

Global poverty estimation using private and public sector big data sources

Rob Marty (Data Scientist, DECDI)

Alice Duhaut (Economist, DECDP)

Introduction: Using low-cost, globally available data sources for poverty estimation

- **Accurate poverty estimates important** for the design, delivery and evaluation of social programs.
- **Obtaining accurate poverty estimates through household surveys is expensive** and nationally representative surveys often fielded infrequently
- **Growing literature uses global, spatially-referenced data** to estimate poverty; use surveys to train ML models using features derived from geospatial data sources
- **Increasing availability of global, spatially referenced datasets (satellite imagery, private sector, etc) suggests promise in capturing socio-economic indicators in low-cost way**



Introduction: Using low-cost, globally available data sources for poverty estimation

- **Jean et al. (2016)**: Uses **daytime and nighttime imagery** to estimate wealth in five countries
- **Fatehkia et al. (2020)**: **Facebook marketing data performs similarly to satellite imagery** for wealth estimation in the Philippines and India; performs better in the Philippines where Facebook usage is higher
- **Pokhriyal and Jacques (2017)**: Tests using **CDR data, climate and environmental variables, and OpenStreetMaps** for wealth estimation in Senegal.
- **Yeh et al. (2020)**: Uses **daytime and nighttime imagery** for 23 countries across sub-Saharan Africa. Model performance not associated with country level indicators like GDP, population, and urban population; model performance lower in countries where within-village variance of wealth is higher
- **Chi et al. (2022)**: Uses features from **satellite imagery, Facebook connectivity data, and OpenStreetMaps** for wealth estimation across 56 countries; mobility connectivity among the most predictive features of wealth

Literature investigates: Which data sources are predictive of wealth/poverty, how do models/data sources perform in different contexts

Article | [Open access](#) | Published: 22 May 2020

Using publicly available satellite imagery and deep learning to understand economic well-being in Africa

RESEARCH ARTICLE



Combining satellite imagery and machine learning to predict poverty



RESEARCH ARTICLE |

Microestimates of wealth for all low- and middle-income countries

Our contribution

Build off of Chi et al. (2022) — test using many data sources across many countries. We test additional/alternate data sources.

- Estimate an asset-based wealth index across 59 countries, spanning:
 - Africa [N countries = 36]
 - Eurasia [N countries = 15]
 - The Americas [N countries = 8]
- Train models on data sources including:
 - Daytime satellite imagery
 - Nighttime satellite imagery
 - Radar imagery
 - Weather and climate indicators
 - Facebook marketing data
 - Roads and points of interest (OpenStreetMaps)

Primary research questions

1. Which data sources are most predictive of levels and changes in wealth?
2. What are the characteristics of countries where models perform best?

Data: Survey Data + Wealth Index

- Rely on Demographic and Health Surveys (DHS)
- **Wealth index:** First principle component of socio-economic variables—assets, floor/roof/wall material, HH has piped water, N people sleeping per bedroom.
- **Estimating levels of wealth**
 - Rely on most recent survey for each country.
 - 63,854 survey clusters
 - 59 countries
- **Estimating changes in wealth**
 - Most recent survey and oldest survey that was implemented closest to 2000.
 - Not panel data; match nearest clusters in different years within 10km of each other
 - 7,714 survey clusters
 - 33 countries

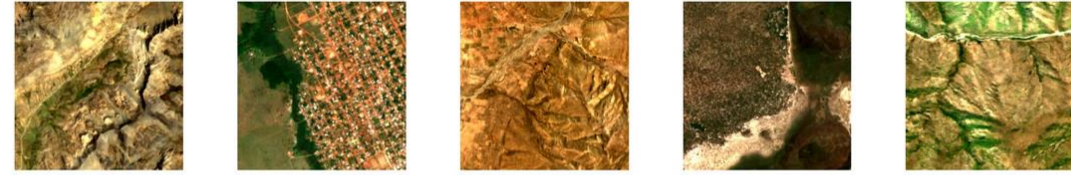


Data: Features from Public/Private Sector Data

Daytime and Nighttime Satellite Imagery

- **Nighttime lights:** Average and standard deviation of NTL; common proxy for economic activity
- **Daytime imagery:** Spectral bands and indices (NDVI and built-up index); captures land use
- **NTL + Daytime:** Features from CNN used to train daytime imagery on NTL; captures features more targeted to wealth estimation
- **MOSAIKS:** Spatial embeddings extracted from high-resolution daytime imagery

Daytime Images from Areas with Low Nighttime Lights



Daytime Images from Areas with Medium Low Nighttime Lights



Daytime Images from Areas with Medium Nighttime Lights



Daytime Images from Areas with Medium High Nighttime Lights



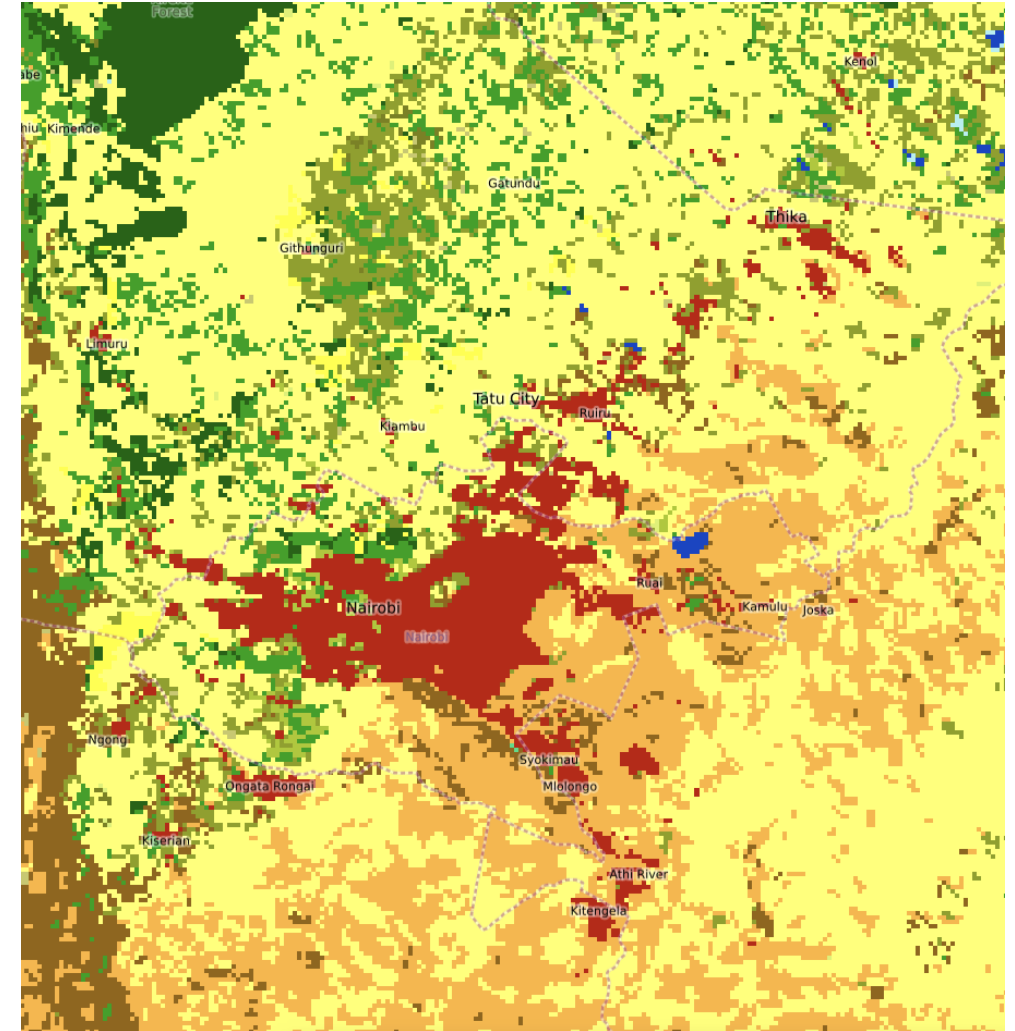
Daytime Images from Areas with High Nighttime Lights



Data: Features from Public/Private Sector Data

Satellite-Derived Data

- **Synthetic Aperture Radar Data (Sentinel-1):** Strength of radar signals; used for different contexts such as vehicle detection, crop classification, urban change monitoring, etc.
- **Land Cover (ESA):** Proportion of land classified according to 36 different land cover classes
- **Elevation & Slope (SRTM):** Average near survey cluster
- **Climate (WorldClim):** 19 bioclimatic variables taking average of 1970-2000—such as temperature, precipitation, mean temperature of wettest quarter, etc.
- **Weather (ERA5):** Average precipitation and temperature
- **Pollution (Sentinel-5P & MODIS):** NO₂, CO, AOD, etc.



Data: Features from Public/Private Sector Data

Other Indicators

- **Roads and points of interest (OpenStreetMap)**
 - Number and distance to points of interests (schools, health facilities, parks, etc)
 - Length and distance to roads of different types
- **Facebook Marketing Data:** Proportion of monthly active Facebook users near survey cluster according to select attributes, such as:
 - Behaviors (eg, frequent travelers, early tech adopter)
 - Education (eg, more than high school)
 - Phone usage (eg, high end vs lower end phone)
 - Interests (eg, restaurants, travel, fitness)
 - Mobile OS (eg, iOS vs android)
 - Network access (eg, 4G vs 2G)

The Relative Value of Facebook Advertising Data for Poverty Mapping

Masoomali Fatehkia,¹ Benjamin Coles,² Ferda Ofli,¹ Ingmar Weber¹

¹Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar

²Princeton University, Princeton, NJ, USA




{mfatehkia, fofli, iweber}@hbku.edu.qa, bcoles@princeton.edu

Using Facebook advertising data to describe the socio-economic situation of Syrian refugees in Lebanon

[Masoomali Fatehkia](#)¹, [Zinnya del Villar](#)², [Till Koebe](#)^{2,3}, [Emmanuel Letouzé](#)^{2,4}, [Andres Lozano](#)², [Roaa Al Feel](#)⁵, [Fouad Mrad](#)⁵, [Ingmar Weber](#)^{1,6,*}

RESEARCH ARTICLE | ✓

Analyzing gender inequality through large-scale Facebook advertising data

David Garcia , Yonas Mitike Kassa, Angel Cuevas, , and Ruben Cuevas  [Authors Info & Affiliations](#)

Correlated impulses: Using Facebook interests to improve predictions of crime rates in urban areas

Masoomali Fatehkia, Dan O'Brien, Ingmar Weber 

Methods

- Use Extreme Gradient Boosting (XGBoost) as well suited to handle high-dimensional data
- Test four approaches (balance sample size vs similar data)
 - **Within country estimation:** Divide country into 5 folds based on ADM2. Train model on 4 folds and predict in remaining fold
 - **Within continent estimation:** Train model on all countries except country i and predict wealth in country i
 - **Other continents estimation:** Train model on other continents and predict wealth in remaining continent
 - **Global estimation:** Train model on all countries except country i to predict wealth in country i
- For changes: Compute changes in wealth and features; directly train on changes

Performance metrics

Squared Pearson correlation: Strength of linear association; how well values move together

$$r^2 = \text{cor}(y, \hat{y})^2$$

Coefficient of determination: How well values match; penalizes bias (predictions too large or too small)

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

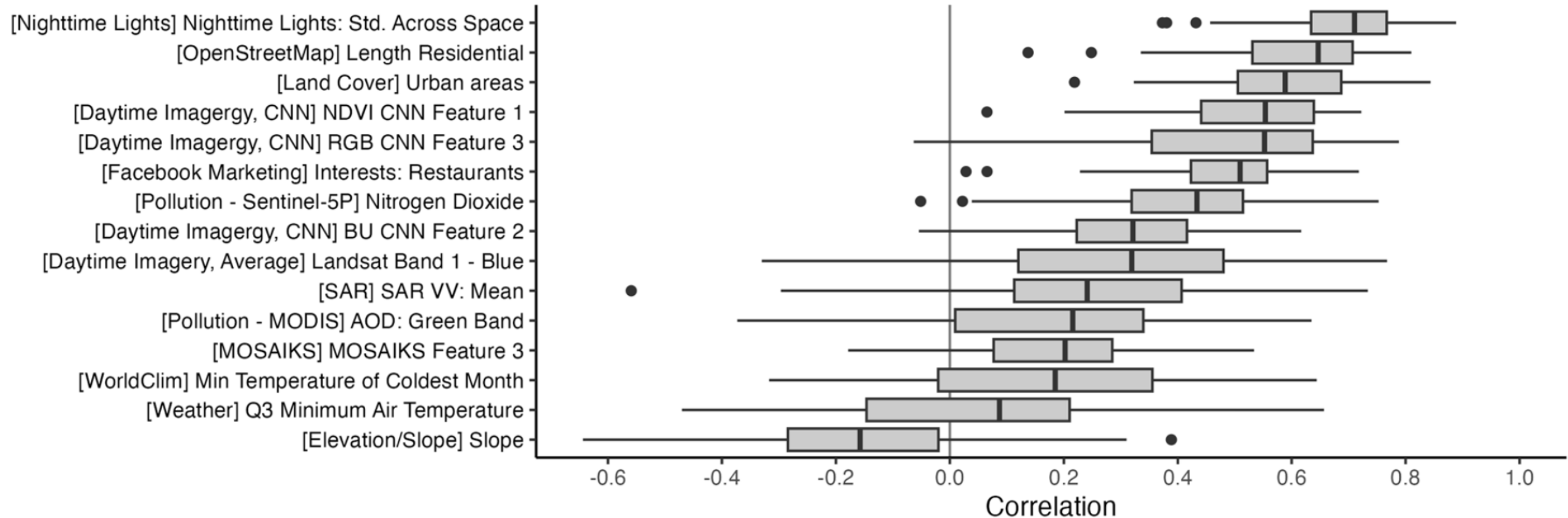
Within Country Correlation [Levels]

Compute correlation for each country and plot distribution

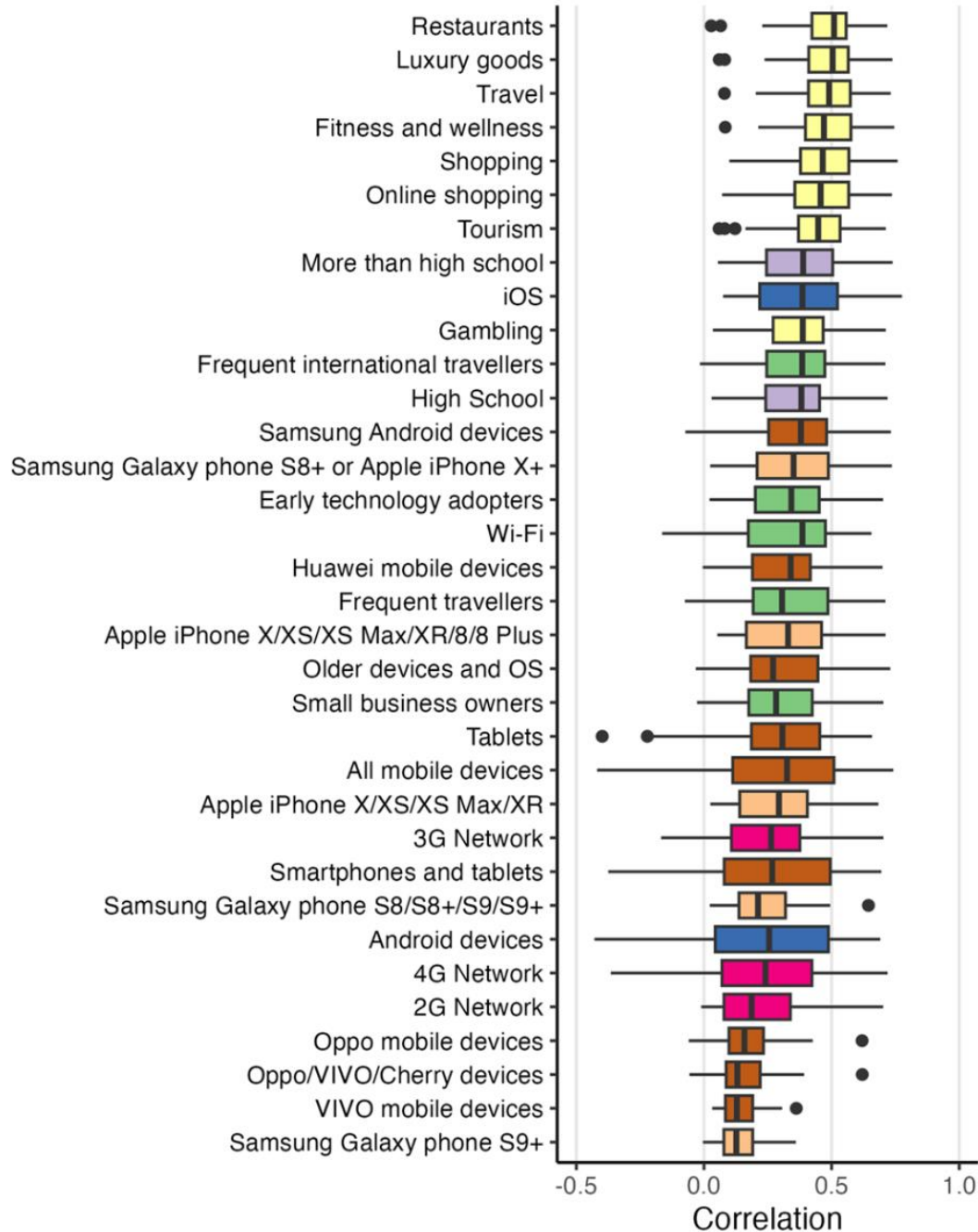
Features from nighttime lights, OSM, and land cover most strongly correlated with wealth. Features from most variable types see high correlations, but distribution varies

A. Correlation of select variables to wealth index across countries

The variable with the highest median correlation for each dataset is shown



B. Correlation of Facebook variables to wealth index



Within Country Correlation [Levels]

Compute correlation for each country and plot distribution

Many "interest" variables have strong association with wealth, although many features across categories correlated with wealth

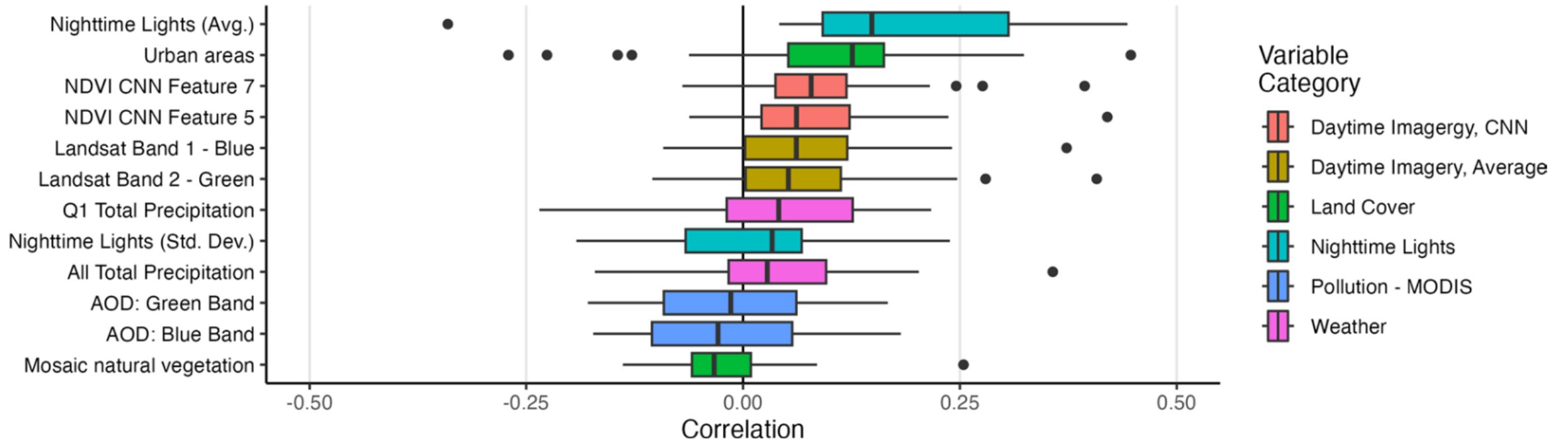


Within Country Correlation [Changes]

Compute correlation for each country and plot distribution

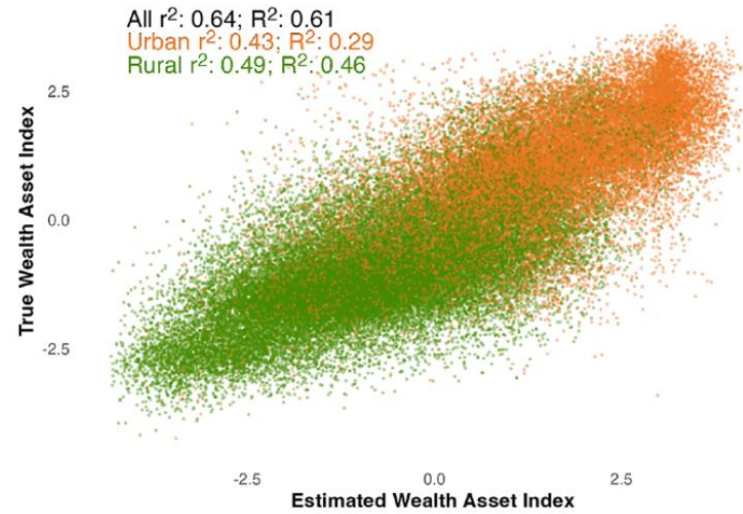
- *As with levels, changes in nighttime lights most strongly correlated with changes in wealth*
- *Correlations for changes lower in magnitude compared to levels*

Correlation between changes in variables and wealth index

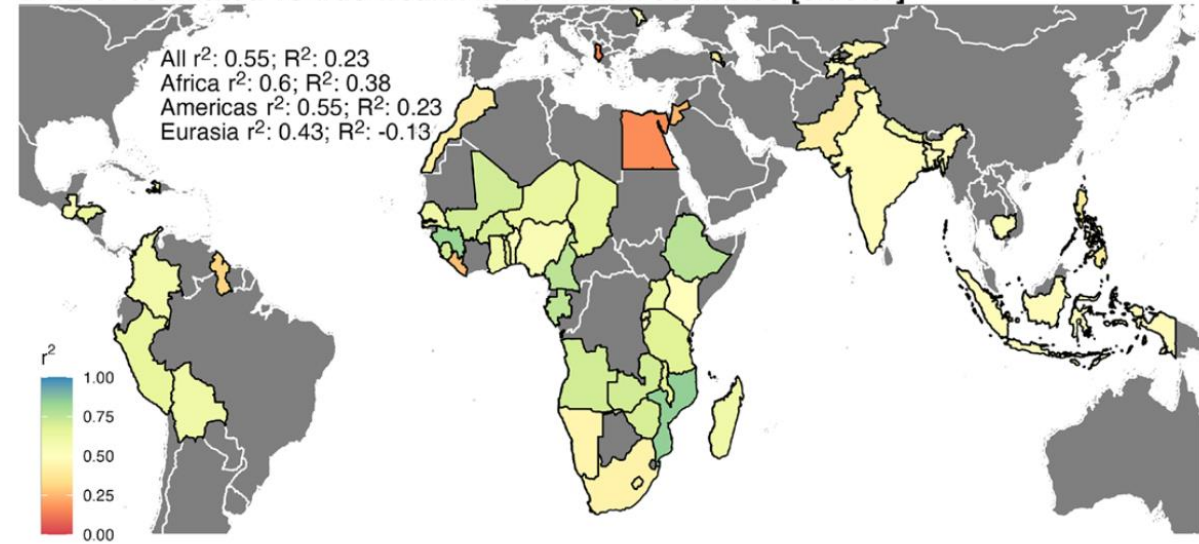


Results [Levels]

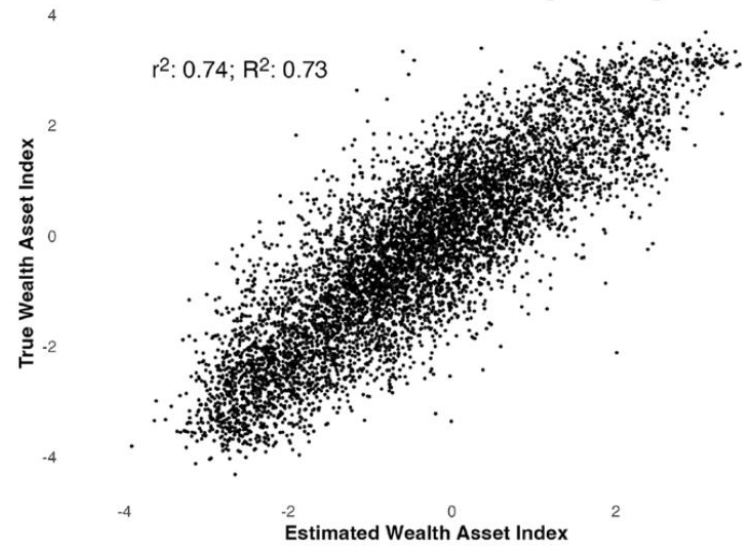
A. Estimated vs. true wealth index [cluster]



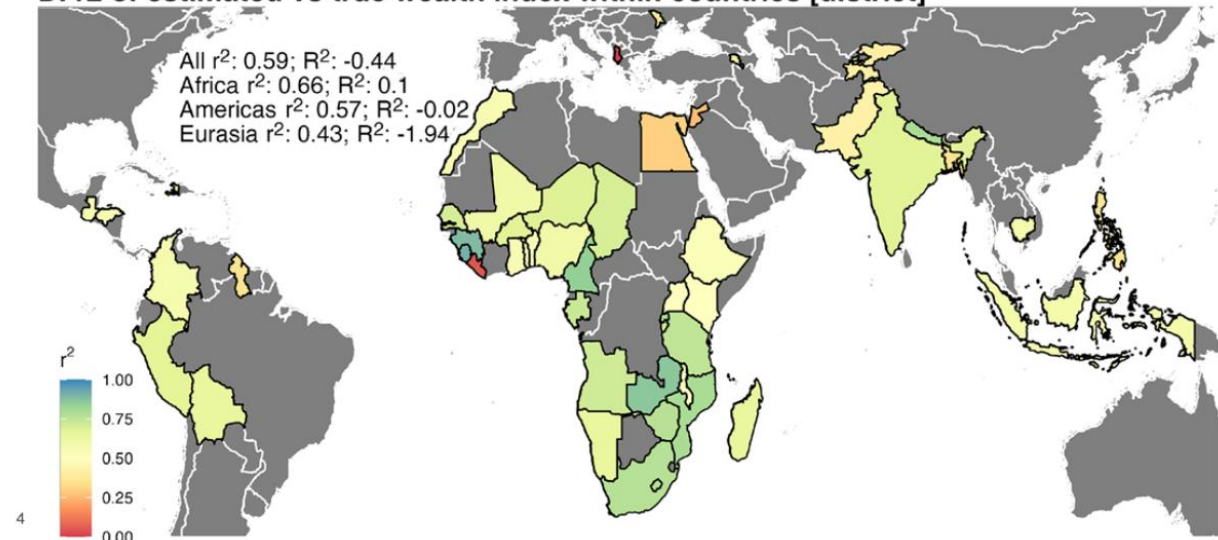
B. r^2 of estimated vs true wealth index within countries [cluster]



C. Estimated vs. true wealth index [district]

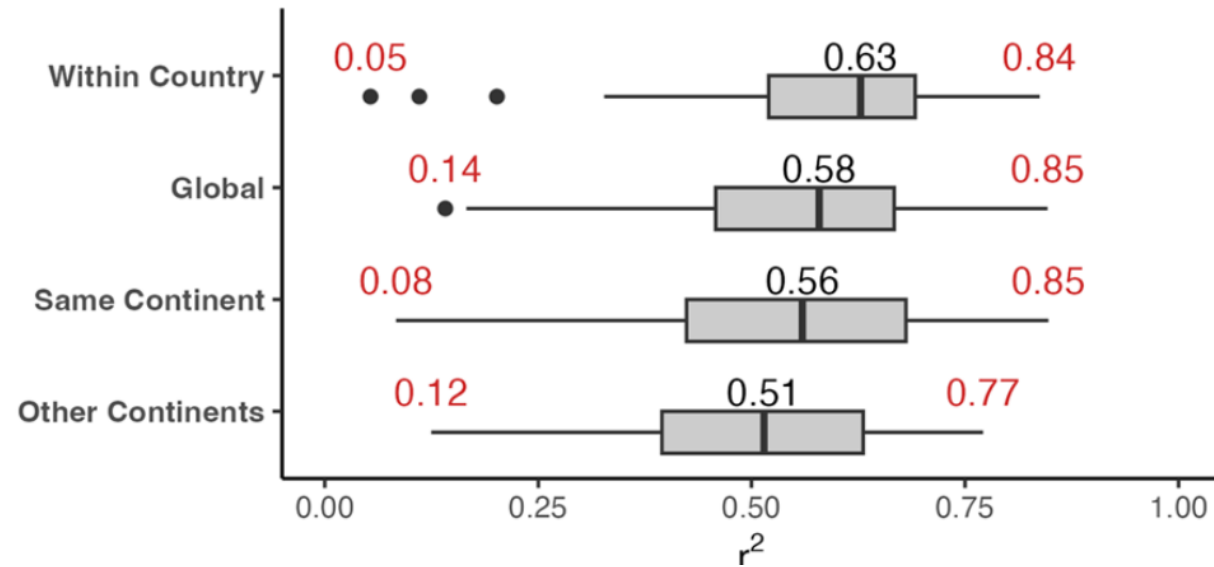


D. r^2 of estimated vs true wealth index within countries [district]



Results [Levels]

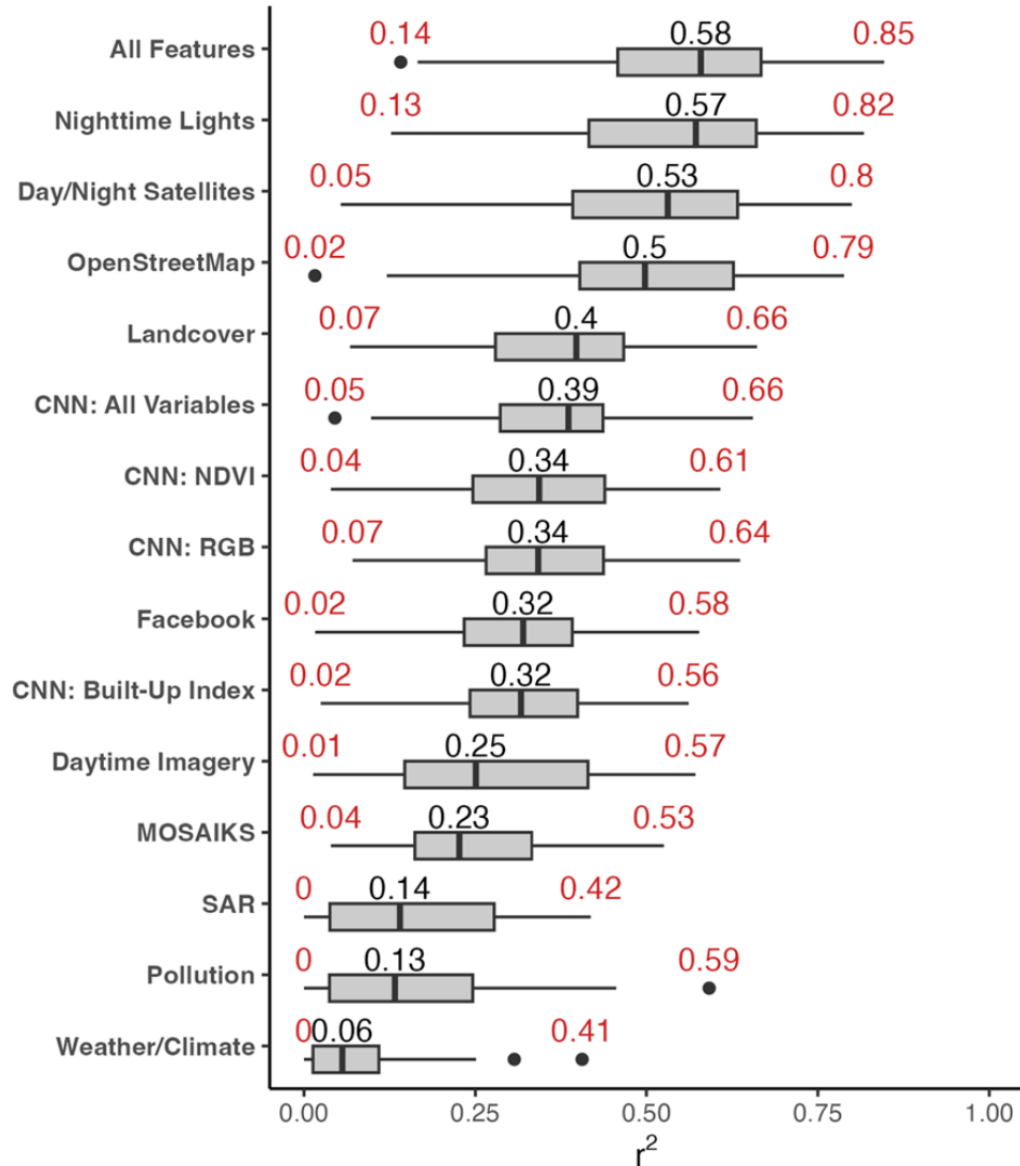
A. Performance by training sample type [cluster]



- Training on data within a country to predict wealth in the same country works the best (train and test sets from different ADMs)
- Training on less similar data (from a different continent) performs the worst, but still decent.

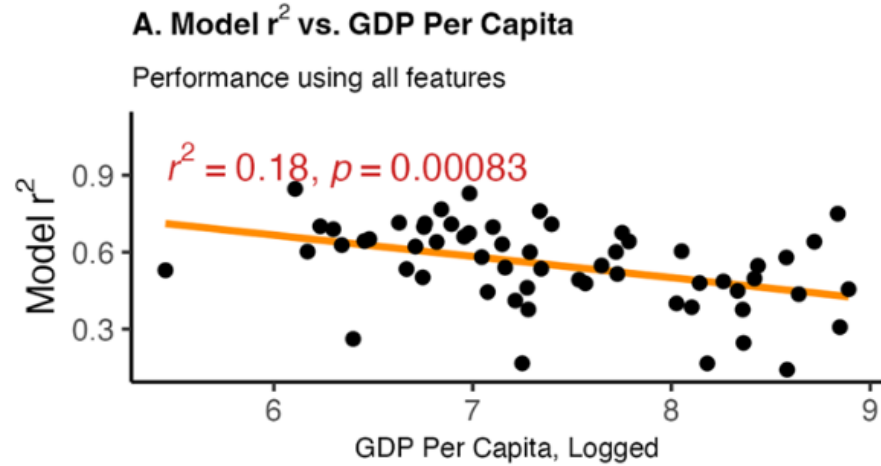
Results [Levels]

B. Performance using different features [cluster]



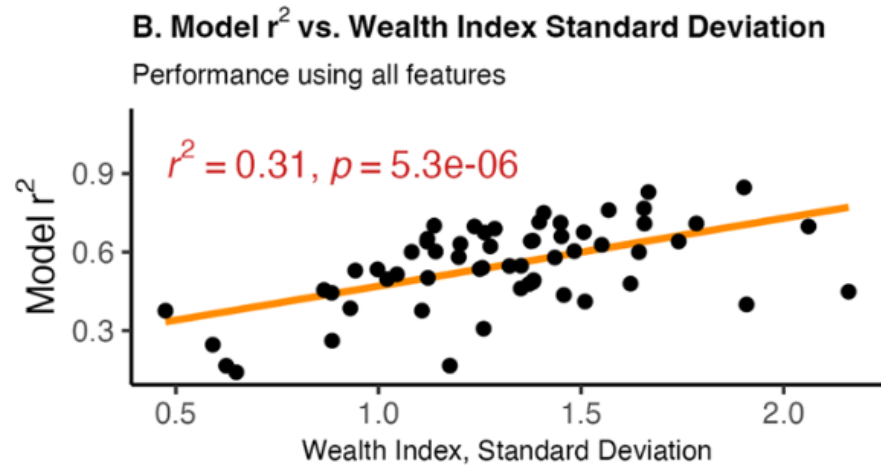
- Training on data using all features works the best—but is closely followed from just using NTL.
- Using only select groups of features: NTL, Day/Nighttime Imagery, and OpenStreepMap perform best

Results [Levels]—Where do models perform best?



Wealth estimation tends to work better in:

- Poorer countries (according to GDP per capita)
- Countries with more variation in wealth (standard deviation in wealth index across survey clusters in a country)

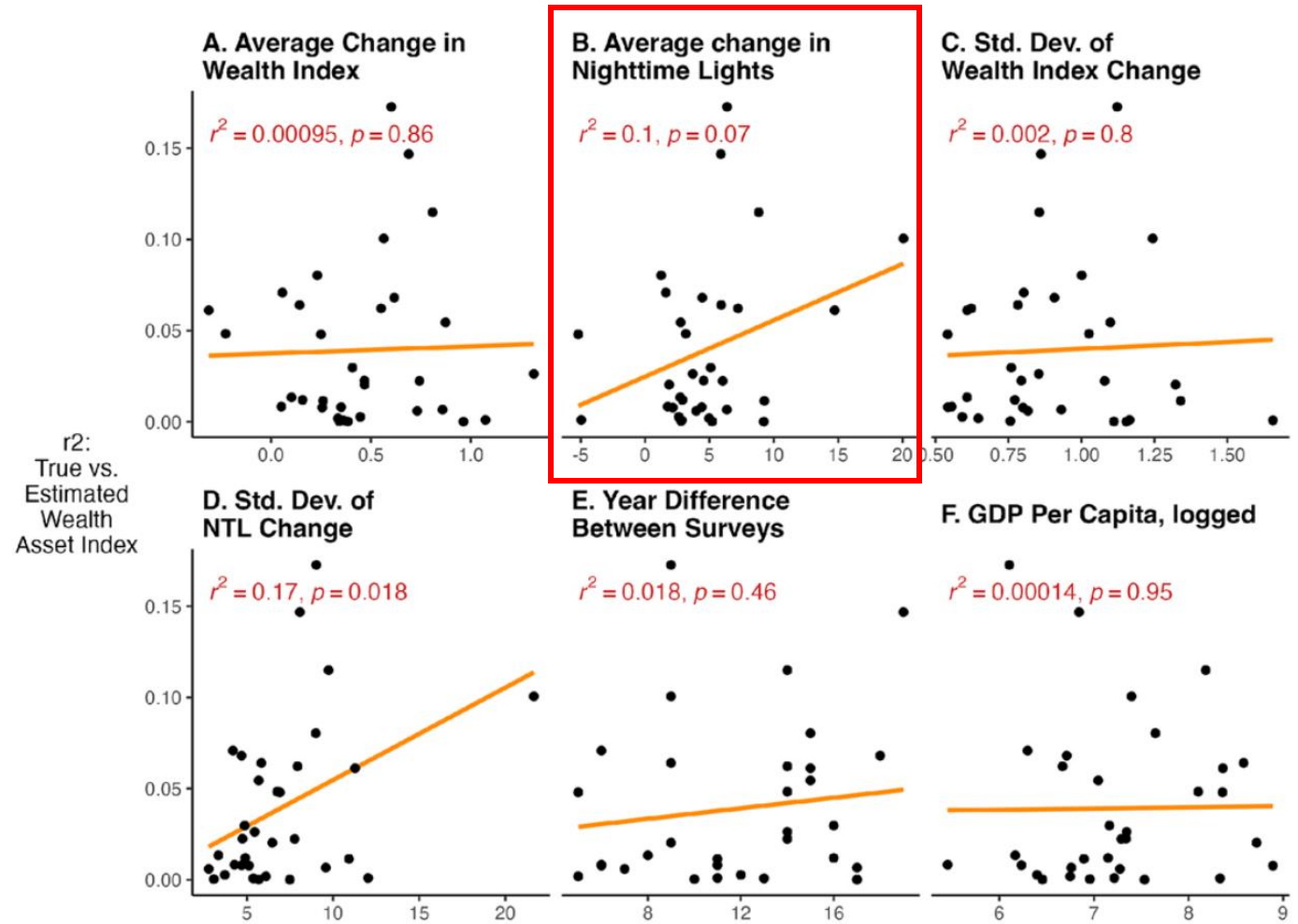
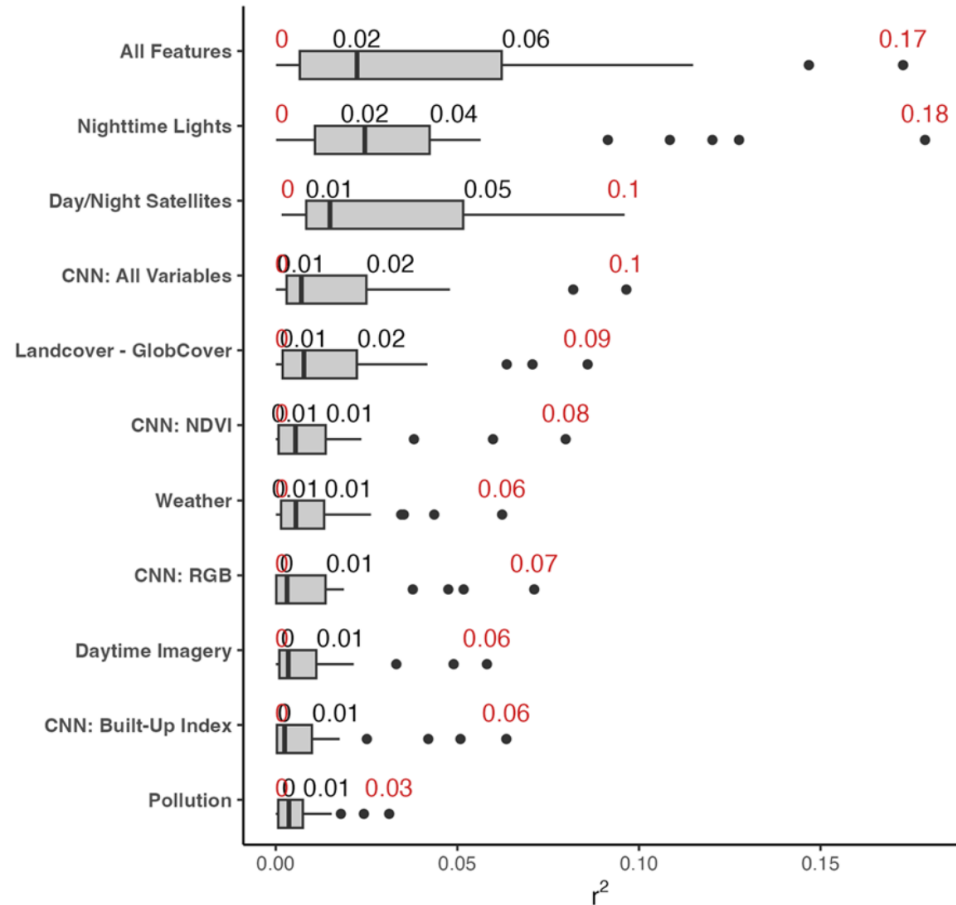


Results [Changes]

Estimating changes in wealth performs less well, but OK for some countries

Models tend to perform better in countries with more visible changes in characteristics (models perform better in countries where NTL grew more)

B. Performance using different features [cluster]



Machine learning vs simple approach?

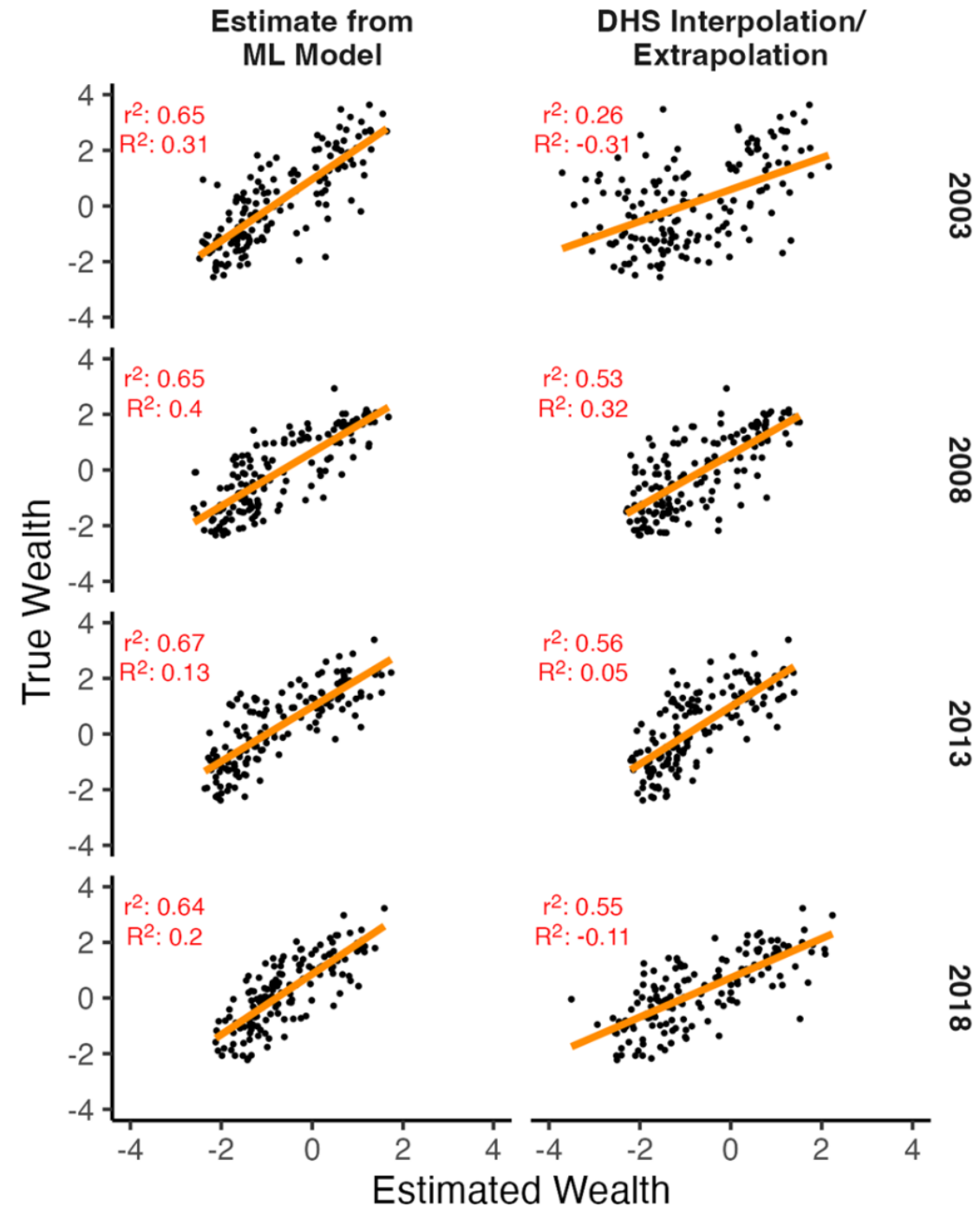
Case Study: Nigeria

Nigeria has four rounds of DHS surveys: 2003, 2008, 2013, and 2018

Compare:

- Train ML data using one year; predict ML in other years
- Interpolate or extrapolate wealth using surveys in multiple years
 - Use data from 2008 and 2013 to extrapolate values to 2018
 - Use data from 2003 and 2013 to interpolate values in 2008

Main result: ML does better for all, but interpolation does OK.



Takeaways

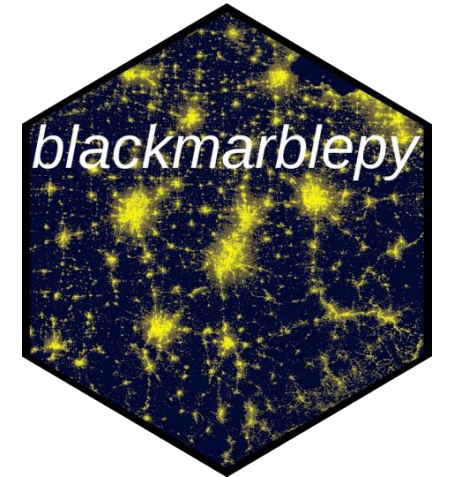
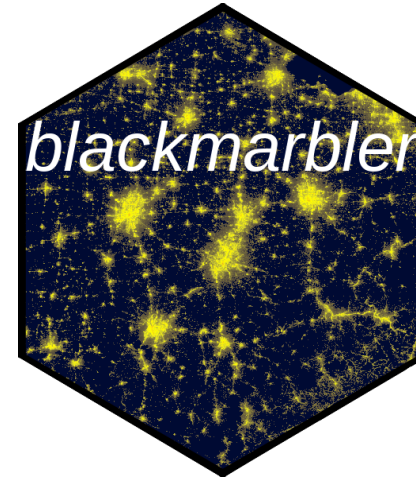
Key results

- Results (for levels) are good but not perfect, but work better in countries where more needed (lower-income)
- Using all features works best, but only using select features performs only slightly worse (balance performance with time to process more data)

Facilitating access to data

- Surveys are gold standard, but are expensive and time consuming.
- Level of effort of ML models should be commensurate with results:
 - ML predictions relying on public/private sector data provide rough (imperfect) estimates of poverty, and therefore should be fast and easy to implement
 - Created public tools to facilitate access to key variables

R & Python packages to access nighttime lights



R package to access Facebook marketing data



Thank You