The Relative Value of Facebook Advertising Data for Poverty Mapping

Masoomali Fatehkia,[1] Benjamin Coles,[2] Ferda Ofli,[1] Ingmar Weber[1]
[1]Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar
[2]Princeton University, Princeton, NJ, USA
{mfatehkia, fofli, iweber}@hbku.edu.qa, bcoles@princeton.edu

Abstract

Having reliable and up-to-date poverty data is a prerequisite for monitoring the United Nations Sustainable Development Goals (SDGs) and for planning effective poverty reduction interventions. Unfortunately, traditional data sources are often outdated or lacking appropriate disaggregation. As a remedy, satellite imagery has recently become prominent in obtaining geographically-fine-grained and up-to-date poverty estimates. Satellite data can pick up signals of economic activity by detecting light at night, it can pick up development status by detecting infrastructure such as roads, and it can pick up signals for individual household wealth by detecting different building footprints and roof types. It can, however, not look inside the households and pick up signals from individuals. On the other hand, alternative data sources such as audience estimates from Facebook's advertising platform provide insights into the devices and internet connection types used by individuals in different locations. Previous work has shown the value of such anonymous, publicly-accessible advertising data from Facebook for studying migration, gender gaps, crime rates, and health, among others. In this work, we evaluate the added value of using Facebook data over satellite data for mapping socioeconomic development in two low and middle income countries – the Philippines and India. We show that Facebook features perform roughly similar to satellite data in the Philippines with value added for urban locations. In India, however, where Facebook penetration is lower, satellite data perform better.

Introduction

The United Nations Sustainable Development Goals (SDGs) represent a global consensus on the world's most pressing challenges ranging from reducing poverty and inequalities to protecting the environment. Unfortunately, high quality data on SDGs is hard to find, particularly on SDG #1 (i.e., "No Poverty") for countries in most need of development. For instance, 5 out of 15 countries in South and South-East Asia do not have poverty data collected since 2015[1].

To overcome these challenges, researchers have investigated the use of non-traditional data sources including nighttime light intensities (Elvidge et al. 2009), daytime satellite imagery (Jean et al. 2016), and mobile phone Call Detail Record (CDR) data (Blumenstock, Cadamuro, and On 2015) to map poverty. In related efforts, audience estimates from Facebook's advertising platform have proven useful for obtaining insights into the devices and internet connection types used by individuals in different locations. Previous work has shown that these audience estimates, which are traditionally used for planning advertising campaigns, can create new pathways for estimating stocks of migrants (Zagheni, Weber, and Gummadi 2017; Spyratos et al. 2019) and measuring digital gender inequalities (Garcia et al. 2018; Fatehkia, Kashyap, and Weber 2018), among others.

In this study, we assess the value that combining such publicly-accessible, anonymous advertising data with satellite imagery can add to mapping poverty. Concretely, we use the Facebook Marketing API to query how many Facebook users match certain criteria to obtain insights into the spatial distribution of Facebook users by type of devices (e.g., iOS vs. Android) and access to connectivity (e.g., 2G vs. 4G). These insights provide signals for the distribution of poverty, and these signals are partly complementary to the signals that satellite imagery provides.

We evaluate various models using Facebook and Satellite data alone and in combination to test the value added by Facebook features overall and across urban vs. rural locations. We do this for two lower middle income countries: the Philippines and India. We find that Facebook features perform roughly similar to satellite features in the Philippines with value added for urban locations. In India, however, where Facebook penetration is lower, satellite features perform better.

Related Work

Researchers have explored non-traditional data sources to estimate poverty rates in a timely manner (Blumenstock 2016). Several studies have shown that night lights, typically linked to electricity usage, correlate with economic activity (Mellander et al. 2015; Chen and Nordhaus 2011; Henderson, Storeygard, and Weil 2012). Other work has in-

[1] http://iresearch.worldbank.org/PovcalNet/povOnDemand.aspx

vestigated the use of daytime satellite imagery for poverty mapping (Jean et al. 2016; Engstrom, Hersh, and Newhouse 2017; Head et al. 2017; Watmough et al. 2019). Besides satellite imagery, mobile phone Call Detail Records (CDR) have been used for predicting aggregate population-level socioeconomic characteristics (Soto et al. 2011; Fernando et al. 2018) and poverty levels in a variety of countries (Fernando et al. 2018; Njuguna and McSharry 2017; Marco Hernandez Lingzi Hong Vanessa Frias-Martinez Enrique Frias-Martinez 2017). Some studies have combined satellite imagery with other data sources such as CDR data (Pokhriyal and Jacques 2017; Steele Jessica E., Sundsøy Pål Roe, and others 2017), geolocated Wikipedia articles (Sheehan et al. 2019), and crowdsourced geographic information from OpenStreetMap (OSM) (Gervasoni et al. 2018; Isabelle Tingzon et al. 2019). These studies demonstrate the value of non-traditional data sources but they do not evaluate the relative value of one data source over another.

Recently, audience estimates from Facebook's advertising platform have proven useful for a variety of research studies ranging from estimating stocks of migrants (Zagheni, Weber, and Gummadi 2017; Spyratos et al. 2019) and measuring digital gender inequalities at national-level (Garcia et al. 2018; Fatehkia, Kashyap, and Weber 2018) as well as subnational-level (Mejova et al. 2018) to predicting crime rates in urban areas (Fatehkia, O'Brien, and Weber 2019), online monitoring of health (Araujo et al. 2017; Mejova, Weber, and Fernandez-Luque 2018), and polling elections (Ribeiro, Kansaon, and Benevenuto 2019). In this work, we evaluate the relative value – compared to the use of available satellite imagery – that Facebook advertising audience estimates can add to the problem of mapping poverty.

Data and Feature Extraction

Demographic and Health Survey

The Demographic and Health Survey (DHS)[2] provides population and health data through surveys in developing countries around the globe. The surveys collect information on assets owned by surveyed households including appliances such as radio and televisions as well as the type of housing material and access to water and sanitation facilities. This information is then used to construct the Wealth Index through a Principal Component Analysis. The Wealth Index, reported for all surveyed households, is a real-valued score with higher values indicating better living standards. We use the Wealth Index as our ground truth measure of socioeconomic well-being as has been done in previous studies (Jean et al. 2016; Blumenstock, Cadamuro, and On 2015).

The DHS survey locations are called clusters with each cluster consisting of a group of surveyed households (there was a median of 23 households per cluster in the Philippines and 21 in India). The geographic location of each cluster is reported in the form of latitude and longitude coordinates to which a random displacement has been added to protect respondent confidentiality. As a result, the reported location coordinates for urban clusters contain up to 2km of error and for rural locations up to 10km of error. There were a total of 1,249 and 28,524 clusters in the DHS data for the Philippines and India respectively. Of these, we use a subset of 1,205 and 28,043 clusters which were geolocated and for which Facebook advertising data could be collected. An aggregated Wealth Index was computed for each cluster by averaging the Wealth Index scores of all households in that cluster. We use the latest available data for both countries: the 2017 DHS for the Philippines and the 2015-16 DHS for India.

Facebook Advertising Data

The Facebook advertising platform provides advertisers with a rich array of targeting criteria. These include targeting users by geographic location, demographics such as age and gender, as well as behaviours such as the type of devices or network connections used to access the platform. For a specified targeting criteria, before an ad is placed the advertiser is provided with an estimate of Monthly Active Users (MAU) matching the specified criteria. For example, there are 570k users aged 18+ living in the city of Delhi, India, who are predominantly using an iOS device to access Facebook.[3] Using Facebook's Marketing API, the data collection for various combinations of targeting criteria can be automated.

We collected these MAU estimates for the DHS cluster locations in the Philippines and India. For each location, data were collected on estimates of users aged 18+ utilizing a range of network/device types. These include type of network access (2G, 3G, 4G and WiFi), the mobile operating system used (iOS, Android and Windows), usage of high-end devices (the latest releases of Apple iPhone and Samsung Galaxy phones) as well as other device types. The fraction of users in each location with access to these different device types were computed and used in the analysis in addition to the Facebook penetration which was computed as the ratio of MAU Facebook users to the estimated population of the cluster.

Data were collected from the Facebook Marketing API using the pySocialWatcher library[4], during Mar-Apr 2019 for the Philippines and Jun-Sep 2019 for India.

Satellite Imagery

We followed a similar approach presented in (Isabelle Tingzon et al. 2019; Jean et al. 2016) to collect and extract features from satellite images. When collecting satellite images, we kept two factors in mind: the location of the DHS clusters and the positioning of the nighttime luminosity (NTL) pixels in VIIRS DNB database.[5] The process involves 'matching' the satellite images to the VIIRS DNB NTL pixels (i.e., 0.25km^2 squares with a given center point) within a specified radius of each DHS Cluster (5km for rural points and 2km for urban points). We used the High Resolution Settlement Layer dataset (Tiecke et al. 2017) to eliminate images that contained no human settlements. Then, us-

[2] https://dhsprogram.com/

[3] As of 15 Jan 2020.

[4] https://github.com/maraujo/pySocialWatcher

[5] https://ngdc.noaa.gov/eog/viirs/download_dnb_composites.html

ing Google Static Maps API, we downloaded images with a zoom level of 17, scale of 1, and pixel resolution of approximately 1.25m. The size of each image is 400×400 pixels and matches the land area covered by a single pixel of nighttime lights data, which typically covers 0.25km^2. We downloaded 134.5k images for the Philippines and 3.2M images for India.

We then used a transfer learning approach to train our image models. We took the convolution neural network architecture described in (Isabelle Tingzon et al. 2019) and fine-tuned it on our daytime satellite images using the corresponding nighttime lights intensity labels (i.e., low, moderately low, medium, moderately high, and high). We used a 90/10 training/validation data split. As we had a significant class imbalance (far fewer moderately high or high nightlight intensities), we upsampled the minority class with replacement and downsampled the majority class. The balanced training data included around 20,000 images per class for the Philippines and 200,000 for India. After following a training procedure similar to (Isabelle Tingzon et al. 2019), we achieved 74% accuracy and 61% macro-averaged F1-score on the validation set of the Philippines, and 62% accuracy and 50% macro-averaged F1-score on the validation set of India.

Finally, we extracted for each image a 4,096-dimensional feature vector from the penultimate layer of the fine-tuned CNN models. Since there are multiple images that belong to a particular DHS cluster, we represented each DHS cluster with the average vector of all image features belonging to that cluster. We then used these cluster-level image features in our experiments.

Additional Data

To capture geographic variation, regional dummy variables were also used. These are variables that take a value of 1 if a given cluster belongs to a specified region and a value of 0 otherwise. For the Philippines and India, there are 17 and 36 Admin Level 1 regions respectively.

Population estimates for the DHS clusters were computed from high resolution population data acquired from WorldPop (Gaughan et al. 2013) for the Philippines and India. The population data was used to compute the cluster Facebook penetration and the population density. Table 1 provides a summary of the different sets of variables that are used in the models.

Table 1: Features used in the analysis.

Feature	# of variables
Satellite feature embeddings	4096
Facebook features	23
Regional dummy	Philippines: 17, India: 36
Population density (log)	1

Experiments

Country-Level GDP Prediction

Table 2 shows the Pearson correlation coefficient between some Facebook features and the 2018 log GDP per capita from the International Monetary Fund (IMF). This is based on data for 135 countries with at least 500k monthly active Facebook users. The overall Facebook penetration as well as the fraction of Facebook users who utilize high-end phones are strongly correlated with country-level income. A linear three variable model to predict the log GDP per capita using the Facebook penetration, fraction of users with high-end devices and users with access to 4G Network achieves a cross-validated R^2 of 0.737.

Table 2: Pearson correlation between the fraction of Facebook users with various device/network types and country-level log GDP per capita. Data for 135 countries.

Feature	Pearson corr.
iOS	0.691
High-end devices	0.711
4G network	0.529
Facebook penetration	0.693

Figure 1 displays the log GDP per capita estimated by the three variable model against the actual data. While the estimates made by the model roughly follow the ground truth, the model predicts a slightly higher income for the lower income countries in the data and predicts a lower income for some countries throughout the income distribution. For example, the model over-predicts in the Democratic Republic of the Congo (true: 6.11, predicted: 6.98) and under-predicts in Switzerland (true: 11.33, predicted: 10.77).

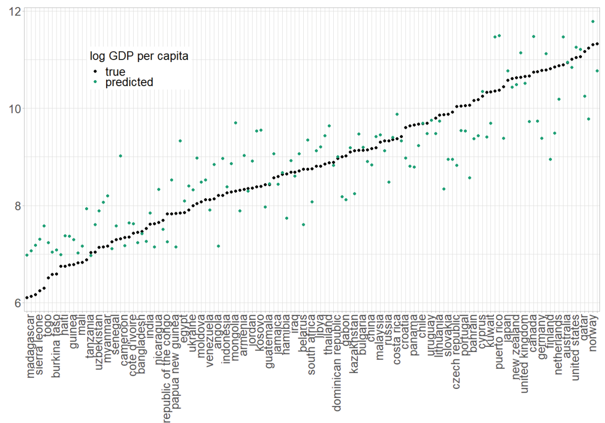

Figure 1: Countries arranged by log (base e) GDP per capita. Plotted are the true log GDP per capita (in black) and the predictions by a model using Facebook features (in green). For readability, the x-axis label shows every other country.

These country-level results suggest that the Facebook features provide predictive signals of income levels. Moreover, using this data it is fairly easy to create estimates for large geographic units (i.e. country level) with simple models which may not be as easy to do using satellite imagery.

Cluster-Level Results for India and the Philippines

Given the large number of features, ridge regression models were fitted to estimate the DHS Wealth Index using different combinations of Facebook and Satellite imagery features. Models were fitted separately for each of the two countries. Each model was fitted and evaluated through 10-fold cross validation. Table 3 reports the performance of the various models. Models were fitted using the Facebook or Satellite features alone (FB and Satellite models) and in combination with the regional dummy variables and log population density (FB+ and Satellite+ models). Finally, a model was fitted using all of these variables (FB&Satellite+ model).

Table 3: Wealth Index prediction performance of various models using different combinations of Facebook, Satellite, regional dummy variables and population density. Reported values are R^2 (MAE) based on 10-fold cross validation.

Model	Philippines	India
FB	0.595 (35,510)	0.483 (45,413)
Satellite	0.602 (34,957)	0.662 (35,862)
FB+	0.631 (33,768)	0.627 (37,886)
Satellite+	0.641 (33,243)	0.705 (33,071)
FB&Satellite+	0.655 (32,663)	0.708 (32,811)

Based on Table 3, in the Philippines, the FB model achieves roughly similar results to the Satellite model (two-sided paired t-test comparing the MAE of FB and Satellite models: MAE difference (FB - Satellite) = 553.1, t = 0.801, p = 0.423, df = 1204). For both models, performance improves with the addition of population density and regional dummy variables. The best performance is achieved by the combined FB&Satellite+ model (paired t-test testing for MAE FB&Satellite+ model $<$ MAE of Satellite+ model: MAE difference (FB&Satellite+ - Satellite+) = -580.36, t = -2.874, p = 0.002, df = 1204).

In India, the FB model achieves decent performance (R^2 0.483) which improves further when including the population density and regional dummy variables (R^2 of 0.627); however, the Satellite and Satellite+ models achieve better performance than the FB and FB+ models.

Rural, Urban Disaggregation

The previous section looked at the performance of the different models when evaluated on the full dataset. This section looks at the relative differences of the models when evaluated separately on urban or rural locations only. Urban and rural locations may differ both in terms of the physical layout of the buildings/landscape as observed by Satellite imagery as well as the use of and access to technology/devices and social media, which could influence the performance of models using these data. Table 4 reports the performance of the models when fitted and evaluated separately on the data for urban and rural clusters.

Table 4: Performance of various models fitted and evaluated separately for Urban and Rural locations. Reported values are R^2 based on 10-fold cross validation.

	Philippines		India	
Model	Urban	Rural	Urban	Rural
FB	0.446	0.451	0.235	0.393
Satellite	0.406	0.488	0.314	0.618
FB+	0.454	0.510	0.338	0.610
Satellite+	0.447	0.531	0.364	0.690
FB&Satellite+	0.464	0.546	0.368	0.695
Number of clusters	437	768	8,429	19,614

Note that in all settings the urban-only and rural-only R^2 is lower than the whole-country results as the two subsets are more homogeneous, with higher and lower wealth indices respectively, so that teasing apart the remaining variance becomes a harder task. In the Philippines, the FB model performs better than the Satellite model for urban locations (FB R^2 of 0.446 vs. Satellite R^2 of 0.406; paired t-test for FB MAE $<$ Satellite MAE: MAE difference (FB - Satellite) = -1,703, t = -1.678, df = 436, p = 0.047) while the Satellite model performs better for rural locations (Satellite model R^2 of 0.488 vs. FB R^2 of 0.451; paired t-test for Satellite MAE $<$ FB MAE: MAE difference (FB - Satellite) = 1,898, t = 2.204, df = 767, p = 0.014). In India, the Satellite model performs better than the FB model for both Urban and Rural locations, consistent with the overall better performance by Satellite models observed earlier.

Conclusion

We evaluated the value of using social media advertising data from Facebook for estimating socioeconomic status across locations in the Philippines and India, comparing the value added with respect to approaches based on Satellite imagery. Overall, using Facebook advertising data performed better in the Philippines achieving similar performance to models with satellite imagery features with value added for urban locations. Based on these results, studying the use of Facebook advertising data for poverty estimation in other countries with high Facebook penetration such as Indonesia appears promising.

Other benefits of using social media advertising data form Facebook are the relative ease of data collection through the platform's advertising API. It is also plausible that sudden changes in socioeconomic status will more quickly be reflected in the types of mobile devices people use, than in the satellite footprint. Hence, a potential avenue for future work is to explore the relative strengths of these approaches for capturing *temporal changes* in socioeconomic well-being. Another potential advantage of Facebook's advertising data which was not explored here is the ability to obtain disaggregated estimates, e.g. by the gender of the user. Such information on device usage for different user groups could be used to create gender or age disaggregated poverty estimates which is more difficult to do with satellite imagery features.

References

Araujo, M.; Mejova, Y.; Aupetit, M.; and Weber, I. 2017. Visualizing Health Awareness in the Middle East. In *ICWSM*.

Blumenstock, J.; Cadamuro, G.; and On, R. 2015. Predicting poverty and wealth from mobile phone metadata. *Science* 350(6264):1073–1076.

Blumenstock, J. E. 2016. Fighting poverty with data. *Science* 353(6301):753–754.

Chen, X., and Nordhaus, W. D. 2011. Using luminosity data as a proxy for economic statistics. *PNAS* 108(21):8589–8594.

Elvidge, C. D.; Sutton, P. C.; Ghosh, T.; Tuttle, B. T.; Baugh, K. E.; Bhaduri, B.; and Bright, E. 2009. A global poverty map derived from satellite data. *Computers & Geosciences* 35(8):1652–1660.

Engstrom, R.; Hersh, J. S.; and Newhouse, D. L. 2017. Poverty from space : using high-resolution satellite imagery for estimating economic well-being. Technical Report WPS8284, The World Bank.

Fatehkia, M.; Kashyap, R.; and Weber, I. 2018. Using Facebook ad data to track the global digital gender gap. *World Development* 107:189–209.

Fatehkia, M.; O'Brien, D.; and Weber, I. 2019. Correlated impulses: Using facebook interests to improve predictions of crime rates in urban areas. *PLOS ONE* 14(2):1–16.

Fernando, L.; Surendra, A.; Lokanathan, S.; and Gomez, T. 2018. Predicting Population-level Socio-economic Characteristics Using Call Detail Records (CDRs) in Sri Lanka. In *DSMM*, 1:1–1:12.

Garcia, D.; Kassa, Y. M.; Cuevas, A.; Cebrian, M.; Moro, E.; Rahwan, I.; and Cuevas, R. 2018. Analyzing gender inequality through large-scale Facebook advertising data. *PNAS* 115(27):6958–6963.

Gaughan, A. E.; Stevens, F. R.; Linard, C.; Jia, P.; and Tatem, A. J. 2013. High Resolution Population Distribution Maps for Southeast Asia in 2010 and 2015. *PLOS ONE* 8(2):e55882.

Gervasoni, L.; Fenet, S.; Perrier, R.; and Sturm, P. 2018. Convolutional neural networks for disaggregated population mapping using open data. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, 594–603.

Head, A.; Manguin, M.; Tran, N.; and Blumenstock, J. E. 2017. Can Human Development Be Measured with Satellite Imagery? In *ICTD*, 8:1–8:11.

Henderson, J. V.; Storeygard, A.; and Weil, D. N. 2012. Measuring Economic Growth from Outer Space. *American Economic Review* 102(2):994–1028.

Isabelle Tingzon; Ardie Orden; Stephanie Sy; Vedran Sekara; Ingmar Weber; Masoomali Fatehkia; Manuel Garcia Herranz; and Dohyung Kim. 2019. Mapping Poverty in the Philippines Using Machine Learning, Satellite Imagery, and Crowd-sourced Geospatial Information. In *AI for Social Good ICML 2019 Workshop*.

Jean, N.; Burke, M.; Xie, M.; Davis, W. M.; Lobell, D. B.; and Ermon, S. 2016. Combining satellite imagery and machine learning to predict poverty. *Science* 353(6301):790–794.

Marco Hernandez Lingzi Hong Vanessa Frias-Martinez Enrique Frias-Martinez. 2017. Estimating Poverty Using Cell Phone Data: Evidence from Guatemala. Technical report, The World Bank.

Mejova, Y.; Gandhi, H. R.; Rafaliya, T. J.; Sitapara, M. R.; Kashyap, R.; and Weber, I. 2018. Measuring subnational digital gender inequality in india through gender gaps in facebook use. In *COMPASS*.

Mejova, Y.; Weber, I.; and Fernandez-Luque, L. 2018. Online health monitoring using facebook advertisement audience estimates in the united states: Evaluation study. *JMIR Public Health Surveill* 4(1):e30.

Mellander, C.; Lobo, J.; Stolarick, K.; and Matheson, Z. 2015. Night-Time Light Data: A Good Proxy Measure for Economic Activity? *PLOS ONE* 10(10):e0139779.

Njuguna, C., and McSharry, P. 2017. Constructing spatiotemporal poverty indices from big data. *Journal of Business Research* 70:318–327.

Pokhriyal, N., and Jacques, D. C. 2017. Combining disparate data sources for improved poverty prediction and mapping. *PNAS* 114(46):E9783–E9792.

Ribeiro, F. N.; Kansaon, D.; and Benevenuto, F. 2019. Leveraging the facebook ads platform for election polling. In *WebMedia*, 305–312.

Sheehan, E.; Meng, C.; Tan, M.; Uzkent, B.; Jean, N.; Burke, M.; Lobell, D.; and Ermon, S. 2019. Predicting economic development using geolocated wikipedia articles. In *KDD*, 2698–2706.

Soto, V.; Frias-Martinez, V.; Virseda, J.; and Frias-Martinez, E. 2011. Prediction of Socioeconomic Levels Using Cell Phone Records. In *UMAP*, 377–388.

Spyratos, S.; Vespe, M.; Natale, F.; Weber, I.; Zagheni, E.; and Rango, M. 2019. Quantifying international human mobility patterns using facebook network data. *PLOS ONE* 14(10):1–22.

Steele Jessica E.; Sundsøy Pål Roe; and others. 2017. Mapping poverty using mobile phone and satellite data. *Journal of The Royal Society Interface* 14(127):20160690.

Tiecke, T. G.; Liu, X.; Zhang, A.; Gros, A.; Li, N.; Yetman, G.; Kilic, T.; Murray, S.; Blankespoor, B.; Prydz, E. B.; and Dang, H.-A. H. 2017. Mapping the world population one building at a time.

Watmough, G. R.; Marcinko, C. L. J.; Sullivan, C.; Tschirhart, K.; Mutuo, P. K.; Palm, C. A.; and Svenning, J.-C. 2019. Socioecologically informed use of remote sensing data to predict rural household poverty. *PNAS* 116(4):1213–1218.

Zagheni, E.; Weber, I.; and Gummadi, K. 2017. Leveraging Facebook's Advertising Platform to Monitor Stocks of Migrants. *Population and Development Review* 43(4):721–734.