24 September 2020

Estimating Socioeconomic Indicators in the Philippines using Machine Learning and Open Geospatial Information

Stephanie Sy
Founder and CEO
stef@thinkingmachin.es

Thinking Machines Data Science

OUR MISSION

Our social impact mission is to empower evidence-based policy and action by:

- Filling critical data gaps
- Making data open and useful
- Innovating with purpose

PHILIPPINE CONTEXT

Can machine learning support development studies with cheap and fast data inference methods?

In a series of studies, we combined **cost-efficient machine learning** with **freely accessible geospatial information** as a **fast, low-cost, and scalable** means of providing **poverty estimates**.

Specifically, we examine the extent to which geospatial data including **remote-sensed data, digital activity, and crowd-sourced information** can be used to estimate socioeconomic well-being in the Philippines.

We look into the viability of using free and openly available satellite images taken from **Google Earth Engine, Facebook Marketing data, and OpenStreetMap data** to estimate poverty indicators derived from the **2017 National Demographic and Health Survey**.

Thinking Machines
Data Science

EVERY 4-5 YEARS, PSA RUNS

The National Demographic and Health Survey

Granularity: Household | **Breadth:** Key demographic and health indicators

Most Recent DHS Surveys

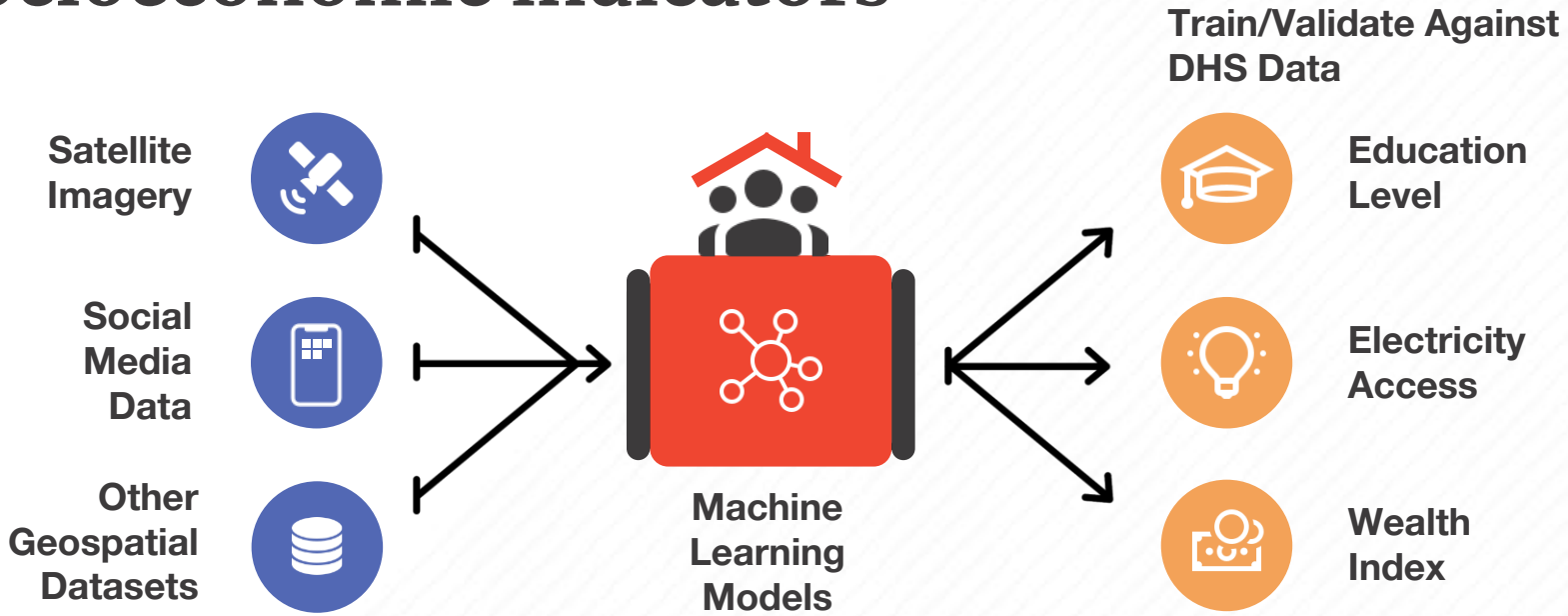
27,496 households

Representative nationwide sample

Thinking Machines Data Science

GLOBAL TREND

Using unconventional data sources to infer socioeconomic indicators



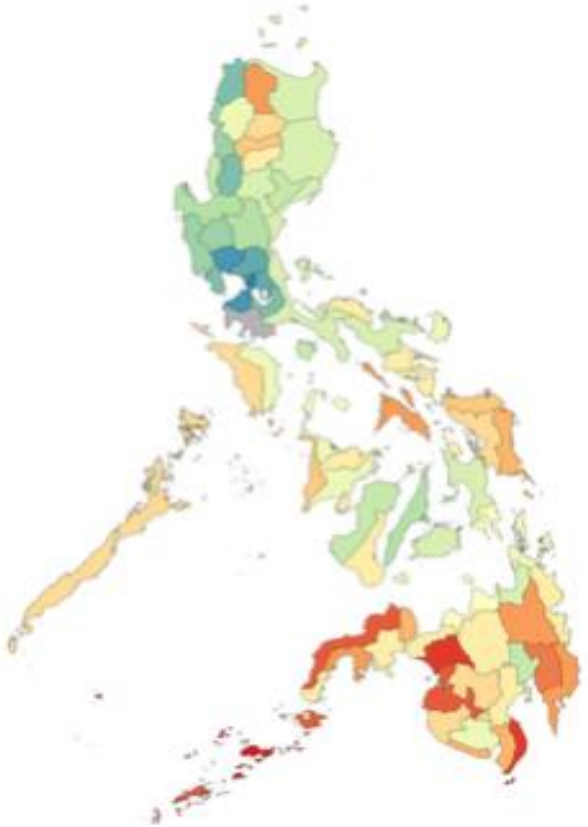
Thinking Machines
Data Science

Our goal is to support surveyors and decision makers by using tech to infer useful data for areas where surveys are not feasible

SOLUTION

1 First Approach, Using Satellite Imagery

- **Goal: Faster**, **cheaper**, and **more granular** reconstruction of poverty measures in the Philippines
- Replicated a study by **Jean et al.*** from the Stanford Sustainability and AI Lab
  - Estimated asset-based wealth for five sub-Saharan African countries
- Crowdsourced geospatial information for poverty prediction



Provincial-level Wealth Index (Ground truth)

*Jean, Neal, et al. "Combining satellite imagery and machine learning to predict poverty." *Science* 353.6301 (2016): 790-794.

APPROACH

Using satellite images to predict wealth

APPROACH

Problem: Data Scarcity

Need **a lot** of labeled training data for an end-to-end deep learning approach

Model

Poverty Predictions

Wealth Index, Education, Access to Water, Electricity, Child Mortality

Data Gap: Not enough labeled training data!

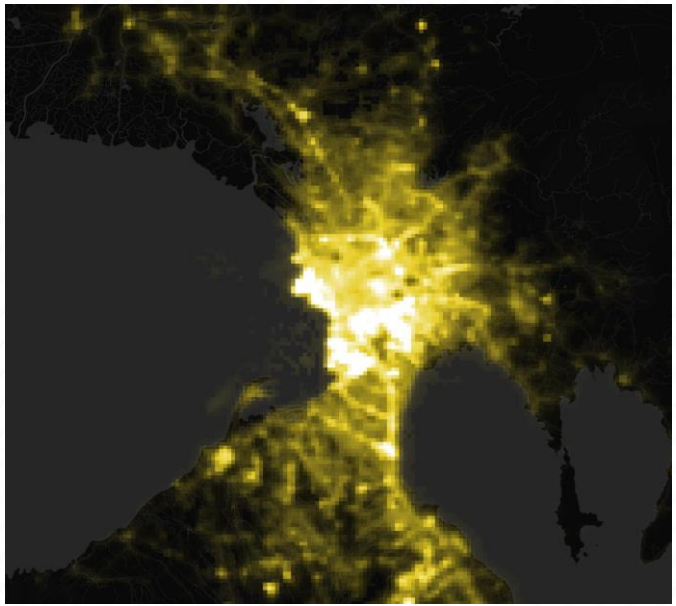
Thinking Machines Data Science

SOLUTION

Nighttime lights can be used as a proxy for economic development

Metro Manila (Daytime)

Metro Manila (Nighttime)



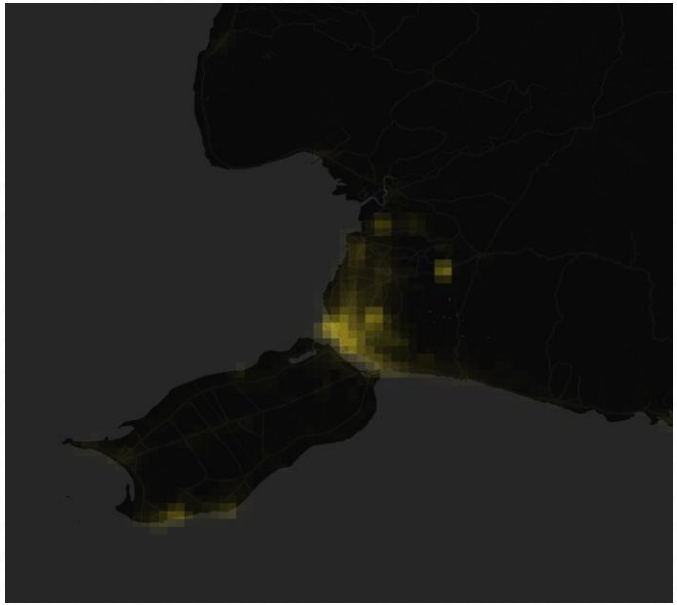
Thinking Machines Data Science

SOLUTION

Nighttime lights can be used as a proxy for economic development

Tagbilaran, Bohol (Daytime)

Tagbilaran, Bohol (Nighttime)

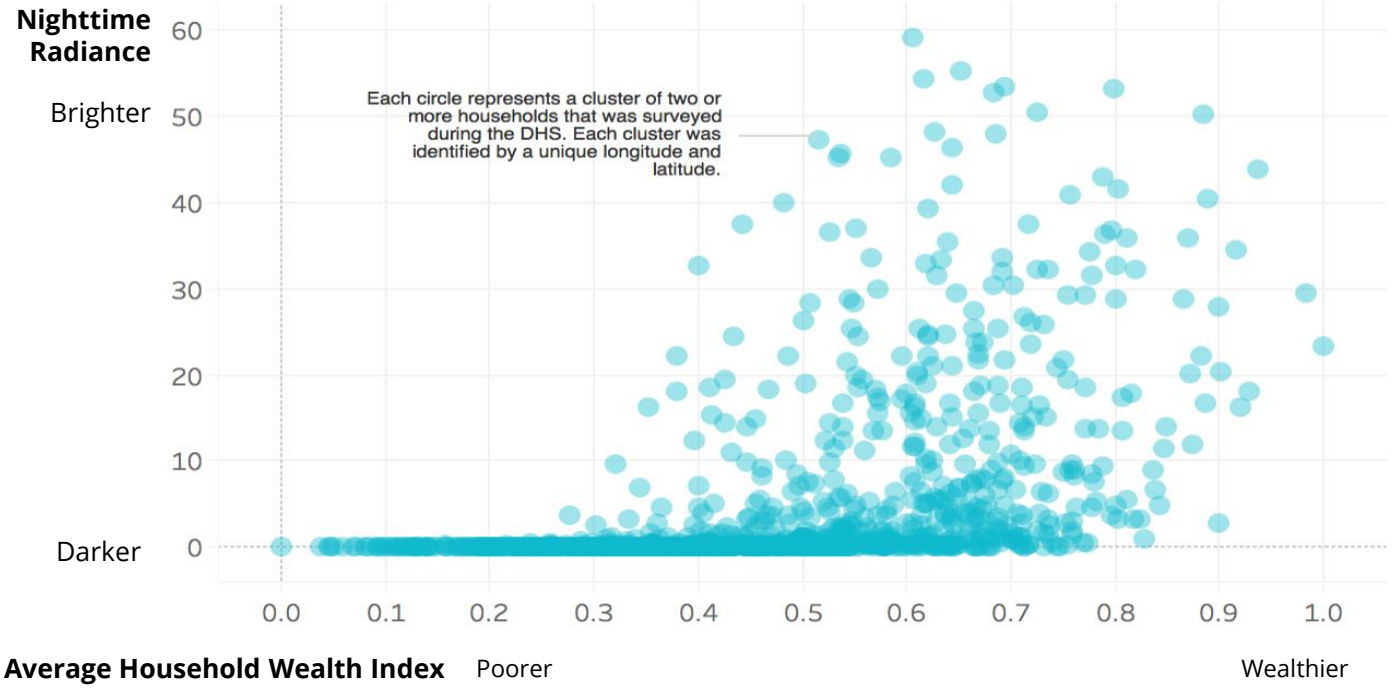


Thinking Machines
Data Science

SOLUTION

Wealthier places are brighter at night

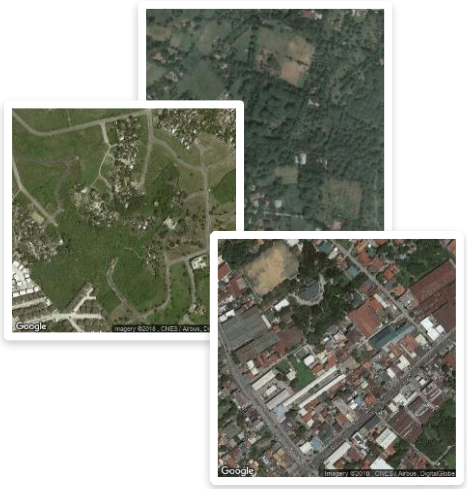
For each of the 1,200+ sampled locations or “clusters” in the 2017 Demographic and Health Survey (DHS), there was a positive correlation between nightlight luminosity and average household wealth index. *(p-0.75, r-0.49)*



Thinking Machines Data Science

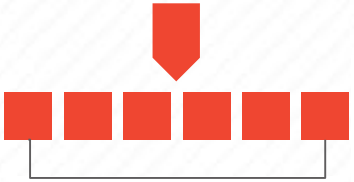
METHODOLOGY

Step 1. Predict nighttime light intensity as a proxy task

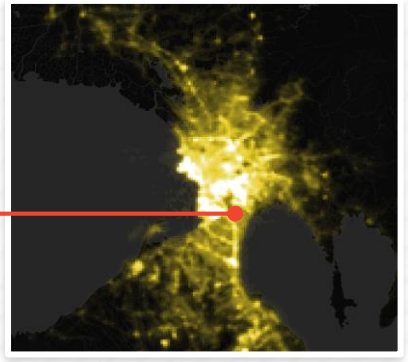


Daytime Satellite Images

Convolutional Neural Network
(as feature extractor)



Low-dimensional feature embeddings



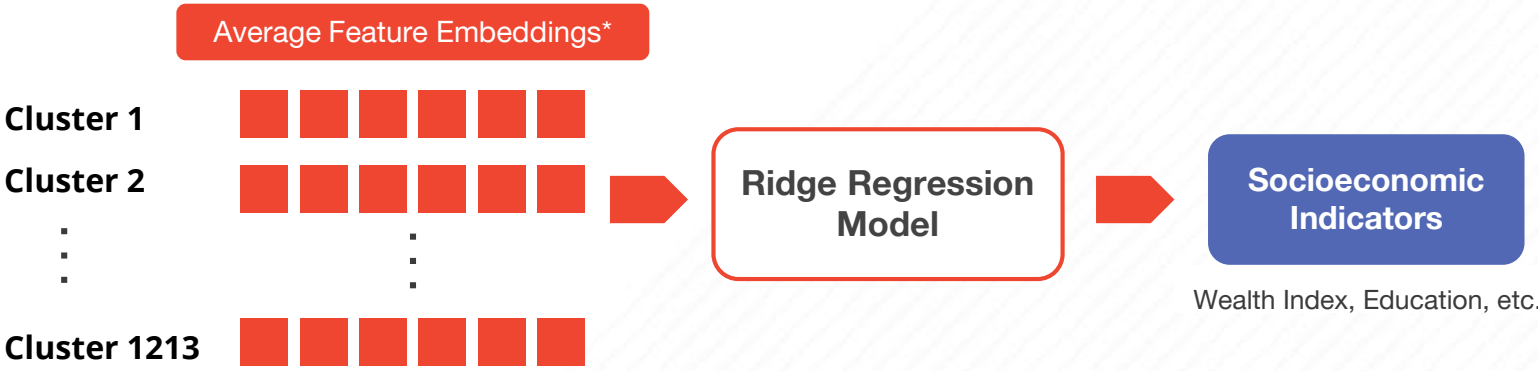
Nighttime Light Intensity Levels

Five Classes: Low, Moderately low, Medium, Moderately High, and High
(based on pixel brightness values)

*Jean, Neal, et al. "Combining satellite imagery and machine learning to predict poverty." *Science* 353.6301 (2016): 790-794.

METHODOLOGY

Step 2. Compute the average feature embeddings per cluster to estimate wealth



*Average feature embeddings of up to 400 image tiles within a 2km (urban) or 5km (rural) radius from the cluster centroid. We removed images containing no human settlements using the HRSL dataset by Tiecke et al. [2]

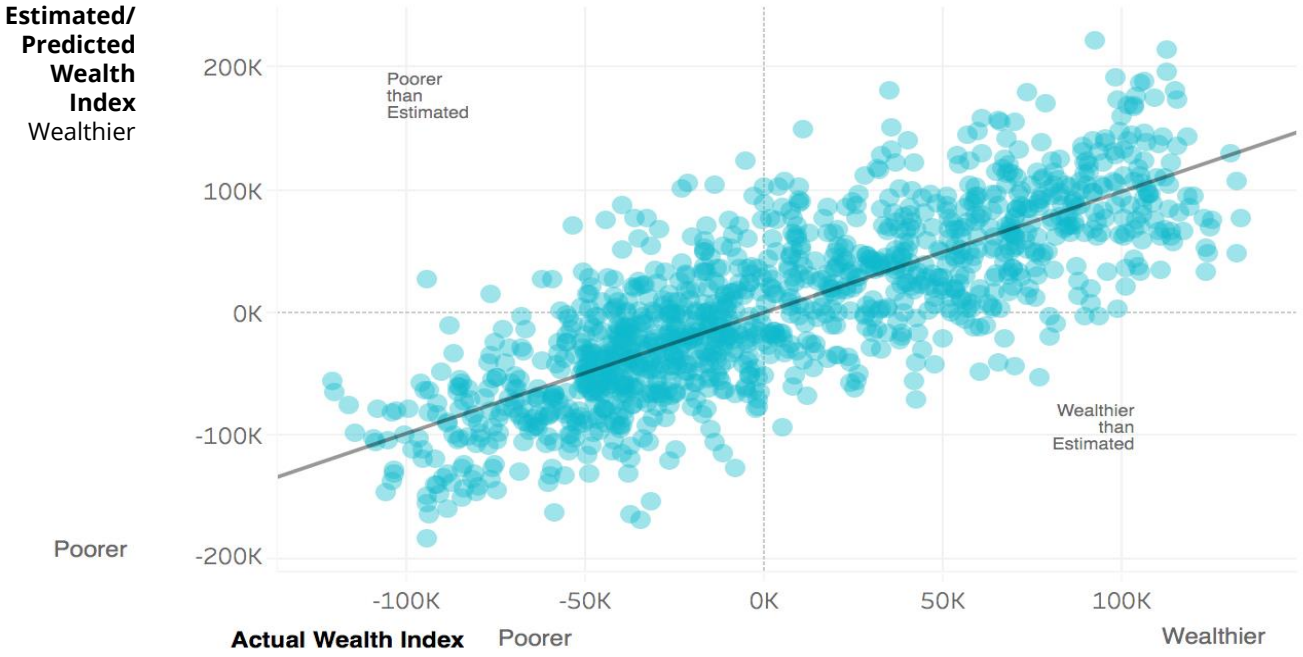
*Jean, Neal, et al. "Combining satellite imagery and machine learning to predict poverty." *Science* 353.6301 (2016): 790-794.

Thinking Machines Data Science

RESULTS

How accurate are our wealth predictions?

The chart below compares the actual versus estimated or predicted average household wealth index for each of the 1,200+ cluster surveyed in the 2017 Demographic and health survey. *(r2=0.625)*



The model is able to explain **62.5%** of the variance

*Using CNN feature embeddings with regional indicators

Predictions and reported r^2 values are from five-fold nested cross-validation.

Thinking Machines Data Science

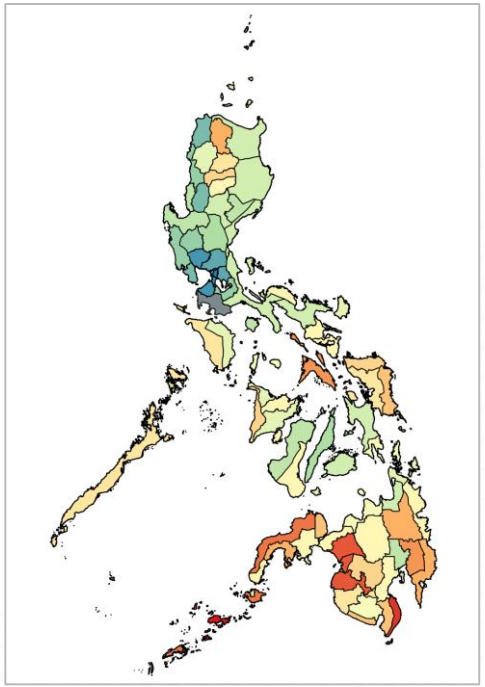
RESULTS

Reconstructing Provincial-level Maps

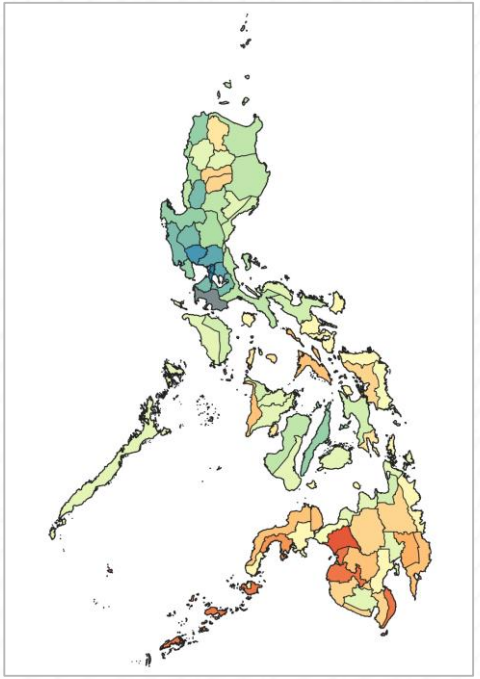
Normalized Wealth Index

0 (poor) 1 (wealthy)

Actual Wealth Index



Predicted Wealth Index

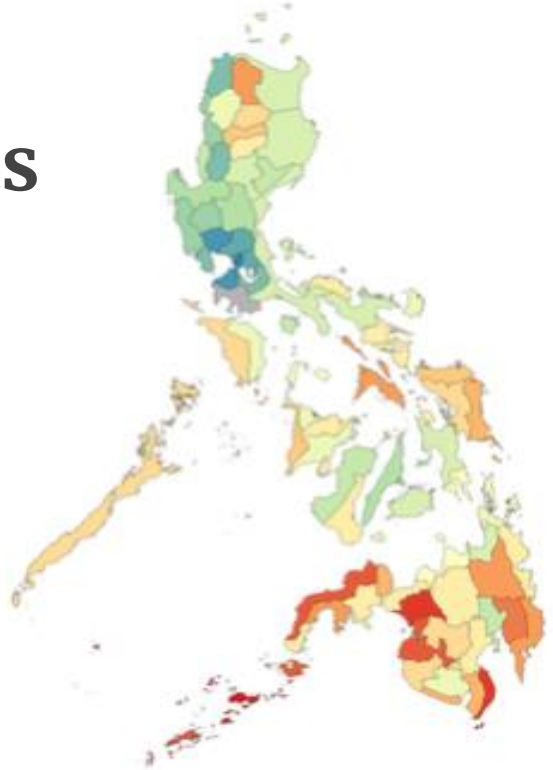


Predictions and reported r^2 values are from five-fold nested cross-validation.

Thinking Machines Data Science

2 Second Approach, Including Unconventional Digital Datasets

- **Goal: Use unconventional datasets** to infer wealth
- Datasets
  - Night-time lights data
  - Facebook marketing data
    - Internet Connectivity, iOS Device, mid-to-high goods
  - OpenStreetMap Data
  - CheckMySchool Data



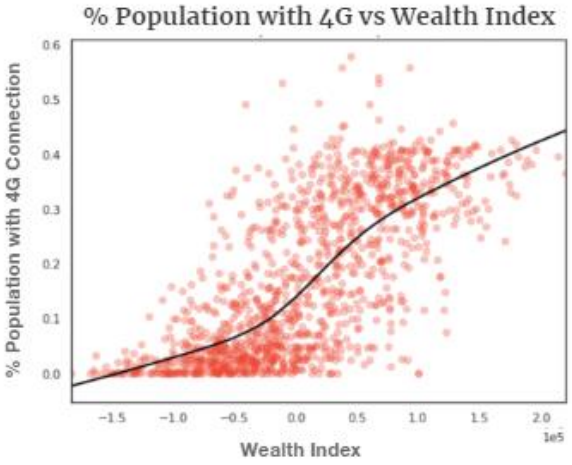
Provincial-level Wealth Index (Ground truth)

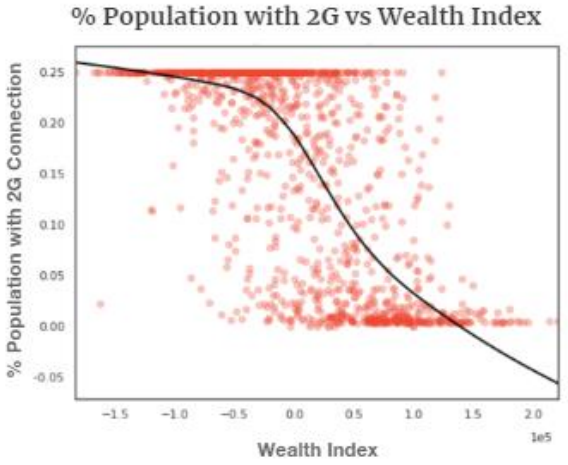
Thinking Machines Data Science

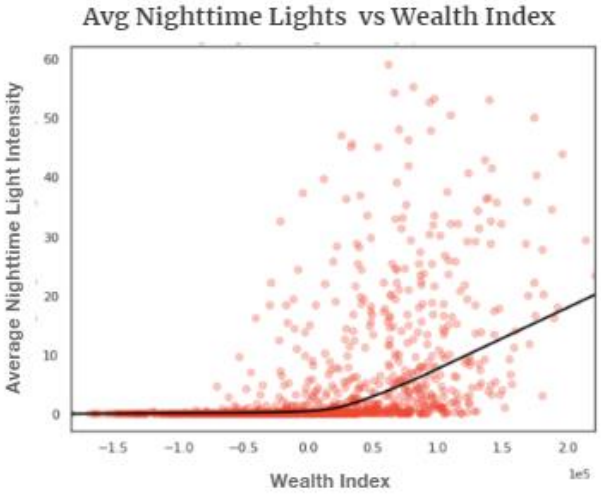
*Jean, Neal, et al. "Combining satellite imagery and machine learning to predict poverty." *Science* 353.6301 (2016): 790-794.

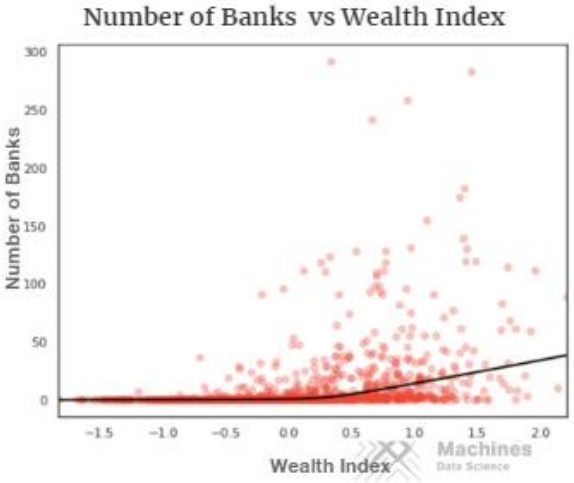
EXPLORATORY ANALYSIS

4G and 2G correlates well with DHS survey data





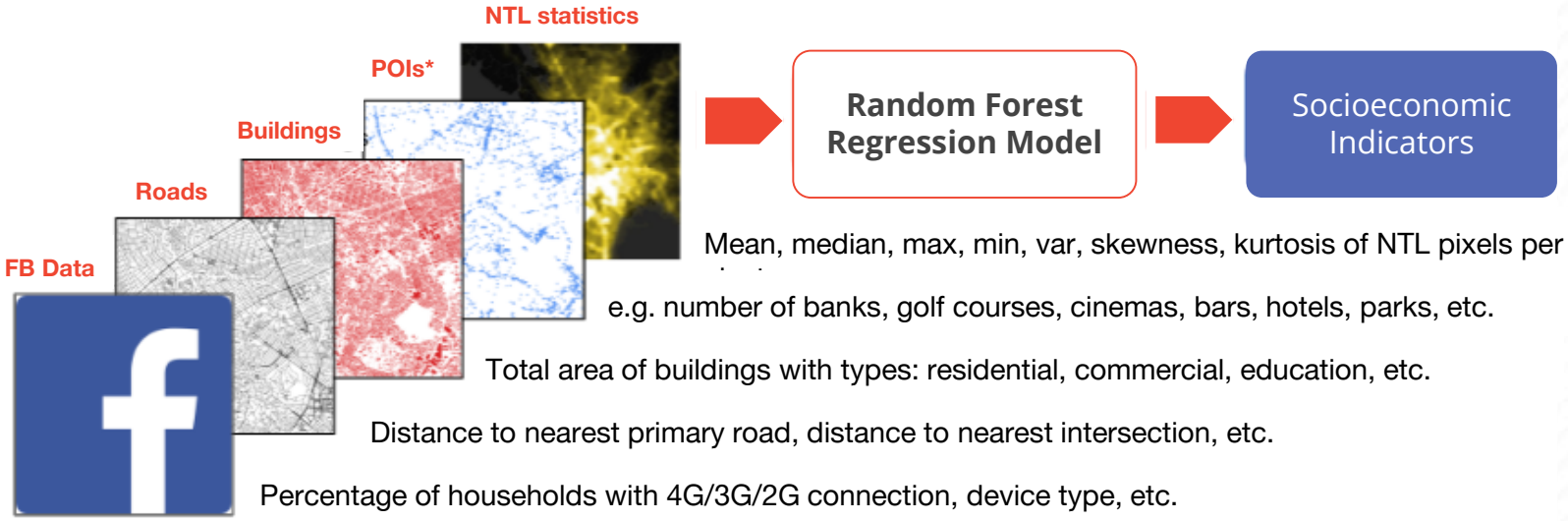




MODEL

Model Building

Specifically, we evaluated the performance of Random Forest Regression, Linear Regression, Lasso Regression, Support Vector Regression, Ridge Regression, and Light Gradient Boosting Machine Regression



Thinking Machines Data Science

RESULTS

Model Evaluation on Wealth Estimation

Five Fold Nested Cross Validation Approach, focus on R^2 to be aligned with similar predictive models run in other regions

- Past studies have published results on using deep learning methods for predicting wealth in sub-Saharan African countries (Jean et al., 2016) as well as non-African countries (Head et al., 2017).
- Predictive models achieved r-squared results ranging from 0.51 to 0.75 (Haiti: 0.51; Malawi: 0.55; Tanzania: 0.57; Nepal: 0.64 Nigeria: 0.68; Uganda: 0.69; Rwanda: 0.75).

Model	r^2
Linear Regression	0.565949
Lasso Regression	0.559921
Ridge Regression	0.547000
Support Vector Regression	0.000426
Random Forest Regression	**0.660433**
LGBM Regression	0.641586

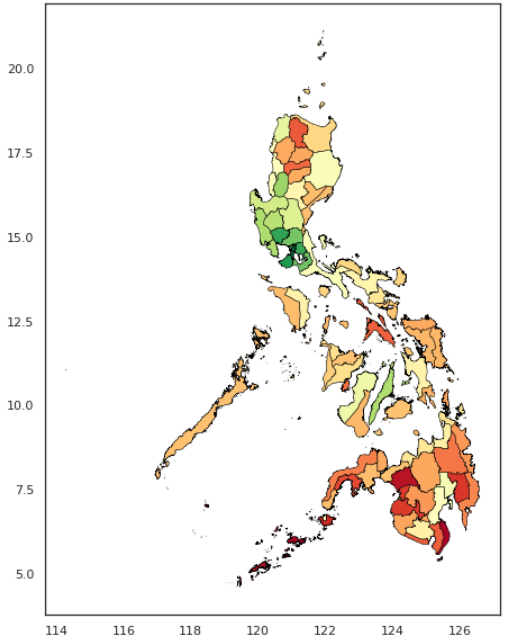
Thinking Machines Data Science

RESULTS

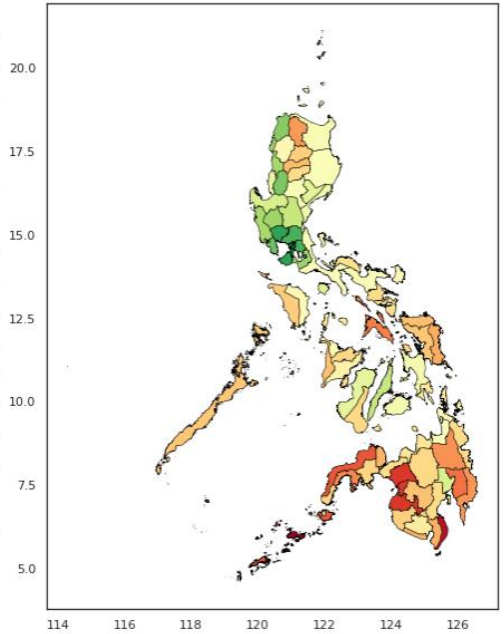
Reconstructing Provincial-level Maps

Normalized Wealth Index

0 (poor) 1 (wealthy)



Predicted Wealth Index using the Machine Learning Model



Actual Wealth Index (Demographic and Health Survey)

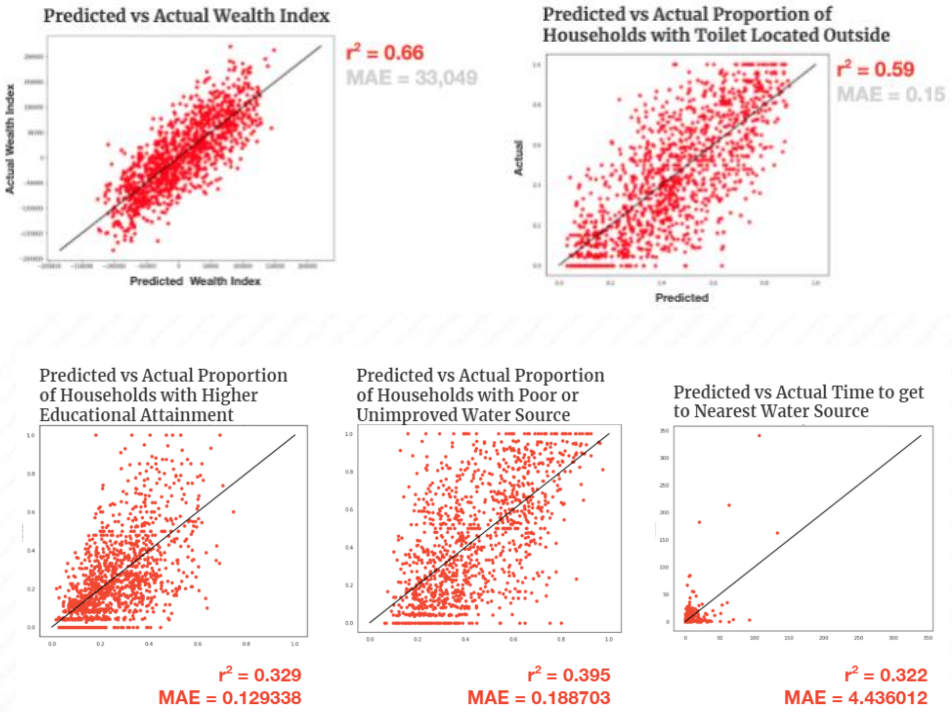
Predictions and reported r^2 values are from five-fold nested cross-validation.

Thinking Machines Data Science

RESULTS

Model doesn't generalize well to other socioeconomic indicators

We note that these results are consistent with the conclusions reached in Head et al., which states that high performance on satellite imagery trained models cannot be expected when there is no clear relationship between the development indicator and nighttime lights (Head et al., 2017).

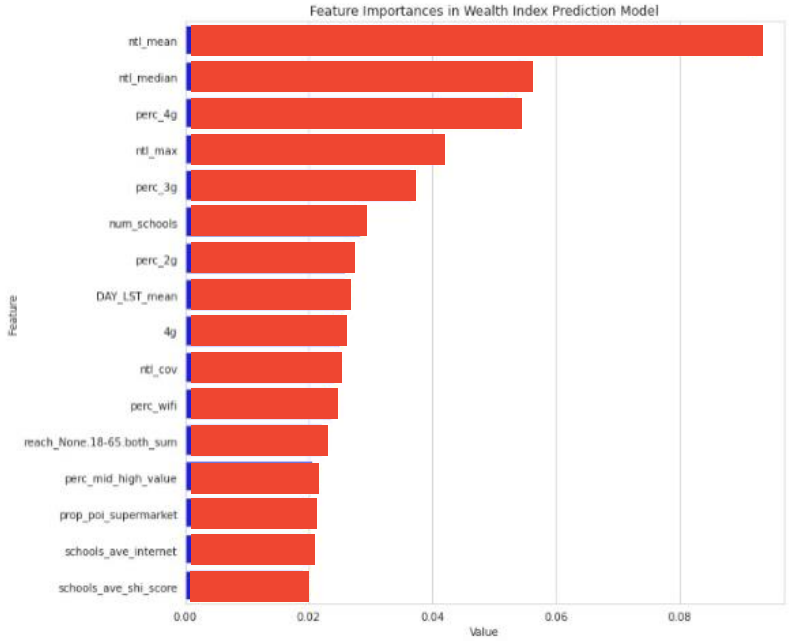


Thinking Machines Data Science

Head, Andrew, et al. "Can human development be measured with satellite imagery?." *ICTD*. 2017.

RESULTS

RF Feature importance indicates that 3G/4G usage should be studied in more depth



Description of features:

- **ntl2016/ntl_mean**: Average value of nighttime light intensity
- **ntl_median**: Median value of nighttime light intensity
- **perc_4g**: Proportion of population with 4G
- **ntl_max**: Maximum value of nighttime light intensity
- **perc_3g**: Proportion of population with 3G
- **num_schools:** Number of schools
- **perc_2g**: Proportion of population with 2G
- **DAY_LST_mean**: Average land surface temperature (daytime)
- 4g: Number of 4G users
- **ntl_cov**: Variance of nighttime light intensity within the cluster
- **percent_wifi:** Proportion of people using Facebook with WiFi
- **reach_None.18-65.both_sum:** Total population of an area as measured by Facebook
- **perc_mid_high_value**: Proportion of population with mid- or high-tier cellular phones
- **prop_poi_supermarket:** Proportion of POIs in a grid that are supermarkets
- **schools_ave_internet:** Proportion of schools with internet
- **schools_ave_shi_score:** Average SHI score for schools in grid

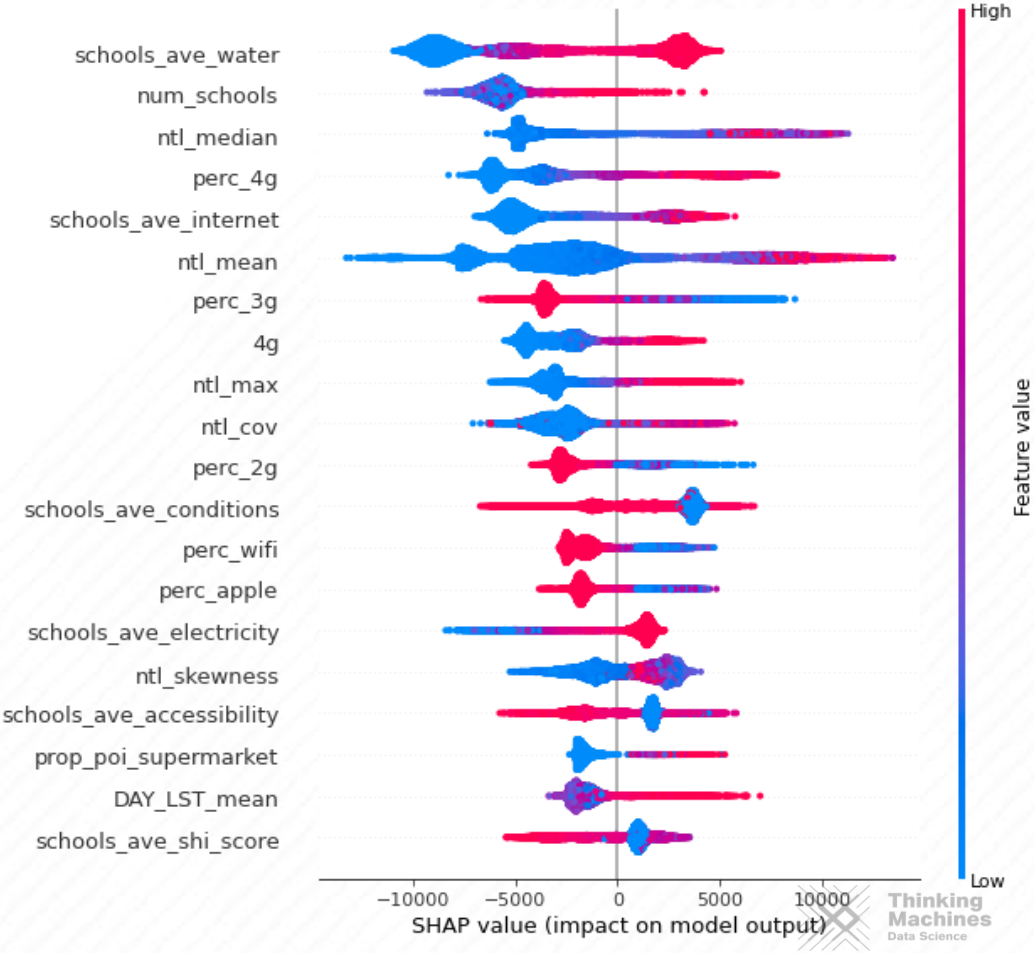
Thinking Machines Data Science

RESULTS

4G interestingness further validated by SHAP

Further research

- Are Telcos particularly good at identifying areas becoming wealthier? Does 4G access cause growth in wealth?



SOLUTIONS

Can we use unconventional data sources to infer socioeconomic indicators?

1 Our first study replicates existing global studies and validates the method's usefulness in the Philippine context

2 While our second method improves on the first.

- This method is highly **explainable** and **interpretable** compared to the first set of models
- Generating nationwide wealth estimates costs ~$1,000 per run of cloud compute using the first set of computationally intensive models. This model runs in 5 mins with a per run cost of ~$20. This is a cost level which enables iteration and experimentation!

Thinking Machines Data Science

SOLUTIONS

We would like to participate in deeper collaboration between industry, academe, and government to apply machine learning and big data methods to support and augment ground truth studies, in support of a stronger Philippines.

Selected Reference Material

- Tingzon, Isabelle, et al. “Mapping Poverty in the Philippines Using Machine Learning, Satellite Imagery, and Crowd-sourced Geospatial Information.” *International Conference for Machine Learning AI for Social Good Workshop*, Long Beach, United States, 2019. URL https://aiforsocialgood.github.io/icml2019/accepted/track1/pdfs/7_aisg_icml2019.pdf.
- Engstrom, R., Hersh, J., and Newhouse, D. Poverty from space: using high-resolution satellite imagery for estimating economic well-being, 2017.
- Head, A., Manguin, M., Tran, N., and Blumenstock, J. E. Can human development be measured with satellite imagery? In Proceedings of the Ninth International Conference on Information and Communication Technologies and Development, pp. 8. ACM, 2017.
- Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., and Ermon, S. Combining satellite imagery and machine learning to predict poverty. Science, 353(6301):790–794, 2016.
- Philippine Statistics Authority PSA, I. Philippines national demographic and health survey 2017, 2018. URL http://dhsprogram.com/pubs/pdf/FR347/FR347.pdf.

Stephanie Sy
stef@thinkingmachin.es

Thinking Machines
Data Science