Patrick Collins
DATA 201 - Final Report
4/27/18

Executive Summary

The purpose of this project was to predict night-time lights (NTLs) in different counties[1] within the United States in the year 2020; more specifically, which counties' NTLs are expected to brighten the most. Variables were selected and projected forward to the year 2016 (when the last NTL data had been collected). Uncertainty was applied in a way that was customized for each different variable, and a model was built on historical data and applied to predicted data for the year 2020. That process was repeated 20,000 times, with different uncertainties applied each time, and aggregate results were collected. The ten counties identified to brighten the most are Los Angeles, California; Maricopa, Arizona; San Diego, California; Riverside, California; San Bernardino, California; Clark, Nevada; King, Washington; Loving, Texas; Middlesex, Massachusetts; and Branch, Michigan.

Methods

Step 1: Selecting data sources to use in predictions

Though a wealth of statistical data is available for the United States, each extra variable incorporated into the model would need to be projected into the future, sometimes twice; to train the historic model and predict the year 2020. Additionally, the uncertainties inherent in how the data is measured, collected, and projected need to be accounted for, and that can easily dilute the predictive value of a variable. Datasets with elevation and distances to coastlines and cities were chosen; all benefiting from their relative stability over time; which makes them less risky to use as predictors, though also less valuable. If they were to change, that could predict changes in NTLs; but as they are now, they just serve to account for a portion of each statistic. Most variables were chosen based on their connection to the urbanization of an area. Higher populations and population densities indicate cities; as, of course, do very low travel-times to cities. Distance to coastline is similarly used for its correlation with urbanization and developed areas. Other variables were chosen for their direct impact on the visibility of NTLs. Areas with higher elevation have less atmosphere to obscure the view of the satellite that collects NTL data. Similarly, concentrations of particulate matter can indicate smog and other occluding pollutants. Historical NTL data is also used to predict future NTL data. Datasets that split their information into more granular locations (counties, etc.) were chosen; in an effort to better measure local trends and patterns in exchange for some extra uncertainty.

Step 2: Projecting known data to the years that we have or want NTL data for

Drawing a straight line between the NTL data for 1992 and 2015, then extending it to 2016, created a very accurate prediction of 2016 night time lights; the same was true for population data. That kind of simple linear projection was used to predict the NTL, population, and air pollution variables into the years 2016 and 2020. 2016 data is needed to train a historical model, and it must be projected for variables whose datasets do not already include data for 2016.[2] The year 2020 was chosen mainly to leverage existing population forecasts; with the assumption that population has a major impact on NTLs.

Step 3: Accounting for uncertainty for each different type of data

Random uncertainty was applied individually to each value in each dataset. Each variable had been assigned a percentage of uncertainty based on research into its validity and collection methodology, as well as whether or not it had been projected into the future or past, and by how much. Elevation and distances to cities and coasts were given low uncertainties, (5%) reflecting their stable situations; though distance to cities was considered more uncertain (15%) due to its reliance on moving populations. NTLs were confirmed to be rather accurate after a bit of research, so they also had somewhat low percentages of uncertainty.[3] (15%) The 2020 NTL projections were of course considered much more uncertain (40%)

---

[1] Some areas were cities/districts/boroughs/municipalities, etc.; not just counties.
[2] And NTL; as the model is being trained to see how a projection of previous data correlates with the actual data.
[3] http://www.mdpi.com/2072-4292/8/1/41/pdf

than the 2016 projections; owing to the larger gap between the prediction and known data. Population values were given quite low levels of uncertainty (3%-6%) - these predictions were based on professionally collected and predicted data, and the population totals for a country generally don't change by more than 1% per year regardless.[4] Air pollution data was given a considerable level of uncertainty, (20-30%) as the data is collected in part by using measurement devices around the country.[5] While there are quite a few of them in the US; they can't cover all areas, and the rest of the data is based on satellite imaging. Uncertainties were made larger the further a variable had been projected into the future. Minimums and maximums were applied as needed, and values were generated in a normal distribution.

Step 4: Building a historic model using known and projected data

The historic model was then built using data for the year 2016 with uncertainty already accounted for. Multiple linear regression models were tested; and ElasticNet was chosen due to it achieving the lowest mean average error (MAE) (5.0) for the historical data it was being trained on.

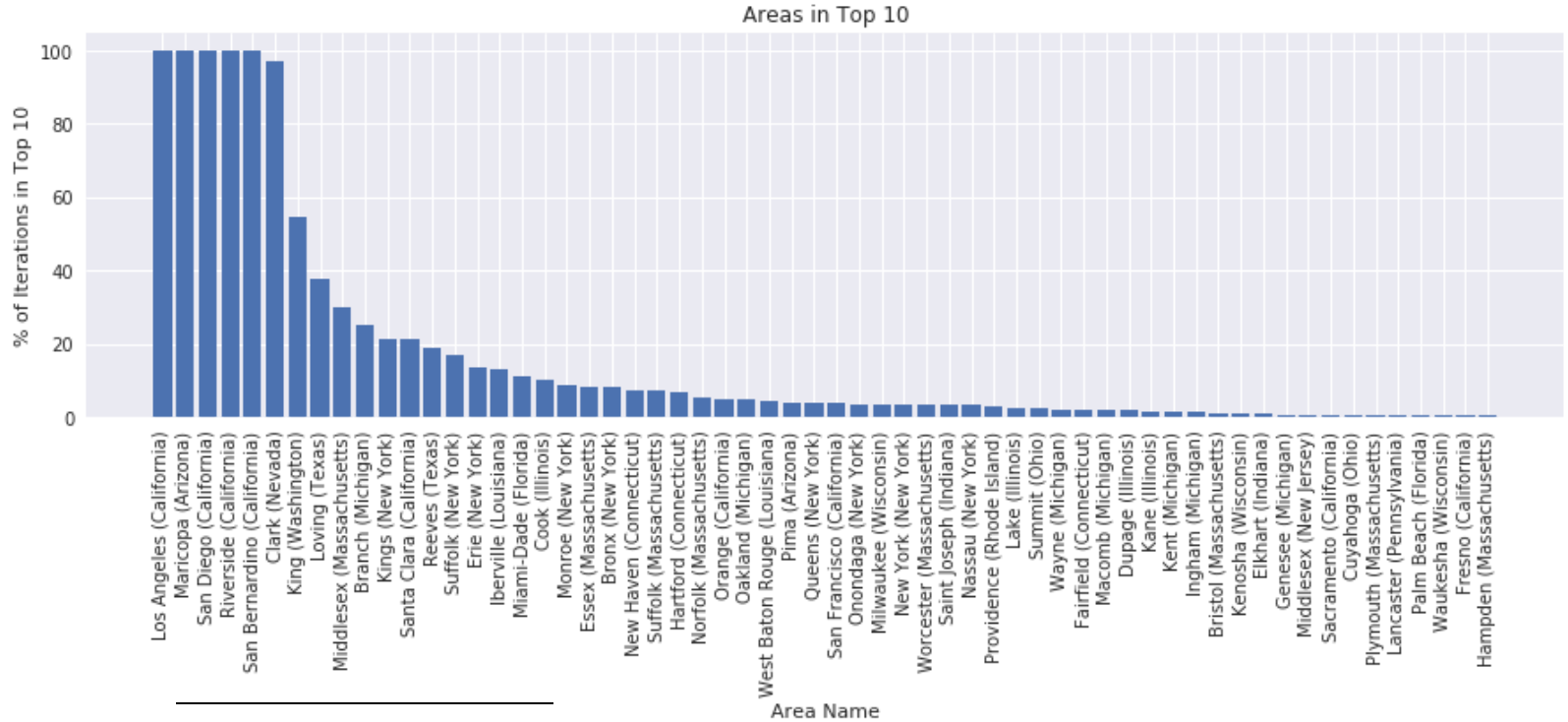
Step 5: Predicting future NTLs; recording the results

The historic model was then applied to data predicted for the year 2020; creating its own predictions for NTLs. The top 10 locations expected to brighten the most were recorded for future aggregation.

Step 6: Repeat steps 3 through 5 20,000 times.

A different set of uncertainties was applied in each iteration; creating enough variance to simulate many possible combinations of true error values for the different types of data. For example, some iterations would test assuming that population was an underestimate while distance to cities was an overestimate, and vice-versa. These differences were enough to change which 10 locations were predicted to brighten the most. These results were aggregated to determine which locations were most often predicted to brighten the most by 2020.

Results



[4] https://www.statista.com/statistics/183481/united-states-population-projection/

[5] http://www.worldbank.org/en/news/feature/2015/07/14/understanding-air-pollution-and-the-way-it-is-measured

Patrick Collins
DATA 201 - Final Report
4/27/18

Across 20,000 iterations, the ten areas identified to brighten the most were Los Angeles, California; Maricopa, Arizona; San Diego, California; Riverside, California; San Bernardino, California; Clark, Nevada; King, Washington; Loving, Texas; Middlesex, Massachusetts; and Branch, Michigan - the first six of which appeared in the top 10 over 90% of the time. In the highest rated units of observation, (Los Angeles, California; Maricopa, Arizona; and San Diego, California,) the model predicts a NTL factor of around 100. This contrasts to a MAE of 5 found during the historic period. Given that the MAE of the historic model is significantly lower than the predicted values, a reasonable degree of confidence can be ascribed to this model if historic trends are similar to future trends; especially when used on areas with brighter NTLs. The predictions were rather accurate; except in cases where counties had a very sharp peak or dip in NTL the previous year.

Limitations

The predictions appear to be heavily reliant on the linear projection of historical NTL. Excluding the NTL projection data from the model achieves results that compare quite favorably to results achieved using random data; but removing all data besides the NTL projection data barely harms the MAE.

Uncertainty is also applied in an un-ideal way. Values for uncertainty are themselves rather uncertain, and could be picked better with a bit more research and expert advice. Values of 0 are not modified when uncertainty is applied, due to concerns about determining what is a reasonable scale.

Some of the data chosen is a cause for concern: total population is heavily tied to population density when measuring within static areas, and both are clearly tied to the distance-to-cities metric. More care needs to be taken to untangle the connections between the chosen data.

Faulty reasoning was used for the level of uncertainty applied to population data. Though the overall population of the country generally changes by less than 1% per year,[5] internal migrations likely mean that population changes for individual counties are much more significant.

Somewhat surprisingly, NTL data was not very stable for many of the areas; sometimes shooting up by large amounts one year and back down the next. That harmed the accuracy of the linear projection data that these predictions rely on.

The MAE for the historical training data is rather large, (5.0;) much larger than the MAE achieved when predicting 2016 data (1.1.) This difference is flipped (causing historical MAE of 0.8 and future MAE of 4.4) when using a model based on linear regression rather than elastic net; creating a situation that prompts further investigation. MAEs around 5 are reaching the mean of the NTL data; thus indicating somewhat meaningless predictions.

Key Findings

- Anchoring predictions to historical values of the same variable can be very effective.
- Mixing highly effective (past NTL data) and highly ineffective (air pollution, elevation) predictors can harm predictions when using certain models.
- NTL data often doesn't follow clean trends; upwards or downwards.