Fatehkia et al. *EPJ Data Science* (2020) 9:22
https://doi.org/10.1140/epjds/s13688-020-00235-w

EPJ Data Science
a SpringerOpen Journal

EPJ.org

REGULAR ARTICLE Open Access

Mapping socioeconomic indicators using social media advertising data

Masoomali Fatehkia[1], Isabelle Tingzon[2], Ardie Orden[2], Stephanie Sy[2], Vedran Sekara[3,4], Manuel Garcia-Herranz[3] and Ingmar Weber[1*]

*Correspondence:
iweber@hbku.edu.qa
[1]Qatar Computing Research Institute, HBKU, Doha, Qatar
Full list of author information is available at the end of the article

Abstract

The United Nations Sustainable Development Goals (SDGs) are a global consensus on the world's most pressing challenges. They come with a set of 232 indicators against which countries should regularly monitor their progress, ensuring that everyone is represented in up-to-date data that can be used to make decisions to improve people's lives. However, existing data sources to measure progress on the SDGs are often outdated or lacking appropriate disaggregation. We evaluate the value that anonymous, publicly accessible advertising data from Facebook can provide in mapping socio-economic development in two low and middle income countries, the Philippines and India. Concretely, we show that audience estimates of how many Facebook users in a given location use particular device types, such as Android vs. iOS devices, or particular connection types, such as 2G vs. 4G, provide strong signals for modeling regional variation in the Wealth Index (WI), derived from the Demographic and Health Survey (DHS). We further show that, surprisingly, the predictive power of these digital connectivity features is roughly equal at both the high and low ends of the WI spectrum. Finally we show how such data can be used to create gender-disaggregated predictions, but that these predictions only appear plausible in contexts with gender equal Facebook usage, such as the Philippines, but not in contexts with large gender Facebook gaps, such as India.

Keywords: Poverty mapping; Facebook advertising data; Remote sensing; Gender data

1 Introduction

The 2030 Agenda for Sustainable Development [1] reflects a unique commitment of the world's countries to work towards a set of Sustainable Development Goals (SDGs). These 17 ambitious goals come with a set of indicators to serve as a kind of scorecard to measure progress against. Furthermore, to aid in outcome-oriented decision making to improve lives, the data on development progress should be up-to-date and disaggregated across various dimensions, including gender.

Unfortunately, especially for those countries in most need of development, high quality and up-to-date data on the SDGs is hard to come by. For example, for SDG #1 "No poverty", of 7 South and 19 South-East Asian countries only 4 and 9 countries respectively have

Springer

poverty data collected since 2015 [2]. Furthermore, poverty data disaggregated by gender is even less available [3].

To overcome challenges related to the timeliness of data, researchers have investigated the use of non-traditional data sources for the purpose of mapping poverty levels [4]. Nightlights data from satellites have been used as a proxy of human well-being [5] and for mapping poverty globally [6, 7] and at sub-national levels [8, 9] as night light, typically linked to electricity usage, correlates with economic activity [10–12]. Other work has examined the use of daytime satellite imagery for poverty mapping [13, 14], tracking human development indicators [15] and for estimating household level poverty for rural locations based on land use information extracted from satellite images [16]. Beyond satellite imagery, mobile phone Call Detail Record (CDR) data have been used in predictive models to map aggregate population level socioeconomic characteristics [17, 18] and poverty levels in a variety of countries [18–20] as well as at the individual level for mobile phone subscribers [21]. Other research has combined satellite imagery with CDR data [22, 23] and with crowd-sourced geographic information from OpenStreetMap (OSM) [24].

In this work, we evaluate the potential value that publicly accessible, anonymous advertising data holds for the mapping of wealth and poverty. Concretely, we use data from Facebook's Marketing API on how many Facebook users match certain criteria. These audience estimates, which are traditionally used for advertising campaign planning purposes, have shown promising results for tasks such as estimating stocks of migrants [25, 26] and generating measures of digital gender inequalities [27, 28].

We test this approach for creating small area estimates (SAE) across Philippines and India. As ground truth we use an asset-based measure of poverty, the Wealth Index (WI), derived from the Demographic and Health Surveys (DHS) for each country. According to PEW surveys, 58% and 24% of adults in Philippines and India respectively use Facebook [29] which enables testing this approach in two countries with relatively high and low penetration of Facebook usage. We generate a dataset containing estimates of the proportion of Facebook users utilizing different internet connection types, mobile operating systems and device types.

We use these audience estimates to obtain insights into the spatial distribution of Facebook users, including information on (i) iOS vs. Android devices usage, or (ii) 2G vs. 4G connectivity. We demonstrate that these insights provide strong signals for the distributions of wealth and poverty.

Furthermore, these audience estimates can be disaggregated by gender, age or self-declared education level, creating opportunities for more disaggregated estimates of asset ownership and wealth. Focusing on the example of gender, we show how in countries with gender equal Facebook usage, such as the Philippines, it seems feasible to derive gender disaggregated models for poverty. However, in India, where the gender selection bias is too strong, our approach fails to provide plausible gender disaggregated poverty estimates.

2 Materials and methods

2.1 The Demographic and Health Survey (DHS)

The Demographic and Health Survey (DHS) collects survey data in many countries around the globe with the aim of providing nationally representative data on health and population. The survey consists of several types of questionnaires including a household questionnaire that collects data for the household unit in addition to individual questionnaires which collect data on eligible women and men from the surveyed households. In

addition to health related information, the household survey also collects data on household ownership of various assets such as televisions and bicycles, housing materials as well as access to water and sanitation facilities. The data on asset ownership is used to compute the Wealth Index for each surveyed household through a Principal Component Analysis (PCA) [30]. The Wealth Index is a real-valued score that takes both negative and positive values with higher values indicating higher wealth. The Wealth Index is the ground truth measure of poverty we use in this study. The data used here are from the 2017 DHS survey for Philippines [31] and the 2015-16 DHS survey for India [32].

In the reported DHS data, households are grouped into units called clusters with geographic location reported for these clusters in the form of the latitude and longitude coordinates of its center. In order to preserve respondent confidentiality, the actual coordinates undergo a spatial perturbation process before being reported; location coordinates are perturbed up to 2 km for urban clusters and up to 5 km for rural clusters with a further 1% of rural clusters displaced up to 10 km.

As the analysis here is done at the cluster level, the Wealth Index values reported for surveyed households were averaged across all households in a cluster to get an aggregated mean Wealth Index value for the cluster. Table 1 provides a summary breakdown of the survey cluster locations from the Demographic and Health Survey (DHS) for each country. Geographic coordinates were not reported for some clusters (36 in the Philippines and 131 in India). These clusters with missing coordinates could not be used in the analysis as Facebook data could not be collected for them. Some clusters had to be excluded due to sparsity of the Facebook data (8 in the Philippines and 350 in India). The row indicated in bold face in Table 1 shows the subset of clusters that were used in the analysis. Data from 1205 survey clusters in the Philippines and 28,043 in India were used in the analysis.

Tables S1 and S2 in the Additional file 1 report the summary statistics of the DHS Wealth Index distribution for different subsets of clusters in both countries. The clusters used in the analysis had on average slightly higher Wealth Index (Philippines: mean = 5599; India: mean = 1346) than among all the clusters (Philippines: mean = 4130; India: mean = 783) but roughly similar spread of the distribution (Philippines: standard deviation for all clusters = 71,532, for clusters used in the analysis = 70,626; India: standard deviation for all clusters = 79,299, for clusters in the analysis = 79,390). The excluded clusters had lower Wealth Index scores on average (Philippines: mean = –36,105; India: mean = –32,035) than the overall group of clusters.

DHS survey datasets can be accessed for research purposes from the DHS website[a] after creating an account and requesting access for the desired surveys.

Table 1 Breakdown of the data for each country for clusters with at least one surveyed household

	Philippines	India
Number of DHS clusters	1249	28,524
Clusters missing geo-location	36	131
Geo-located DHS clusters	1213	28,393
Clusters with <100 FB users 18+	8	350
Clusters with ≥100 FB users 18+	**1205**	**28,043**
Clusters with >1000 FB users 18+	1043	25,316
Median number of households surveyed (DHS)	23	21

2.2 Facebook's marketing platform

Facebook's marketing platform makes a rich array of targeting options available to advertisers. Using this platform, advertisements can be targeted based on various user characteristics including geographic location, demographics such as age and gender as well as the type of devices and networks that are used to access the social media platform. To enable advertisers with budgeting their ads, the platform provides an estimate of aggregate number of users (called the Monthly Active Users (MAU)) matching a given targeting criteria. For example, in the Philippines there are an estimated 63 million Monthly Active Users on Facebook who are aged 18+.[b]

In this study we investigate how data collected from this platform on the types of networks/devices used by the Facebook users in a given location can be used to predict the socioeconomic situation in that location. For each of the geo-located DHS clusters, we collected data on estimates of Monthly Active Users using a variety of network and device types for the 18+ Facebook user population. Since DHS cluster locations are reported as spatially perturbed latitude and longitude coordinates, we collected data for a given radius around the reported coordinates so that the original location is included in the area for which data is collected. In the Philippines we collected data for a 2 km radius around urban clusters and a 5 km radius around rural clusters. In India we used a radius of 5 km and 10 km for urban and rural clusters respectively; this was done to alleviate data sparsity issues due to the lower Facebook penetration in India. The Additional file 1, Sect. 1.2 provides more details on the choice of the radius of data collection.

Table 2 provides a list of network and device types for which data were collected. These include various Network types, mobile operating systems, high-end Apple and Samsung devices plus a variety of other device types. For the high-end devices, the Apple and Samsung devices released in the last two years prior to the data collection were targeted.[c] For the list of network/device types, features were generated by computing the fraction of Facebook users who used that network/device type to access Facebook. These are the features used in the predictive models to predict the Wealth Index. In addition to the above-mentioned features, we also include the Facebook penetration as a feature in the model. This variable is the number of Monthly Active Facebook users aged 18+ as a fraction of the total population in a given cluster location where the cluster population was computed using high-resolution population estimates from WorldPop [33].

For clusters where the number of estimated Monthly Active Facebook users exceeded the estimated offline population, the Facebook penetration values were set to 1. There are two possible reasons why the Facebook user population may exceed the offline population. First, the offline population of a cluster may be under-counted as we used high-resolution gridded population estimates to calculate the cluster population. In a study evaluating the methodology that was used to generate these population estimates [34], relative Root Mean Squared Error (as a percentage of the mean population size of the respective census units) ranging from 39% in Cambodia to 91% in Kenya were reported when comparing the high-resolution population estimates aggregated to the level of census units to census populations. Second, the Facebook user population may be over-counted as about 10% of Facebook accounts are estimated to be duplicate accounts (such as pet accounts, duplicate for-my-family vs. for-my-private friends accounts) and some fraction of fake accounts [35].

Table 2 List of features derived from the Facebook advertising audience estimate data. All features, with the exception of Facebook penetration, are the fraction of Facebook users in the targeted location who use a given network/device type to access Facebook. All data are for users aged 18+. The Facebook penetration is the number of users divided by the total population of the location; where there were more estimated users than the estimated population the value was capped at 1. Note that according to the Facebook audience estimates, of all users who use a smartphone, the percentage who do not use either of the three specified Mobile OS types (Android, iOS, Windows) are 61% (India) and 51% (Philippines); of all users, the percentage who do not use either of the four specified network types (2G, 3G, 4G, WiFi) to access Facebook are 25% (India) and 37% (Philippines)

Feature type	Feature Description	Correlation with cluster Wealth Index	
		Philippines	India
	Facebook penetration	0.664	0.555
Network access	2G Network	0.115	0.346
	3G Network	–0.378	0.296
	4G Network	0.693	0.003
	WiFi	0.740	0.524
Mobile OS	Android	0.449	0.510
	iOS	0.663	0.567
	Windows phones	0.387	0.357
High-end phones	Apple iPhone X	0.573	0.435
	Apple iPhone X/8/8 Plus	0.628	0.454
	Samsung Galaxy phone S9+	0.540	0.391
	Samsung Galaxy phone S8/S8+/S9/S9+	0.643	0.499
	Samsung Galaxy phone S8/S8+/S9/S9+ or Apple iPhone X/8/8 Plus	0.669	0.524
Other device types	All mobile devices	0.264	–0.061
	Feature phones	0.096	0.163
	Smartphone and tablets	0.217	–0.072
	Tablet	0.492	0.423
	Cherry mobile	–0.275	–
	VIVO mobile devices	0.539	0.024
	Huawei mobile devices	0.534	0.292
	Oppo mobile devices	0.499	0.129
	Oppo/VIVO/Cherry devices	0.184	0.013
	Samsung Android devices	0.123	0.087

For locations and targeting criteria with low number of users, the marketing platform does not return estimates of monthly active users below 1000. For such instances, to alleviate data sparsity, we attempted to estimate the number of users following the approach in [36] which gives an estimate in the hundreds $(0, 100, 200, \ldots, 900)$ for such locations. Using this data augmentation approach resulted in a small improvement in modeling performance. Details of this data augmentation approach as well as its effect on model performance are explained in the Additional file 1, Sect. 1.6.

The data used in the main analysis is for the age 18+ user demographic on Facebook. Data were also collected for different age brackets, by gender and by self-declared education status to test the potential to produce demographically disaggregated estimates. With the exception of the age-disaggregated data collections, all other data collections (disaggregated by gender/education) were for the 18+ age group. Data for the Philippines were collected over the period March-April 2019 and data for India were collected over the period June-September 2019. Data collection was done using 'pySocialWatcher,'[d] a Python based wrapper library that automates the data collection process by using Facebook's Marketing Application Programming Interface (API) [37].

2.3 Population data

Population data were acquired for the DHS cluster locations using population estimates released by Worldpop [33, 38]. Worldpop provides high-resolution population estimates for countries around the world. The population data are provided for an approximately 100 m resolution grid of the entire country for the year 2015. For each cluster, the estimated population living in that cluster was computed by adding together the population counts for all grid cells that fell within a given radius of the cluster coordinates, matching the radius for which the Facebook data were collected. The population data were used to compute (i) the Facebook penetration and (ii) the log of population density for each cluster. These variables were used as predictive features in the models predicting the Wealth Index.

2.4 Regional indicators

In addition to the Facebook features and population density, regional indicator variables were used as additional features in the models. These are binary variables that indicate whether a given DHS cluster falls within a given administrative region in the country. We used the level 1 administrative division that were reported in the DHS data. Including these features allows a model to account for regional level variations. There were a total of 17 administrative regions in the Philippines and 36 in India. As both India and the Philippines are large countries, different regions may exhibit different dynamics of poverty. The addition of regional indicator variables can enable models to account for possible region specific trends in the data. Generally, the inclusion of the regional indicator variables resulted in improved model performance.

2.5 Models for predicting the Wealth Index

We evaluated the performance of (i) linear regression models selected using LASSO and (ii) tree based regression models to predict the Wealth Index using data from the available set of covariates. The distribution of Wealth Index for the clusters used in the analysis is reported in Tables S1 and S2 in the Additional file 1. The Wealth Index is a real-valued score ranging from negative to positive values with higher values being better. The linear LASSO models were fitted using 'glmnet'[e] and the tree models were fitted using 'gbm'[f] package in the R programming language; the 'gbm' package fits regression trees using gradient boosting. Models were fitted and evaluated separately for each country using data from that country.

Model parameters were tuned using cross validation. For the tree models, the optimal number of trees was chosen through cross validation for up to a maximum of 5000 trees. Each model was fit and evaluated using 10-fold cross validation. The predictions over the cross validation folds were then used to evaluate the cross-validated R^2 which captures the proportion of the variation in the Wealth Index that is explained by the model predictions. In addition to R^2 values, we also compute and report the Root Mean Squared Error (RMSE) metric for all models using the cross-validated predictions.

3 Results

3.1 Performance of models for estimating the Wealth Index

Our general approach of modeling poverty in this work is one of supervised machine learning or, more specifically, of building regression models. For this we use the Wealth Index (WI) of a given DHS survey location (DHS cluster) as ground truth and train a model

that estimates the WI. The features that we use for this task include a number of Facebook-derived features. Concretely, for all geo-located DHS survey locations, data was collected on estimates of total Facebook users as well as the number of Facebook users accessing Facebook using different types of Networks, mobile operating systems, high-end devices as well as a variety of other device types. Using this data we then compute the proportion of Facebook users in a particular location who utilize a given network/device type. These features were used as input variables to build models for the DHS WI. A complete list of Facebook derived features used in the models as well as their correlation with the DHS Wealth Index is provided in Table 2. In addition to the Facebook features, data was collected on other variables such as population density as well as the Wealth Index and poverty incidence data from past surveys. These additional data were used to predict the Wealth Index both individually as baseline models and in combination with the Facebook features.

As a preliminary step, the correlations in Table 2 demonstrate that features pertaining to the overall Facebook adoption, access to WiFi networks, iOS and high-end device types are most strongly correlated with the Wealth Index. Additional file 1, Table S5 reports the performance of the various models that were fitted to predict the Wealth Index using data from the Facebook features in combination with other covariates, namely log population density and regional indicator variables which indicate the administrative region to which a given location belongs. We experimented both with linear models (LASSO) [39] as well as regression trees [40]. 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. The cross-validated R^2 is reported for all models.

Table 3 reports the performance of regression tree models using various combinations of the predictive features. A full table of results can be found in the Additional file 1, Table S5. As shown in Table 3, regression tree models using Facebook features achieve an R^2 of 0.608 for Philippines and 0.563 for India respectively. This further improves when incorporating the regional indicators and log population density variables into the model: R^2 of 0.627 for Philippines and 0.691 for India. The result for Philippines is comparable to the R^2 of 0.63 in prior work [24] that predicts the DHS Wealth Index using features extracted from day-time satellite imagery, night-time light intensities and crowd-sourced geospatial information from OpenStreetMap. We leave the combination with additional features for future work, as those do not easily permit a disaggregation by gender or other demographic attributes.

Note that our models achieve an improvement over simple baseline models (reported in Additional file 1, Table S4) such as using past DHS surveys (Philippines (2008 DHS): R^2

Table 3 Performance of regression tree models using various features to predict the DHS Wealth Index for Philippines and India. The table reports cross-validated R^2 and RMSE values

Model features					
Interpolated DHS Wealth Index		X			X
Facebook features			X	X	X
Log population density				X	X
Regional indicators				X	X
Philippines ($N = 1205$)	R^2	0.480	0.608	0.627	0.630
	RMSE	50,983	44,218	43,099	42,965
India ($N = 28{,}043$)	R^2	0.652	0.563	0.691	0.728
	RMSE	46,810	52,502	44,149	41,394

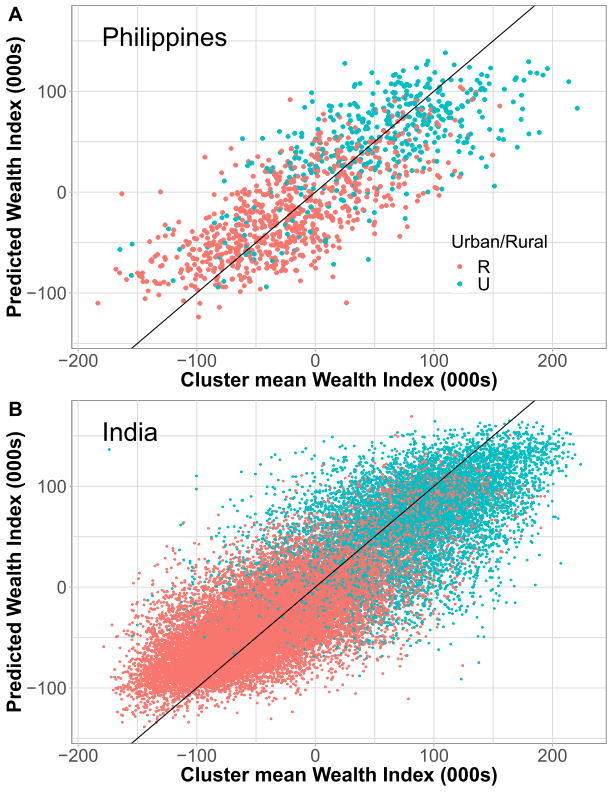

Figure 1 Scatter plot of DHS Wealth Index against predictions. Scatter plot of ground truth Wealth Index from the DHS against the predictions from the 10-fold cross-validation for (**A**) Philippines (1205 DHS clusters) and (**B**) India (28,043 DHS clusters); all values are in thousands. Plotted line is the diagonal. Predictions are for the regression tree models with Facebook features, log population density and regional indicator variables for each country

of 0.444; past DHS surveys for India were not geo-located), regional indicator variables (Philippines: R^2 of 0.378; India: R^2 of 0.334) or log population density (Philippines: R^2 of 0.448, India: R^2 of 0.180). Figure 1 shows a scatter plot of the predicted Wealth Index against the survey data for DHS cluster locations in the Philippines (panel A) and India (panel B).

3.2 Considering sources of noise in the ground truth data

To put the reported results into perspective, it is also good to have a sense of the best imaginable performance one can expect to attain regardless of the data/model used. As the Wealth Index is a noisy ground truth measure, even the best model (which does not overfit the data) can not achieve a perfect R^2 of 1.0. Put simply, if one was to collect ground truth data for the same locations independently twice on the same day, then the two measures of ground truth would not be in perfect agreement with each other.

The two main sources of noise in the measurement of the DHS Wealth Index are the noise due to (i) sampling variation and (ii) the geographic perturbation of survey geo-locations. 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. Using bootstrap and simulation

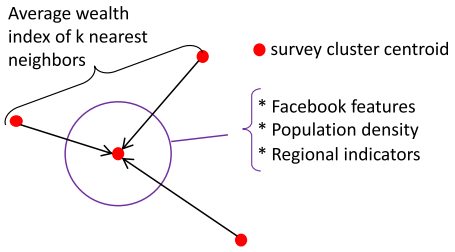

Figure 2 Combining interpolated ground truth data with other features. The various types of features that can be combined in predictive models including Wealth Index values interpolated from the survey itself as well data from Facebook features, log population density and regional indicators

methods we estimate these sources of noise in order to establish the best achievable R^2 (details provided in the Additional file 1, Sect. 2). Based on this analysis we establish an expected best model performance as R^2 of 0.85 and 0.84 for Philippines and India respectively (Additional file 1, Table S18). Note that these are not strict upper bounds as overfit models that simply output the training data as predictions could trivially achieve an R^2 of 1.0.

3.3 Interpolating Wealth Index from spatial neighbours

The models reported above use covariates from outside the DHS survey such as Facebook features for predicting the Wealth Index. In practical settings one could use the data on Wealth Index from the DHS survey itself in combination with external data sources in order to create poverty estimates for locations throughout a country [41]. To test this approach, we interpolated the DHS Wealth Index values using a nearest neighbour approach where for each survey location, the average Wealth Index values of the survey locations closest to it were computed. See Fig. 2 for an illustration. These interpolated values were then used as features in the regression tree models and combined with the other variables. Table 3 demonstrates the results. The model using only the interpolated DHS Wealth Index values attains a cross-validated R^2 of 0.480 for Philippines and R^2 of 0.652 for India. These results indicate how well we would expect to be able to estimate the Wealth Index for non-surveyed locations if we simply used interpolated values from the nearest surveyed locations. The model performance improves when the interpolated DHS Wealth Index values are combined with the additional Facebook, population density and regional indicator variables: R^2 of 0.630 for Philippines and R^2 of 0.728 for India. Detailed results can be found in Additional file 1, Table S7. Overall, these findings suggest that the predictive performance is best when combining interpolated poverty estimates from the survey together with other covariates so that in practical settings one can augment traditional survey data with non-traditional data sources to achieve the best results.

3.4 Model performance across the distribution of Wealth Index

The previous results demonstrate that Facebook data provides a signal on the distribution of asset-based wealth and poverty. However, beyond simply maximizing the overall model performance, it is also important to respect the SDGs vision to "leave no one behind". In other words, a model that works well in general but does not work well for the poorest elements of a population might not be desirable. Figure 3 demonstrates, for both countries, the mean absolute rank difference (as a fraction of the total number of clusters) between the model predictions and the ground truth DHS Wealth Index for each decile of the Wealth Index. For each cluster the rank difference is the difference between its ranking

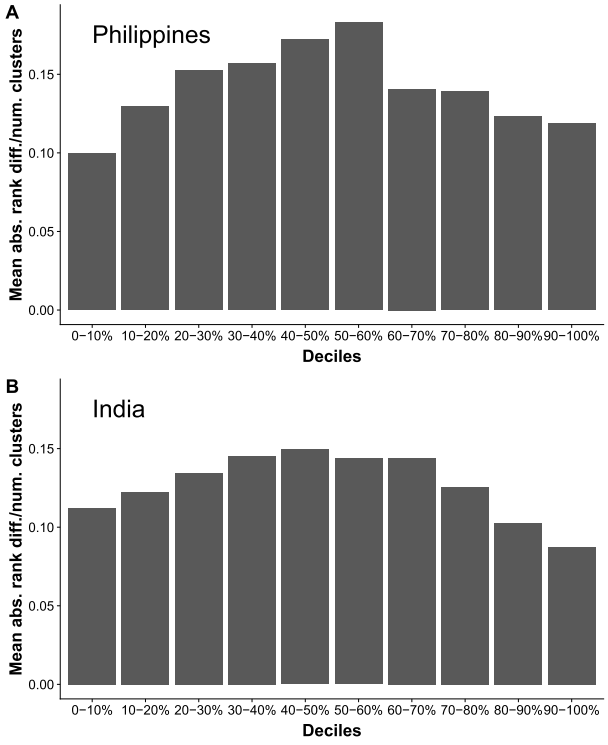

Figure 3 Performance of the models across the distribution of Wealth Index. Performance of the models for Philippines (**A**) and India (**B**) across the distribution of Wealth Index. For each cluster, the difference between its ranking when ordered according to the DHS wealth index versus when ordered according to predicted Wealth Index was calculated. The data was then split into deciles according to the DHS Wealth Index and the average absolute rank difference computed for each. The y-axis shows the mean absolute rank difference for clusters in that decile as a fraction of the total number of clusters in the dataset. The figures are based on cross-validated predictions from the tree model with Facebook features, log population density and regional indicators

when ordered according to its DHS Wealth Index or when ordered according to its predicted Wealth Index, so a lower value indicates better model performance. As can be seen in the figures, the rank difference tends to be lower for both the lowest (= poorest) and highest (= richest) deciles. Though this effect can be partly explained due to the one-sided nature of errors at the boundaries—it is impossible to under-predict the rank of the poorest location, or to over-predict the rank of the richest location—the results still provide evidence that models derived from Facebook data do not break down at the extreme ends of the wealth distribution.

3.5 Demographically disaggregated predictions

A further aspect of "leaving no one behind" relates to reducing poverty for men, women and children of all ages [42]. However, monitoring the progress of such a goal necessitates the availability of demographically disaggregated poverty maps. A potential advantage of social media data is the ability to acquire data on user groups broken down by various demographic traits such as gender, age and education levels. Such data could then be used in the models to make demographically disaggregated poverty predictions. We test this approach here by applying the models fitted above to demographically disaggregated social media data in order to make predictions for specific demographic groups. That is, here

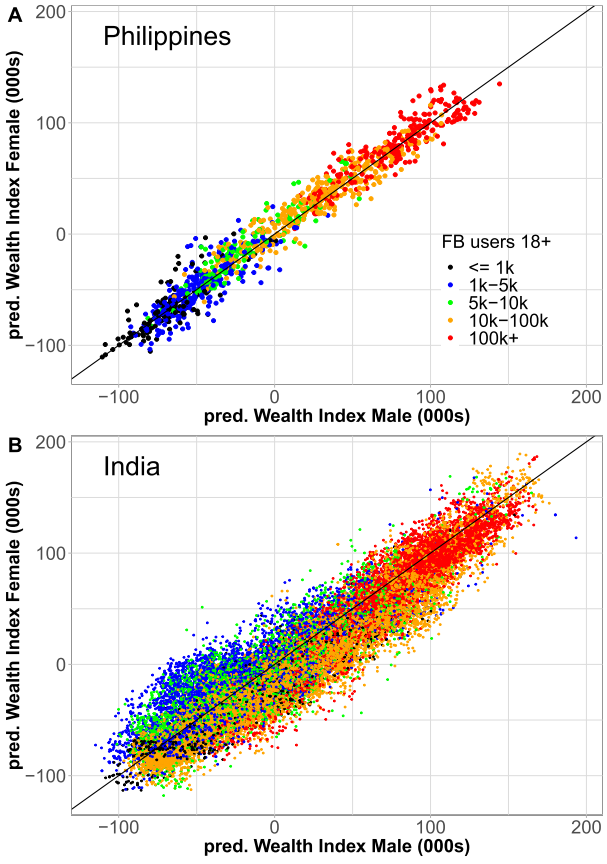

Figure 4 Gender disaggregated predictions. Gender disaggregated predictions (male vs. female) for DHS clusters in the Philippines (**A**) and India (**B**); plotted line is the diagonal. Predictions were made by applying the tree model with Facebook features, log population density and regional indicator variables (fitted on data for the 18+ user group) to disaggregated data collected separately for each gender. Predictions shown for the 1175 clusters in the Philippines and 27,271 clusters in India where there was an estimated non-zero Facebook user population for both males and females aged 18+

we apply the models that were trained in a gender oblivious setting to data, i.e., Facebook audience estimates, that were collected for women and men separately.

Figure 4 shows plots of the gender disaggregated predictions (female vs. male) that were made for Philippines (Panel A) and India (Panel B). The model used to make the predictions in Fig. 4 is the model combining Facebook features, log population density and regional indicators that was fitted using data for the 18+ user demographic. See the Additional file 1, Figures S3 and S4 for gender disaggregated Wealth Index predicted using different choices of models. Whereas for the Philippines all choices of model give similar overall trends, for India the model choice greatly affects the results.

In order to create the gender disaggregated predictions, the gender specific Facebook features were input to the model (for all features such as the fraction of users with iOS devices, the fraction of female users with that device/network type was input to the model to generate the female Wealth Index predictions and likewise for the male Wealth Index predictions). For the Facebook penetration variable, the gender specific Facebook penetration was computed by assuming an equal gender split in the offline population of the clusters. The gender-specific Facebook penetration was then the number of female/male

Facebook users in the cluster divided by half the offline population of the cluster. Note that the population density and regional indicator variables were the same for both genders as these represent the location specific characteristics. Similar plots for age and education can be found in the Additional file 1, Figures S1 and S2.

As survey data, such as the data on asset-ownership from the DHS, includes *household level* information, rather than individual level information, a common approach is to disaggregate poverty measures by the gender of head of household in an effort to obtain gender disaggregated poverty estimates. However, comparison of male and female headed households is unlikely to provide an accurate picture of gender poverty gaps [43, 44]. The fact that no gender disaggregated poverty estimates exist both motivates our attempts to create these, but also limits the possibility for validating estimates.

Despite the lack of ground truth, some observations concerning our predictions can be noted. In the Philippines (Fig. 4 Panel A), the predicted male and female values are generally close to each other, i.e. close to the diagonal line, with slightly higher predictions for women than for men on average. This result may be plausible as the Philippines has small gender gaps in economic participation, even *exceeding* gender parity on senior, managerial, professional and tech work [45].

In India (Fig. 4 Panel B), the predictions are also close to the diagonal line with, on average, slightly higher predictions for men than for women. However, in India the gender disparities in economic opportunities are considerable [45], making these predictions implausible. Moreover, unlike in the Philippines, Facebook usage in India is much lower among women than men (According to PEW surveys [29], 14% of women and 34% of men in India use Facebook, compared to 59% of women and 57% of men in the Philippines; see Additional file 1, Table S15 for more details). This combined with the low overall Facebook penetration in the country, means that the sample of female Facebook users in India is likely to be biased towards women from the upper socioeconomic strata. Hence the case of India presents a major caveat of our approach with regards to representation of different demographic groups on the social media platform. A similar observation concerning the case of fewer but higher status women being active on social networks in less gender equal countries was also reported by other researchers [46].

On the positive side, the predictions for the Philippines, where for most locations the prediction for men and women are similar, are plausible. According to data from the Global Gender Gap Report,[g] women outnumber men in the Philippines as both "legislators, senior officials and managers" (f/m ratio 1.06) and as "professional and technical workers" (f/m ratio 1.39). The same report ranks the Philippines 8 out of 149 countries in terms of gender gaps.

4 Discussion

Our results demonstrate the potential of social media advertising data from Facebook's marketing platform to capture geographic variations in wealth and poverty levels. The analysis indicates that the types of devices and network connections accessed by the Facebook user population act as proxies for socioeconomic status of a given location. Such an approach can be used to estimate the levels of socioeconomic well-being at high spatial resolutions. The results from India where just about a quarter of the population use Facebook suggest that this approach could be useful even in countries with low penetration of Facebook users.

The analysis here looked at data from a single snapshot. Furthermore, the DHS ground truth data was not aligned in terms of collection period with the Facebook data. For the purpose of long term monitoring of poverty for the Sustainable Development Goals, it is important to understand the temporal stability of the models as well as whether and how changes in the device types accessed by Facebook users reflects changes in the socioeconomic situation of a particular location. This would be a potential area for future exploration as more data, both in terms of ground truth and in terms of social media, becomes available.

Beyond aggregate estimates of the geographic variation in socioeconomic well-being, the potential to use demographically disaggregated social media data to create disaggregated estimates such as by gender, age and education was explored as well. While it was not possible to directly validate these estimates due to lack of ground truth, as shown by the results for Philippines and India, one must take into account potential selection biases for different demographic groups when interpreting such predictions.

Selection bias also affected a small number of DHS clusters that were dropped from the analysis due to data sparsity (see Sect. 2.1 and Tables S1 and S2 in the Additional file 1). These clusters had lower than average Wealth Index.

Especially for sparsely populated areas, social media data could be further combined with data from other sources, in particular satellite data, for the purpose of monitoring socioeconomic well-being. Such an approach can combine the strengths of different data sources to boost predictive accuracy. In particular, it combines satellite data's spatial resolution and truly global coverage with Facebook's data's demographic disaggregation capabilities and the direct links to a particular type of asset ownership—a mobile phone. Such a combination provides an interesting avenue for exploration in future work.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1140/epjds/s13688-020-00235-w.

Additional file 1. Supplementary Information (PDF 9.6 MB)

Funding
The publication of this article was funded by the Qatar National Library.

Availability of data and materials
The datasets supporting the conclusions of this article are available in the Harvard Dataverse repository https://doi.org/10.7910/DVN/7I70FA. Data from the Demographic and Health Survey (DHS), including the location information for the DHS clusters, can be requested at https://dhsprogram.com/data/new-user-registration.cfm.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
MF, IT, AO, SS, VS, MG-H and IW designed research; MF and IW performed research; MF analyzed data; MF, IT, AO, VS and IW wrote the paper. All authors read and approved the final manuscript.

Author details
[1]Qatar Computing Research Institute, HBKU, Doha, Qatar. [2]Thinking Machines, Manila, Philippines. [3]UNICEF Innovation, New York, USA. [4]Department of Computer Science, IT University, Copenhagen, Denmark.

Endnotes
[a] List of available survey data: https://dhsprogram.com/data/available-datasets.cfm
[b] As of January 23, 2020.
[c] Samsung Galaxy S10 does not appear in this list as it was not available for targeting on Facebook's marketing platform by 31 August 2019, in time for the data collection.

[d] https://github.com/maraujo/pySocialWatcher
[e] https://cran.r-project.org/web/packages/glmnet/glmnet.pdf
[f] https://cran.r-project.org/web/packages/gbm/gbm.pdf
[g] http://reports.weforum.org/global-gender-gap-report-2018/data-explorer/#economy=PHL

Publisher's Note

Received: 26 January 2020 Accepted: 18 June 2020 Published online: 29 July 2020

References

1. United Nations: (2015) Transforming our World: The 2030 Agenda for Sustainable Development. Technical report. https://sustainabledevelopment.un.org/post2015/transformingourworld/publication. Accessed 2019-09-29
2. World Bank (2019) PovcalNet. http://iresearch.worldbank.org/PovcalNet/povOnDemand.aspx. Accessed 2019-09-29
3. Open data Watch: (2019) Bridging Gender Data Gaps in Africa. Technical report. https://opendatawatch.com/publications/bridging-gender-data-gaps-in-africa/. Accessed 2019-09-29
4. Blumenstock JE (2016) Fighting poverty with data. Science 353(6301):753–754. https://doi.org/10.1126/science.aah5217. Accessed 2019-06-25
5. Ghosh T, Anderson SJ, Elvidge CD, Sutton PC (2013) Using nighttime satellite imagery as a proxy measure of human well-being. Sustainability 5(12):4988–5019. https://doi.org/10.3390/su5124988. Accessed 2019-05-01
6. Elvidge CD, Sutton PC, Ghosh T, Tuttle BT, Baugh KE, Bhaduri B, Bright E (2009) A global poverty map derived from satellite data. Comput Geosci 35(8):1652–1660. https://doi.org/10.1016/j.cageo.2009.01.009. Accessed 2019-06-25
7. Pinkovskiy M, Sala-i-Martin X (2014) Lights, camera, ... income!: estimating poverty using national accounts, survey means, and lights. Working Paper 19831, National Bureau of Economic Research. https://doi.org/10.3386/w19831. http://www.nber.org/papers/w19831. Accessed 2019-05-01
8. Noor AM, Alegana VA, Gething PW, Tatem AJ, Snow RW (2008) Using remotely sensed night-time light as a proxy for poverty in Africa. Popul Health Metr 6(1):5. https://doi.org/10.1186/1478-7954-6-5. Accessed 2019-06-25
9. Wang W, Cheng H, Zhang L (2012) Poverty assessment using DMSP/OLS night-time light satellite imagery at a provincial scale in China. Adv Space Res 49(8):1253–1264. https://doi.org/10.1016/j.asr.2012.01.025. Accessed 2019-06-25
10. Mellander C, Lobo J, Stolarick K, Matheson Z (2015) Night-time light data: a good proxy measure for economic activity?. PLoS ONE 10(10):0139779. https://doi.org/10.1371/journal.pone.0139779. Accessed 2019-06-25
11. Chen X, Nordhaus WD (2011) Using luminosity data as a proxy for economic statistics. Proc Natl Acad Sci 108(21):8589–8594. https://doi.org/10.1073/pnas.1017031108. Accessed 2019-05-01
12. Henderson JV, Storeygard A, Weil DN (2012) Measuring economic growth from outer space. Am Econ Rev 102(2):994–1028. https://doi.org/10.1257/aer.102.2.994. Accessed 2019-06-25
13. Jean N, Burke M, Xie M, Davis WM, Lobell DB, Ermon S (2016) Combining satellite imagery and machine learning to predict poverty. Science 353(6301):790–794. https://doi.org/10.1126/science.aaf7894. Accessed 2019-02-03
14. Engstrom R, Hersh JS, Newhouse DL (2017) Poverty from space: using high-resolution satellite imagery for estimating economic well-being. Technical Report WPS8284, The World Bank. http://documents.worldbank.org/curated/en/610771513691888412/Poverty-from-space-using-high-resolution-satellite-imagery-for-estimating-economic-well-being. Accessed 2019-05-01
15. Head A, Manguin M, Tran N, Blumenstock JE (2017) Can human development be measured with satellite imagery? In: Proceedings of the ninth international conference on information and communication technologies and development. ICTD '17. ACM, New York, pp 8–1811. https://doi.org/10.1145/3136560.3136576. event-place: Lahore, Pakistan. Accessed 2019-05-01
16. Watmough GR, Marcinko CLJ, Sullivan C, Tschirhart K, Mutuo PK, Palm CA, Svenning J-C (2019) Socioecologically informed use of remote sensing data to predict rural household poverty. Proc Natl Acad Sci 116(4):1213–1218. https://doi.org/10.1073/pnas.1812969116. Accessed 2019-05-01
17. Soto V, Frias-Martinez V, Virseda J, Frias-Martinez E (2011) Prediction of socioeconomic levels using cell phone records. In: Konstan JA, Conejo R, Marzo JL, Oliver N (eds) User modeling, adaption and personalization. Lecture notes in computer science. Springer, Berlin, pp 377–388
18. Fernando L, Surendra A, Lokanathan S, Gomez T (2018) Predicting population-level socio-economic characteristics using Call Detail Records (CDRs) in Sri Lanka. In: Proceedings of the fourth international workshop on data science for macro-modeling with financial and economic datasets. DSMM'18. ACM, New York, pp 1–1112. https://doi.org/10.1145/3220547.3220549. event-place: Houston, TX, USA. Accessed 2019-05-01
19. Njuguna C, McSharry P (2017) Constructing spatiotemporal poverty indices from big data. J Bus Res 70:318–327. https://doi.org/10.1016/j.jbusres.2016.08.005. Accessed 2019-06-25
20. Hernandez M, Hong L, Frias-Martinez V, Frias-Martinez E (2017) Estimating poverty using cell phone data: evidence from Guatemala. Technical report, The World Bank. https://doi.org/10.1596/1813-9450-7969. Accessed 2019-06-25
21. Blumenstock J, Cadamuro G, On R (2015) Predicting poverty and wealth from mobile phone metadata. Science 350(6264):1073–1076. https://doi.org/10.1126/science.aac4420. Accessed 2019-02-05
22. Pokhriyal N, Jacques DC (2017) Combining disparate data sources for improved poverty prediction and mapping. Proc Natl Acad Sci 114(46):9783–9792. https://doi.org/10.1073/pnas.1700319114. Accessed 2019-05-01
23. Steele JE, Sundsøy PR, Pezzulo C, Alegana VA, Bird TJ, Blumenstock J, Bjelland J, Engø-Monsen K, de Montjoye YA, Iqbal AM, Hadiuzzaman KN, Lu X, Wetter E, Tatem AJ, Bengtsson L (2017) Mapping poverty using mobile phone and satellite data. J R Soc Interface 14(127):20160690. https://doi.org/10.1098/rsif.2016.0690. Accessed 2019-05-01
24. Tingzon I, Orden A, Go KT, Sy S, Sekara V, Weber I, Fatehkia M, García-Herranz M, Kim D (2019) Mapping poverty in the Philippines using machine learning, satellite imagery, and crowd-sourced geospatial information. In: ISPRS—international archives of the photogrammetry, remote sensing and spatial information sciences XLII-4/W19, pp 425–431. https://doi.org/10.5194/isprs-archives-XLII-4-W19-425-2019

25. Zagheni E, Weber I, Gummadi K (2017) Leveraging Facebook's advertising platform to monitor stocks of migrants. Popul Dev Rev 43(4):721–734. https://doi.org/10.1111/padr.12102. Accessed 2019-02-03
26. Spyratos S, Vespe M, Natale F, Weber I, Zagheni E, Rango M (2019) Quantifying international human mobility patterns using Facebook network data. PLoS ONE 14(10):1–22. https://doi.org/10.1371/journal.pone.0224134
27. Garcia D, Kassa YM, Cuevas A, Cebrian M, Moro E, Rahwan I, Cuevas R (2018) Analyzing gender inequality through large-scale Facebook advertising data. Proc Natl Acad Sci 115(27):6958–6963. https://doi.org/10.1073/pnas.1717781115. Accessed 2019-02-05
28. Fatehkia M, Kashyap R, Weber I (2018) Using Facebook ad data to track the global digital gender gap. World Dev 107:189–209. https://doi.org/10.1016/j.worlddev.2018.03.007. Accessed 2019-02-03
29. Pew Research Center (2019) Mobile connectivity in emerging economies. Technical report. https://www.pewinternet.org/2019/03/07/mobile-connectivity-in-emerging-economies/. Accessed 2019-06-20
30. Rutstein SO, Johnson K (2004) The DHS Wealth Index, ORC Macro, Calverton. http://dhsprogram.com/pubs/pdf/CR6/CR6.pdf
31. Philippine Statistics Authority, ICF (2018) The DHS Program—Philippines: Standard DHS, 2017 [Dataset] Quezon City, Philippines, and Rockville, Maryland, USA. https://dhsprogram.com/what-we-do/survey/survey-display-510.cfm. Accessed 2019-06-20
32. International Institute for Population Sciences—IIPS/India and ICF (2017) The DHS Program—India: Standard DHS, 2015-16 [Dataset] Mumbai, India: IIPS and ICF. https://dhsprogram.com/what-we-do/survey/survey-display-355.cfm. Accessed 2019-09-04
33. School of Geography and Environmental Science, University of Southampton, Department of Geography and Geosciences, University of Louisville, Departement de Geographie, Universite de Namur, Center for International Earth Science Information Network (CIESIN), Columbia University (2018) WorldPop—global high resolution population denominators project. https://www.worldpop.org/. Accessed 2019-06-03
34. Stevens FR, Gaughan AE, Linard C, Tatem AJ (2015) Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. PLoS ONE 10(2):1–22. https://doi.org/10.1371/journal.pone.0107042
35. Nicas J (2019) Does Facebook really know how many fake. Accounts it has? The New York times. Chap. Technology. Accessed 2020-04-16
36. Rama D, Mejova Y, Tizzoni M, Kalimeri K, Weber I (2020) Facebook Ads as a demographic tool to measure the urban-rural divide. In: The Web Conference (WWW)
37. Araujo M, Mejova Y, Weber I, Benevenuto F (2017) Using Facebook ads audiences for global lifestyle disease surveillance: promises and limitations. In: ACM web science. ACM, New York
38. Gaughan AE, Stevens FR, Linard C, Jia P, Tatem AJ (2013) High resolution population distribution maps for southeast Asia in 2010 and 2015. PLoS ONE 8(2):55882. https://doi.org/10.1371/journal.pone.0055882. Accessed 2019-06-03
39. Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc, Ser B, Methodol 58(1):267–288. Accessed 2018-07-30
40. Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning—data mining, inference, and prediction, 2nd edn. https://www.springer.com/gp/book/9780387848570. Accessed 2019-09-29
41. Gething P, Tatem A, Bird T, Burgert-Brucker CR (2015) Creating spatial interpolation surfaces with DHS data. Technical report. https://dhsprogram.com/publications/publication-SAR11-Spatial-Analysis-Reports.cfm. Accessed 2019-06-20
42. UN General Assembly (2015) Transforming our world: the 2030 Agenda for Sustainable Development. Technical report. https://sustainabledevelopment.un.org/post2015/transformingourworld/publication. Accessed 2019-06-27
43. International Bank for Reconstruction and Development/The World Bank (2018) Poverty and Shared Prosperity 2018: Piecing together the poverty puzzle Washington DC, USA. https://www.worldbank.org/en/publication/poverty-and-shared-prosperity. Accessed 2019-09-22
44. Munoz Boudet AM, Buitrago P, Leroy de la Briere B, Newhouse D, Rubiano Matulevich E, Scott K, Suarrez-Becerra P (2018) Gender differences in poverty and household composition through the life-cycle: a global perspective. Technical Report WPS8360, World Bank Group, Washington DC. http://documents.worldbank.org/curated/en/135731520343670750/Gender-differences-in-poverty-and-household-composition-through-the-life-cycle-a-global-perspective. Accessed 2019-09-22
45. World Economic Forum (2018). Global Gender Gap Report. Technical report, World Economic Forum (2018). wef.ch/gggr18. Accessed 2019-06-26
46. Magno G, Weber I (2014) International gender differences and gaps in online social networks. In: Social informatics, pp 121–138. https://doi.org/10.1007/978-3-319-13734-6_9