Using Night Time Light data and Machine Learning to predict Population.

By: Keith Carson

University of Colorado, Denver  
Geospatial Big Data, Programing & A.I.  
Spring, 2020

Abstract:

Using Night time light imagery in combination with population data, population was access in the United States at the County level. The goal of this project was to gain experience working with python techniques to create spatial data tables that were further assessed with machine learning analysis to create predictive models. Utilizing two statistical machine learning methods the results of each machine learning model were compared to assess the reliability of each statistical method employed. The results of this data analysis were used to make two statements; Yes, night time light data can be used to accurately predict night time light and yes, there are differences in predictive accuracy. Lastly, model refinement is suggested that could increase the usefulness and accuracy of the machine learning model employed.

Introduction:

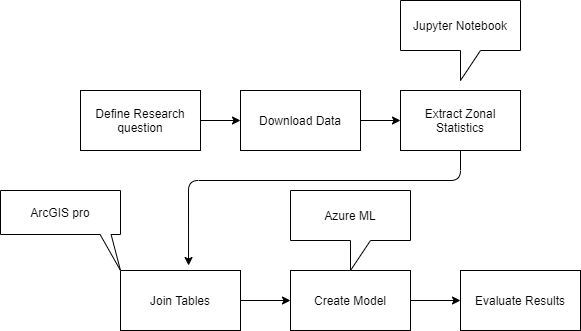
Night time light is a by product of development that is observable using imagery of the earth’s surface obtained from satellite sensors (Bennet & Smith, 2017). The image obtained by satellite can be used with advanced analysis methods to produce usable information and data tables compatible with machine learning software.

The intent of this project is to measure how accurately night time light can be used to assess population across a large scale to answer the following research questions:

* Can night time light data be used to predict population at the county level?
* Does the linear regression method produce more accurate results than the Decision Forest method when assessing population using night time light data?

The process of answering these research question took two forms. First, extracting data table from satellite imagery of night time light at the county level and extracting data tables containing population data at the county level. Second, joining these two data tables and loading them into a machine learning model under the parameters of two different machine learning analysis algorithms.

To perform the analysis to answer the research questions, three programs are used at various stages of data processing. The Work flow to complete this can be seen in Figure 1, that outline the key steps and programs utilized at each phase.



*Figure 1.*

Methods:

The methods to perform this research project were perform in three phases: data acquisition, data preprocessing and machine learning data analysis.

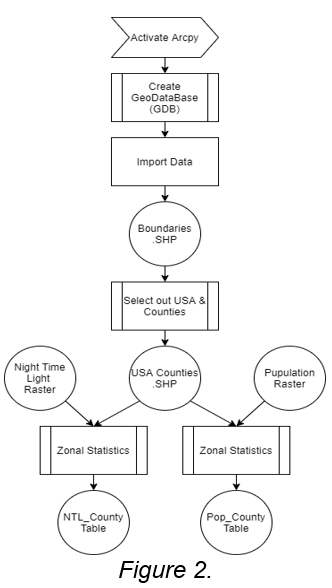
*Data acquisition*

To answer the research questions data sets containing night time light, population data and county boundaries were sourced from data repositories. Night time light data was acquired from the National Oceanic & Atmospheric Administration (NOAA), the United States government agency charged with the stewardship of the environment and disseminating knowledge and data resources with the public. The dataset acquired from NOAA contained raster data with global coverage that depicts the network of night time light across the world. Population data was acquired from the Global Human Settlement Layer (GHS), which is a project hosted by the European Union whose mission is to produce global spatial information on population and on the physical size of settlements across the globe. The population data acquired from GHS was in raster format and contained global coverage. The last data set required to answer the research questions is administrative boundaries. The boundaries data was acquired from the University of California Davis (UCD), data repository. The University of California maintains this data for public distribution and to support academic productions. The boundary data acquired from UCD contained boundaries with almost complete global coverage, for this application, data in the form of shapefiles depicting the United States was utilized.

* + - 1. Used to Calculate NTL Zonal Stats with boundaries.

*Data preprocessing*

Data preprocessing was performed utilizing Jupyter Notebook. To perform spatial data processing utilizing python in Jupyter notebook, the first step is activating Arcpy. The first step of actual data preprocessing was to create a Geo-database (GDB) to host the processed geospatial data. With these two operations performed we can then proceed with data processing.

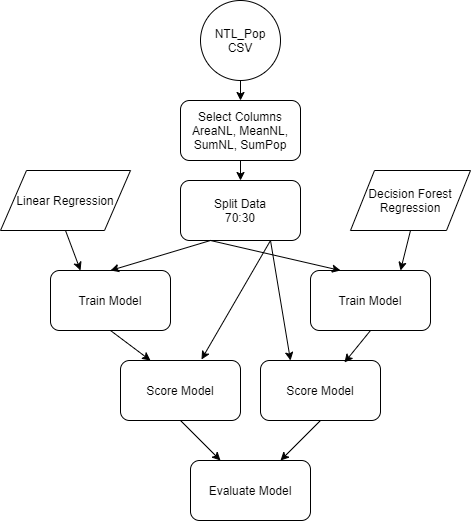
To answer the research questions, data needs to be processed into tables at the county level. Utilizing python in Juptyer Notebook, county boundaries in the united states were selection out of the data set and retained. With county boundaries in the United States isolated in the dataset, zonal-statistics functions was programmed in the Jupyter notebook with the night time light raster. This process creates a zonal-statistics table that contains values that show the amount of night time light for each county in the United States.

To create the second table required to answer the research question, we perform the same process stated above however, instead of utilizing night time light raster, we substitute the population raster acquired from GHS. The output of this process is a zonal-statistics table that contains values that define the amount of population for each county in the United States. The process to create the zonal-statistics tables is displayed in Figure 2. The last step of data preprocessing is to join the two tables, this step was performed in ArcGIS Pro and data was exported as a .CSV file.

*Machine Learning Data Analysis*

With the data in the CSV table format containing all the data required to answer the research questions we are now ready to perform the Machine Learning analysis phase. The Machine Learning analysis (ML) was performed in Microsoft Azure ML Studio. The first step of the ML analysis is to import the CSV file containing the preprocessed data into the Azure Studio. With the data loaded into the workspace, variables from the table are selected utilizing the columns selection tool. The variables selected for this ML experiment were: Area of night time light, Mean of night time light, Sum of night time light and Sum of population since these variables in the table contain the information required to answer the research questions. After selecting the variables from the CSV, the next process is to split the dataset into two group.

The first of the two groups contain 70% of the observations. This larger portion of the data is used to train the model which means the model will statistically assess this dataset to discover patterns in the selected variables. Training the model requires input of a statistical method that is used to explore the patterns in the data set. For this project, two methods were employed to compare differences in outputs to help answer research question 2. The first method is a basic linear regression model. This method was selected because it is straight forward and since it is a basic method the goal is that it will highlight the advantages of using a method that has more depth. The training method the linear regression output is compared to is the Decision Forest method (8 tree). The decision Tree method is more complex then the linear regression method because it iterates on its self, utilizing 8 levels of statistical analysis to extrapolate probability of an outcome occurring.

The second grouping of data, representing 30% of the total data set, is used to test the model for accuracy based on what the model learning from analysis of the larger dataset. The results from the test data set are used to statistically compare the Machine Learning performance. In Figure 3, the architecture of this machine learning experiment is shown in detail.

*Figure 3.*

Results:

After performing the data analysis with Azure ML studio, the results are displayed that compare the two Machine Learning methods. From these results, the research questions driving this project can be answered.

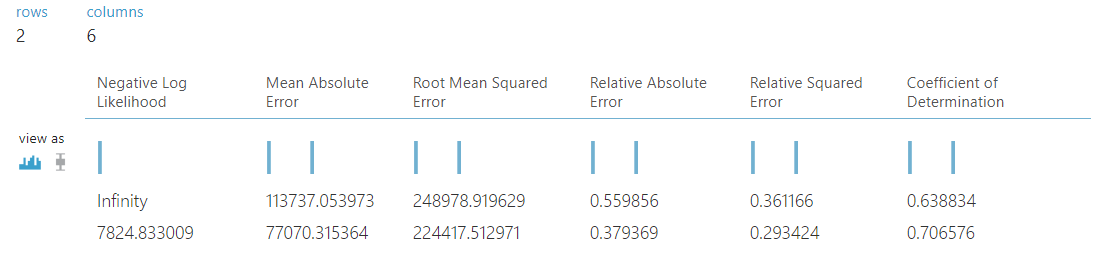
* Yes, we are able to use Night time light data to predict Population over an area.

We can tell that we can accurately utilizing these data inputs to predict population by reviewing the coefficient of determination. With the values of .6388 and .7065 for the linear regression method and Decision Forest Regression, we know that 63% of the time the Linear Regression method is able to accurately predict the population utilizing night time light data. Likewise, we can also see that the Decision Forest Regression is able to accurately predict the population utilizing night time light data 70.65% of the time. Since both of these values are above .50, we can say that night time light is useful to predict population because these methods are more reliable than a simple guess which probability of being correct is 50/50.

By comparing the coefficient of determination, we can also answer our second question, that inquires if one of the train methods is more reliable than the other.

* The Coefficient of determination was higher with the Decision Forest Regression method over the Linear regression method which suggest the decision forest is a better method to utilize when modeling Night time light to predict Population.

Below, Figure 4 displays the output results from the Azure ML studio experiment using the linear regression method and the decision forest method.



*Figure 4.* (Top row, Linear Regression output; Bottom row, Decision Forest Regression output)

Discussion:

During the production of this experiment, numerous issues presented themselves. I had was not able to export tables created in Jupyter Notebook to CSV format. Python does have functionality to perform this task, I was not able to produce python code to perform this task successfully. To overcome this issue, tables created utilizing python were loaded into ArcGIS Pro, which has the ability to read said files and has a graphical user interface that makes manipulating tables and exporting them into CSV much more intuitive than python.

The second issues that occurred during this analysis happened in the first iteration of the Azure ML model. During the initial run of the model, all variables from the CVS table were selected to be utilizing in the model. The multitude of variables included in the model had issues with multicollinearity, where too many of the variables loaded into the model overlapped with each-other. Essentially, 3-4 variables loaded into the equation had the same effect on the respond variable. Since these variables were so similar, they produce in accurate results and the coefficient of determination when included is unrealistically high. By removing these variables and reducing the dimensionality of the model, the coefficient of determination went down however the amount of error went down significantly more. With this step of reducing the about of variables in the model, the model became more trustworthy by increasing model the model’s precision.

*Refining the project in the future*

Predicting population with Night time light data is logical, very rarely is infrastructure built in an area without habitation. To use night time light data to predict population, is like predicting a fish will be in water. If I were to reproduce this project, I would incorporate economic productivity into the analysis to better understand the relationship between NTL and Economic activity since economic activity is not so closely related to night time lighting. Incorporating economic activity into the model in the future, researchers would be able to answer the question: Is night time lighting useful for predicting economic productivity for an area?

Reference:

Bennet MM. & Smith LC. 2017. Advances in using multitemporal night-time lights satellite imagery to detect, estimate, and monitor socioeconomic dynamic. Remote Sensing of Environment (192):176-197; <https://www.sciencedirect.com/science/article/abs/pii/S0034425717300068>

GitHub: <https://github.com/CarK3/NTL_K.Car.>