Fighting Global Poverty with Data

Group 1 - Alexandra DeKinder, Ameya Karnad, Kulkanya Lekhyananda and Nitasha Nair

Introduction

Surveys data of household income and consumption have been traditionally used by researchers to measure economic activities, wealth, and poverty. However, there are many limitations of this methodology. Generating national survey needs a very large sample, and doing so is very expensive. This results in a major data gap as many countries do not conduct surveys continuously, which makes it harder to conduct analysis with[1].

Big data in the form of night lights data, other satellite data, social media data, and mobile phone patterns offers an alternate route to understand socio-economic factors in data scarce regions. In the past few decades, researchers has developed alternative techniques to estimate poverty using information provided through satellites. The advantage of using big data is that they are abundant, can enable real-time analysis, it does not depend on the government action to collect (eg. survey), and with the publicly available data like the nightlights, it comes out more economical.

Various research predicting socio-economic activities has been done using night-time light data. The research that we are focusing this time is Provill et al. (2017). The research uses linear regression model to find the correlations between nightlights and several economic indicator such as GDP, electricity consumption, and CO2 emission.

In this research we are trying to explore whether night-time lights can be used to predict various economic development indicators and if we can improve upon the basic regression model applied in the paper .

Dataset

We use the night-time light data - the Defence Meteorological Satellite Program (DMSP) dataset – as used in Proville et al. (2017) – together with the economic indicator data from the World Bank to find factors that correlates with nightlights data to better predict poverty.

In order to highlight another way that the DMSP data can be used, we built a simple classification model.The logistic regression model uses 10 variables, including DMSP, to classify observations as either being high GDP (top 50 of all observations) or not.

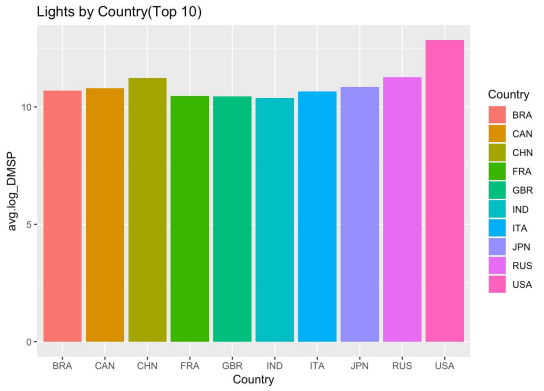
As the confusion matrix below shows, the model does relatively well at classifying the data. But due to strong correlations of some of the variables, the model might not be useful

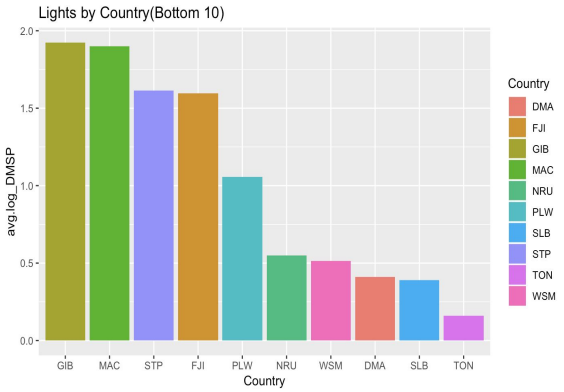
	False	True
0	365	6
1	5	63

[1] Blumenstock, “Fighting poverty with data,” *Science* 353 (6301), 753-754.

Some basic exploratory analysis can be found as follows -

Top 10 and bottom 10 countries with respect log_DMSP values

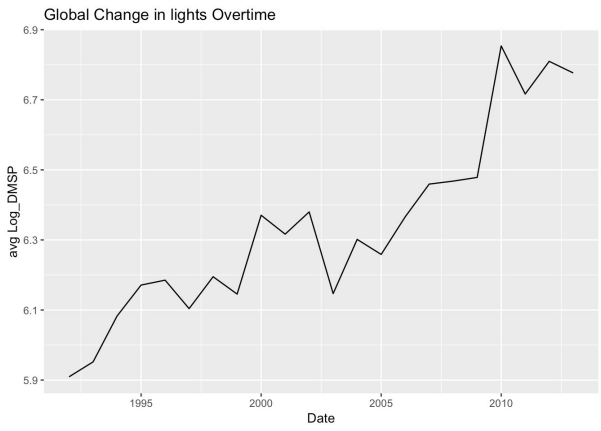




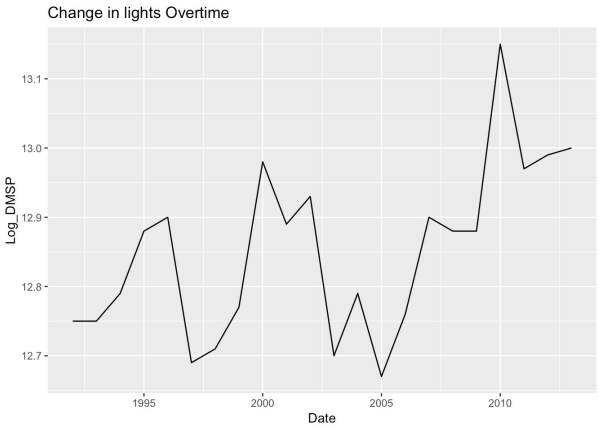
As expected, USA is the Country with the highest log_DMSP value, while the country with the least log_DMSP value is Tonga

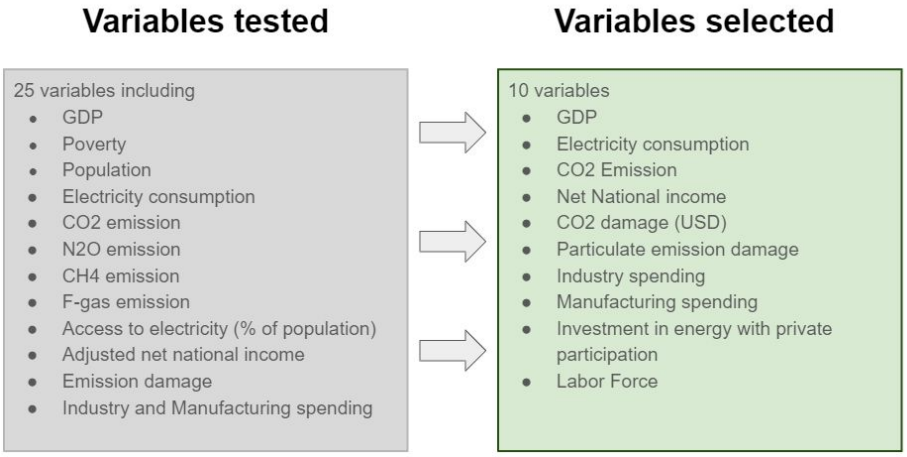
Changes in lights overtime

Global

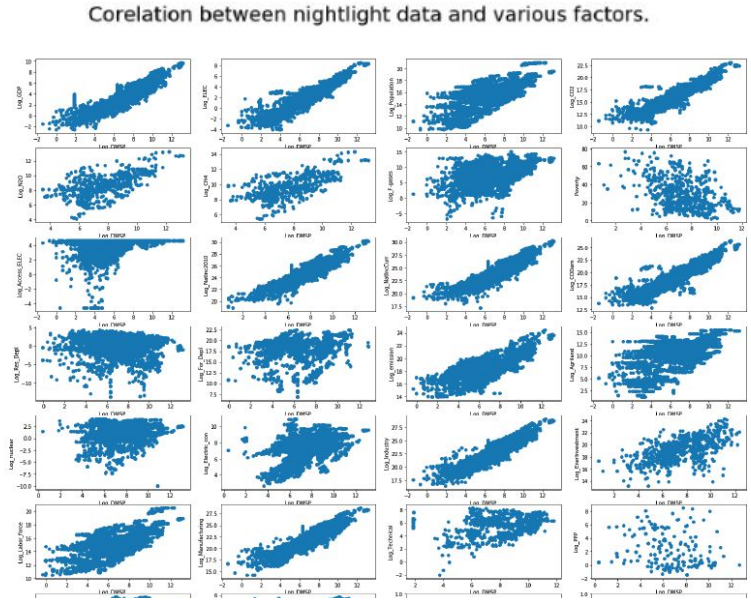


USA





Out of the 25 variables, 10 variables were ultimately selected for analysis by looking at their correlations with log_DMSP



A graph showing the correlation of log_DMSP and various other factors like electricity, population, GDP, CO2 emission, national income, nuclear power generation, manufacturing

Limitations

Using linear models violates the Independent and identically distributed (IID) random variables assumption. The IID assumption states that each element is independent of the variables that came

before it which are also random. However, variables like GDP and emissions cannot be independent of the previous values and are based on the previous values.

Methodology

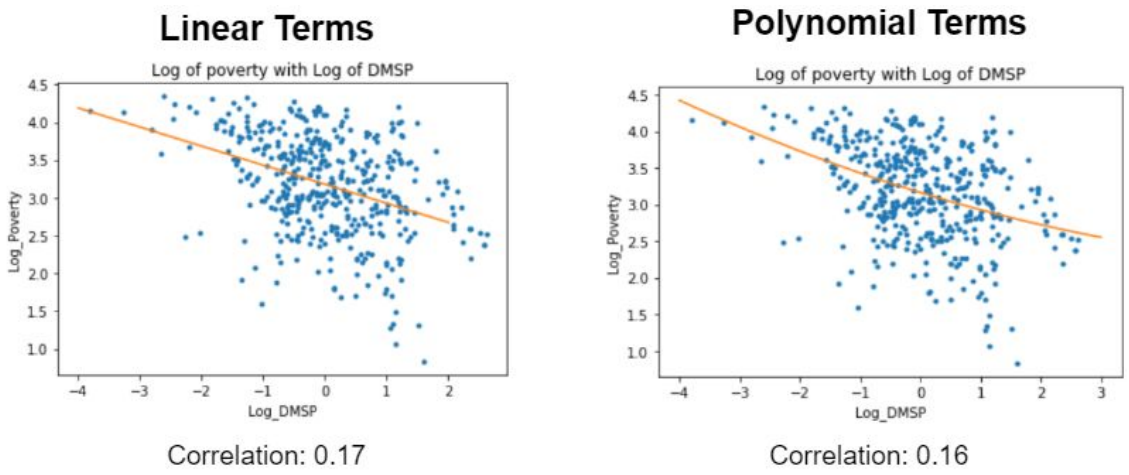
We challenge the linear regression model by using the following approaches and keeping in mind the limitation listed above

- Linear regression models (linear and polynomial terms) to find correlation between night lights and poverty
- Lasso Model to statistically fit night lights data with the other variables.
- Linear models for Individual variables
- Time series model to relate other features with Night lights data
    - Vector Autoregression (VAR)
    - Autoregressive-Moving Average with Exogenous Terms (ARMAX)

In this research, we try to test different models including linear regression (linear and polynomial terms), lasso regression and Time series models such as vector autoregression (VAR), and autoregressive-moving average with exogenous terms (ARMAX) to test the ability to predict various economic development indicator using night-time lights.

Findings

1. **Linear Regression** models that try to directly correlate poverty and nightlights fail drastically



We tried to build the models for Log of Poverty by using a linear regression with linear and polynomial (Quadratic) independent (Log of DMSP) term. It was noticed that both the models drastically failed to fit the data with correlations being 0.17 and 0.18 respectively

2. A **Lasso regression** model was used to try to fit the Night lights (Log_DMSP) with the attributes that were correlated with it. The Lasso model selects a subset of

Variable (log)	Coefficient
GDP	0.4974
Electricity Consumption	0.5942
CO2 emission	0.4867
Industry	0.2830

covariates which simplifies the final model. One of the features of the lasso model is that it controls for multicollinearity.

It was seen that the Log_DMSP could be fit using the Logs of GDP, Electricity consumption, CO2 emission and Industry. Note that lasso model does not give the important features in the model, but selects one among many correlated features present in the data to build the model.

3. **Linear models for** were also applied on the **individual factors** to check if the Log_DMSP had strong correlation with the factors.

While factors like GDP, Electricity consumption, CO2 emission, Industrial and manufacturing investment showed signs of high correlations, factors like Investment in energy and Labor force did not show strong correlations

Variable (log)	R2
GDP	0.84
Electricity Consumption	0.88
CO2 emission	0.88
Net National income	0.85
CO2 damage	0.87
Particulate emission damage	0.6
Investment in energy with private participation	0.12
Industry	0.82
Labor force	0.49
Manufacturing	0.84

4. **Time Series**

As mentioned before, with regards to the dataset that we were working on, it was a time series data from 2002 to 2012. So using of linear regression models to model this data violates the condition that linear models should in IID form (Independent and Identically Distributed) form. So we tried to fit the model with Time series algorithms.

a. **VAR - Vector Autocorrelation Model**

   This model captures the linear interdependencies among multiple time series.

   It is noted that the CO2 emissions, CO2 Damage and Labor force fit good with this time series model.

Variable (log)	n.r.m.s. error
DMSP	0.05036
GDP	0.08453
Electricity Consumption	0.03884
CO2 emission	0.00461
Net National income	0.02243
CO2 damage	0.00979
Particulate emission damage	0.02422
Investment in energy with private participation	0.14019
Industry	0.02824
Labor force	0.00455
Manufacturing	0.03324

b. ARMAX - Autocorrelation Moving Average with Exogenous Terms

The response variable is a function of exogenous inputs and their appropriate lags as well as lagged values of Y and error terms

It is noted that the Electricity consumption, CO2 Damage and Labor force fit good with this time series model.

Variable (log)	r.m.s. error
GDP	0.01324
Electricity Consumption	0.00338
Net National income	0.01356
CO2 damage	0.00214
Particulate emission damage	0.01135
Industry	0.02008
Labor force	0.00263

Conclusion and Policy Implementation

- There is no direct correlation between nightlights and poverty, however, there are other variables that correlate and can be used to indirectly predict poverty. In addition to GDP, CO2 emission studied in the research papers, our analysis found that nightlights can be a good predictor for manufacturing, national income and industrial spending.
- The classification model shows that it is relatively accurate, however, we need to keep in mind that this may change due to multicollinearity.
- The government can use nightlights data to help to make investment choice in areas where there is low economic activities.
- More data is required for time series model versus the other linear model.
- Even though predictions made through economic data collected by World Bank are quite accurate, a limitation is that the analysis would be dependent on the scale of data available, which in most cases are country-level. As the country level data is already available, big data methods like using nightlights can be useful as a proxy for indicators at a more disaggregated level.

Caveats

- Poverty is complex. It is important to understand the different dynamics and avoid generalizations. Hence, features selected as indicators of poverty have to be carefully selected
- Values are Log scale. So a small error in calculations can lead to large distortions. For example, an error of 0.1 at values of 20 at log scale may be actually be an error in millions!
- Time series analysis generally requires lot of training data. The data collinearity might affect the time series analysis
- Night-time lights are unlikely to provide added value as a proxy in countries with good statistical systems, due to the high measurement error as compared to national inventories[2]

[2] Proville et al., “Night-time lights: A global, long term look at links to socio-economic trends”

- Nightlights data are less effective at differentiating between regions at the bottom end of the income distribution, where satellite images appear uniformly dark[3]

[3] Jean et al., “Combining satellite imagery and machine learning to predict poverty”