Using Air Pollution to Predict Economic Activity

Lila Cardell, Manny Kim, Frederick Nyanzu, Pablo Ordoñez, Gowthami Venkateswaran

Draft as of May 13, 2021

The use of different measures of economic activity is important for both policy makers and researchers. However, the availability of these data at a subnational level in developing countries is uncommon, and when available, they come with significant problems in terms of quality (Johnson et al 2009). To address this, Henderson et al. (2012) use night lights to predict income growth, and show that it is a useful proxy. In this paper, we aim to build upon this work, by using air pollution, in addition to night lights, to predict economic activity at a subnational level. Our paper expands this previous work in two directions. First, current measures of air pollution based on satellite and ground monitor data exist for the whole world. Additionally, air pollution might be better at capturing changes in short-term economic activity, since it is correlated with other activities that would not necessarily result in changes in night lights intensity, but might result in changes in average measures of air pollution (eg. increased traffic, industrial and agricultural activities). Finally, we believe that this is a prediction exercise, and we propose the use of machine learning (ML). This allows us to take an agnostic approach in terms of the functional form for the data generating process, and allows us to explore highly dimensional models with the objective of generating the most accurate prediction possible. We find that the random forest model performs the best, and that total population is the most important feature in predicting per capita income in Mexico.

Nightlights and economic activity

Existing studies devoted to estimating economic activities for subnational level analysis and specifically in developing countries, over a long period of time have been polarized due to data limitation. Kim (2008) and Chen and Nordhaus (2010) indicate that among the main reasons for such polarization has mainly been the lack of reliable and consistent sub-national income data. Despite the polarized data limitation, a sample of studies have attempted to predict economic activities in many folds using nightlight data. In an earlier review of satellite data on economic activity such as urbanization, city dynamics, population movements, economic growth, development indicator (Chen and Nordhaus, 2010), Henderson et. al. first show that nightlight has been used as a proxy to measure economic activity. In a similar study, Nordhaus adds that about 3,000 empirical studies have employed nightlights to measure economic activity since the early 2000s. In an attempt to investigate how well nightlight measures economic activity, Mveyange (2015) uses nightlight to investigate regional inequality in income in Africa. The paper found a positive association between nightlights and income, evidence of increasing regional inequality trends between early years (1992 and 2003), and declining regional inequality trends in later years (2004 and 2012). Other existing studies have shown a strong positive relationship between nightlights and GDP at national and at sub-National levels for an economy over a range of spatial scale (Doll et al. 2006, Bundervoet et al. 2015, Pinkovskiy and Sala-i Martin 2014, Bhandari & Roychowdhury 2011, Singhal et al 2020, Weidmann and Schutte 2017).

Pollution and economic activity

Exposure to air pollution can have a myriad effect on economic activity. Smulders and Gradus investigate pollution abatement and long-term growth and explore conditions under which sustained economic growth and preservations of environmental quality are optimal, and to what extent economic growth is affected by environmental policy. The study concludes that failure to take into account pollution cost (private abatement activity) on environmental quality and productivity can negatively affect economic growth. Zhu et. al (2019) use a panel data of 73 cities from the period of 2013 to 2017 to investigate whether there is a bicausal relationship between air pollution and economic activity. The authors' result indicates that there is a unidirectional causality between pollution measured by $PM_{2.5}$ and economic activities such as economic growth, foreign direct investment, and industrial structure in the long-term. Similarly, Liang and Yang (2019) find environmental pollution to have a negative and a significant effect on urbanization and economic growth. Davis et al (2010) uses a long-term economic activity and related it to air pollution from vehicle related particulate matter (PM) over a 30-year period to offer understanding into probable historical surrogate markers of air pollution. The authors use a mixed-modeling approach and conclude that higher concentration of pollution in the long-run can adversely impact economic activities. Koop et. al (2010) report evidence of air pollution, economic activity and respiratory illness using historical data from Canadian cities. They note that variation exists in previous studies that report a relationship between urban air pollution levels and respiratory health problems and the type of model used. By comparing two estimation approaches: model selection and Bayesian model averaging, the authors conclude that the impact of air pollution on respiratory illness could be dependent on socioeconomic covariates.

Climate and GDP

Climate change has many economic impacts, including on agricultural production, human health, and labor productivity, and the degradation of natural resources. Spatial heterogeneity has made it difficult to measure these impacts, although panel methods allow for some control over region-specific effects. The economic literature has focused on measuring the effect of changes in weather, i.e. rainfall and temperature. A few papers have linked these variables and other weather shocks to growth in gross domestic product (GDP). Dell et.al. (2012) found that economic growth in low income countries was affected by rising temperatures and Barrios et. al. (2010) found similar effects for rainfall trends in sub-Saharan Africa. Burke et. al. (2015) found that both high and low income country economic growth was affected by climate change, with an estimated reduction in average GDP by 23% by 2100. Others model the effect of temperature on GDP level (Auffhammer, 2018). The two major issues in estimation of economic effects are that assumptions must be made about adaptation to climate change, and there is little economic theory driving the relationship between climate and economic outcomes. (Newell et. al., 2021) The literature has developed "damage functions" or Integrated Assessment Models (IAMs) that use current trends to predict future impacts of climate change on economic outcomes. (Dell et. al.,2014) A systematic review of climate change-GDP models found that the relationship between temperature and economic growth is generally non-linear, and that the effect of temperature is significant for predicting economic levels, but not growth, and there is

more sampling uncertainty for growth models. Temperature lags are commonly included, as the economic impact may not be contemporaneous. (Newell et. al., 2021)

Machine learning

The tools to process and classify unstructured data such as satellite imagery are relatively new, and application is limited by available sources of data that can be used to test and train models. Supervised machine learning is often used for prediction, by using a subset of the data to evaluate a set of possible regression models, of which the model with the lowest out-of-sample prediction error is selected and then tested on the remaining data. Two common supervised ML methods in applied economics are shrinkage methods such as LASSO and tree-based methods such as random forest. LASSO adds a penalty term to a linear regression model based on the size of the coefficients, and is useful for identifying variables which contribute to prediction. Random forests are a type of decision trees that bootstrap classification using random sets of explanatory variables ("features") to identify the explanatory variables that are highly predictive while reducing out-of-sample error. (Storm et. al., 2020) The predictions that result from the ML approaches can be compared to "ground truth" measurements, if available, to evaluate the validity of the method. A recent study using machine learning techniques to estimate micro level wealth and poverty using multiple sources of remote sensing data and validated their predictions using household level surveys in 18 countries. (Chi et. al., 2021)

Data and Methods

The aim of this study is to use publicly available remote sensing data, together with data from national statistics, to predict changes in local economic activity. First, we use data from Mexico's population censuses in 2009 and 2014, which collects data on income, as well as other socioeconomic characteristics. Here, we use real per capita income by municipality, together with a shapefile of all the municipalities in Mexico. Second, we use remote sensing data on air pollution, nightlights, rainfall and temperature. The air pollution data we use come from Hammer et al. (2020) and includes data on average particulate matter PM2.5, derived using both satellite data and ground pollution monitors. This is yearly data from 1998 to 2018 for the whole world. The rainfall data come from the Climate Hazards Group Infrared Precipitation with Stations (CHIRPS) dataset, which is built around a 0.05° climatology using satellite information, with data at a daily, pentadal, monthly, and yearly frequency, from 1981 to the present day (Funk et al., 2015). We use the average monthly precipitation data. For temperature, we use the monthly mean of the maximum temperature from, which has temperature data for the whole world at a 0.05° resolution, from 1983 to 2016 (Funk et al., 2019). Finally, we use the harmonized series of nightlights data from Li et al (2020), which is a data series that shows the average luminosity in a given area for every year, and is constructed from different satellite measures.

Given that this is an exercise in prediction, we will use two different machine learning (ML) models: Lasso and random forest. As a benchmark, we will use a standard OLS estimation:

$$Y_{it} = \beta_1 PM2.5_{it} + \beta_2 NL_{it} + \sum_{m=1}^{12} \beta_{3m} Rain_{imt} + \sum_{m=1}^{12} \beta_{4m} Temp_{imt} + \varepsilon_i \qquad (1)$$

It is important to note several things. First, all the variables in (1) are the levels for municipality *i* in year *t* (2014 and 2009), so that Y_{it} is the real GDP per capita in year *t*, in municipality *i*. The coefficients of interest are β_1 and β_2,which will capture the correlation between changes in economic activity and changes in air pollution as measured by PM2.5 and nightlights, respectively. Additionally, all the variables are standardized by removing their mean and then dividing them by the standard deviation, so that they are in the same scale. Finally, despite the panel data notation, we will not use the temporal aspect of the data, in the sense that all the explanatory variables included in the models are in the same time period as the dependent variable.

This is the model that we use as a benchmark against which we compare the performance of our other models. Our Lasso model is based on (1), but we also include interactions between all the explanatory variables. This yields a very high-dimensional model, but the advantage of the Lasso model is that it will select those variables that increase the predictive power of the model. The Lasso estimator minimizes the mean squared error subject to a restriction, given by the absolute size of the coefficient estimates:

$$\widehat{\beta}_{lasso}(\lambda) = arg\,min \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{x}_i\beta)^2 + \frac{\lambda}{n}\sum_{j=1}^{p}\psi_j|\beta_j| \qquad (3)$$

We will perform k-fold cross-validation to find the lambda (in 3) that minimized the root mean squared error.

The random forest regression is an ensemble model that puts together the prediction built using a given number of decision trees. Decision trees tend to overfit the training data, and so by using an ensemble of these trees, which have slight variations between them, we can reduce the risk of overfitting (Muller and Guido, 2017).

To find the models with the best out of sample performance, we first need to find the hyperparameters that minimize the out of sample prediction error. For this, we divide the data between a testing (90% of our sample) and a training dataset (10%). Then, using the training dataset, we do k-fold cross validation with 5-folds, to find the hyperparameters that minimize the out-of-sample prediction error. What this cross validation methods does is that it divides the training data into 5 folds (groups), then picks a hyperparameter from a set of possible values, estimates the model using that hyperparameter, and then tests the performance of the prediction against the held out fold, using the root mean squared error (RMSE). The process is repeated 20 times, each holding out a different fold. Then, the hyperparameter that achieves the lowest RMSE is the one that is chosen. The process is the same for both Lasso and Random Forest, but the difference is in the number of hyperparameters in each case. For Lasso, we only need to tune one hyperparameter (lambda in (3)). For the Random Forest model, we need to tune two hyperparameters: the number of trees in the forest and the maximum number of levels in each decision tree.

Results

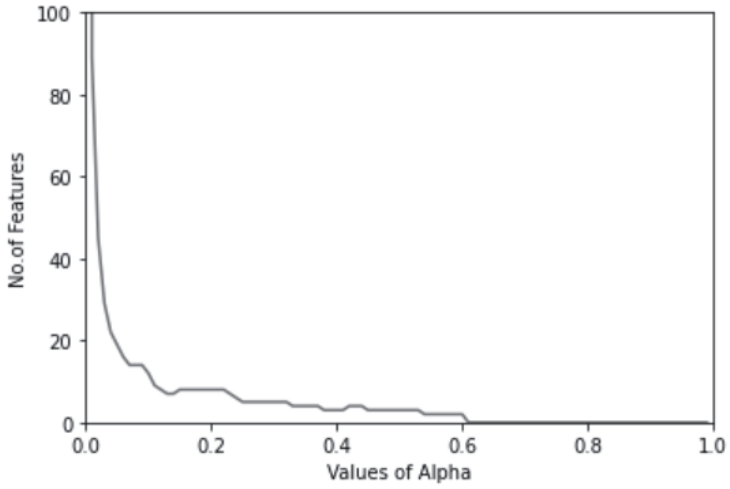
We consider the RMSE and R-squared values for the prediction models to select the best model and report the most important features used by the best model.

Table 1: RMSE and R square values estimated from prediction models

Model	Parameters	RMSE	R square
Linear	-	0.400	-0.014
LASSO 10-fold cross-validation	α = 0.01	0.397	-0.004
Random Forest 5-fold cross-validation	n_estimators = 100 max_depth = 100 bootstrap = True	0.217	0.700

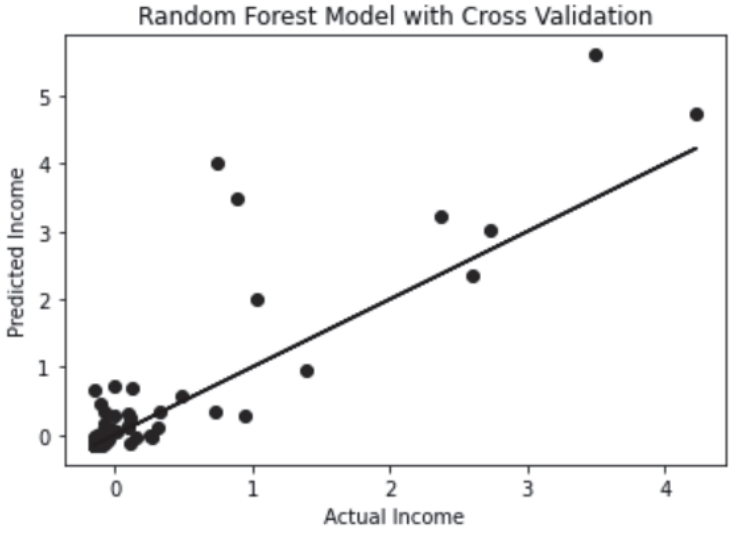
Table 1 presents results from all the methods implemented to predict our outcome variable - per capita GDP. The linear model is estimated with OLS without any interaction terms. A negative R square can be interpreted as the model doing worse than the mean value. The second best model LASSO was implemented with interaction terms. A grid search using sklearn's LassoCV resulted in the selection of α = 0.01. However, estimates of RMSE and R square do not change significantly from the OLS model. Using α = 0.01, the model selects 89 features from 5778 features. Figure 1 depicts the number of features selected by LASSO based on the value of α chosen. The number of features drop to 0 after α = 0.6. As mentioned, our model selected 0.01 with 89 features.

Figure 1: No. of Features selected by LASSO based on α value



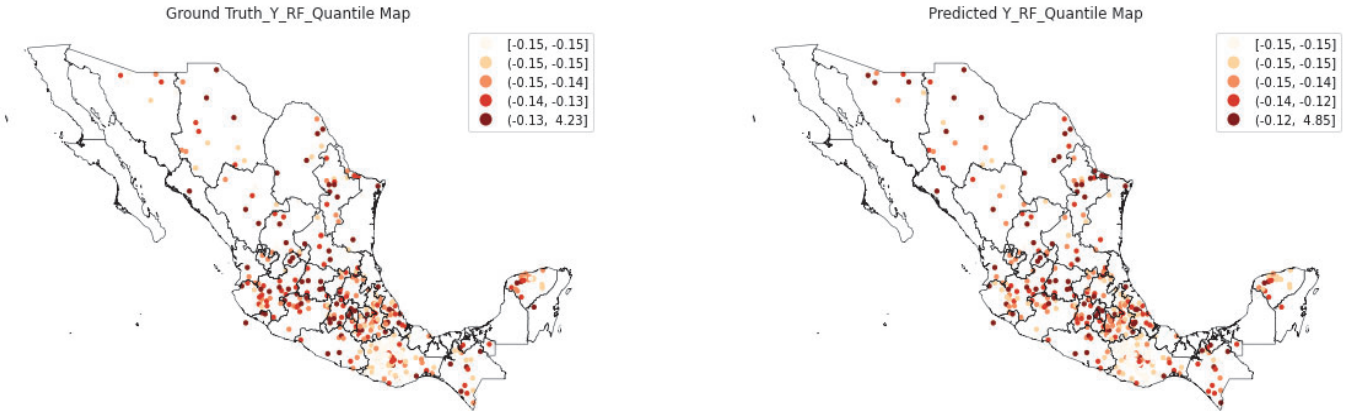
The lowest RMSE and R square values are estimated from the Random Forest model with 5 fold cross validations and both the hyperparameters - number of trees in the forest and the maximum number of levels in each decision tree set at 100. Best results are obtained when data points are bootstrapped. Figure 2 shows a scatter plot of predicted and test data.

Figure 2: Predicted vs Actual Per Capita Income using Random Forest



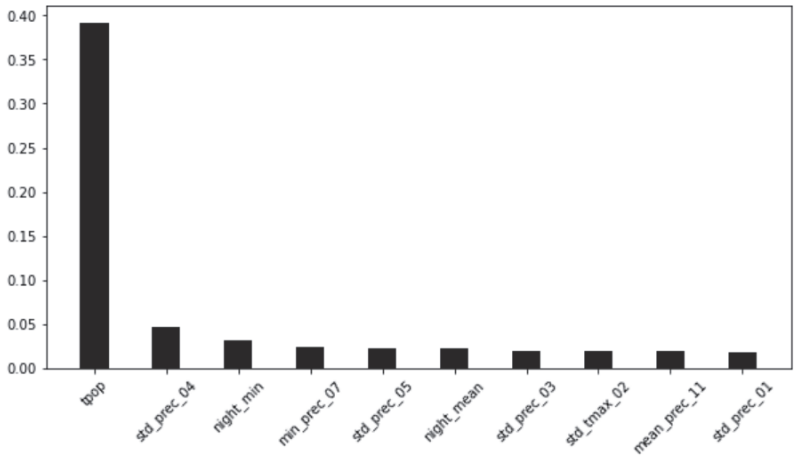
The following Figure 3 plots predicted and actual income for test data using a map of our area of study, Mexico.

Figure 3: Predicted and Actual Income at Municipality



Each point in Figure 3 depicts a municipality and the level of economic activity. From the ground truth and predicted quantile maps, we see that predictions are close to the actual test data. Next, we explore the most important features that were picked up by our model. The following Figure 4 shows the top 10 most important features selected by the model.

Figure 4: Top 10 Important Features from Random Forest



From Figure 4, we see that total population in each municipality is the most important feature for the model. The next most important features include precipitation during the month of April followed by minimum night light luminosity. We note that air pollution variables do not appear in the top 10 important features.

Conclusions and Caveats

We believe this is an important first step towards developing a predictive model for economic activity, as measured by real income per capita. We predict economic activity using remotely sensed data at municipality level in Mexico for the years 2009 and 2014. We find that the random forest with 5-fold cross validation is the best model with the lowest RMSE and highest R-squared values. Despite using remotely sensed weather, pollution and nightlights data, the most important feature selected by the model is population. It is critical to note that air pollution is not one of the most important features.

There are several caveats that are important to note as we move forward. First, even if the models we estimate here are very high-dimensional, it is important to bear in mind that we are essentially including three types of variables: air pollution, nightlights and weather. These variables are likely to capture some of the variation in real per capita GDP, but clearly we need to include other variables that capture both time varying characteristics, like land use, and time invariant characteristics like elevation, slope, and agricultural suitability. Second, the computational requirements needed to tune the hyperparameters are already high and will increase as we include more variables, so that we need to be aware of the tradeoff between improvements in model performance and the computational costs. Third, we need to understand where the measurement error in both the measures of income and the remote sensing data come from, so that we can better explain where we are more likely to have larger prediction errors, and what we can do to improve the predictions for those regions.

Finally, this analysis was a pilot of what we hope will be a much more ambitious project. We would like to expand these to other countries, with the hope that by including areas where the ‘structural’ relationship between per capita GDP and all the other variables is different, will allow us to develop a model that does a better job at predicting per capita GDP for other regions of the world.

References

Auffhammer, M. (2018). Quantifying economic damages from climate change J. Econ. Perspect., 32 (4).

Barrios, S. ,Bertinelli L., Strobl, E. (2010). Trends in rainfall and economic growth in Africa: a neglected cause of the African growth tragedy. Rev. Econ. Stat., 92 (2)

Bhandari, L., & Roychowdhury, K. (2011). Night lights and economic activity in India: A study using DMSP-OLS night time images. Proceedings of the Asia-Pacific advanced network, 32(0), 218.

Bundervoet, T., Maiyo, L., Sanghi, A. (2015). Bright lights, big cities: measuring national and subnational economic growth in Africa from outer space, with an application to Kenya and Rwanda. The World Bank

Burke, M., Hsiang S.M. , Miguel E. (2015). Global non-linear effect of temperature on economic production. Nature, 527 (7577)

Chi, G, Fang, H., Chatterjee, S. & Blumenstock, J. (2021) Micro-Estimates of Wealth for all Low- and Middle-Income Countries. Working paper.

Chen, X. and W. D. Nordhaus (2011). Using luminosity data as a proxy for economic statistics. Proceedings of the National Academy of Sciences 108(21), 8589–8594.

Davis, M. E., Laden, F., Hart, J. E., Garshick, E., & Smith, T. J. (2010). Economic activity and trends in ambient air pollution. Environmental Health Perspectives, 118(5), 614-619.

Dell M., Jones B.F., Olken B.A. (2012). Temperature shocks and economic growth: evidence from the last half century. Am. Econ. J. Macroecon., 4 (3).

Dell M., Jones B.F.,Olken B.A. (2014). What do we learn from the weather? The new climate–economy literature. Journal of Economic Literature, 52 (3)

Doll, C.N., Muller, J.P., Morley, J.G. (2016): Mapping regional economic activity from night-time light satellite imagery. Ecological Economics 57(1), 75

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A., & Michaelsen, J. (2015). The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. Scientific Data, 2, 150066.

Funk, C., P. Peterson, S. Peterson, S. Shukla, F. Davenport, J. Michaelsen, K.R. Knapp, M. Landsfeld, G. Husak, L. Harrison, J. Rowland, M. Budde, A. Meiburg, T. Dinku, D. Pedreros, and N. Mata, 2019: A High-Resolution 1983–2016 Tmax Climate Data Record Based on Infrared Temperatures and Stations by the Climate Hazard Center. J. Climate, 32, 5639–5658,

Hammer, M. S.; van Donkelaar, A.; Li, C.; Lyapustin, A.; Sayer, A. M.; Hsu, N. C.; Levy, R. C.; Garay, M. J.; Kalashnikova, O. V.; Kahn, R. A.; Brauer, M.; Apte, J. S.; Henze, D. K.; Zhang, L.; Zhang, Q.; Ford, B.; Pierce, J. R.; and Martin, R. V., Global Estimates and Long-Term Trends of

Fine Particulate Matter Concentrations (1998-2018)., Environ. Sci. Technol, doi: 10.1021/acs.est.0c01764, 2020

Henderson, J. Vernon, Adam Storeygard, and David N. Weil. 2012. "Measuring Economic Growth from Outer Space." American Economic Review, 102 (2): 994-1028. DOI: 10.1257/aer.102.2.994

Johnson, Simon, William Larson, Chris Papageorgiou, and Arvind Subramanian. 2009. “Is Newer Better? Penn World Table Revisions and Their Impact on Growth Estimates.” National Bureau of Economic Research Working Paper 15455.

Kim, S. (2008). Spatial inequality and economic development: Theories, facts, and policies. Urbanization and Growth, 133–166.

Koop, G., McKitrick, R., & Tole, L. (2010). Air pollution, economic activity and respiratory illness: evidence from Canadian cities, 1974–1994. Environmental Modelling & Software, 25(7), 873-885.

Li, Xuecao; Zhou, Yuyu; zhao, Min; Zhao, Xia (2020): Harmonization of DMSP and VIIRS nighttime light data from 1992-2018 at the global scale. figshare. Dataset.

Liang, W., & Yang, M. (2019). Urbanization, economic growth and environmental pollution: Evidence from China. Sustainable Computing: Informatics and Systems, 21, 1-9.

Muller, A.C & Guido, S.. Introduction to machine learning with Python. O'Reilly Media, 2017.

Mveyange, A. (2015). Night lights and regional income inequality in Africa (No. 2015/085). WIDER Working Paper.

Newell, R., Prest, B., Sexton, E. (2021) The GDP-Temperature relationship:Implications for climate change damages. Journal of Environmental Economics and Management 108 (\

Pinkovskiy, M., Sala-i Martin, X. (2016). Lights, camera, income! illuminating the national accounts-household surveys debate. The Quarterly Journal of Economics
131(2), 579{631

Singhal, A., Sahu, S., Chattopadhyay, S., Mukherjee, A., & Bhanja, S. N. (2020). Using night time lights to find regional inequality in India and its relationship with economic development. Plos one, 15(11), e0241907.

Smulders, S., & Gradus, R. (1996). Pollution abatement and long-term growth. European Journal of Political Economy, 12(3), 505-532.

Storm, H., Baylis, K., & Heckelei, T. (2020). Machine learning in agricultural and applied economics. European Review of Agricultural Economics, 47(3), 849-892.

Weidmann, N. B., & Schutte, S. (2017). Using night light emissions for the prediction of local wealth. Journal of Peace Research, 54(2), 125-140.

Zhu, L., Hao, Y., Lu, Z. N., Wu, H., & Ran, Q. (2019). Do economic activities cause air pollution? Evidence from China’s major cities. Sustainable Cities and Society, 49, 101593.