`

[3] `https://developers.facebook.com/docs/marketing-api/audiences/reference/basic-targeting/#location`

Table S3: Results on clusters with sparse Facebook data for different choices of radius of data collection. The radius of data collection that was ultimately chosen for each country and cluster type (urban/rural) is indicated in bold. Data on clusters with $<$ 100 FB users is not available for India as the data augmentation approach was not applied in the initial, exploratory collection for India.

Country	Urban or Rural (# clusters)	Radius (km)	Clusters with <= 1000 FB users (%)	Clusters with < 100 FB users (%)
Philippines	Urban (437)	**2**	**27 (6.18%)**	**0 (0%)**
		5	5 (1.14%)	0 (0%)
	Rural (776)	**5**	**143 (18.4%)**	**8 (1.03%)**
		10	32 (4.12%)	1 (0.13%)
India	Urban (8,473)	2	1,357 (16.01%)	-
		5	**526 (6.12%)**	**44 (0.52%)**
	Rural (19,920)	5	8,345 (41.89%)	-
		10	**2,551 (12.81%)**	**306 (1.54%)**

1.3 Performance of baseline models

Table S4 presents the performance of various simple baseline models in predicting the DHS Wealth Index. Model B1 uses the Wealth Index from a past DHS survey to predict the recent DHS Wealth Index. The Wealth Index values from the past survey were geographically interpolated by setting the Wealth Index at a given location as the average of the Wealth Index for the five closest clusters in the previous survey; these values were then used to predict the Wealth Index values for the clusters from the most recent survey. The 2008 DHS was used for Philippines as the 2013 DHS did not report geographic coordinates for cluster locations. None of the past DHS surveys in India recorded cluster geographic coordinates.

Model B2 predicts the Wealth Index using the regional indicator variables that were selected by LASSO; this model shows how much of the variation in the Wealth Index is accounted for by regional level variation. Models B3 and B4 demonstrate the predictive performance of single variable models that simply use the population density or the Facebook penetration respectively.

For the models reported here and throughout the paper, all evaluations were done in a 10-fold cross validation where, across 10 iterations, a model is trained on 9/10 of the data and then evaluated on the remaining 1/10. A cross-validated R^2 value was then computed between the survey data and the predictions that were generated during the cross validation. The R^2 value was computed as the proportion of the variation in the ground truth survey data that is explained by the predicted values, that is, $1-RSS/TSS$ where RSS is the sum of squared residuals (difference between ground truth and prediction) and TSS is the total sum of squares (difference between ground truth and its average value). The cross-validated R^2 is reported for all models. In addition to R^2 we also report the Root Mean Squared Error (RMSE) based on the cross-validated predictions.

Table S4: Performance of various baseline models. Past DHS surveys for India were not geolocated and hence could not be used for model B1. Reported values are R^2 and RMSE based on 10-fold cross validation.

	Variable	B1	B2	B3	B4
	Interpolated past DHS data	X			
	Regional indicators		X		
	Log population density			X	
	Facebook penetration				X
Philippines	R^2	0.444	0.378	0.448	0.439
	RMSE	52,647	55,669	52,448	52,866
India	R^2	-	0.334	0.180	0.309
	RMSE	-	64,790	71,869	66,014

1.4 Models for predicting the wealth index

Table S5 reports the performance of various models using the Facebook features individually (model T1) and in combination with log population density and regional indicator variables (models T2 and T4). Model T3 reports the performance of a model using population density and regional indicators.

Table S5: Results of linear and regression tree models using various combinations of variables from the Facebook features, log population density and the regional indicator variables. Reported R^2 and RMSE values are based on 10-fold cross validation.

	Variables	T1	T2	T3	T4
	Facebook features	X	X		X
	Log population density		X	X	X
	Regional indicators			X	X
Philippines	R^2 - linear model	0.595	0.596	0.595	0.630
	RMSE - linear model	44,905	44,898	44,928	42,921
	R^2 - tree model	0.608	0.613	0.600	0.627
	RMSE - tree model	44,218	43,901	44,616	43,099
India	R^2 - linear model	0.479	0.489	0.581	0.623
	RMSE - linear model	57,322	56,743	51,391	48,721
	R^2 - tree model	0.563	0.630	0.627	0.691
	RMSE - tree model	52,502	48,305	48,459	44,149

1.5 Interpolating from the DHS Wealth Index

Using data from the DHS survey, it is possible to interpolate the Wealth Index observations from the surveyed locations to other locations in the country. This allows the Wealth Index values to be estimated for locations of interest where no survey data are available. Moreover, the value interpolated from the survey could be combined with data from other sources such as the Facebook features. Here, the Wealth Index values from the DHS clusters were interpolated using a nearest neighbours approach whereby for each survey cluster, the average of the k nearest clusters is used to estimate the Wealth Index value at that cluster. Table S6 reports the Pearson correlation coefficient between the Wealth Index at a given cluster and its averaged value from the closest k clusters for various values of k.

Using the interpolated Wealth Index values, regression tree models were fitted to estimate the Wealth Index for the survey clusters using combinations of the interpolated variables and the Facebook variables. Table S7 reports the performance of these regression models.

Table S6: Pearson correlation of the interpolated DHS data (average of the k nearest neighbours for various values of k) with the DHS Wealth Index at a given cluster.

	$k = 1$	$k = 3$	$k = 5$	$k = 10$
Philippines	0.597	0.686	0.687	0.681
India	0.739	0.793	0.796	0.788

Table S7: Results of various regression tree models for predicting the DHS Wealth Index using different combinations of data interpolated from the DHS survey or from other sources such as the Facebook variables. Model performance is reported with cross-validated R^2 and RMSE.

	Variables	M1	M2	M3	M4	M5
	Interpolated DHS variables	X	X	X	X	X
	Facebook features			X	X	X
	Log population density		X		X	X
	Regional indicators		X			X
Philippines	R^2	0.480	0.600	0.625	0.629	0.630
	RMSE	50,983	44,530	43,269	43,087	42,965
India	R^2	0.652	0.692	0.679	0.721	0.728
	RMSE	46,810	44,055	45,012	41,972	41,394

1.6 Data augmentation for locations with sparse data

As explained in the main paper, the lowest estimate returned by the Facebook marketing API is a value of 1000, which is hence indistinguishable from 0. Our approach to deal with data sparsity was to augment the data using the "exclusion query" method from (5) and described below, to get approximate estimates, in hundreds, for locations with sparse data.

Suppose for a given location A and specified targeting criteria the number of users M_A is sparse ($M_A <= 1000$) for which the API will return a value of 1000. We use the following approach to estimate the value of M_A in the hundreds. We find another geographically non-overlapping location B where for the same targeting criteria the number of users M_B is non-sparse ($M_B > 1000$); as a result the number of users meeting the specified targeting criteria in location A or B, M_{AB}, will also be non-sparse. Note that since the marketing API returns rounded estimates (values in the thousands are rounded to the nearest hundredth, values in the tens of thousands are rounded to the nearest thousand and so on), location B is chosen such that M_B is bigger than 1000 but less than 10,000 as values in this range are rounded to the nearest hundredth and so changes in the magnitude of hundreds can be detected. Once the value M_B and M_{AB} are known, the number of users M_A is then estimated as $M_A = M_{AB} - M_B$. This will result in an estimate of M_A in the hundreds (0, 100, 200, ... , 900).

As an example, suppose for location A the number of users aged 18+ is sparse $M_A <= 1000$. We find another non-overlapping location B where the number of users aged 18+ (M_B), is 2,200. We then request an estimate for M_{AB}, the number of users aged 18+ who are either in location A or in location B; the API returns an estimate of 2,400 for M_{AB}. We then estimate number of users aged 18+ in location A to be the difference between the number of users aged 18+ who are in either location A or B and those who are in location B only ($M_A = M_{AB} - M_B = 2,400 - 2,200 = 200$). As a result, the number of users aged 18+ in location A is estimated to be 200.

We compare this approach to two other possible ways of tackling data sparsity namely by (i) not making any changes (leaving values of 1000 as they are) and (ii) changing all values of 1000 to 0 (assuming there are no users when the API returns an estimate of 1000). Table S8 compares the performance of the regression tree models using the Facebook data when following these different approaches. The first two columns correspond to approaches (i) and (ii) above and the last column is copied from Table S5 for comparison. There is a small improvement in model performance when using the data augmentation approach described here for locations with sparse data, over the more simpler approaches of leaving the values of 1000 as they are or setting them to 0. However, with the approaches in the last two columns of Table S8, some of the survey locations in the data have to be dropped from the analysis due to lack of Facebook users.

Table S8: Performance of regression tree models with FB features, for the different approaches to handling data sparsity. Reported values are cross-validated R^2 and RMSE. Number of data points (N) used to fit the models is indicated in brackets.

		No changes	Set 1000 to 0	Data augmentation
Philippines	R^2	0.589	0.555	0.608
	RMSE	45,510	44,144	44,218
	N	1,213	1,043	1,205
India	R^2	0.522	0.547	0.563
	RMSE	54,897	53,339	52,502
	N	28,393	25,316	28,043

1.7 Combining different data sources

Combining different data sources and approaches can be useful in improving the accuracy of predictive models for poverty mapping. This section explores the combination of Facebook features with data from other surveys in the Philippines for predicting the Wealth Index.

Given the lack of spatially granular poverty data, the Philippines Statistics Authority periodically releases estimates of poverty incidence at the level of municipalities and cities in the Philippines. These estimates are produced through Small Area Estimation (SAE) techniques using census and survey data. Here, we use the municipality/city level poverty incidence estimates produced for the year 2012 (6). Each DHS cluster was assigned the poverty incidence value of the municipality/city in which its coordinates were located. Given that DHS cluster coordinates are geographically displaced to preserve confidentiality, this provides an approximate matching of cluster to poverty incidence values as some cluster coordinates may have moved into adjacent locations. The models in this section use the subset of 1,173 clusters for which both this data and data on Facebook features is available.

Table S9: Results of linear and regression tree models using various combinations of variables from the Facebook features, log population density and the regional indicator variables combined with the past survey data on municipality level poverty estimates in the Philippines. Reported R^2 and RMSE values are based on 10-fold cross validation.

	Variables	T1	T2	T3	T4	T5
	Facebook features			X	X	X
	2012 poverty estimate	X	X	X	X	X
	Log population density		X		X	X
	Regional indicators		X			X
Philippines	R^2 - linear model	0.513	0.625	0.631	0.634	0.647
	RMSE - linear model	49,464	43,393	43,047	42,852	42,103
	R^2 - tree model	-	0.631	0.634	0.637	0.642
	RMSE - tree model	-	43,037	42,884	42,694	42,418

1.8 Models with various subsets of Facebook features

The different Facebook network/device type features could potentially be measuring different characteristics of the offline population such as the level of electrification or access to telecommunication infrastructure versus access to and use of high-end device types. In order to understand better how these individual sets of features can be predictive of the Wealth Index, Table S10 presents the performance of regression trees using individual subsets of features to predict the Wealth Index.

Table S10: Performance of various subsets of the Facebook features in predicting the Wealth Index. Results are reported for regression tree models using cross-validated R^2 and RMSE values.

		Network access	Mobile OS	High-end phones	Other devices
Philippines	R^2	0.593	0.539	0.486	0.548
	RMSE	45,023	47,934	50,591	47,473
India	R^2	0.406	0.487	0.335	0.437
	RMSE	61,180	56,912	64,774	59,582

1.9 Correlations of Facebook features and ownership of individual assets

The DHS Wealth Index is calculated based on survey data on household asset ownership. Tables S11 and S12 report the correlations of the different Facebook features with the levels of ownership of individual assets (the fraction of surveyed households in a cluster who possessed a given asset) for Philippines and India respectively.

Table S11: Correlation between various Facebook features and the proportion of surveyed households who possess a variety of different asset types. Data are for Philippines.

	Electricity	Radio	TV	Refrigerator	Mobile	Computer	Washing Machine	Air Conditioner	Cable Service
Cluster Wealth Index	.626	.462	.842	.898	.68	.848	.888	.735	.382
Facebook Penetration	.383	.291	.583	.556	.479	.521	.614	.438	.082
Log Population Density	.374	.347	.557	.553	.482	.552	.610	.527	-.095
2G Network	.162	.083	.176	.111	.15	.041	.057	.046	.02
3G Network	-.077	-.251	-.254	-.317	-.113	-.336	-.388	-.287	.047
4G Network	.455	.277	.606	.572	.505	.529	.637	.439	.114
Wifi	.419	.396	.625	.617	.487	.582	.687	.538	.007
Android	.412	.186	.455	.365	.412	.304	.376	.262	.101
iOS	.392	.381	.568	.567	.439	.509	.613	.479	.011
Windows Phones	.169	.236	.28	.322	.232	.363	.391	.315	-.016
Apple iPhone X	.255	.315	.425	.451	.352	.495	.554	.489	-.02
Apple iPhone X/8/8 Plus	.289	.349	.481	.504	.389	.534	.603	.524	-.024
Samsung S9+	.243	.282	.401	.427	.328	.477	.527	.431	-.03
Samsung S8/S8+/S9/S9+	.295	.338	.489	.514	.395	.553	.615	.51	-.015
Samsung S8/S8+/S9/S9+ Apple iPhone X/8/8 Plus	.312	.343	.523	.553	.415	.557	.649	.51	.024
All Mobile Devices	.276	.086	.286	.22	.289	.165	.204	.128	.092
Feature Phones	.087	.076	.106	.116	.145	.09	.01	.107	.011
Smart Phone and tablets	.25	.078	.253	.181	.268	.126	.161	.099	.06
Tablet	.411	.243	.485	.425	.401	.335	.432	.329	.039
Cherry Mobile	-.089	-.086	-.175	-.207	-.162	-.251	-.28	-.187	.075
VIVO Mobile Devices	.408	.197	.514	.424	.425	.402	.438	.323	.082
Huawei Mobile Device	.423	.206	.518	.465	.442	.413	.457	.31	.115
Oppo Mobile Device	.453	.154	.496	.392	.406	.352	.424	.267	.149
Oppo/VIVO/Cherry	.243	.022	.222	.139	.22	.094	.132	.077	.164
Samsung Android devices	.185	-.024	.146	.128	.149	.069	.105	.063	-.013

Table S12: Correlation between various Facebook features and the proportion of surveyed households who possess a variety of different asset types. Data are for India. Note that some of the asset types which were available in the Philippines data were not available in the survey data for India.

	Electricity	Radio	TV	Refrigerator	Mobile
Cluster Wealth Index	0.596	0.153	0.860	0.894	0.604
Facebook Penetration	0.277	0.109	0.446	0.500	0.310
Log Population Density	0.070	0.036	0.256	0.384	0.334
2G Network	0.181	0.117	0.292	0.308	0.213
3G Network	0.193	0.081	0.294	0.260	0.177
4G Network	-0.046	-0.070	-0.022	-0.004	0.057
Wifi	0.256	0.068	0.432	0.524	0.261
Android	0.352	0.007	0.508	0.434	0.302
iOS	0.252	0.026	0.430	0.584	0.288
Windows Phones	0.160	0.070	0.274	0.316	0.199
Apple iPhone X	0.173	-0.004	0.300	0.452	0.206
Apple iPhone X/8/8 Plus	0.183	0.001	0.317	0.472	0.218
Samsung S9+	0.154	0.013	0.266	0.371	0.195
Samsung S8/S8+/S9/S9+	0.212	0.010	0.365	0.494	0.249
Samsung S8/S8+/S9/S9+ Apple iPhone X/8/8 Plus	0.226	0.005	0.386	0.541	0.255
All Mobile Devices	-0.081	-0.029	-0.082	-0.064	0.031
Feature Phones	0.101	0.038	0.117	0.149	0.143
Smart Phone and tablets	-0.089	-0.050	-0.094	-0.073	0.021
Tablet	0.255	0.044	0.383	0.374	0.291
VIVO Mobile Devices	-0.007	-0.138	0.009	-0.001	0.048
Huawei Mobile Device	0.109	0.040	0.249	0.264	0.218
Oppo Mobile Device	0.011	-0.108	0.075	0.120	0.155
Oppo/VIVO/Cherry	-0.020	-0.180	-0.013	0.007	0.018
Samsung Android devices	-0.020	0.040	0.043	0.083	0.131

1.10 Demographic disaggregation of predictions

One advantage of social media advertising data, such as the data from Facebook's marketing API used here, is the ability to demographically disaggregate the data by various attributes. Such disaggregated data could be used to make predictions of the Wealth Index for various demographic groups such as by gender (male vs. female), age (young vs. old) and education (high school graduate vs. college graduate). Note that education statuses are self-declared by Facebook users and not all users may specify a given education status. Disaggregated data was collected by gender (female and male users aged 18+), age (users aged 13-34 and users aged 55+) and self-declared education status (users aged 18+, high school graduates vs. more than high school) in order to test the potential for demographically disaggregating predictions of Wealth Index.

In order to generate the demographically disaggregated predictions, the Facebook penetration (which was one of the features in the prediction model) was estimated for each demographic group as follows. The offline population of individuals in each age/gender/education group was estimated by assuming that the proportion of the population of each group in each cluster was the same as the country level proportion of the population in that group[4]. This proportion was then multiplied by the cluster population to provide an estimate of the offline population of each demographic group in that cluster. The Facebook penetration for each demographic group was then computed as the ratio of Facebook users in that group to the estimated offline population of that group; as before, where the number of Facebook users exceeded the estimated offline population, the Facebook penetration was capped at 1. Table S13 reports the average value across all clusters of the estimated Facebook penetration for each demographic group.

Table S13: The mean and median estimated Facebook penetration (expressed as a percentage), across all clusters, for the different demographic groups.

Country		Female	Male	Young (ages 13-34)	Old (ages 55+)	High school grad.	More than high school
Philippines	mean	59	57	72	38	34	76
	median	60	54	99	28	26	100
India	mean	12	34	37	9	18	43
	median	3	18	20	2	5	24

Demographically disaggregated predictions were estimated as follows. Predictions were made using the tree model with Facebook features, log population density and regional indicators that was trained using data for the 18+ demographic group. In order to generate a prediction for a given demographic group, the Facebook features for the desired group were provided as input to the model (eg: fraction of users aged 13-34 with a given device/network type such as an iOS device, 4G network etc. and the age 13-34 Facebook penetration were provided as input to the model to predict the age 13-34 Wealth Index; likewise for other demographic groups); the log population density and regional indicators were location-specific information so they remained constant when making predictions for different demographic groups. Table S14 reports summary statistics of predicted Wealth Index for various demographic groups. Predictions were made for clusters with a non-zero population of users for both demographic groups.

[4] The following proportions were used for each demographic group in both countries: 50% female, 50% male, 46% young aged 13-34 and 12% old aged 55+, based on UN population demographic breakdown: `https://unstats.un.org/unsd/demographic-social/products/dyb/documents/DYB2018/table07.pdf`. For the education statuses the proportion high school graduate was 42% in the Philippines, 18% in India and the proportion with more than high school was 16% in the Philippines, 9% India, based on World Bank education attainment data: `https://data.worldbank.org/indicator/SE.TER.CUAT.BA.ZS?end=2018&start=1970&view=chart&year=2018` and `https://data.worldbank.org/indicator/SE.SEC.CUAT.LO.ZS?end=2018&start=1970&view=chart&year=2018`.

Table S14: Summary statistics of the disaggregated predictions for various demographic groups by gender, age and education status for Philippines and India. The figures reported here are for the clusters where the number of estimated Monthly Active Facebook users was greater than zero for both demographic groups in each category; the number of clusters for which a prediction was made is reported in the second column.

Philippines	clusters	Min.	Q1	Median	Mean	Q3	Max
Gender							
Female	1,175	-110,497	-44,058	5,355	6,470	53,642	134,938
Male	1,175	-110,497	-45,704	206	4,641	50,819	144,287
Age							
Young (ages 13-34)	1,162	-108,597	-38,660	-3,286	10,709	68,824	126,089
Old (ages 55+)	1,162	-100,516	-61,441	-14,756	1,519	66,637	120,197
Education status							
High school grad.	1,128	-112,199	-53,439	-9,870	3,348	66,671	120,121
More than high school	1,128	-97,050	-18,358	43,183	33,743	89,110	130,561

India	clusters	Min.	Q1	Median	Mean	Q3	Max
Gender							
Female	27,271	-117,897	-35,081	-2,171	10,423	57,711	189,130
Male	27,271	-114,787	-32,576	9,550	17,404	70,439	193,397
Age							
Young (ages 13-34)	21,688	-101,503	-22,206	25,784	29,928	82,204	238,726
Old (ages 55+)	21,688	-73,531	20,416	55,826	63,928	109,436	254,714
Education status							
High school grad.	23,789	-82,687	17,094	51,166	60,040	107,934	231,247
More than high school	23,789	-58,625	54,364	83,540	87,263	121,800	269,090

Figures S1 and S2 plot the age and education disaggregated predictions for both countries; a plot for the gender disaggregated predictions was presented in the main paper.

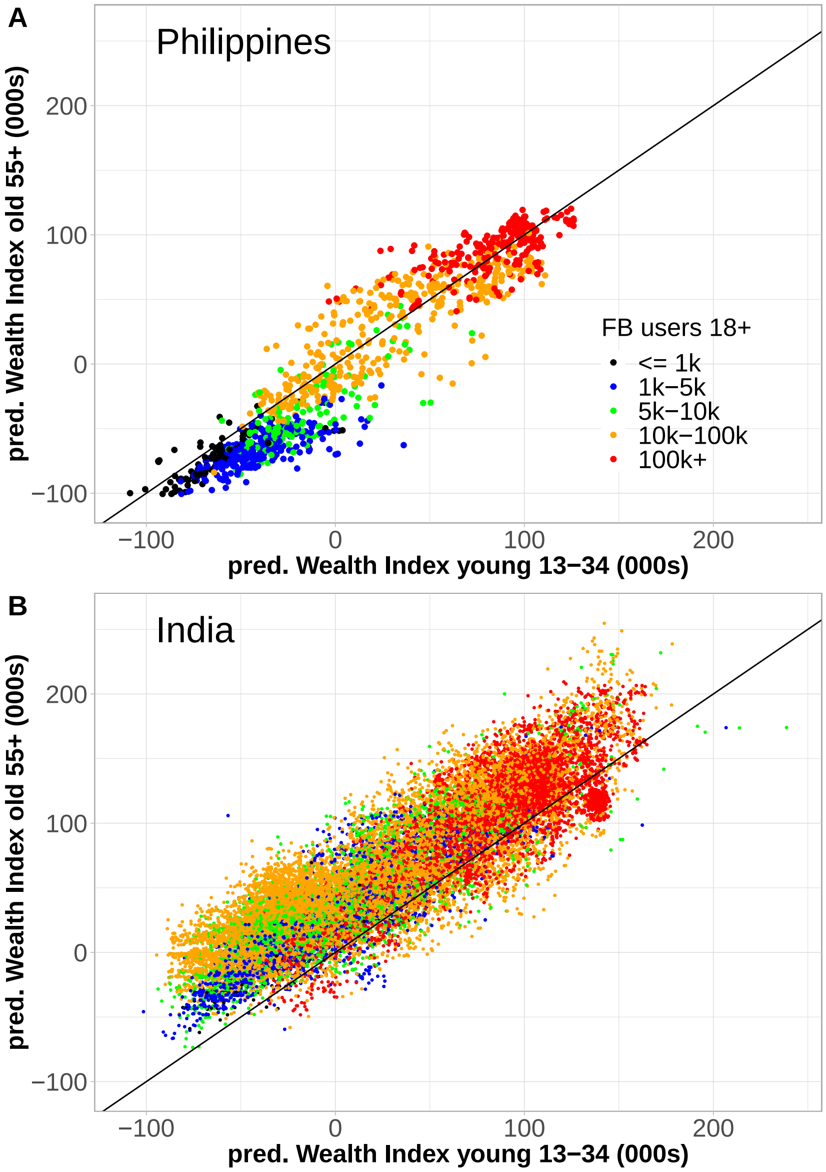

Figure S1: Age disaggregated predictions (young aged 13-34 vs. old aged 55+) for DHS clusters in the Philippines (A) and India (B); plotted line is the diagonal. Predictions were made by applying the tree model with FB features, log population density and regional indicators (fitted on data for the 18+ user group) to disaggregated data collected separately for each age group. Predictions shown for the 1,162 clusters in the Philippines and 21,688 clusters in India where the number of estimated Monthly Active Facebook users was greater than zero for both age groups.

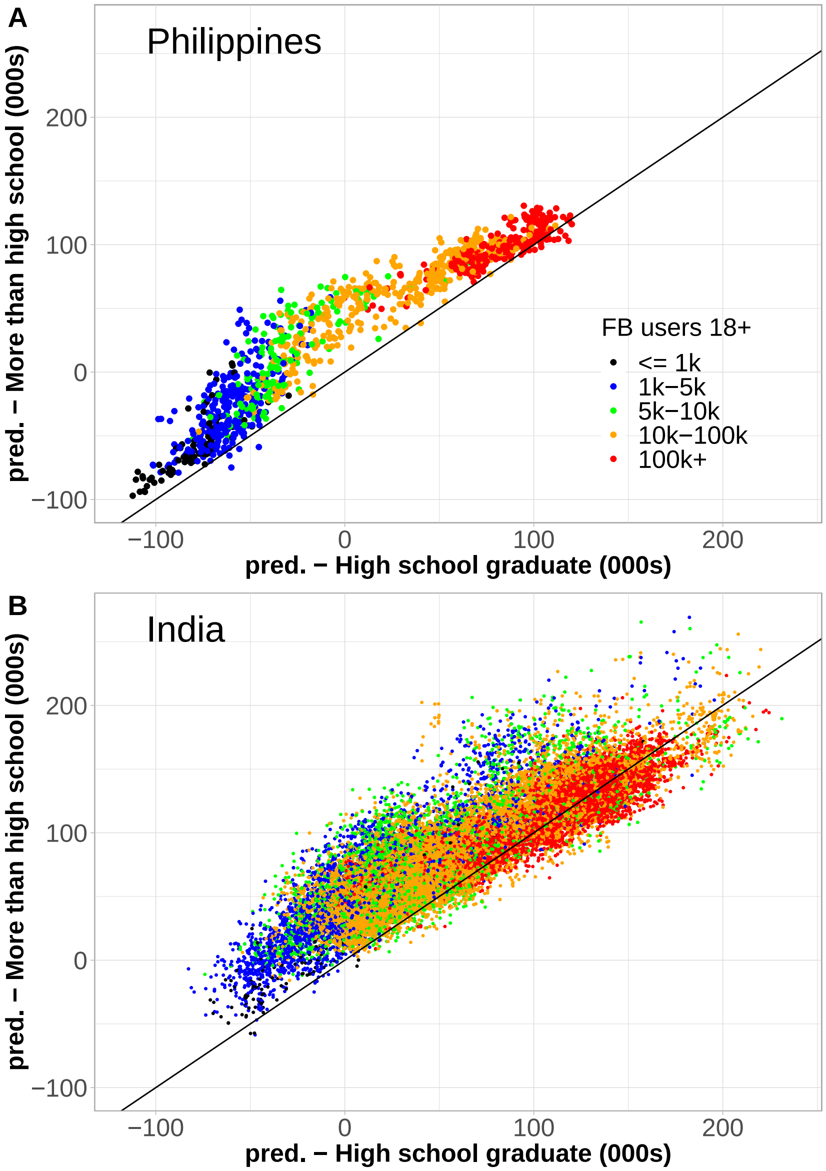

Figure S2: Education status disaggregated predictions (high school graduate vs. more than high school) for DHS clusters in the Philippines (A) and India (B); plotted line is the diagonal. Predictions were made by applying the tree model with FB features, log population density and regional indicators (fitted on data for the 18+ user group) to disaggregated data collected separately for each education status. Predictions shown for the 1,128 clusters in the Philippines and 23,789 clusters in India where the number of estimated Monthly Active Facebook users was greater than zero for both educational statuses.

1.11 Gender disaggregated Wealth Index predictions across different models

Figures S3 and S4 show plots of gender disaggregated predictions for different choices of models. Figure S3 shows the gender disaggregated predictions made using single-variable linear models using different Facebook features namely, Facebook penetration, fraction of WiFi users and fraction of iOS device users.

Figure S4 shows plots of predictions made with tree models using the previously mentioned single Facebook features in combination with the log population density and regional indicators. As explained previously, all models were fitted using data for the 18+ user demographic; gender-specific predictions were then generated by giving the model the gender specific Facebook feature as input (log population density and regional indicators were constant across genders) to make a prediction.

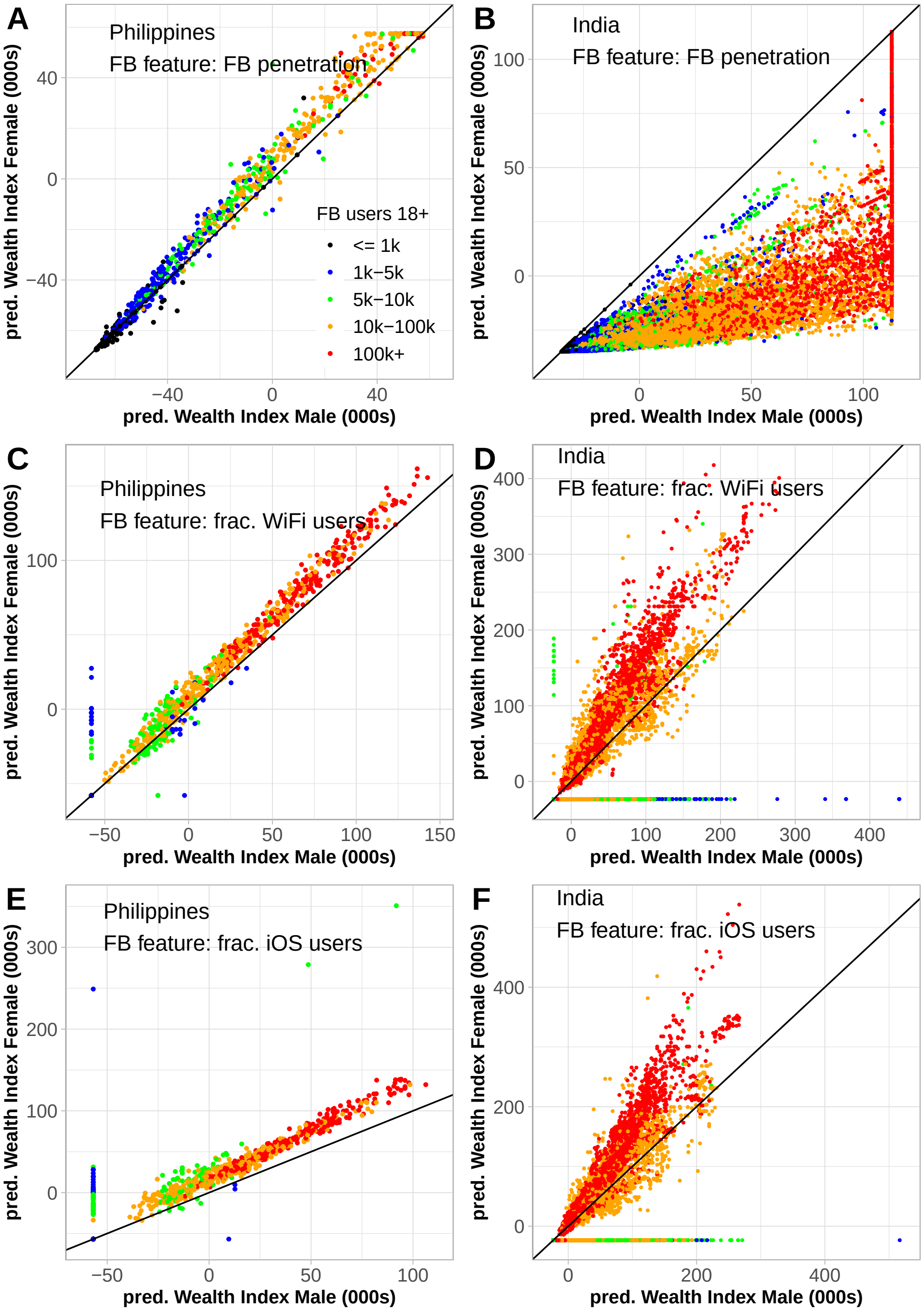

Figure S3: Gender disaggregated predictions made using linear single variable models for different choices of Facebook features. Plots for the Philippines are on the left and for India on the right. The Facebook feature used for the prediction is indicated in the plot. The plotted lines are the diagonal lines where the female and male predictions are equal. Data shown for clusters with non-zero female and male Facebook users (1,175 in the Philippines; 27,271 in India).

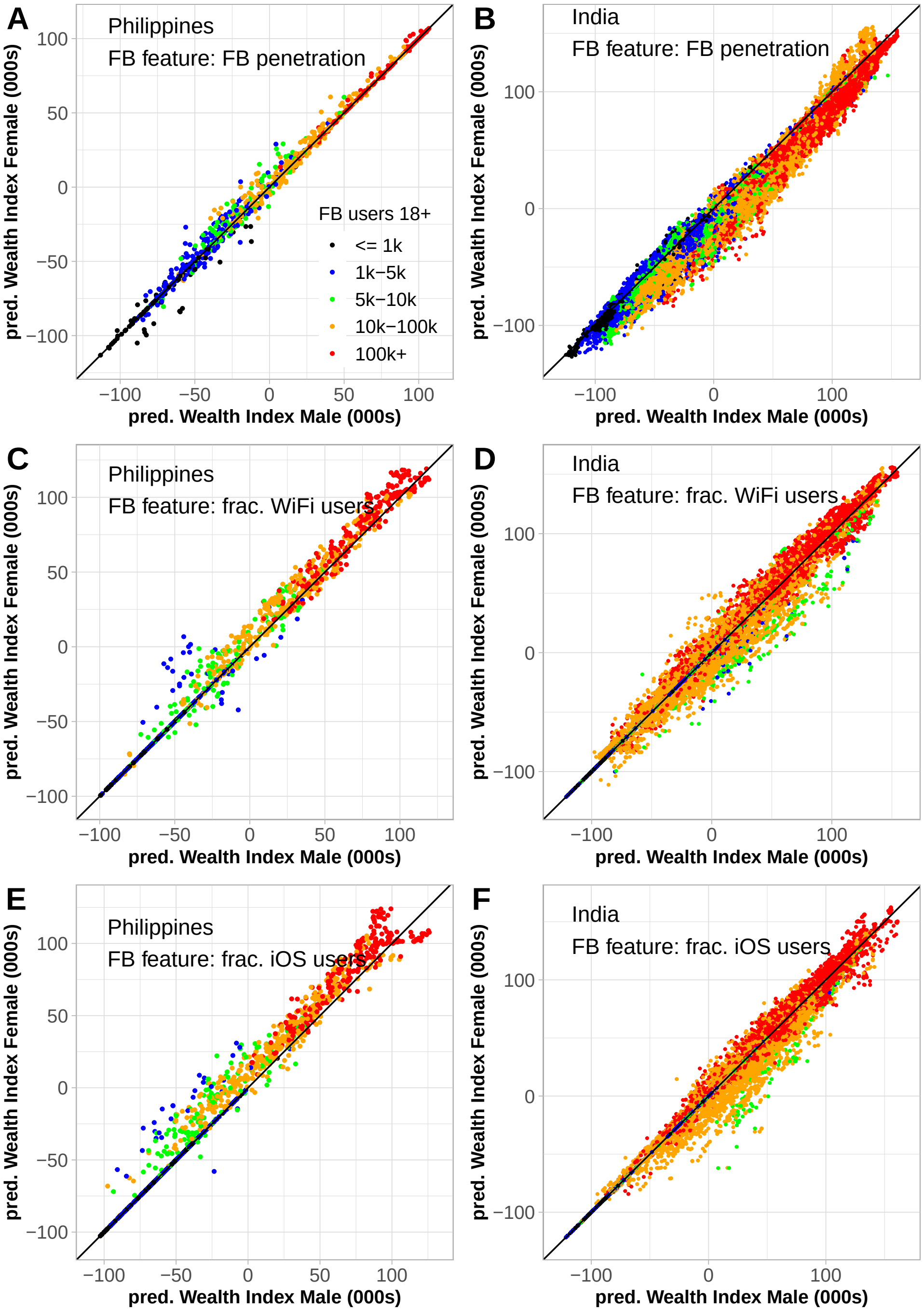

Figure S4: Gender disaggregated predictions made using tree models with log population density, regional indicators and different choices of Facebook features (indicated on the plot). Plots for the Philippines are on the left and for India on the right. The Facebook feature used for the prediction is indicated in the plot. The plotted lines are the diagonal lines where the female and male predictions are equal. Data shown for clusters with non-zero female and male Facebook users (1,175 in the Philippines; 27,271 in India).

1.12 Facebook penetration across demographic groups

The adoption of a social media platform can vary across different demographic groups. Table S15 provides data from the PEW research center surveys on the percentage of surveyed adults who say that they use Facebook (7). As seen in the table, the level of Facebook penetration varies widely across demographics. While usage is roughly equal across gender in the Philippines there is a large gender disparity in India. For both countries, Facebook usage is highest among younger and higher educated individuals.

Table S15: Percentage of adults who say they currently use Facebook by gender, age and education. Data from the PEW research center survey (7). The lower education category is below secondary level and the higher category is secondary or above.

		Age			Gender		Education	
	Total	18-29	30-49	50+	Female	Male	Less educ.	More educ.
Philippines	58	88	57	21	59	57	30	76
India	24	41	21	9	14	34	12	51

2 Sources of noise in the ground truth data

The analysis here focuses on understanding and quantifying two main sources of noise in the DHS Wealth Index. These are noise due to (i) sampling variation and (ii) spatial displacement as shown in Figure S5. The first source of noise is due to sampling as the DHS is a survey of the population and not an exhaustive enumeration, i.e. census. The second source of noise is introduced due to the displacement procedure used by the DHS whereby the data are reported at a slightly perturbed location from their true location. Analyzing these two sources of noise will enable us to estimate a best attainable performance value which is the highest R^2 that we would expect to attain from any model that predicts the Wealth Index without overfitting the noise in the data.

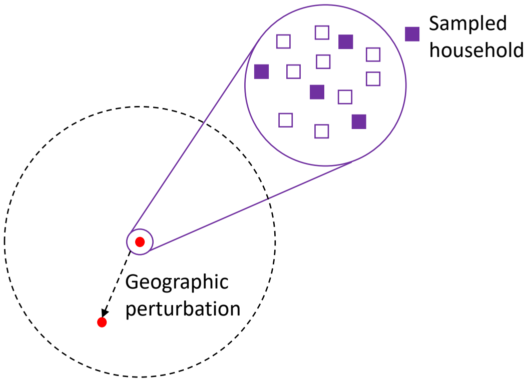

Figure S5: The two sources of noise in the DHS data: (i) sampling noise and (ii) noise due to spatial displacement. By estimating these two sources of noise, we can get an expected R^2 value for the best attainable performance for any model/data predicting the Wealth Index. Given that a model does not overfit the survey data, we would not expect it to attain R^2 exceeding this as the remaining variation in the data would be due to noise.

2.1 Noise due to sampling variation

As with any survey, the DHS uses a sample of households to collect data. This data collection procedure introduces sampling noise whereby if the survey were to be repeated again using a different sample of households, we expect to observe some variation in the resulting data. As a result, noise due to sampling is one source of noise in the observed Wealth Index values. A bootstrap approach was used to investigate the noise due to sampling. For each DHS cluster that had more than 1 surveyed household, the surveyed households were sampled uniformly with replacement to create a new DHS sample. The Wealth Index values of the households from this bootstrapped sample were then averaged to get the cluster Wealth Index. This procedure was repeated multiple times to create a total of 1,000 bootstrap samples of the cluster Wealth Index.

For each of the bootstrapped samples, the R^2 (square of the correlation value) between the bootstrapped sample and the original Wealth Index values were computed. This is a measure of how strongly the bootstrapped samples predict the original DHS survey data. In other words, if the DHS survey were to be repeated multiple times using a different sample of households each time, these are how well we would expect the data from one survey to predict the data from the other surveys. Table S16 presents a summary statistic of the results.

Table S16: Distribution of R^2 between the bootstrapped data samples and the original cluster Wealth Index from the DHS data.

	2.5%	25% (1st Quartile)	50% (Median)	Mean	75% (3rd Quartile)	97.5%
Philippines	0.950	0.953	0.955	0.955	0.956	0.958
India	0.972	0.973	0.973	0.973	0.973	0.974

2.2 Noise due to spatial displacement

Since the early 2000s the collection and reporting of GPS coordinates for survey locations has been standard practice of the DHS (8). In order to preserve survey respondent confidentiality, the survey location coordinates for each cluster of households are displaced randomly within a given radius of the true location. The resulting displaced coordinates are then reported alongside the survey data. This procedure introduces another source of noise as the data are reported at a different location from the one where they were originally collected. This can influence various types of analysis such as those that rely on distance based analysis (9). Here we are interested in understanding the noise introduced to the DHS Wealth Index from the spatial displacement procedure. To do so, the geographic displacement procedure was simulated using the DHS data. Below are the steps undertaken in the simulation. A total of 1000 simulations were run.

1. To begin, we take the cluster geographic coordinates reported by the DHS to be the true before-displacement coordinates. The coordinates were then displaced following the DHS methodology (as explained in Chapter 1 of the 8th DHS spatial analysis report (8); we used the R code provided in appendix B.1 of the report)[5]. This resulted in a new set of coordinates that are random perturbations of the original points; call these the after-displacement coordinates. For each cluster, the DHS would report its after-displacement coordinate together with its Wealth Index value; call these values Y^{disp}, as they are Wealth Index values for each cluster reported at a different location from their true location.

[5] with the exception that the displaced points were not restricted to stay within the geographic boundaries of the second administrative levels of the country. However, this should only have a negligible effect on the amount of displacement for each coordinate (10).

2. Using the before-displacement coordinates, the DHS Wealth Index data were interpolated to the locations of the after-displacement coordinates by using a nearest neighbour approach. That is, for each after-displacement location it was assigned a Wealth Index value from the before-displacement location closest to it in distance. Call these values Y^{interp} as they are the interpolated DHS Wealth Index values.

3. The R^2 (square of the correlation) between Y^{disp} and Y^{interp} was then computed. This value indicates how strongly the reported DHS Wealth Index value from the geographically displaced dataset, predicts the original Wealth Index values (interpolated from the survey data) at each location.

4. Steps 1-3 above were repeated an additional four rounds. Each time, for Step 1, the after-displacement points from the previous round were used, instead of the DHS locations, as the true locations to be displaced.

Table S17: Distribution of R^2 between the after-perturbation and before-perturbation cluster Wealth Index values for each round of the simulation.

Round	2.5%	25% (1st Quartile)	50% (Median)	Mean	75% (3rd Quartile)	97.5%
			Philippines			
0	?	?	?	?	?	?
1	0.861	0.877	0.885	0.885	0.894	0.906
2	0.864	0.882	0.890	0.890	0.898	0.912
3	0.865	0.883	0.892	0.892	0.901	0.917
4	0.866	0.885	0.894	0.894	0.904	0.918
5	0.868	0.887	0.897	0.896	0.905	0.921
			India			
0	?	?	?	?	?	?
1	0.856	0.860	0.861	0.861	0.863	0.867
2	0.854	0.859	0.861	0.861	0.862	0.866
3	0.854	0.858	0.860	0.860	0.862	0.866
4	0.854	0.857	0.860	0.860	0.862	0.866
5	0.853	0.857	0.859	0.859	0.861	0.865

Table S17 reports, for each round, the distribution of the R^2 values computed in Step 3 above across the 1,000 simulations. Notice that each round began by using the displaced coordinates from the previous round as if they were the true DHS locations. The table reports the results from rounds 1-5 of the simulation with round 1 using the reported (already perturbed) DHS cluster coordinates. Round 0 indicates the results we would have had if we performed the simulation starting with the original, unperturbed DHS cluster coordinates; the results for this round are unknown as DHS only reports the geographically displaced DHS locations. The R^2 values from round 0 are what we would like to know; these are estimates of the variation in the DHS Wealth Index that are not due to noise introduced by the spatial displacement process. However, looking at the first 5 rounds of the simulations for Philippines and India, we can see that the R^2 values do not change significantly from one round to the next. As a result, the results from round 1 should provide a reasonable estimate for round 0. Based on these simulation results, we would expect that a best performing model predicting the wealth index could achieve an R^2 value as high as 0.89 for Philippines and 0.86 for India with the remaining variation in the wealth index value being due to the spatial displacement noise.

Combining sampling and spatial displacement noise

The previous sections investigated, separately, the sources of noise from (i) sampling and (ii) spatial perturbation in the DHS wealth index. For each source of noise, simulations of the Wealth Index were generated and R^2 values capturing the strength of the correlations between these simulations and the reported Wealth Index from the DHS data were computed. We are interested to estimate the R^2 values after taking into account both sources of noise together. Since the noise introduced by sampling is independent of the noise introduced by the spatial perturbation, we would expect the R^2 when taking into account both sources of noise to be a product of the individual R^2 values. Below is a formal explanation for this intuition.

Let Y be a random variable that represents the Wealth Index value at a given location. Whenever a survey is run on a sample of households, the computed cluster Wealth Index from that survey can be represented as the random variable $Y_s = \epsilon_{s1} * Y + \epsilon_{s2}$ where ϵ_{s1} and ϵ_{s2} are the independent noise observed due to sampling. The cluster coordinates are then displaced in order to protect survey respondents' confidentiality. As a result, the Wealth Index value observed for each location can be represented by the random variable $Y_p = \epsilon_{p1} * Y_s + \epsilon_{p2}$ where ϵ_{p1} and ϵ_{p2} are the independent noise due to spatial perturbation. Based on the results of the previous sections we already know R_s^2 which measures how strongly Y_s predicts Y and R_p^2 which measures how strongly Y_p predicts Y_s. We are interested in R_{ps}^2 which measures how strongly Y_p predicts Y.

The value of R_{ps}^2 can be derived from R_s^2 and R_p^2 as follows. It is assumed that ϵ_{s1} and ϵ_{s2} are independent of Y (the sampling noise is independent of the Wealth Index value at a given location), ϵ_{p1} and ϵ_{p2} are independent of Y_s (the noise due to spatial perturbation is independent of the value of the Wealth Index in a given location) and that ϵ_{s1}, ϵ_{s2}, ϵ_{p1} and ϵ_{p2} are independent of each other (independent sources of noise). Furthermore, $E[\epsilon_{p1}] \neq 0$ and $E[\epsilon_{s1}] \neq 0$.

The following are known:

$$R_s^2 = \frac{Cov(Y, Y_s)^2}{Var(Y) * Var(Y_s)}, R_p^2 = \frac{Cov(Y_p, Y_s)^2}{Var(Y_p) * Var(Y_s)} \tag{1}$$

$$\begin{aligned} Cov(Y_p, Y) &= Cov(\epsilon_{p1} * Y_s + \epsilon_{p2}, Y) = Cov(\epsilon_{p1} * Y_s, Y) + Cov(\epsilon_{p2}, Y) \\ &= Cov(\epsilon_{p1} * Y_s, Y) = E[\epsilon_{p1} * Y_s * Y] - E[\epsilon_{p1} * Y_s] * E[Y] \\ &= E[\epsilon_{p1}](E[Y_s * Y] - E[Y_s] * E[Y]) \\ &= E[\epsilon_{p1}] * Cov(Y_s, Y) \end{aligned} \tag{2}$$

$$\begin{aligned} Cov(Y_p, Y_s) &= Cov(\epsilon_{p1} * Y_s + \epsilon_{p2}, Y_s) = Cov(\epsilon_{p1} * Y_s, Y_s) + Cov(\epsilon_{p2}, Y_s) \\ &= Cov(\epsilon_{p1} * Y_s, Y_s) = E[\epsilon_{p1} * Y_s^2] - E[\epsilon_{p1} * Y_s] * E[Y_s] \\ &= E[\epsilon_{p1}](E[Y_s^2] - E[Y_s]^2) \\ &= E[\epsilon_{p1}] * Var(Y_s) \end{aligned} \tag{3}$$

We can then compute R_{ps}^2 as follows using the above results:

$$\begin{aligned} R_{ps}^2 &= \frac{Cov(Y_p, Y)^2}{Var(Y_p) * Var(Y)} = \frac{(E[\epsilon_{p1}] * Cov(Y_s, Y))^2}{Var(Y_p) * Var(Y)} \\ &= \frac{(Cov(Y_p, Y_s) * Cov(Y_s, Y))^2}{Var(Y_s)^2 * Var(Y_p) * Var(Y)} \\ &= \frac{Cov(Y_p, Y_s)^2}{Var(Y_p) * Var(Y_s)} * \frac{Cov(Y_s, Y)^2}{Var(Y_s) * Var(Y)} \\ &= R_p^2 * R_s^2 \end{aligned} \tag{4}$$

Therefore, we have $R^2_{ps} = R^2_p * R^2_s$.

Table S18 presents the R^2 values from taking into account the two sources of noise in the DHS data due to (i) sampling and (ii) spatial perturbation. This is the distribution of R^2 values observed across the 1000 simulations and they indicate how strongly we would expect different replications of the DHS Wealth Index data to be predictive of each other. As a result, the R^2 values here can be interpreted as an estimate of the highest performance that could be expected to be achieved in predicting the DHS Wealth Index for a model that does not overfit the data, with the remaining variation in the Wealth Index being due to noise.

Table S18: Distribution of R^2 values.

	2.5%	25% (1st Quartile)	50% (Median)	Mean	75% (3rd Quartile)	97.5%
Philippines	0.821	0.837	0.845	0.845	0.853	0.867
India	0.833	0.836	0.838	0.838	0.840	0.843

References

[1] Philippine Statistics Authority, ICF: Philippines National Demographic and Health Survey 2017. Technical report, Quezon City, Philippines, and Rockville, Maryland, USA: PSA and ICF (October 2018). https://dhsprogram.com/publications/publication-FR347-DHS-Final-Reports.cfm Accessed 2019-06-20

[2] International Institute for Population Sciences (IIPS), ICF: India National Family Health Survey NFHS-4 2015-16. Technical report, Mumbai: IIPS (2017). https://dhsprogram.com/publications/publication-FR339-DHS-Final-Reports.cfm Accessed 2020-04-14

[3] Tingzon, I., Orden, A., Go, K.T., Sy, S., Sekara, V., Weber, I., Fatehkia, M., García-Herranz, M., Kim, D.: MAPPING POVERTY IN THE PHILIPPINES USING MACHINE LEARNING, SATELLITE IMAGERY, AND CROWD-SOURCED GEOSPATIAL INFORMATION. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences **XLII-4/W19**, 425–431 (2019). doi:10.5194/isprs-archives-XLII-4-W19-425-2019

[4] Jean, N., Burke, M., Xie, M., Davis, W.M., Lobell, D.B., Ermon, S.: Combining satellite imagery and machine learning to predict poverty. Science **353**(6301), 790–794 (2016). doi:10.1126/science.aaf7894. Accessed 2019-02-03

[5] Daniele Rama, Yelena Mejova, Michele Tizzoni, Kyriaki Kalimeri, Ingmar Weber: Facebook Ads as a Demographic Tool to Measure the Urban-Rural Divide. In: The Web Conference (WWW), p. (2020)

[6] PSA Releases the 2012 Municipal and City Level Poverty Estimates | Philippine Statistics Authority. https://psa.gov.ph/content/psa-releases-2012-municipal-and-city-level-poverty-estimates Accessed 2019-06-03

[7] Pew Research Center: Mobile Connectivity in Emerging Economies. Technical report (March 2019). https://www.pewinternet.org/2019/03/07/mobile-connectivity-in-emerging-economies/ Accessed 2019-06-20

[8] Perez-Haydrich, C., Warren, J.L., Burgert, C.R., Emch, M.E.: Guidelines on the Use of DHS GPS data. ICF International, Calverton, Maryland, USA (2013). http://dhsprogram.com/pubs/pdf/SAR8/SAR8.pdf

[9] Warren, J.L., Perez-Heydrich, C., Burgert, C.R., Emch, M.E.: Influence of Demographic and Health Survey Point Displacements on Distance-Based Analyses. Spatial Demography **4**(2), 155–173 (2016). doi:10.1007/s40980-015-0014-0

[10] Burgert, C.R., Colston, J., Roy, T., Zachary, B.: Geographic Displacement Procedure and Georeferenced Data Release Policy for the Demographic and Health Surveys. ICF International, Calverton, Maryland, USA (2013). http://dhsprogram.com/pubs/pdf/SAR7/SAR7.pdf