Students: Vid Chan, Anderson Monken, Doug Neumann, Nicole Yoder
Group Name: NOAA Nighttime Lights

Executive Summary

Utilizing satellite photos and available datasets containing country-specific statistics, our research compares changes to countries over time in order to predict gross domestic product (GDP). The types of models constructed were linear regression, random forest, and gradient boosting machine with varying degrees of accuracy.

Introduction

Fundamental economic statistics like GDP and quality of life measurements like average life span have long been used to measure a country's level of development. Our research develops another metric. We utilized satellite images from the National Oceanic and Atmospheric Administration's (NOAA) Nighttime Lights Time Series. We compared the changes in those images from 1993 to 2013 to historical data from the International Monetary Fund (IMF), World Bank (WB), and Organisation for Economic Co-operation and Development (OECD) over the same period with the goal of predicting GDP per capita.

Code Files

https://github.com/AndersonMonken/ANLY502_Spring2020_Project

Methods

Light Data Preprocessing

The night light data from NOAA was downloaded and unzipped using a python script ("*light_data_parallel_download.py"*); the total size of the unzipped files totaled 79 GB. Data was moved to S3. It was determined that geopyspark was the best program to work with the geotiff format of the raw night light data. Due to concerns with cluster bootstrapping on AWS, a personal workstation (24-core CPU, 96GB RAM, 1TB PCIe 4.0 SSD) was used in spark local mode for geopyspark. This rivaled and exceeded performance on smaller AWS clusters for normal spark tasks. Programs were written to be scalable on AWS cluster for future use. Geotiff files were read in using geopyspark. There were issues with reading in the files using the spacetime layer due to missing data info. Instead, a spatial layer was used and each geotiff file was processed individually. There were multiple methods to read in the tiff file, including reading in the tiff as an RDD then transforming into a raster spatial layer or reading in the geotiff directly to a raster spatial layer. The first method using the RDD would not work due to java heap space errors. A variety of spark executor configurations were tested, and even with 12 GB RAM per worker the same problems persisted. Using geotiff worked properly. After making the geotiff file into a raster layer, the data is "tiled" so that spark can effectively divide work across the map of data. Care must be taken in choosing the appropriate map projection for the data, as one projection is for mercator projections while a local layout is for doing analytics. Geotiff files were processed and a polygonal mean calculated for each country at every time period and camera.

Economic Data Preprocessing

We also created another dataset of economic and various other country attributes to combine with the light data. The most important variable in the second dataset was GDP per capita, since it (or its indexed version) was used as the target variable in our models. The others were collected as ideas for possible predictors, but we did not end up using them all. Table 1 summarizes all of those variables and their data sources. Many of the variables were missing some years for certain countries. Countries were removed entirely if they were missing more than 20% of their values. Otherwise, the missing values in each column were imputed by country using the pandas function interpolate. “Interpolate” filled in missing values between two available years linearly, while missing values at the beginning or end were filled in with the next closest value. Table 2 shows an example of the imputation. After cleaning was completed, we were left with data for 99 countries over 20 years (1993-2013) each.

EDA

The light data is applied to a world map animation using Plotly to show changes over the 20-year timespan. Camera differences were briefly considered, and the newer camera was selected when there were multiple available for a year. The data values range from 0-64, but most of the distribution is below 10, which meant that the mean light on its own did not show changes well. This is due to some smaller island countries having very concentrated light sources without empty land to offset it. We also then created a mean light index value (1993=100) for each country so the relative changes over time can be visualized in Figure 1. The GDP per capita variable, which is our response variable, was also turned into an index. The relationship between these two variables is plotted in Figure 2. For code on eda, see code file (*“light_EDA.ipynb”*).

Modeling

We first divide our dataset into train and test sets (75%-25% respectively). We use the train set to train the model and then, test it with the test data. Ten different models such as linear regressions, random forest (RF), and gradient boosted trees (GBM) were run in our study. The main dependent variable for all of this model is GDP per capita (raw and indexed value). Different combinations of independent variables are GDP per unit of CO2, import and export percentages, government revenue, agricultural percentage of GDP, international tourist arrival number, mobile subscription, total life expectancy, total population, year, and average brightness (from NOAA). For details of each model, please go to our code repository (*“models.ipynb”*).

In our study, we used root mean squared error for the test set (test RMSE) as our metric for model evaluation. Table 3 shows the results. The ten models can be grouped into two different sub-groups. Those that are highlighted in red have GDP per capita index as their target variable, while those that are NOT highlighted use raw GDP per capita. Comparing the models that used all of the predictors (models 1-4), we see that random forest worked the best for both variable types--raw or indexed. Those two models were also the best overall. However, if it is necessary to use light data to predict a country's GDP per capita due to unreliable or unavailable reporting, then those other country attributes used as predictors are also probably unreliable or unavailable. Therefore, we ran models that only used year and mean light as predictors (models 5-10), even though they were not as accurate. For this group, random forest worked best again for the raw variables, but linear regression was slightly better for the indexed variables.

Conclusions

Our research successfully identified a positive relationship between the light data and GDP data. However, the relationship was weaker than we had hoped to see. The main lessons learned were overcoming java heap space errors by properly adjusting the configuration of clusters and different ways to project images on a mercator map.

Future Work

The Covid-19 crisis is an unprecedented event in terms of the modern world economy. Its long-term and short-term impacts are largely unknown and have yet to be studied. As of April 2020, the United States had energy use drop to a 16-year low (DiSavino, 2020). Further examination of trends, like energy use, during a worldwide pandemic may prove to be highly beneficial. Furthermore, continued additions to the datasets we used in our research could be used to shed light on some of our conclusions.

Bibliography

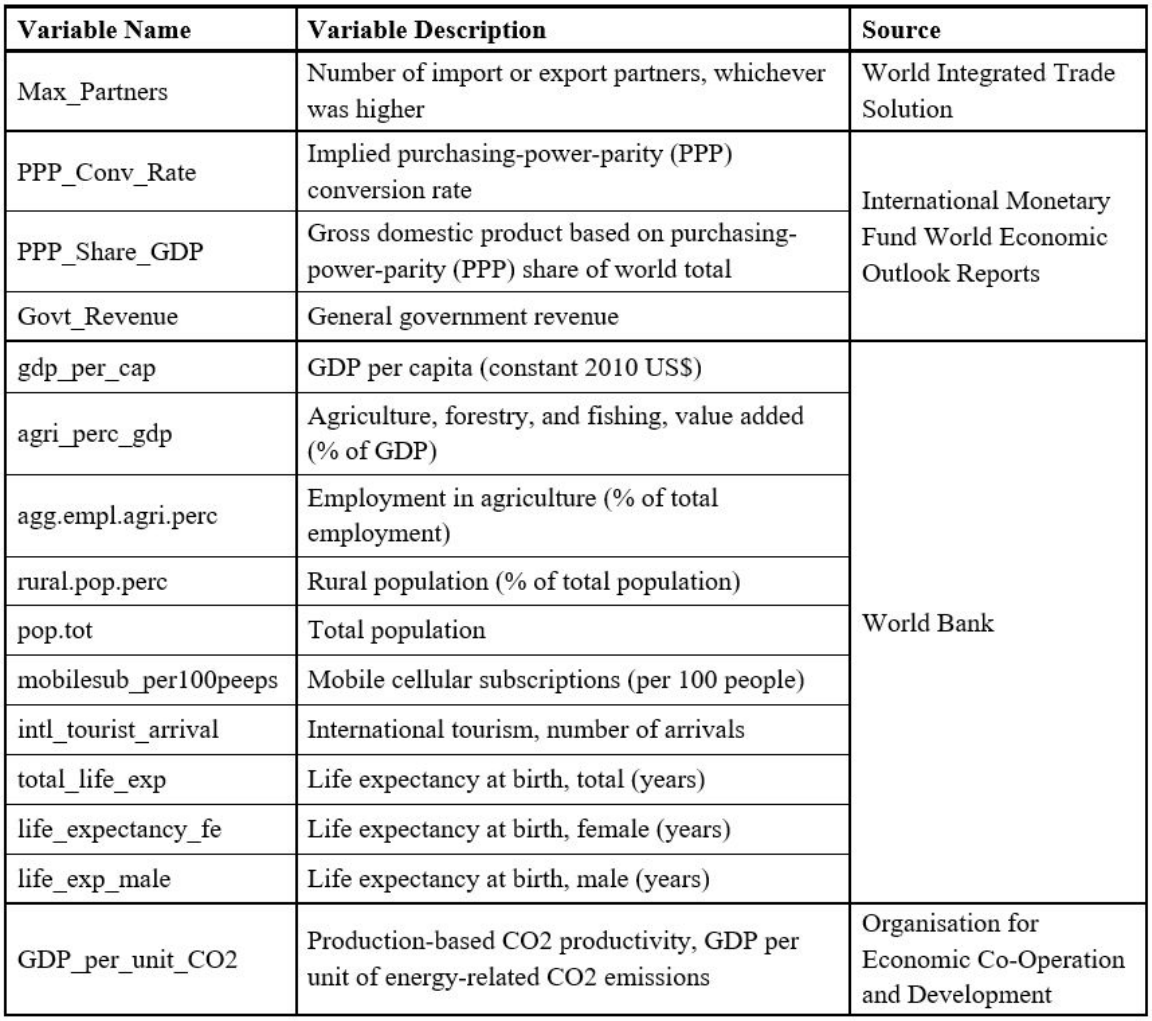
DiSavino, S., & Reuters, T. (2020, April 14). COVID-19: America hasn't used this little energy in 16 years. Retrieved from https://www.weforum.org/agenda/2020/04/united-states-eneregy-electricity-power-coronavirus-covid19/

Division of Labor

Development Data Collection & Cleaning from World Bank, IMF, OECD: Vid, Doug, Nicole
Image Data Processing: Anderson
EDA & Modeling: Everyone
Project Write-Up: Everyone

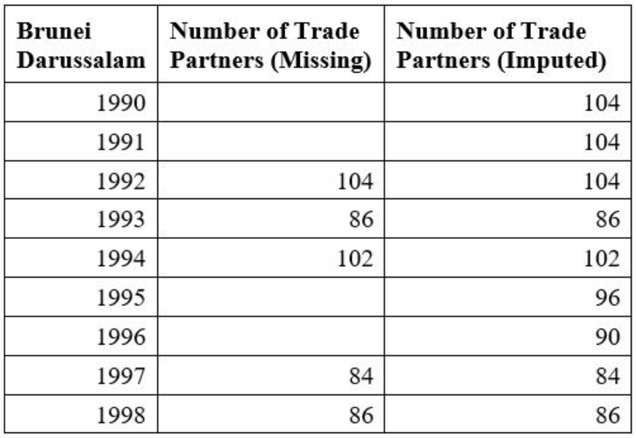
Appendix

Table 1: Variable Descriptions and Sources (Econ Data)



Variable Name	Variable Description	Source
Max_Partners	Number of import or export partners, whichever was higher	World Integrated Trade Solution
PPP_Conv_Rate	Implied purchasing-power-parity (PPP) conversion rate	International Monetary Fund World Economic Outlook Reports
PPP_Share_GDP	Gross domestic product based on purchasing-power-parity (PPP) share of world total	
Govt_Revenue	General government revenue	
gdp_per_cap	GDP per capita (constant 2010 US$)	World Bank
agri_perc_gdp	Agriculture, forestry, and fishing, value added (% of GDP)	
agg.empl.agri.perc	Employment in agriculture (% of total employment)	
rural.pop.perc	Rural population (% of total population)	
pop.tot	Total population	
mobilesub_per100peeps	Mobile cellular subscriptions (per 100 people)	
intl_tourist_arrival	International tourism, number of arrivals	
total_life_exp	Life expectancy at birth, total (years)	
life_expectancy_fe	Life expectancy at birth, female (years)	
life_exp_male	Life expectancy at birth, male (years)	
GDP_per_unit_CO2	Production-based CO2 productivity, GDP per unit of energy-related CO2 emissions	Organisation for Economic Co-Operation and Development

Table 2: Example of Data Imputation



Brunei Darussalam	Number of Trade Partners (Missing)	Number of Trade Partners (Imputed)
1990		104
1991		104
1992	104	104
1993	86	86
1994	102	102
1995		96
1996		90
1997	84	84
1998	86	86

Table 3: Test RMSE for Various Models

	Model	Test RMSE	Note
Model 1	Linear Regression 1	14075.2	All predictors
Model 2	Linear Regression 2	31.9452	All predictors. Indices of mean light and GDP per capita are used instead of raw values.
Model 3	Random Forest 1	6225.35	All predictors
Model 4	Random Forest 2	26.2339	All predictors. Indices of mean light and GDP per capita are used instead of raw values.
Model 5	Linear Regression 3	18020.8	Only year and mean light as predictors
Model 6	Linear Regression 4	32.7086	Only year and mean light index as predictors. GDP per cap index as target
Model 7	Random Forest 3	16848.6	Only year and mean light as predictors.
Model 8	Random Forest 4	34.3648	Only mean light index and year as predictors. GDP per cap index as target.
Model 9	GBM 1	17463	Only mean light and year as predictors.
Model 10	GBM 2	39.6207	Only mean light index and year as predictors. GDP per cap index as target.

Figure 1: Snapshot of light index map animation (see light_index.html)

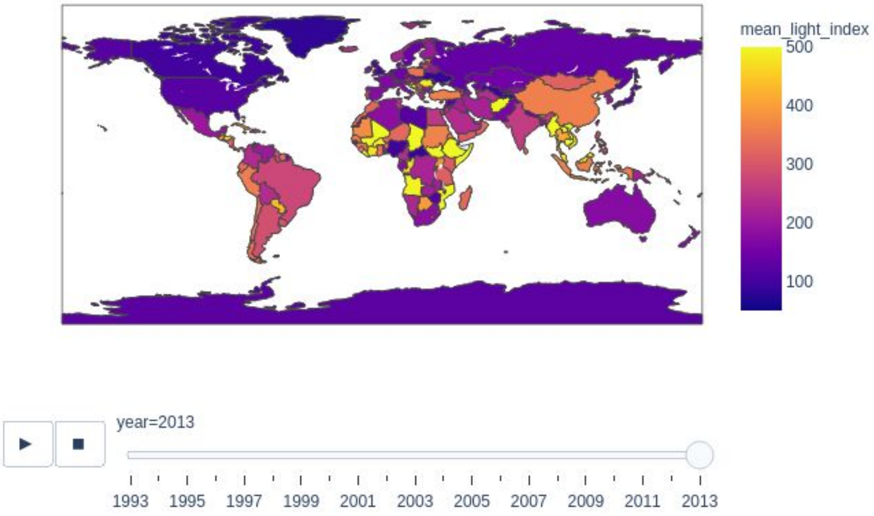
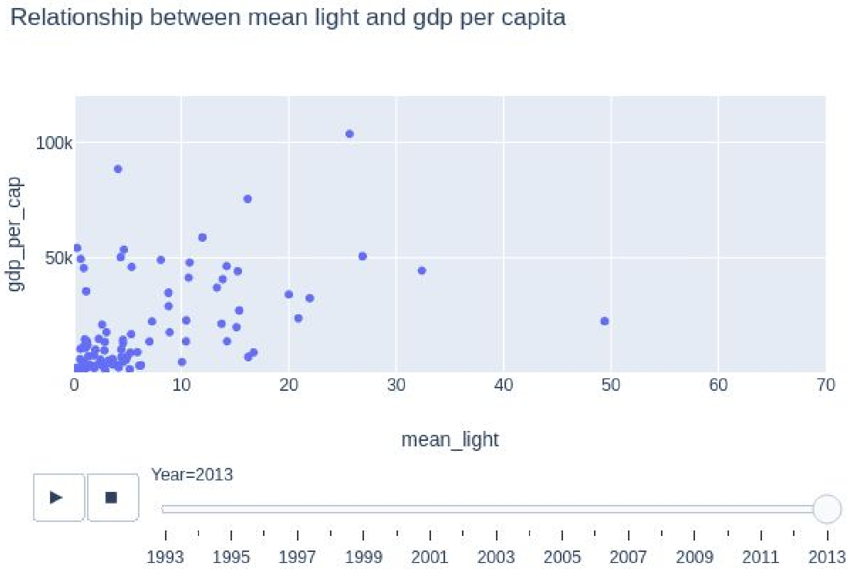


Figure 2: Snapshot of GDP per capita and mean light animation (see light_gdp.html)


Relationship between mean light and gdp per capita
gdp_per_cap
100k
50k
mean_light
▶
■
Year=2013