Measuring Economic Activity Using Open Data

Klaus Ackermann Paul Raschky Miethy Zaman

Introduction

- Africa is the home of most of the least developed economies in the world and in particular, poorer economies are associated with low quality statistical data.
- On a survey of the national income, Jerven (2013) analyses the three major sources of national income data: the world development indicators, Penn world tables and the data sets of Angus Maddison. The range of variation among the data sets for the African countries is much larger than the rest of the world which goes beyond the different methods used to represent the data.
- Prospective, high-quality national and subnational GDP data remain rare in sub-Saharan Africa.
- Inaccurate data and inaccurate primary observations:

-Poor statistical capacity, paired with the possibility of politically motivated misreporting of indicators of economic performance led to widespread concerns about the reliability of GDP figures, in African countries.

-The countries don’t revise the national accounts. Some countries still follow the 1968 standards and some the 1993 ones. Revision of data shows unrealistic changes: ex. a 50% increase in GDP in 2011 for Nigeria (Jerven, 2013).

Alternative data Source

- Asset Index - DHS Survey data

-based on data collected in the Household Questionnaire on household assets

-relies on opinions of standardized questions

-Problem of false negative and false positive

-Difficulty in ascertaining asset quality

- Passively collected data - Night Lights

-Night Lights data is pioneered in the field of economics by Henderson et al (2012).

-Show that night-lights data can supplement measures of economic activity in countries where national statistics are poor.

-Independent of the measurement or reporting error by statistical agencies.

-A fast growing literature in economics uses night-time lights as a proxy for local economic activity (e.g. Hodler and Raschky, 2014; Alesina et al., 2016; Lessmann and Seidel, 2017).

Limitations of Night-lights

- top-coding problem (Bluhm and Krause, 2017)

-Fails to capture the brightness of large and densely populated cities.

- Poor reading of low luminosity

-Nightlights have problem distinguishing between the poor, densely populated areas.

-luminosity levels are generally also very low and show little variation, making nightlights potentially less useful for studying and tracking the livelihoods of the very poor.

-Many zero valued city-years appear in Africa

Limitations of Night-lights

Night-time Lights Composite, Source: https://ngdc.noaa.gov/eog/dmsp/downloadV4composites.html

Limitations of Night-lights

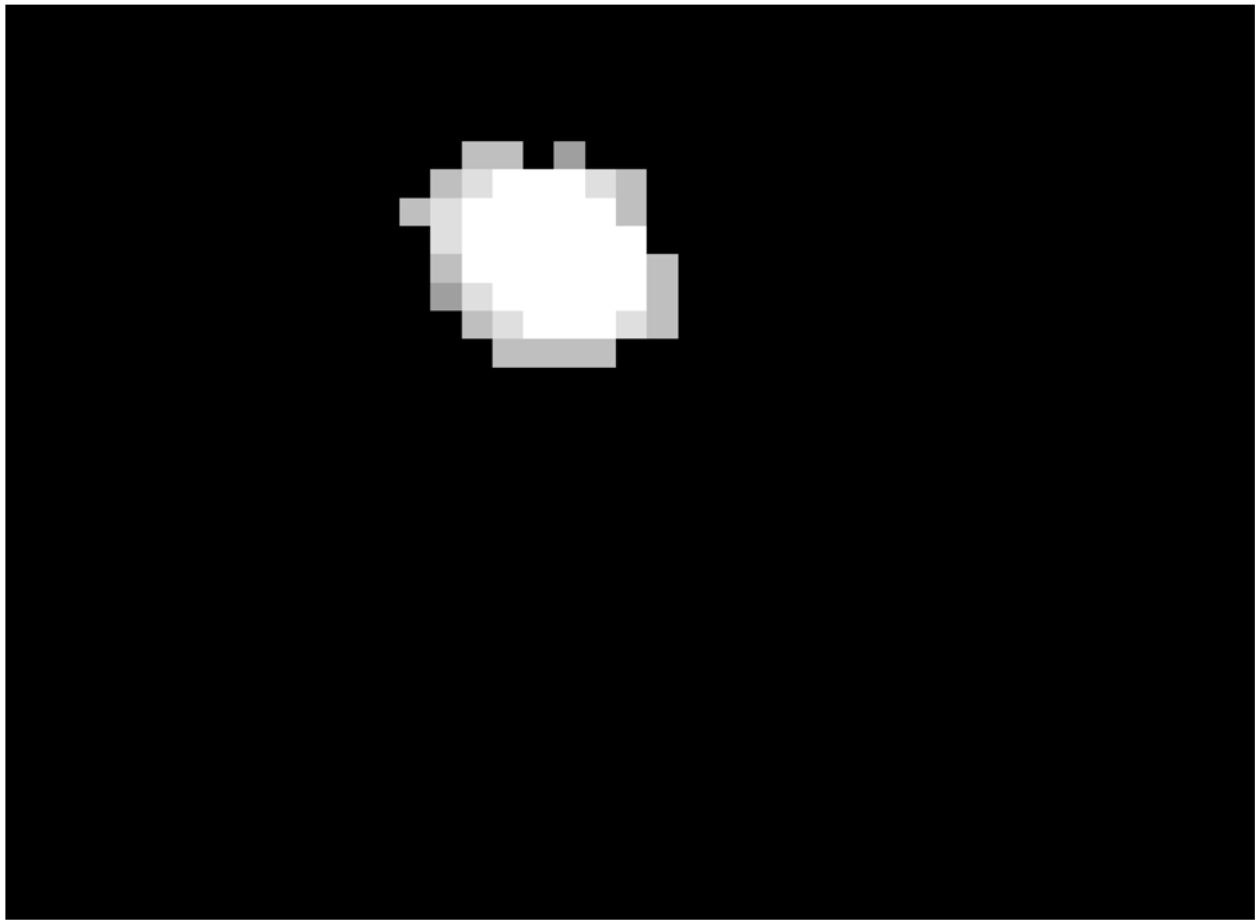


Fig: Nightlight- Kanzenze, Rwanda

Limitations of Night-lights

Fig: Google Daytime Image-Kanzenze, Rwanda

Limitations of Night-lights

Fig: Google Daytime Image-Kanzenze, Rwanda

Limitations of Night-lights

Fig: Google Daytime Image(scale1:4000) -Kanzenze, Rwanda

Limitations of Night-lights

Fig: Google Daytime Image-Mugina, Rwanda

Limitations of Night-lights

Fig: Nightlight-Kayonza, Rwanda

Objective

- The goal of our study is to go beyond night-light data as an indicator of economic activity or wealth.

- Create the wealth index at subnational level using only passively collected open sourced data.

- Motivated from the science paper ‘Combining satellite imagery and machine learning to predict poverty’ by Jean et. al (2016) .

  -Use a machine learning algorithm to predict economic measure.

Objective

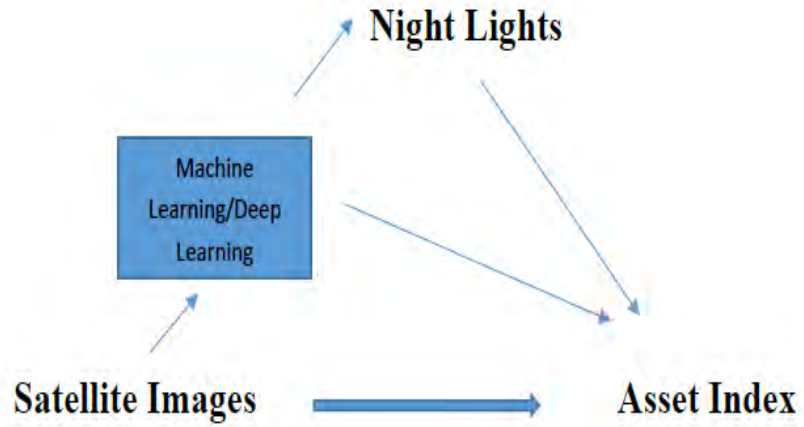
- Three step machine learning/transfer learning procedure used in Jean et. al(2016):

-Use a pre-trained deep neural network on the satellite images to extract 4096 filtered features that includes information on road, water, urban ways, buildings etc.

-Uses night-lights to explain the features extracted

-Predicts an asset score using the DHS survey data and along with the night lights.

-Implements out of sample training for five African countries (Nigeria,Tanzania, Uganda, Malawi, and Rwanda)to estimate Asset score using the daytime images along with night-lights information.

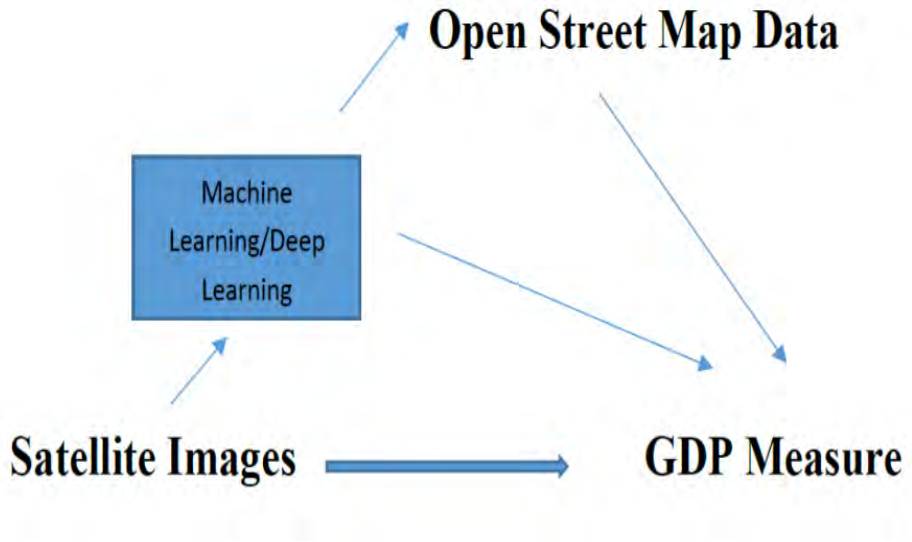


Our Approach

-Use the Satellite Images of 400X400 pixel at zoom level 16 which covers roughly 1sq Km area along with the pre-trained CNN model to extract the 4096 features.

-Use the filtered measures along with the OSM data to predict 18 different OSM's categories (roads, railways, buildings and sub-categories etc.) using European countries with good coverage of the OSM data.

-Use the OSM predicted values to conduct out of sample training to construct a GDP measure for African countries.



Open Street Map

- OpenStreetMap (OSM) is a collaborative project to create an editable map of the world.
- Collects geospatial data and provide geospatial data for anybody to use and share for free.
- OpenStreetMap represents physical features on the ground (e.g., roads, railways, paths, waterways) including total counts and sub-categories.
- Data is also gathered on features along the roads, such as buildings (private and public), parks and natural areas, land use, cultural resources, and recreational facilities.
- Most of Europe especially western Europe have a good coverage of the geodata for the different categories.

Open Street Map

Contrary to jean et al. (2016), we use OSM data instead of night lights in the second stage for our prediction. The intermediate step using night lights is done to learn the image features which are correlated with economic well being (asset index and consumption expenditures). Also both night light day time images have the same layout and thus aids in summarizing the high dimensional input day-time images to a concise vector representation. Even though the second stage does not depend on night lights distinguishing among the poor and estimating the final output, we still refrain from using nightlights in the intermediate stage as our objective is to go beyond night lights.

OSM coverage

Fig: Grid Id 8600400_5549200, Rotterdam

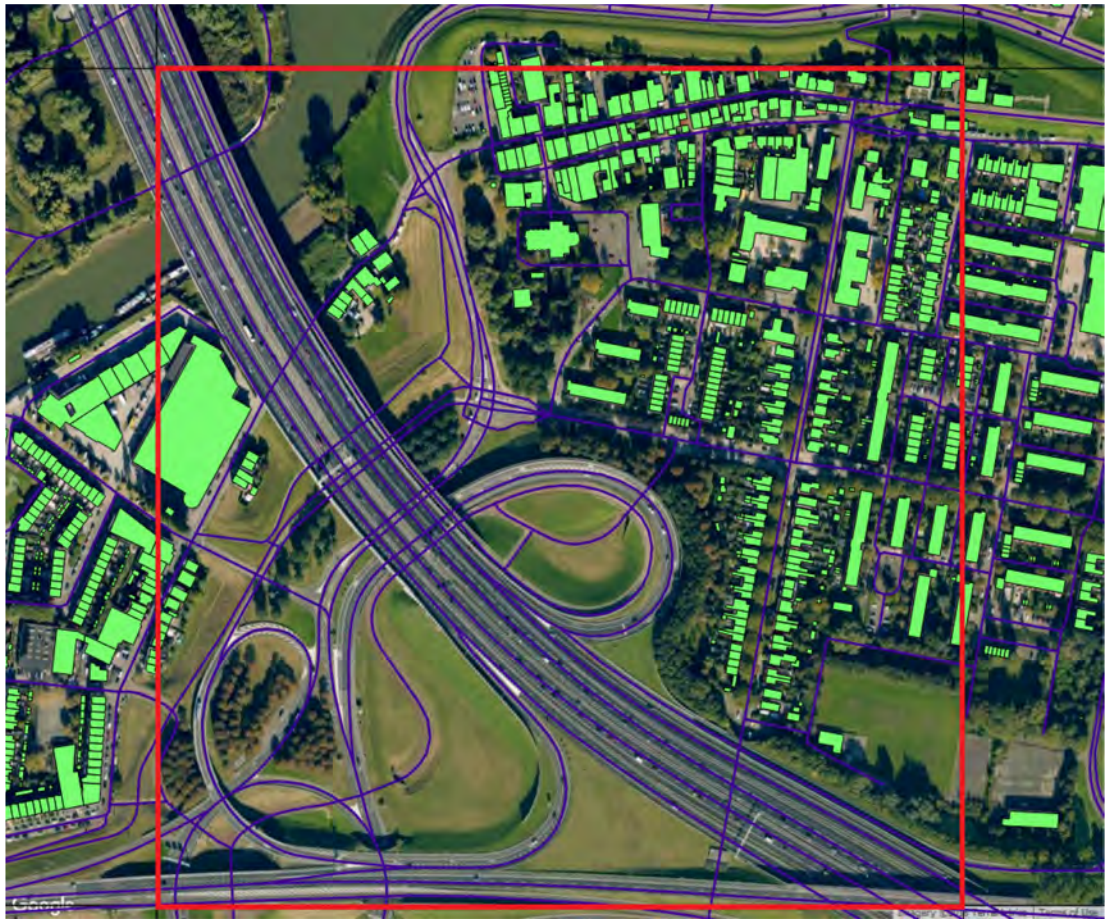


Fig: Grid Id 8600400_5549200, Rotterdam overlayed with OSM categories-Buildings and roads

OSM Data

- **Waterway_all = waterway_canal, waterway_drain, waterway_river and waterway_stream**
- **Railway_stations = railway station and railway halt**
- **Railway_rail**; **Railway_other**
- **Buildings are divided into five categories: accommodation, religious, commercial, amenity/civic and others:**

-Buildings_accomodation = buildings_apartments + buildings_farm + buildings_hotel + buildings_house + buildings_residential + buildings_detached + buildings_dormitory + buildings_terrace + buildings_houseboat + buildings_bungalow + buildings_cabin + buildings_villa + buildings_hostel + buildings_semidetached_house

-Buildings_religious = buildings_church + buildings_chapel + buildings_temple + buildings_cathedral + buildings_shrine + buildings_wayside_chapel + buildings_wayside_shrine + buildings_monastery + buildings_basilica + buildings_convent

- **Buildings_commercials**= buildings_commercial + buildings_office + buildings_industrial + buildings_retail + buildings_supermarket + buildings_warehouse + buildings_kiosk + buildings_store + buildings_shop + buildings_manufacture

-**Buildings_amenity_civic** = buildings_civic + buildings_hospital + buildings_transportation+ buildings_school + buildings_stadium + buildings_kindergarten + buildings_government + buildings_train_station + buildings_university + buildings_grandstand + buildings_public + buildings_bank+ buildings_substation + buildings_parish_hall+ buildings_community_centre+ buildings_sports_hall + buildings_townhall + buildings_railway_station + buildings_college + buildings_service + buildings_hall + buildings_public + buildings_transportation

-**Buildings_others** = buildings_barn + buildings_farm_auxiliary + buildings_bridge + buildings_bunker+ buildings_hangar + buildings_ruins + buildings_construction + buildings_farm + buildings_shed + buildings_stable + buildings_cowshed + buildings_sty + buildings_storage_tank + buildings_collapsed + buildings_abandoned+ buildings_tower + buildings_transformer_tower + buildings_container + buildings_pavilion + buildings_prison + buildings_winery + buildings_water_tower + buildings_slurry_tank + buildings_glasshouse + buildings_parking + buildings_chimney + buildings_tent + buildings_shelter + buildings_carport + buildings_detached + buildings_castle + buildings_roof + buildings_garage + buildings_greenhouse + buildings_service

OSM Data Categories

- **Roads are divided into eight categories:**

-Highway_main:

Highway_motorway + highway_trunk + highway_primary + highway_secondary + highway_tertiary

-Highway_pedestrian

-Highway_road

-Highway_unclassified

-Highway_residential

-Highway_route_ferry

-Highway_track

-Highway_cycleway

- **Aeroway_others**

OSM data, GDP per capita and capital prediction

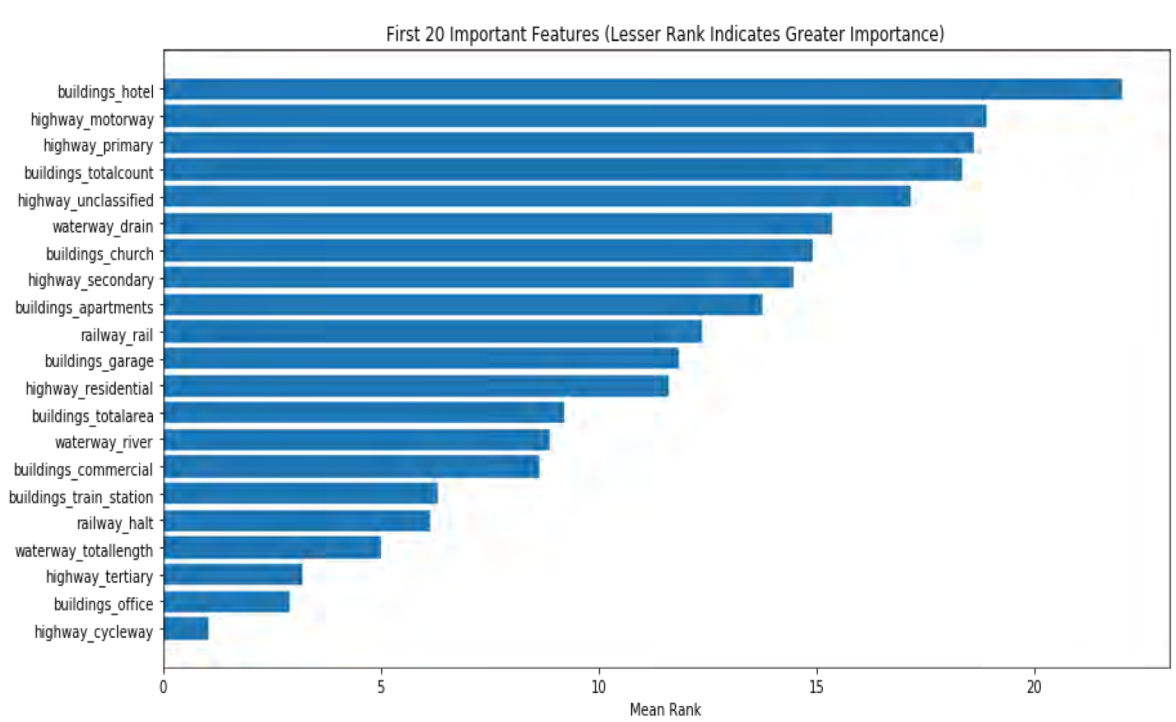
- We attempt two different approaches:

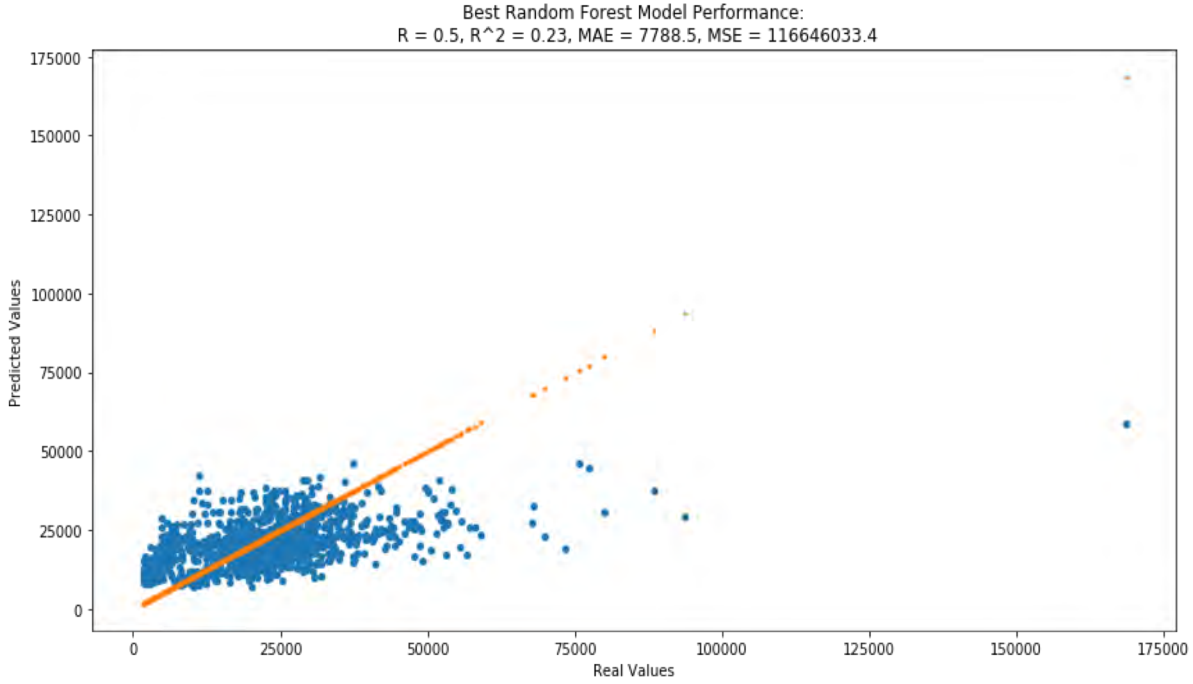
1) predicting GDP per capita using the OSM variables along with the features extracted from the satellite images and

2) predicting capital using the OSM variables along with the features extracted from the satellite images and then calculate gdp per capita based on a simple Cobb-Douglas production function. The second method is implemented to see if predicting capital using OSM variables gives better results as the variables are indicative of economic assets which are more or less reflective of capital.

- The features have been converted to Nuts3 and Nuts2 level (admin level 2 and 3 at the sub-national level) to conduct the out of sample prediction at the respective admin levels.
- We train different supervised learning prediction techniques: ridge, lasso , decision tree and random forest on the set of input features to predict for GDP per capita and capital. In our analysis, random forest performs best.

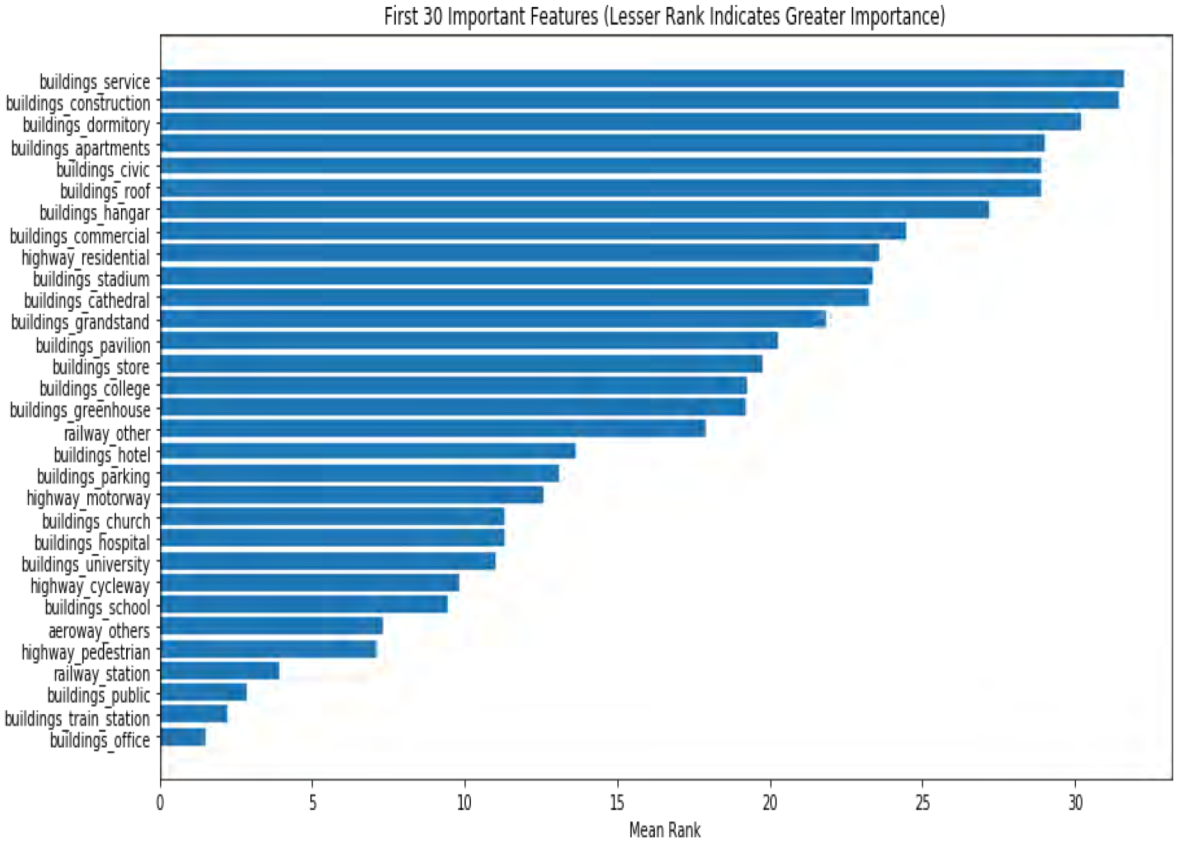
Prediction Results-GDP per capita

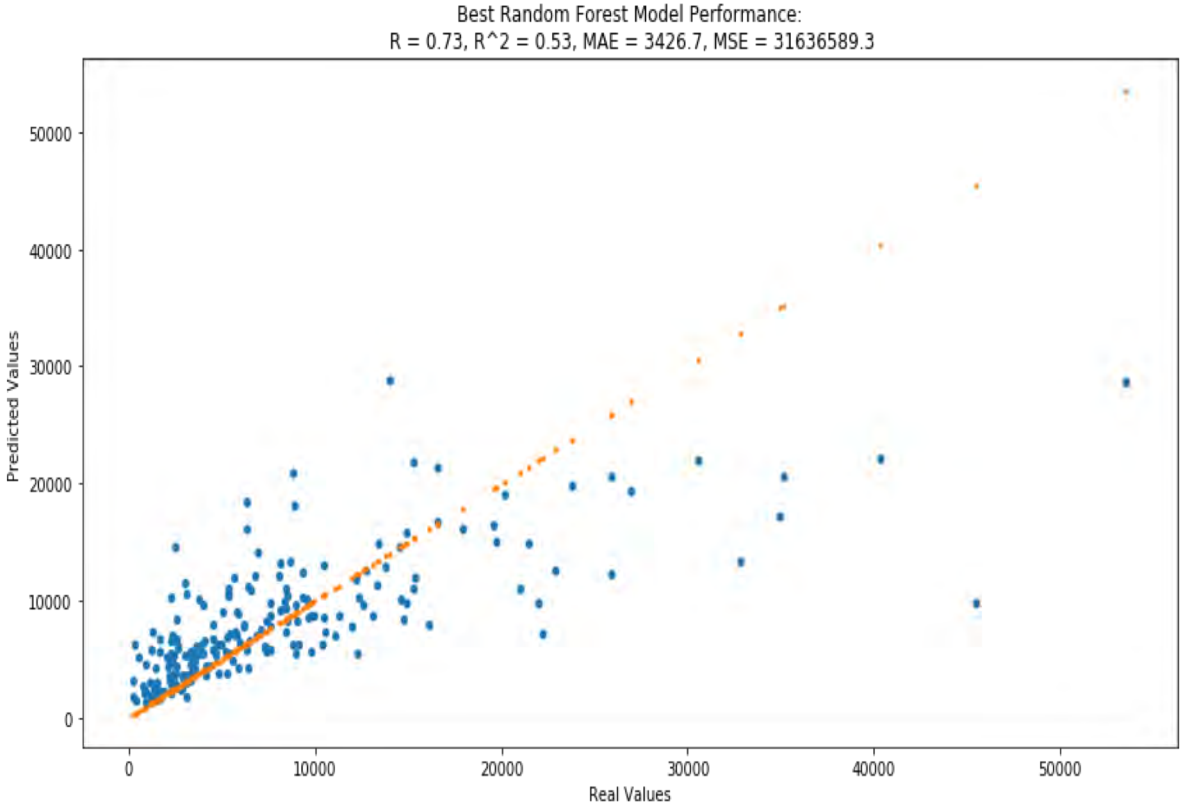




Top OSM features and the predicted gdp per capita from Transfer Learning

Prediction Results-Capital





Calculating GDP using the predicted capital

- $Y = AK^{\alpha}L^{\beta}$
- In Linear form: $\ln Y = \ln A + \alpha \ln K + \beta \ln L$

VARIABLES	lngdp_nuts
lnpop_nuts	0.240***
	(4.825)
lncapital_total	0.807***
	(21.93)

Predicted GDP= = $\alpha.predicted\ capital + \beta.population + uhat$

Calculated GDP

	(1)
	lngdp_nuts
lngdp_predicted2	0.864
lngdp_predicted3	0.873
lngdp_predicted4	0.882
lngdp_predicted5	0.871
lngdp_predicted6	0.863
lngdp_predicted7	0.873
lnlight_mean	0.616
lnlight_sum	0.502

Predicted capital with Light and OSM

	(1)
	fixed_capital_total
predicted_capital_osm_rf	0.733
predicted_capital_light_rf	0.265

	(1)
	fixed_capital_total
predicted_capital_light_rf	0.265
predicted_capital_light_linear	0.261
predicted_capital_light_lasso	0.249
predicted_capital_light_rigde	0.261
predicted_capital_light_decision	0.222

Previous Empirical Result

Table 1 : Pre-colonial institutions and regional development using the new proxy data

	1	2	3	4
Jurisdictional Hierarchy	0.4312***	0.3101***	0.3001***	0.2471***
	(0.0713)	(0.0722)	(0.0441)	(0.0402)
Country Fixed Effect	Yes	Yes	Yes	Yes
Location control	No	Yes	No	Yes
Geographic Control	No	Yes	No	Yes
Population Density	No	No	Yes	Yes

All the results are reported with country fixed effects. The dependent variable is log gdp per capita estimated from our machine learning method. The pre-colonial ethnic institutions is Murdock's (1967) jurisdictional hierarchy. The double-clustered standard errors are in parenthesis. . ***,** and * signify statistical significance at 1% level 5% and 10% respectively.

Conclusion

- In this study, we go beyond the nightlight data to construct a new proxy measure for economic activity.
- Using a multi-step machine learning process, we extract information from the day-time images and utilize the key information along with Open Street Map(OSM) data to construct Gross Domestic Product.
- Our novel approach show that high resolution daytime images can be used to construct measures for economic activity. Notably, using a model built on Europe maybe able to identify determinants of economic activity in Africa despite the differences in economic settings.
- This approach can be used for broader application to produce granular cross-sectional data across different socioeconomic studies and also can be extended over time for particular locations using only publicly available data.