Global poverty estimation using private and public sector big data sources

Robert Marty[1] and Alice Duhaut*[1]

[1]World Bank

October 23, 2023

*Corresponding author: aduhaut@worldbank.org

Source	Time Span	Level/Change	Features for Poverty Estimation
Daytime and Nighttime Satellite Imagery			
VIIRS	2012- Present	Level	Nighttime lights: Average, standard deviation over time, and standard deviation over space
Harmonized DMSP-OLS and VIIRS	1992 - 2021	Changes	Nighttime Lights: Average and standard deviation over space
Landsat 7	1999 - 2021	Both	Spectral bands and indices (NDVI and build-up index): Average, standard deviation over time, and standard deviation over space
-	-	Changes	Convolutional neural network used to train daytime imagery on nighttime lights; features from CNN extracted
Sentinel-2	2015 - Present	Levels	Convolutional neural network used to train daytime imagery on nighttime lights; features from CNN extracted
MOSAIKS	2019	Levels	Features extracted from high resolution daytime imagery
Synthetic Aperture Radar Data			
Sentinel-1	2014 - Present	Levels	Synthetic aperture radar data, measuring average and standard deviation of VV and VH signals, and the ratio of the two—VV/VH. VV indicates vertical transmit, vertical receive, and VH indicates vertical transmit, horizontal receive.
Facebook Marketing Data			
Facebook	Present	Levels	Proportion of monthly active Facebook users according to select attributes (e.g., proportion of Facebook users with an iPhone)
Roads and Points of Interest			
OpenStreetMap	Present	Levels	(1) Number of points of interests (POIs) near survey (all and by type—e.g., restaurants, schools, health facilities, etc), (2) distance to nearest POI (all and by type), (3) Length of roads near survey (all and by type—e.g., trunk roads, primary roads, etc), and (4) distance to nearest road (all and by type)
Land Cover and Type			
ESA-GlobCover	1992-2018	Both	Proportion of area near survey classified according to 36 different land cover classes
Shuttle Radar Topography Mission (SRTM)	Time-Invariant	Levels	Average elevation and slope
Weather and Climate			
WorldClim	Average of 1970-2000	Levels	19 bioclimatic variables, including annual mean temperature, annual precipitation, mean temperature of wettest quarter, etc.
European Centre for Medium-Range Weather Forecasts: ERA5	1979-2020	Both	Average annual precipitation and temperature
Pollution			
Sentinel-5P	2018 - Present	Levels	Average pollution levels from six metrics: nitrogen dioxide, carbon monoxide, sulphur dioxide, ozone, formaldehyde, and an aerosol index.
MODIS	2000 - Present	Both	Aerosol optical depth

Table 1: Summary of data sources for poverty estimation.

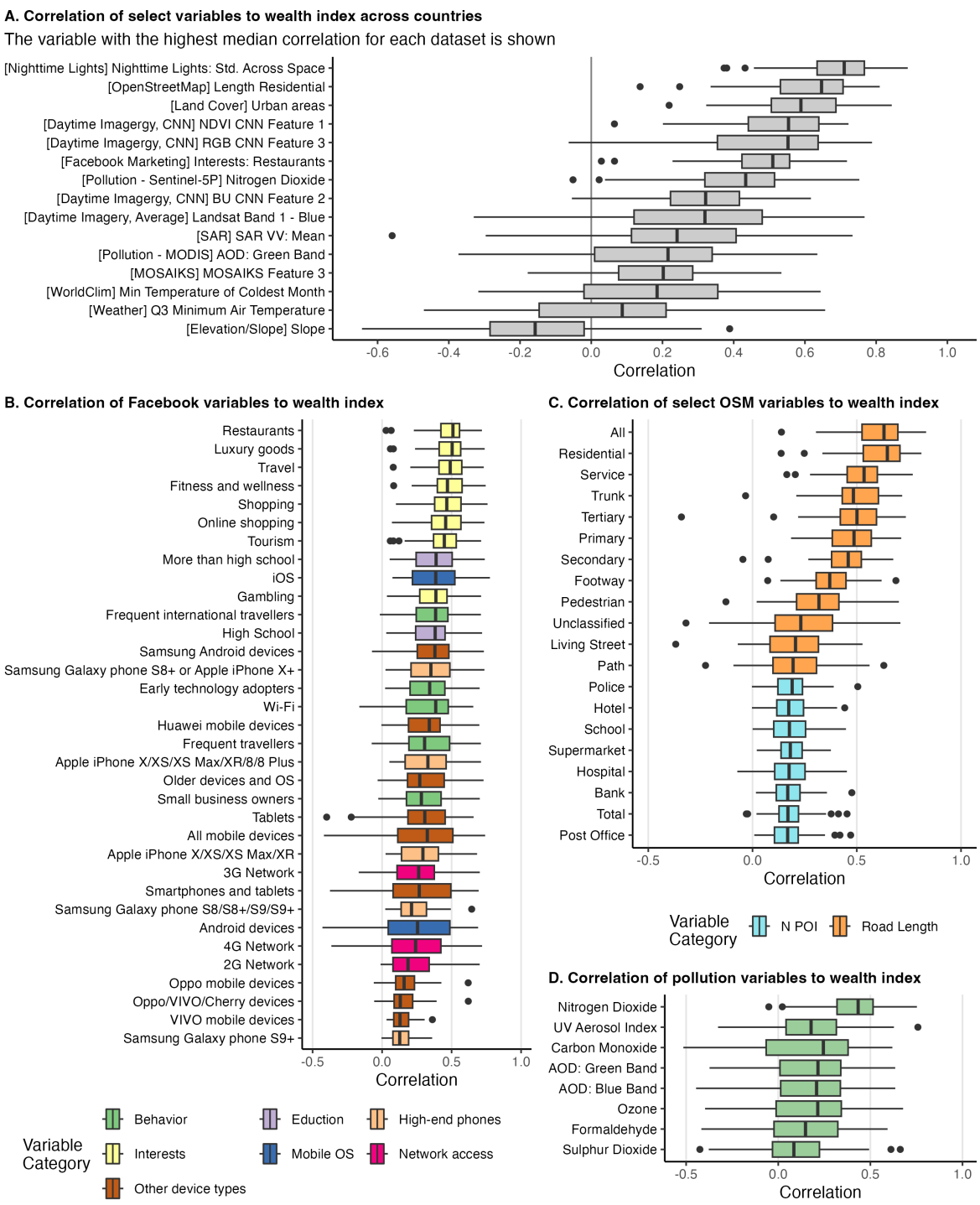

Figure 1: Distribution of within-country correlations of select variables to levels of wealth. Correlations are computed at the cluster level using the latest survey year for each country. **Panel A** shows the distribution of within-country correlations of the feature with the highest median correlation across countries for each dataset. **Panel B** shows the correlation of all variables from the Facebook marketing data. **Panel C** shows select variables from OpenStreetMap data; the panel shows all variables of (1) the length of different classes of roads and (2) the number of different points of interests (POIs) near survey clusters. **Panel D** shows the correlation of all pollution variables from Sentinel-5P and MODIS; AOD variables are from MODIS and the other variables are from Sentinel-5P. The boxplots include: center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; points beyond whiskers, outliers.

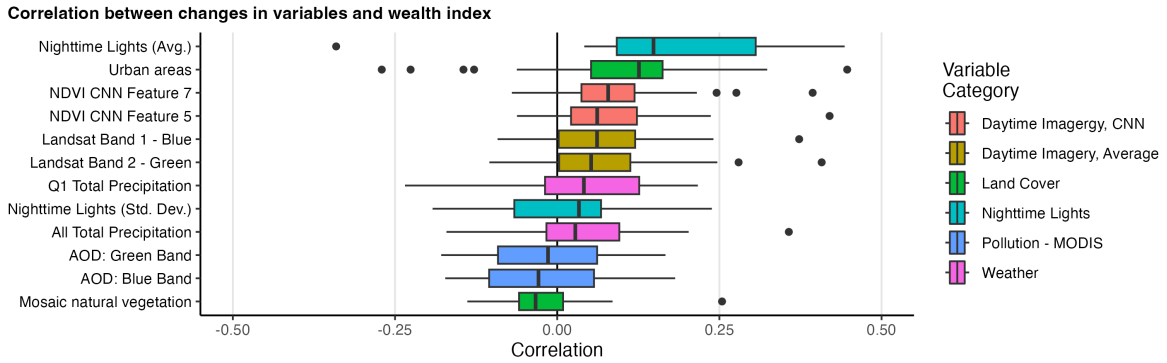

Figure 2: Distribution of within-country correlation of changes in select variables to changes in wealth, using clusters as the unit of analysis. We show the two variables with the highest median correlation for each dataset. The boxplots include: center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; points beyond whiskers, outliers.

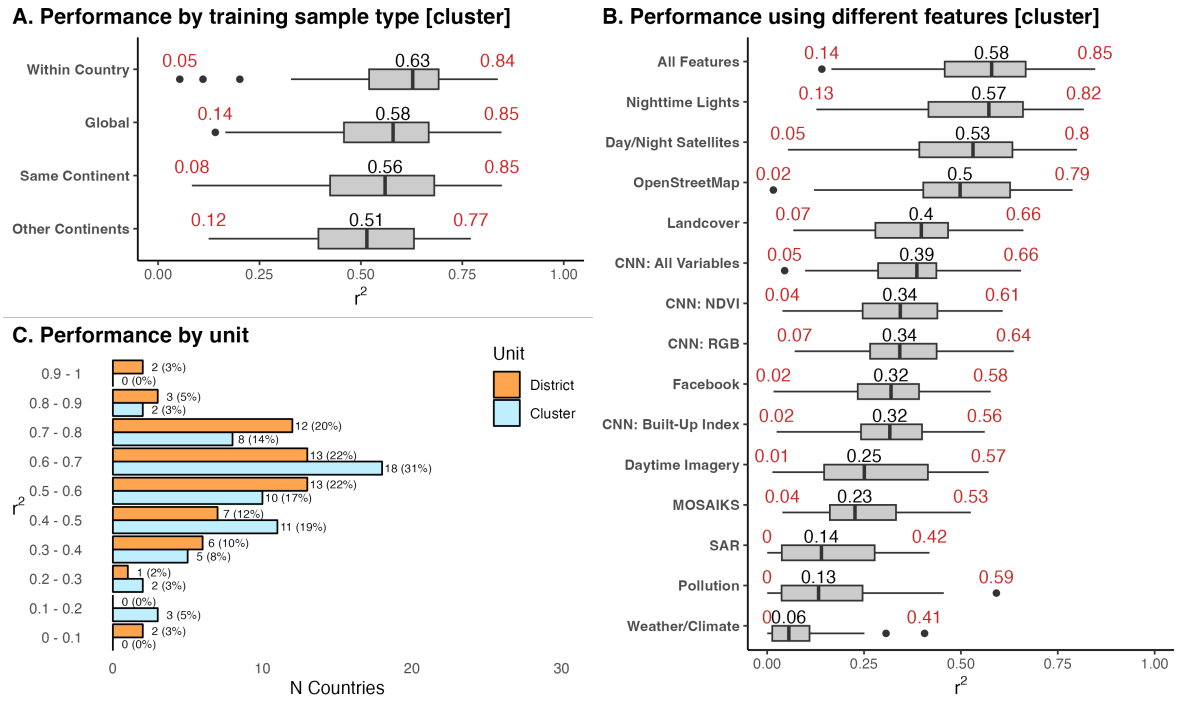

Figure 3: Distribution of model performance across countries, explaining levels of wealth. The black number shows the median and the red numbers show the minimum and maximum r^2. **Panel A** shows model performance by the sample used to train countries. **Panel B** shows model performance when using different sets of features to train models. **Panel C** shows the distribution of model performance at the village and district level. Panels B and C show results where models for each country are trained on data from all other countries (the global training sample approach). The boxplots include: center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; points beyond whiskers, outliers.

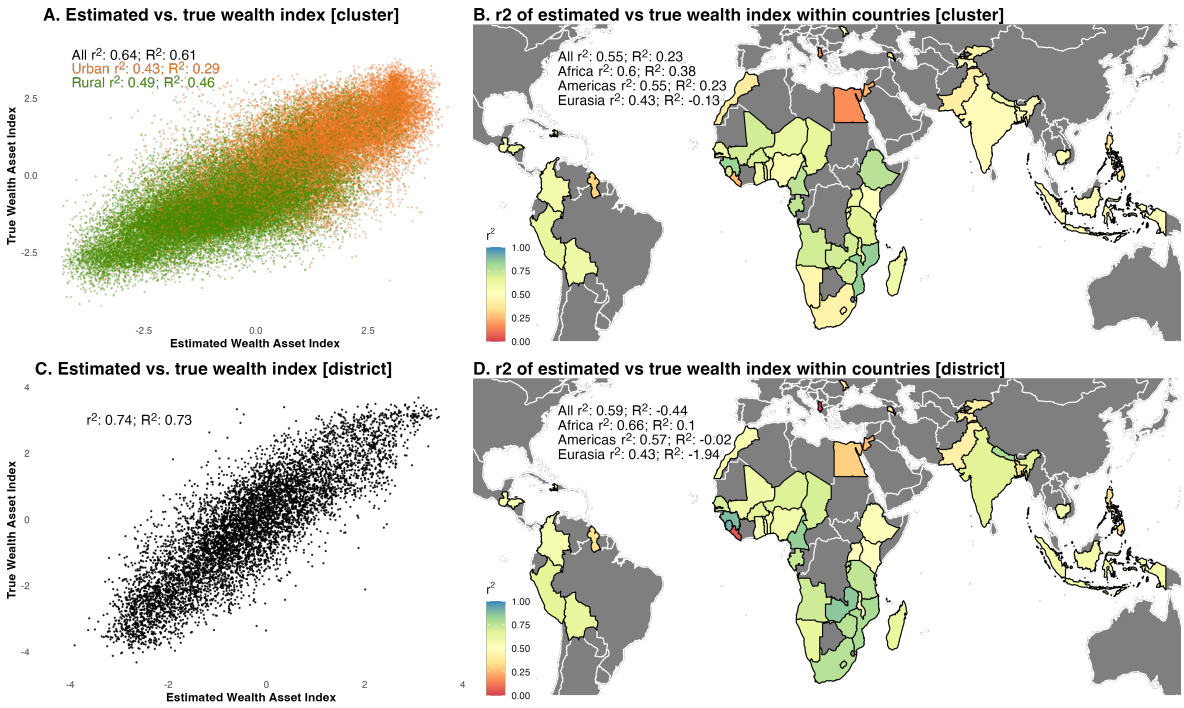

Figure 4: Performance of models predicting levels of wealth index. **Panel A** shows a scatterplot of the estimated and true wealth indices. **Panel B** shows model performance when considering individual countries. **Panel C** and **Panel D** are similar to panels A and B, but using results aggregated at the district level. All panels use results where models for each country are trained on data from all other countries (the global training sample approach). r^2 is the squared Pearson correlation coefficient, and R^2 is the coefficient of determination. The maps in panels B and D were produced using R, version 4.2.2 (`https://www.r-project.org/`); data to produce the country-level basemaps come from Natural Earth (`https://www.naturalearthdata.com/`).

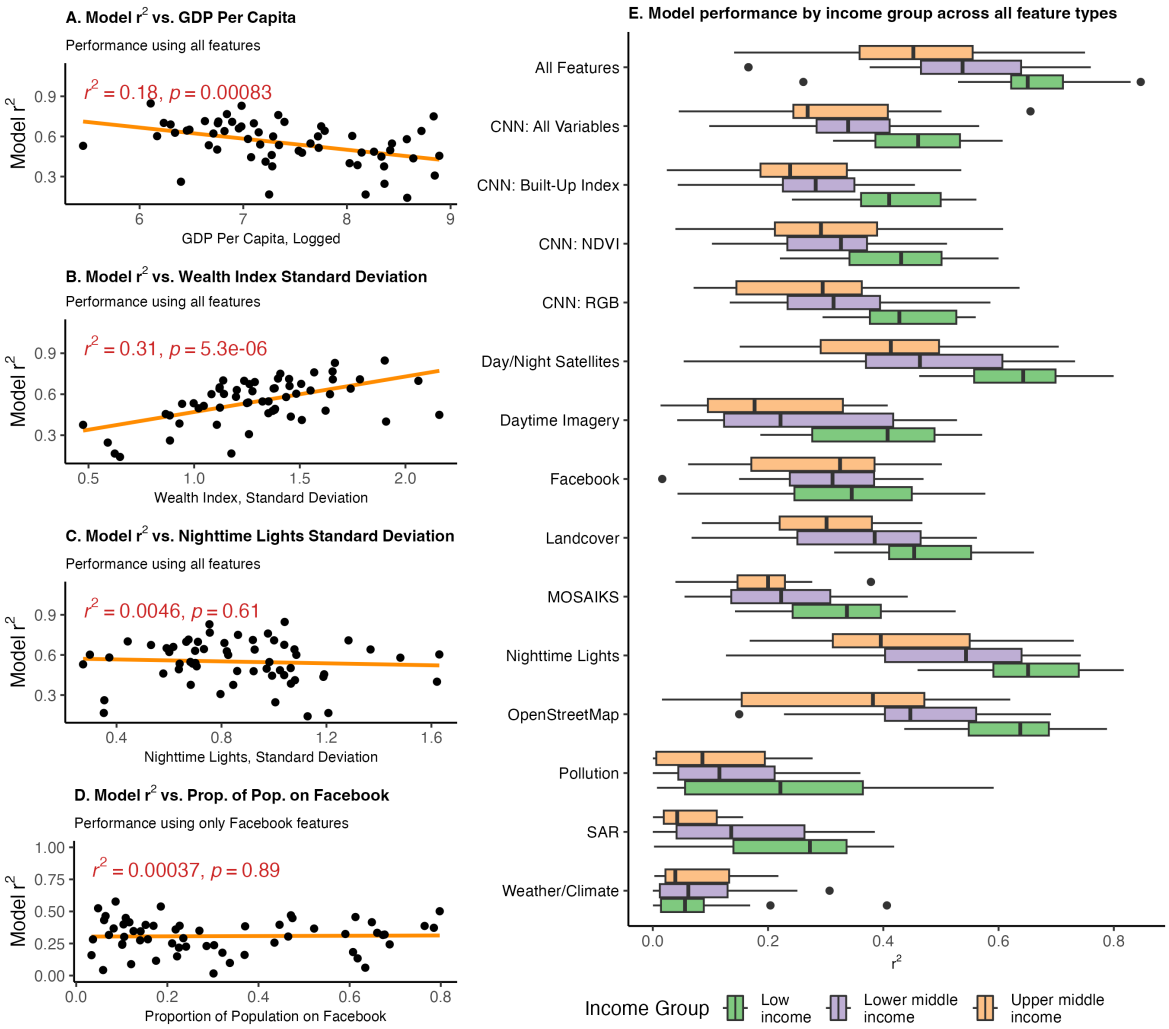

Figure 5: Determinants of the variation in model performance across countries, explaining levels of wealth. **Panel A** shows the association between GDP per capita and model performance. **Panel B** shows the association between the wealth index standard deviation and model performance. **Panel C** shows the association between the nighttime lights standard deviation and model performance. **Panel D** shows the association between the proportion of the population on Facebook—measured using monthly active users divided by a country's population—and model performance using only Facebook features to train the model. **Panel E** shows the distribution of model performance by income level using models trained across different feature sets. The boxplots include: center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; points beyond whiskers, outliers. All panels use results where models for each country are trained on data from all other countries (the global training sample approach), and where the unit of analysis is clusters.

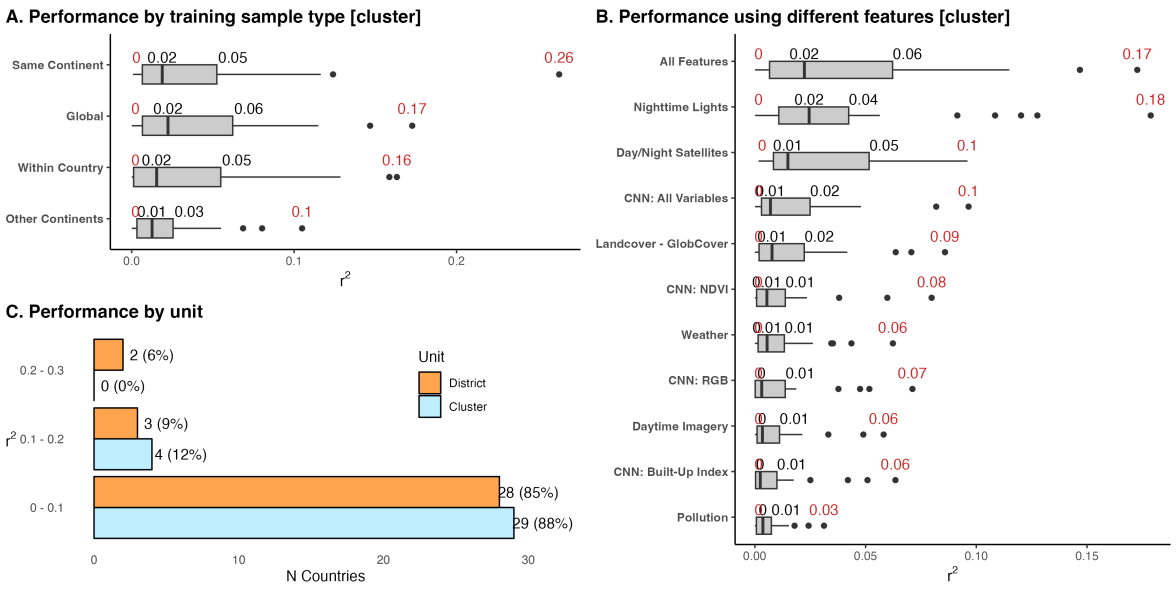

Figure 6: Distribution of model performance across countries, explaining changes in wealth. The black number shows the median and the red numbers show the minimum and maximum r^2. **Panel A** shows model performance by the sample used to train countries. **Panel B** shows model performance when using different sets of features to train models. **Panel C** shows the distribution of model performance at the village and district level. Panels B and C use results where models for each country are trained on data from all other countries (the global training sample approach). The boxplots include: center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; points beyond whiskers, outliers.

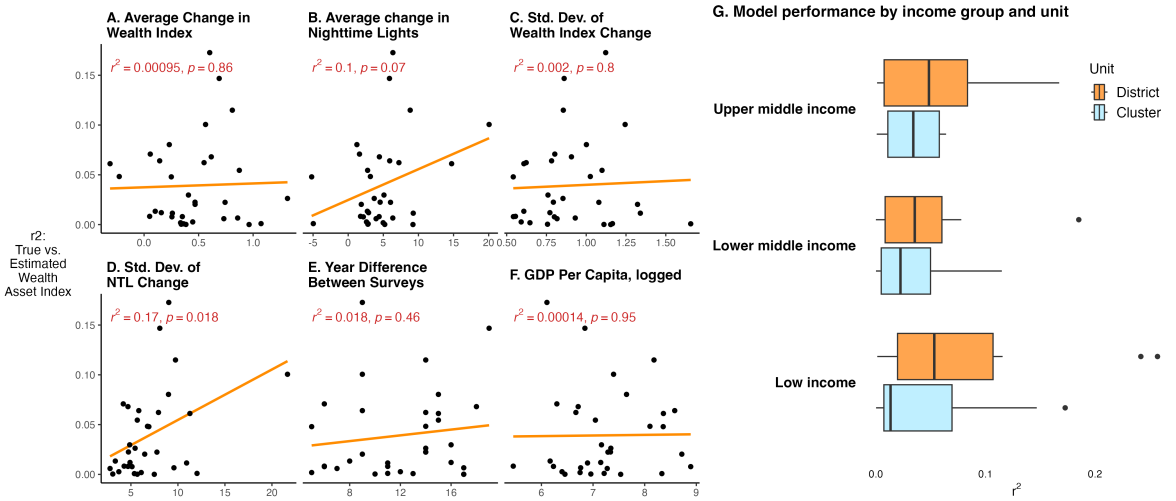

Figure 7: Explaining variation in model performance across countries, explaining changes in wealth. The figure shows the association between model performance and average changes in wealth (**Panel A**), average changes in nighttime lights (**Panel B**), the standard deviation of the change in wealth (**Panel C**), standard deviation of the change in nighttime lights (**Panel D**), the years between between surveys (**Panel E**), and current GDP per capita (**Panel F**). **Panel G** shows the distribution of model performance by income level, using villages and when aggregating to the district level. The boxplots include: center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; points beyond whiskers, outliers.

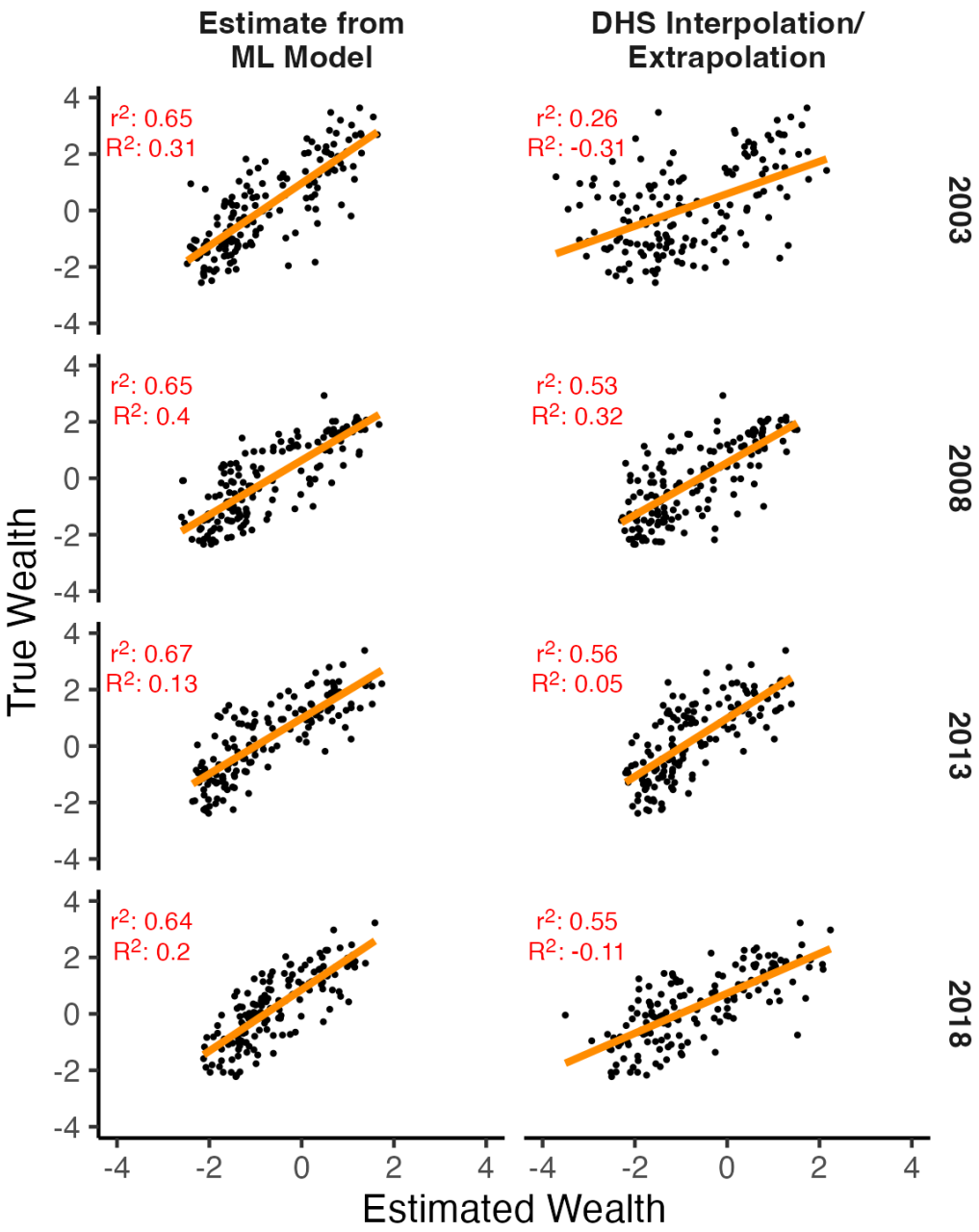

Figure 8: Comparison of true wealth estimates versus estimates from machine learning model and using DHS data to interpolate and extrapolate wealth estimates; data is extrapolated for 2003 and 2018 data, and interpolated for 2008 and 2013. For each year, the two closest DHS rounds are used to estimate wealth; for example, to data in 2003 and 2013 is used to interpolate wealth for 2008. Wealth estimates are at the second administrative division level. The r^2 is the squared Pearson correlation coefficient, and R^2 is the coefficient of determination.

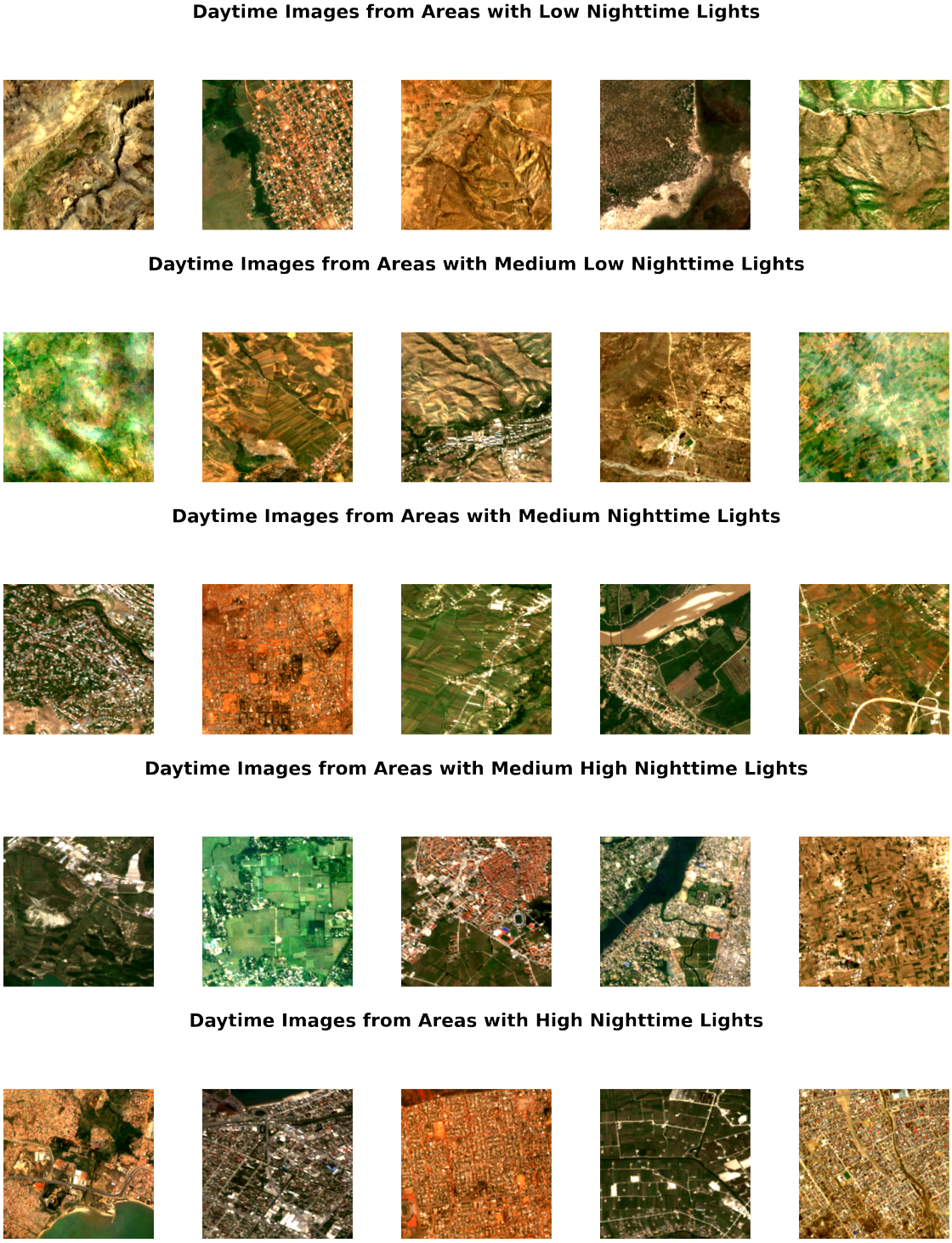

Figure 9: Example daytime images from Sentinel-2 for different nighttime lights groups. The figure was produced using Python, version 3.9.13 (`https://www.python.org/`). Sentinel-2 data was queried using Google Earth Engine.