Authors
Robert Crimi, Elanor Hoak, Ishita Srivastava, Jesse Wisniewski, Jonathan Blackstock, Keerthi Chikalbettu Pai, Melissa Bica, Michelle Bray, Mikhail Chowdhury, Monal Narasimhamurthy, Nika Shafranov, Sachin Muralidhara, Satchel Spencer, Saurabh Sood, and Caleb Phillips

Projects in Geospatial Data Analysis: Spring 2016

Editor:
Caleb Phillips

Authors:
*Robert Crimi**
Elanor Hoak
*Ishita Srivastava**
Jesse Wisniewski
*Jonathan Blackstock**
*Keerthi Chikalbettu Pai**
*Melissa Bica**
Michelle D. Bray
Mikhail Chowdhury
*Monal Narasimhamurthy**
Nika Shafranov
*Sachin Muralidhara**
Satchel Spencer
*Saurabh Sood**

University of Colorado
Department of Computer Science
Boulder, CO, 80309-0430 USA

August 5, 2016

* Graduate Students

Foreword

This document contains semester projects for students in CSCI 4380/7000 Geospatial Data Analysis (GSA). The course explores the technical aspects of programmatic geospatial data analysis with a focus on GIS concepts, custom GIS programming, analytical and statistical methods, and open source tools and frameworks.

Students were given the final 4 weeks of class (reduced from 6 weeks in 2015) to work on a project that applies skills and concepts from the class to a problem of their choosing. This project is a focused sprint demonstrating how skills developed during the course might be applied to problems of the students' personal interests. Aside from some readings, lectures and class discussion, students were allowed to work on their projects exclusively during that time and were supported with meetings, peer-discussion and copyediting. In terms of the scope of the final product, undergraduate students were asked to perform a research or engineering task of some complexity while graduate students were additionally required to perform a survey of related work, demonstrate some novelty in their approach, and describe the position of their contribution within the broader literature. All students who performed at or above expectation were offered the opportunity to contribute their paper for publication in this technical summary.

The diversity of the papers herein is representative of the diversity of interests of the students in the class. There is no common trend among the papers submitted and each takes a different topic to task. Students made use of open, public GIS data from federal and environmental agencies, as well as research data sources. Several students pivoted their projects early on due to limitations and difficulties in data access --- a real-world challenge in geospatial research and GIS. The projects herein range from analyzing snowpack, lightening strikes, and light pollution to wireless security, disaster informatics, and human gut flora. Analysis approaches are similarly varied: geovisualization, geostatistical modeling, multiple regression, graph analysis, etc.. Most papers can be understood as exploratory data analysis, although some emphasize interactive visualization and others emphasize statistical modeling and prediction aimed at testing a well-defined research question. To inform the style of the their approach, students read papers from a broad sampling of geospatial research. They used these readings to build an understanding of approaches to presentation and analysis in the modern scientific literature. Two research projects developed during the class (both by Ph.D. students) have been henceforth submitted to peer-reviewed academic venues for publication.

Please direct questions/comments on individual papers to the student authors when contact information has been made available.

Table of Contents

An Analysis of Diet Trends and Potential Effects on American Health
Nika Shafranov
6 pages

Predicting transportation modes through GPS logs
Sachin Muralidhara
5 pages

Demographic Access to Wireless Security
Satchel Spencer
5 pages

Geoanalysis of the 2016 US Presidential Elections using Twitter Data
Saurabh Sood
6 pages

Towards High-Resolution Real-Time Avalanche Forecasting

Robert Crimi
University of Colorado, Boulder
robert.crimi@colorado.edu

Abstract

In an effort to improve the current state of avalanche prediction, this research aims to create a framework for avalanche forecasting based on Machine Learning and Data Mining techniques. Historical avalanche data, collected from the Colorado Avalanche Information Center (CAIC), a Digital Elevation Model, and a binary machine learning classifier will be used in an attempt to predict avalanche danger at a high-resolution in real-time. This classification system will be an experimental analysis to show whether Machine Learning techniques could be a viable option in avalanche forecasting. This will improve the current state of avalanche forecasting techniques that rely heavily on expert knowledge.

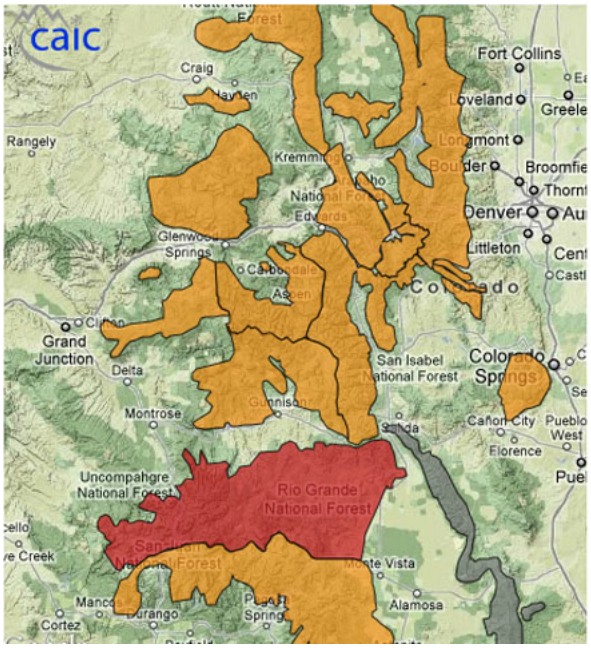
I. Introduction

With recent advances in backcountry skiing and snowboarding technology, more and more people are moving away from ski areas into the backcountry to experience the thrill of skiing. Due to this high influx of backcountry skiers, we are seeing a large rise in the number of avalanche accidents over the past few years. The average number of fatalities due to avalanches in the US has doubled since 1991[1].

In an effort to improve the safety of skiers and snowboarders entering the backcountry, this research aims to establish a framework for predicting avalanches. Currently, there are a few online resources that aim to give insight into the current avalanche danger in particular areas. These prediction systems are flawed in a few ways. For one, their predictions are made by human observations. While the people making these predictions are usually experts in the field, there are no "set-in-stone" statistical models for avalanche forecasting. Thus, we rely heavily on the forecaster's expertise in the field when looking at avalanche predictions.

Another pitfall of the current state of avalanche forecasting is the lack of high-resolution prediction systems. This is mostly due to the lack of avalanche forecasters. Since there are very few of these forecasters, predictions are done on a large scale. Forecasters will use observational measurements at several locations and interpolate to make their predictions. For example, below is CAIC's forecast map for the Colorado Rockies region.

Figure 1: *CAIC Forecasts*



The forecasts on this map are done at a mountain range size resolution. The entire

Front Range of Colorado is grouped together and given the same forecast. While this type of prediction is useful in giving relative avalanche danger, it does not let a user "hone-in" on a particular area to get a finer resolution prediction. This research will attempt to allow users to get high-resolution forecasts in specific areas of interest. Users will be able to explore an interactive map where they will be able to zoom into a particular area for a high-resolution forecast.

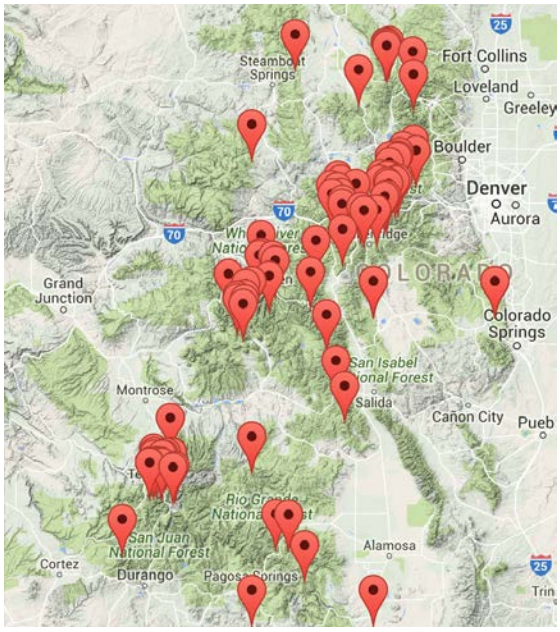
Unfortunately, due to the lack of an online community for backcountry skiing and snowboarding, there is little data about backcountry skiing. The only readily available data comes from online resources that document avalanche occurrences. Currently, the biggest accessible resource for this data is the Colorado Avalanche Information Center's database of avalanche observations. While their dataset is large (over 2,000 records), the set only contains cases of avalanche occurrences. To successfully and accurately train a Machine Learning classifier, one must obtain cases where there were no avalanches. Currently, there are no resources for obtaining "successful" (no avalanche occurred) backcountry trip records.

II. Data

I. Avalanche Reports

In order to train a binary classifier, one must obtain feature vectors of positive $\vec{X_+}$ and negative $\vec{X_-}$ examples. Throughout this paper, positive avalanche cases will represent times when an avalanche occurred due to a skier or snowboarder while negative cases will represent times when there were no avalanches triggered. Positive examples were acquired from the Colorado Avalanche Information Center (CAIC)[2]. Figure 2 shows the locations of reports in the Colorado region acquired from the CAIC. CAIC's online database has reports containing data about the terrain's elevation, slope, aspect, and characteristic and additional metadata including date, time of day, and size of avalanche.

Figure 2: *CAIC Avalanche Reports*



Unfortunately, these reports are the only examples available. In order to train a Machine Learning classifier, we must obtain these reports ourselves. To do this, assumptions about backcountry areas with high traffic will be made in order to generate negative examples. We assumed that on a given date when there were no reports of an avalanche in a highly visited backcountry area, negative test cases could be generated based on the features of the location. For example, a few negative training points were acquired after observing no avalanche cases during a weekend on Loveland Pass, which is one of the most highly visited backcountry areas in Colorado[3]. Samples of these areas were acquired from the Digital Elevation Model and similar models.

II. Digital Elevation Model

A Digital Elevation Model (DEM) is a representation of continuous elevation values over a topographic surface. The DEM data used in this research was acquired from ColoradoView at a resolution of 10 meters and will be the foundation for the forecasting mechanisms[4]. Aspect and slope data was generated from the

DEM data and GDAL tools. Due to the large size of the data, these data sources were downsampled to a resolution of 50 meters. Thus, all forecasts will have a resolution of 50 meters.

III. Methods

Supervised Machine Learning algorithms for binary classification attempt to learn from a set of labeled positive and negative training examples. They build statistical models from the training data and use these models to predict the classification of new data. In this research, a binary classifier to distinguish between avalanche prone and non-avalanche prone areas was trained.

There are many binary classification algorithms in the field of Machine Learning. When choosing between different binary classifiers, it is important to weigh the strengths and weaknesses of each one. For example, a standard linear SVM will only be accurate if the data is linearly separable.

Throughout this research, Python's Scikit-Learn package was used to train all classifiers[5]. This package allows data scientists to easily build statistical Machine Learning models with ease. Scikit-Learn implements a multitude of Machine Learning techniques including SVMs, Naive Bayes classifiers, and Logistic Regression.

In this research, the results of two particular Machine Learning algorithms will be explored. These include a Support Vector Machine (SVM) and a Naive Bayes classifier. An SVM attempts to draw a linear hyperplane between two classes that maximizes the space between the hyperplane and the "support vectors". In its standard form, this is a linear hyperplane. While a linear SVM is useful for a lot of classification systems, it is only accurate when classes can be linearly separable. To separate classes that may not be linearly separable, one can use a kernel. A kernel transforms the space that the data lives in hopes of finding a linear hyperplane. The most common kernels for SVM's are polynomial, string, and Radial Based Functions (RBF). A polynomial kernel attempts to build polynomial relationships between the features in order to transform the space into a higher dimension[6]. In this research, accuracy will be tested on a linear, polynomial, and RBF SVM.

In addition, a Naive Bayes classifier will also be trained and results will be compared to those of the SVM. A Naive Bayes classifier is a useful algorithm when a dataset is sparse. In this research, the data was acquired from skiers and snowboarders in the field. Unfortunately, when one must rely on individuals to report data, the data is usually very sparse. For example, a skier caught in an avalanche may only report information about the slope and elevation of the area. Other skiers may only report about the snow characteristics and time of day. In most binary classification algorithms, it is very difficult to interpret missing data. For example, an SVM must have all features present to be successfully trained and predict data. However, a Naive Bayes classifier is able to make assumptions about the data even when features are missing[7]. For this reason, a Naive Bayes classifier will most likely be the most effective at making predictions of avalanches with the acquired data.

IV. Results

I. Support Vector Machine

The first task was to train a linear SVM. The starting hypothesis was that a linear classifier would not be very accurate as the provided dataset was fairly sparse and lived in very high dimensions. However, as a baseline, a linear SVM was trained using 80% of the training data. Testing with the remaining 20%, 37% accuracy was achieved. This result was consistent with the hypothesis about using a linear classifier. Test points near the linear hyperplane were consistently misclassified.

The next step was to introduce a kernel function into the SVM training process. A few kernel functions were tested including polynomial and RBF methods. With a polynomial kernel 82% accuracy was achieved. This was much

higher accuracy than the linear SVM results, which agreed with the hypothesis. Like-wise, when using a RBF kernel, 80% accuracy was achieved. These results show that the relationships between features in the data cannot be described linearly and a higher-order kernel must be used.

II. Naive Bayes Classifier

The hypotheses was that a Naive Bayes classifier would give me much better results. Fortunately, Scikit-Learn's library makes it very easy to switch between different classifiers without much change in one's code. Switching to a Gaussian Naive Bayes classifier, 90% accuracy was achieved. This increase in accuracy was due to the sparseness of the data. In the experiments with SVM's, a good portion of my training data was thrown away due to missing features. However, when using a Naive Bayes classifier, all of the training data was able to be used. This not only allowed the classifier to be trained on more data, but allowed the classifier to build models based on sparse data.

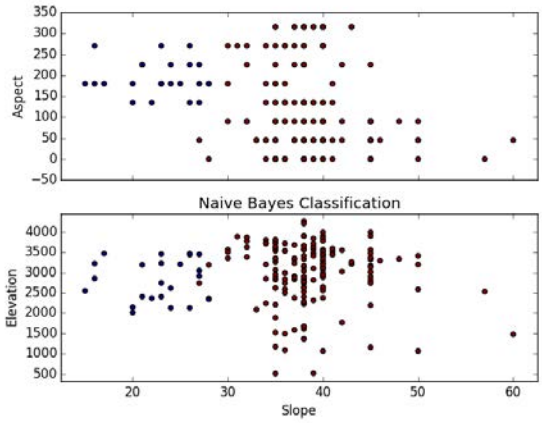
Below is a summary of the results of the different classification techniques:

ML Algorithm	Kernel	Accuracy
SVM	Linear	37%
	Polynomial	82%
	RBF	80%
Naive Bayes	Gaussian	90%

These results proved that a Naive Bayes classifier was the most accurate. As a result, all future classifications were made using this algorithm.

Figure 3 shows the classification of test data when using the Naive Bayes Classifier. It appears as though the classifier is using the slope of the terrain as the biggest indicator of whether or not there will be an avalanche.

Figure 3: *Model Forecasts*



III. DEM Classifications

After concluding that a Naive Bayes Classifier was the best algorithm for the data, the DEM data was classified. Due to the immense size of the DEM, the model was averaged from 10 m to 50 m resolution. The classifications were done in the area surrounding Vail, CO. This area is another commonly visited area in Colorado by backcountry skiers and snowboarders[3].

Figure 4: *Model Forecasts*

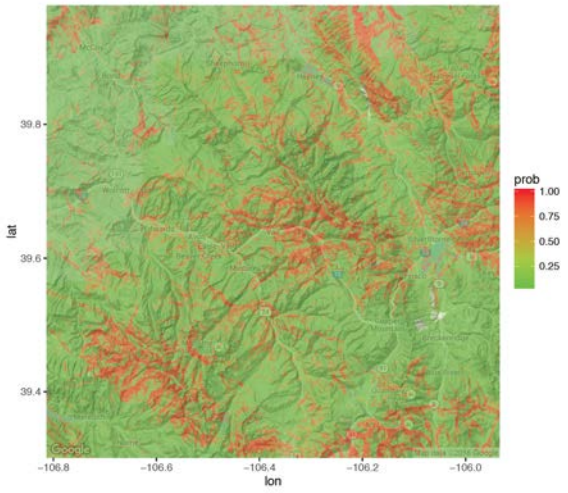
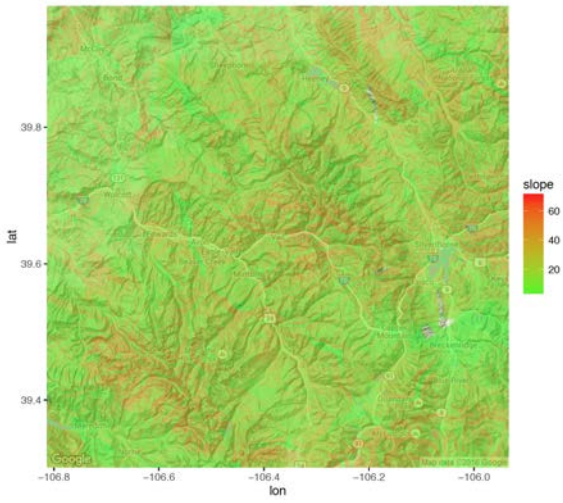


Figure 4 shows the results of this classification. High elevation locations with moderate

slopes appear to be the areas where the classifier predicted high avalanche danger. This observation is made clearer when observing figure 5, which shows the average slope of the predicted area over 50 meter resolution.

Figure 5: *DEM Slope Data*



Area's the classifier predicted had high avalanche danger were also areas where the slope was higher than 30 degrees. These results agree with the test results of the Naive Bayes Classifier. They also agree with statistical data about avalanche occurrences, which show that more than 75% of avalanche cases occur when the slope is higher than 30 degrees[8].

V. Conclusions

Overall, this research aims to be a proof-of-concept for avalanche forecasting using Machine Learning algorithms. Results of this research showed that avalanche predictions were most accurate when using a Naive Bayes classifier, due to the sparseness of avalanche data. Future work in this field should use such a classifier to attain high accuracy. Future work will aim to use more data about avalanche occurrences to train the classifiers. For example, weather data could be incorporated to give real-time forecasts. This was originally a goal of this research; However due to data limitations, weather data was not able to be used.

Overall, more data would ultimately build a more advanced model and improve the Naive Bayes classifier's accuracy. Due to the lack of online resources and data, an online community for backcountry skiing and snowboarding would greatly help in this data retrieval. This community would allow skiers and snowboarders to upload trip reports of their backcountry trips. This framework should allow backcountry skiers and snowboarders to log information about slope, aspect, elevation, snowpack, etc. as well as whether or not there was an avalanche. Relevant data will be mined into a large-scale database in order to build a statistical model for avalanche forecasting. In order to make logging this information as easy and streamlined as possible, the online community would be accompanied by a mobile app.

There are many benefits to using a mobile app to log these reports. As mobile devices are becoming more and more advanced, the amount of data they can collect is ever increasing. For example, most mobile devices are able to track one's GPS location. From this GPS data and our DEM data, the slope, aspect, and elevation of skier's and snowboarder's runs will be automatically collected by the mobile app. The GPS data will also allow the app to automatically collect weather information. An easy-to-use entry system will allow users to log other information about their runs that cannot be tracked through GPS. For example, the user would be able to enter information about the snow-pack and wind speed as well as whether or not an avalanche occurred. This streamlined app will allow users to easily track their runs while simultaneously advancing the avalanche model and improving the safety of backcountry skiers and snowboarders.

References

[1] "Statistics and Reporting." Colorado Avalanche Information Center. CAIC, n.d. Web. 18 Apr. 2016.

[2] "Colorado Data." ColoradoView. Colorado State University, 28 Feb. 2016. Web. 01 Mar. 2016.

[3] "POPULAR COLORADO BACKCOUNTRY SKIING ROUTES." Colorado Ski Authority. Colorado Ski Authority, 2016. Web. 15 Apr. 2016.

[4] "DEM." Def. 1. GIS Dictionary: An Alphabetical List of the Terms and Their Definitions Relating to Geographic Information and Its Associated Technologies. London: Association for Geographic Information, Standards Committee, 1991. Print.

[5] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[6] Manning, Christopher D., and Prabhakar Raghavan. "Nonlinear SVMs." Nonlinear SVMs. Cambridge University Press, 2008. Web. 26 Apr. 2016.

[7] Mitchell, Tom M. "Generative and Discriminative Classifiers: Naive Bayes and Logistic Regression." (n.d.): n. pag. 15 Feb. 2016. Web. 2 May 2016.

[8] Weed, Toby. "Steepness." Utah Avalanche Center. N.p., 27 Mar. 2013. Web. 1 May 2016.

Striking Changes: Trends in Lightning Frequency and Magnitude in the United States

Elanor Hoak

University of Colorado
elanor.hoak@colorado.edu

Abstract

Climate change is a phenomenon proven to influence the majority of Earth's natural cycles. Research has given focus to its most human-centric impacts such as temperature, precipitation, and sea-level rise. Extreme temperatures have been correlated with tropical storms and more intense weather world-wide. This paper investigates how lightning storms have been affected over time as its contributing factors change. The National Oceanic and Atmospheric Administration (NOAA) provides the data for the number of lightning events in the years 1996 to 2015. Rather than using the whole database for the entire United States, three states have been selected for comparison based on pas statistics for yearly lightning frequency: low frequency, moderate frequency, and high frequently. These chosen states are California, Colorado, and Florida respectfully. Lightning activity trends over time should be good indicator for determining a rough overall developments in lightning storms and an estimate for future storm frequency and position.

I. Introduction

Change is defined as a difference between two things, and for climate change, this means differences from one time compared to another. In order to understand what changes will come in the future, we must compare event trends that have happened at different times in the past. Earth is made of countless complex, interconnected systems, and lightning is one of the less understood phenomenon. Causes of lightning are largely unknown, but many other changing systems were mysteries until researchers investigated what external components influenced them.

This paper is not asking what is causing lightning events to change over time, but if it is yet another system that has been impacted over recent years. If there is a noticeable trend, then further research may be able to connect lightning frequency and possibly geographic position to climate change. Most lightning is not a concern for people, remaining high in clouds. However, if storms become more severe, damage to property, crops, and electricity could have a much wider impact on people, so it would be useful to know if we should prepare for this kind of change in the future.

Though there has been evidence indicating overall storms are becoming less frequent yet more severe (Voiland), not all storms yield lighting. Lightning storms overlap classifications of other storms such as hail, rain, and even dust. The National Oceanic and Atmospheric Administration (NOAA) has recorded lighting strike data since 1996 for every state in the U.S. Because this database is so vast, the states of California, Colorado, and Florida have been selected as the main focus of this paper.

California is known to have some of the fewest lightning storms, and Florida has the most (King). Colorado is also in the top ten most lightning-prone states, however it has a high geographic contrast to its maritime fellows and is located half way between both. If there is an indication that lightning behavior has changed over the past two decades, then comparing the lightning frequency of these states will reveal it.

II. Data

Recording when and where lightning strikes is relatively simple but has not been extensively recorded by NOAA until 1996. NOAA's dataset is available through 2015, giving nearly 20 years of lightning data for the United States. The data is in in CSV format containing an immense selection of storm details such as azimuth and duration, but this paper is only concerned with the fields of latitude, longitude, year, and state. The latitude and longitude used are the ones where the storm began because some traveled long distances before dissipating.

Global coordinates were not specifically recorded until late 2006, providing only the city name as a reference point, so the coordinates of these early storms had to be added in order to produce a complete map of the data. The most data is attributed to Florida and the least to California. ColoradoâĂŹs lightning data fell between to two with less than half the storm frequency of Florida.

III. Methods

Before the data could be used, the empty fields of latitude and longitude had to be entered for 2006 and earlier. Most of the cities experienced a storm at a later date, which then had its coordinates recorded, and those that were left hand to be found by hand. Because only the city name was given, the latitude and longitude for that city as defined by Google Earth was used as a general point of origin. The California data had a shift in some and a couple of points that fell outside the state boundary fields that had to be corrected. One Colorado city could not be located and therefore had to be omitted from the study.

With the data in order, the states could be compared. R was used for calculations and creating graphs for each state as shown in Figure 1. New CSV files were created to easily store the lightning statistics for several comparisons: the total number of storms per year, per state; box plots for average latitude; and box plots for average latitude. Three graphs were produced to show the trends for each state.

Also using R and GoogleâĂŹs ggplot2 package, a map was made to show the dispersal of storms for the entire period (Figure 3). This method was also used to make the box plots for each state's latitude and longitude, Figure 3 is the example for Colorado's latitude. These plots allow us to see spatial change on a more specific scale.

IV. Results

Three aspects of storm change were analyzed: storm frequency, latitude position, and longitude position. Concerning storm frequency, there is no overall trend that goes for all three states. Most noticeable is the fact that the number of storms decreases steadily according to Colorado's plot in Figure 1. Florida also exhibits a slight decrease overall, but not as clearly. California and Florida spike and fall in storm count at about the same times. These high and low points, though illustrated in Colorado's graph, are seen more as times of leveling off and further decline. As rough conclusion, the number of storms is decreasing but not evenly across the country.

As for spatial distribution, all states exhibit a sinusoidal tendency in latitude and longitude. Florida has the greatest fluctuation in mean latitude particularly for recent years. Whereas most storms took place between 27°N and 29°N, there is a slight shift northward toward 30°N from 2008 onward. This could correlate to warmer tropical storms being able to travel farther north.

Change in longitude varies greatly for all three states and no tendency can be found over time. The data for Florida is particularly difficult concerning this attribute because the data only concerns the land-based storms. All storms occurring on the panhandle are marked as outliers. In general, it appears that the longitude distribution has remained normal over the past nineteen-year period.

V. Discussion

Change in lightning frequency over the past couple decades shows no concrete trends. If anything, there is a decrease in frequency, as shown significantly by Colorado and slightly by Florida, but the United States as a whole may not bare the same tendency. The geographical placement of lightning storms appears to change according to five- to eight-year cycles, and these cycles show little deviation. This study focused on only a small subset of the country's area and over a relatively short time span. The intensity of the storms and their duration were also not examined but may also be factors connected to change. Further research using data from more area and going further back in time may give a clearer idea about how lightning is changing and what causes it to change.

VI. Figures

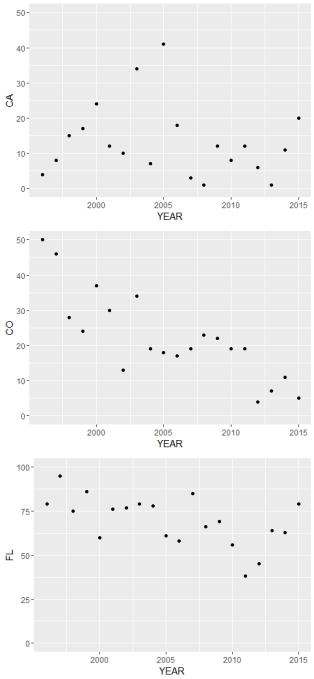


Figure 1. Number of storms per year for each state. From top to bottom: California, Colorado, Florida.

Figure 2. Map of all points used.

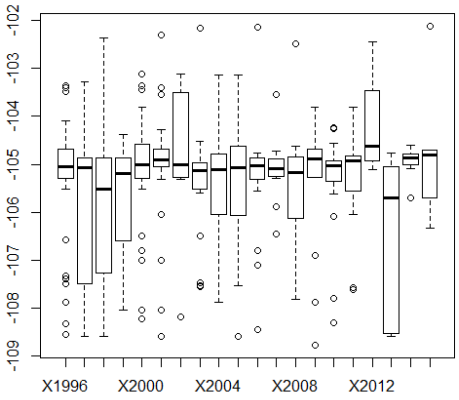


Figure 3. Box plot for Colorado's latitude coordinate of storms showing north-south variance in storm position over the years.

References

[1] D. Kahle and H. Wickham. ggmap: Spatial Visualization with ggplot2. The R Journal, 5(1), 144-161. URL http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf

[2] King, Hobart. "World Lightning Map." World Lightning Strikes Map. Geology.com, 24 Apr. 2015. Web. April 2016.

[3] Severe Weather Data, NOAA National Centers for Environmental Information, Version 3.0, http://www.ncdc.noaa.gov/stormevents/ (accessed Apr 2016)

[4] Voiland, Adam. "In a Warming World, the Storms May Be Fewer But Stronger : Feature Articles." In a Warming World, the Storms May Be Fewer But Stronger : Feature Articles." NASA Earth Observatory, 5 Mar. 2013. Web. April 2016.

Predicting Markets for Hyperloop Technology

Ishita Srivastava*

University of Colorado Boulder
ishita.srivastava@colorado.edu

Abstract

Technology and innovation has grown by leaps and bounds over the past decade. Elon Musk, CEO of Tesla motors and SpaceX, envisioned the fifth mode of transportation that would be not only a high-speed transportation system with an average speed twice that of a typical jet but would also have immunity to weather, ability to never experience crashes and low power requirements.This conceptual high speed transportation system is known as Hyperloop which has been open-sourced by Elon Musk and SpaceX. Several companies and student-led teams are working to advance this technology. This paper presents an analysis on various factors responsible for predicting the potential markets for this technology. As per the analysis, Median Household income comes out as the strongest factor affecting the prediction followed by mindset of people-Liberal or Conservative and population. Based on further analysis, New York out-stands as the most favorable market followed by Los Angeles and Chicago.

I. Introduction

Hyperloop technology has the potential to breakthrough as the fastest, safe and convenient mode of transportation in the coming years. According to Elon Musk, CEO and founder of Tesla motors and SpaceX- "A one way trip between San Francisco and Los Angeles on the Hyperloop is projected to take about 35 minutes." This means that science fiction of tele-porters and jetpacks over the years can finally step into reality with Hyperloop technology by not only allowing mass transit but short travel times and reduced damage on the environment. Why hyperloop- the need stems from the drawbacks of the current means of transportation. Road travel is sometimes convenient for smaller distances but poses serious environmental damages with carbon emissions and worsens energy consumption. This makes mass transit very crucial in the years to come. Rail transport, although relatively better environmental friendly option, proves to be slow and expensive for mass travel. Air travel allows us to travel long distances in less time but for places which are are approximately 900-1000 miles apart, supersonic travel becomes a hassle. Hyperloop eliminates the problem of environmental damage by facilitating tubes with solar panels installed on the roof, allowing for a clean and self-powering system. In addition to this, according to the white paper provided by Elon Musk on Hyperloop-"The Hyperloop (or something similar) is, the right solution for the specific case of high traffic city pairs that are less than about 1500 km or 900 miles apart." Since this technology calls for specific audience, it is important to analyze what factors would contribute to predict the possible markets for this technology to be installed. These factors can then be tuned to present an analysis with comparison to data on current means of transportation: Air transport data, train transport data. In this paper, a similar analysis is presented.

II. Data

The prediction of markets for hyper-loop technology needs a base model against which fac-

*A thank you or further information

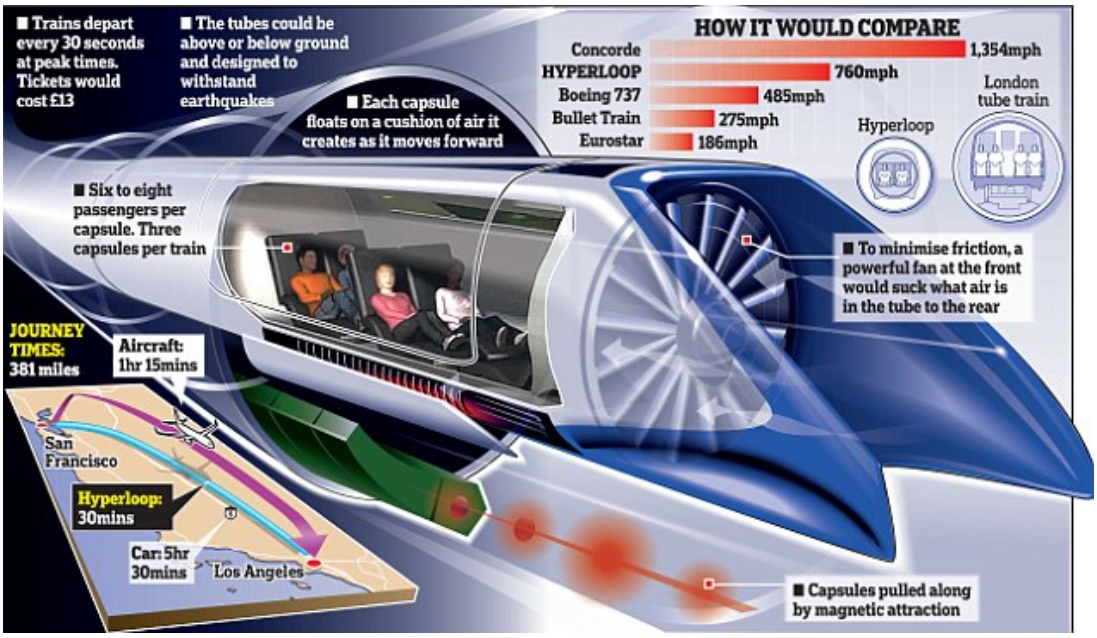


Figure 1: *Hyperloop Technology and its comparison with current means of transportation*

tors possibly affecting this prediction can be analyzed. In order to to this, current means of transportation like air travel and train travel is used. Number of passengers from each city airport and number of train stations in each city is used as a metric to prioritize cites based on its popularity in terms of travel. The data for this base model is collected from "data.gov" website. Distance between two cities is identified as one of the primary features affecting the prediction of hyperloop technology. This is because, according to the white paper provided by Elon Musk on Hyperloop-"The Hyperloop (or something similar) is, the right solution for the specific case of high traffic city pairs that are less than about 1500 km or 900 miles apart." This data has been collected from "infoplease.com" where the distance between cities is recorded as an NxN matrix(N=city). In addition to this, population and median household income are also identified as factors affecting this prediction and the data for each city is collected from "Wikipedia". Lastly, nature of the people-liberal or conservative is also considered as one of the factors assuming that people who are more liberal are more open to technology. This determines the acceptance of hyper-loop technology in each city and is determined using publicly available political donation data on "crowdpac.com". This data is organised and used for further analysis.

III. Methods

Firstly, using the distance data between two cities, those city pairs in USA are extracted which are less than 900 miles apart thereby making distance between cities as one of the primary features to predict markets for hyperloop technology. Secondly, economic status of the cities is identified by collecting data on the median household income of various cities. This factor is taken into consideration assuming that people with higher income would be able to afford such transportation and hence those cities with higher median household income will prove to be a market for hyperloop technology. Apart from this, population in cities is also identified as a factor in determining whether a city is a potential market. It is assumed that higher the population, more are the chances for people to use the technol-

ogy.In addition to this, the mindset of people in cities- Liberal or Conservative is assumed to be correlated to the fact that liberal people are more open to new technology. The current high traffic cities are identified using data like number of air passengers and number of train stations in each city. This is used as a base model to rank each city as a candidate for the technology. Statistical methods like pearson correlation and plots are used to determine whether features like population, median household income, and mindset of the people in the cities- Conservative or Liberal form any relationship among themselves or with the base model. These relationships can either not exist or exist in a positive or a negative direction. Based on these results, parameters are determined to predict the city pairs for hyperloop technology. The following equation is used to rank each city as a possible market for hyperloop technology:

Ax + By + Cz where x=Population of the city y=Median household income z=Nature of the people and A,B,C are the Pearson coefficients denoting the extent to which the factors are correlated to current air and train travel data.

IV. Results

The air travel and train travel data are combined to form a base model and on evaluating the aforementioned factors against it, the following results were found.

I. Base Model v/s Population

The plot in Figure 2 shows a positive relationship of population with the base model. This also proves the assumption that higher the population, higher is the rate of transportation. However, the relationship is not very sharp and this is confirmed through Pearson Correlation test: 0.3201581 as the Pearson coefficient. The Pearson test shows that there is some positive correlation however not a very strong one.

II. Base Model v/s Median Household Income

The plot in Figure 3 shows a positive relationship of median household income with the base model. This also proves the assumption that higher the median household income, higher is the rate of transportation. However, the relationship is sharp relative to population and this is confirmed through Pearson Correlation test: 0.6047431 as the Pearson coefficient. The Pearson test shows that there is some positive correlation and much sharper than population which had the Pearson coefficient as 0.3201581. This establishes the fact that median household income affects the prediction of markets for hyperloop technology more in comparison to population.

III. Base Model v/s Mindset of People: Liberal or Conservative

The plot in Figure 4 shows a positive relationship of mindset of people with the base model. This also proves the assumption that more liberal the mindset of the people, higher is the rate of transportation. The relationship is sharp relative to population but slightly less relative to median household income and this is confirmed through Pearson Correlation test: 0.6047431 as the Pearson coefficient. The Pearson test shows that there is some positive correlation and much sharper than population which had the Pearson coefficient as 0.3201581. This establishes the fact that mindset of people affects the prediction of markets for hyperloop technology more in comparison to population but less in comparison to median household income.

IV. Final Prediction

Figure 5 shows a list of cities ranked as a possible market for hyperloop technology. Taking Pearson coefficients as parameters for each of the factors: Population, Median Household Income and Mindset of People, a metric with the following equation is calculated: res= Ax + By + Cz where x=Population of the city y=Median

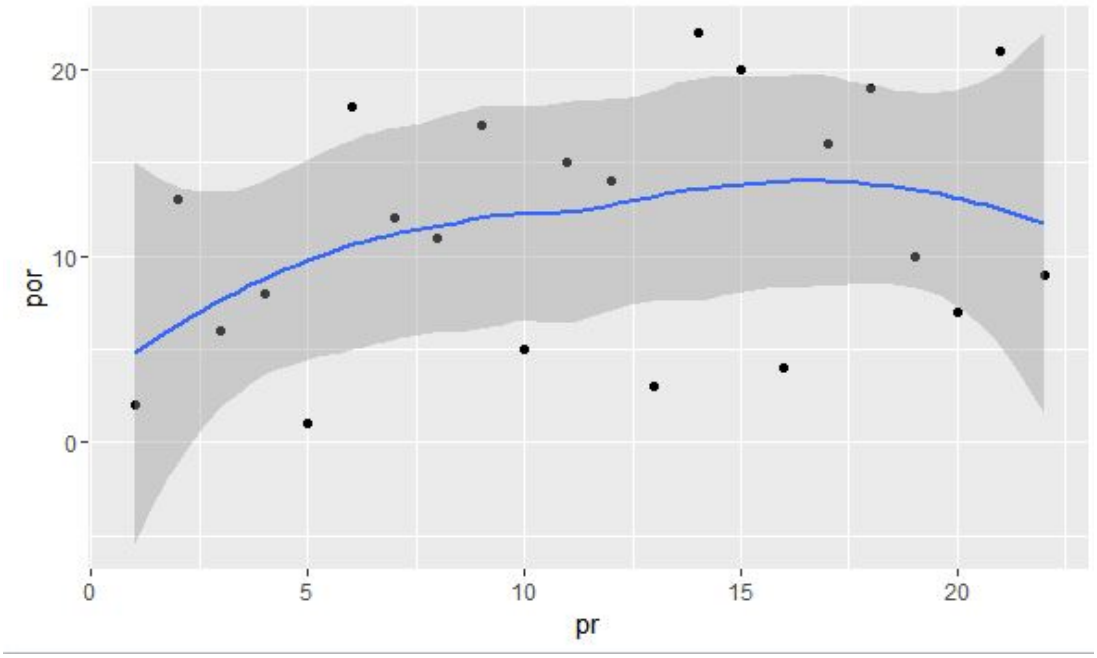


Figure 2: *plot of base model with population*

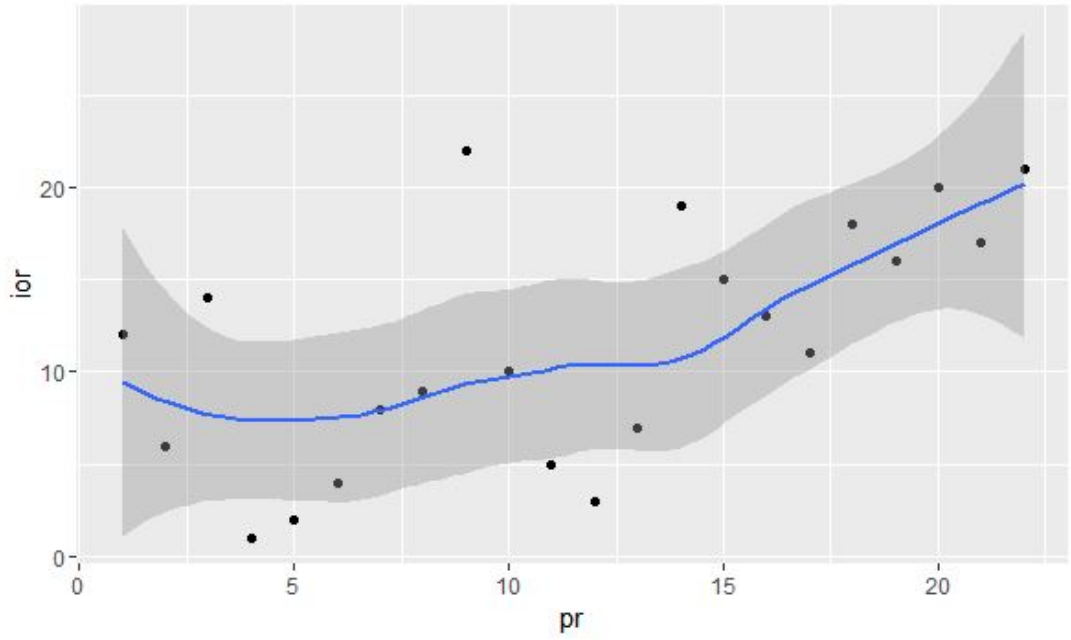


Figure 3: *plot of base model with median household income*

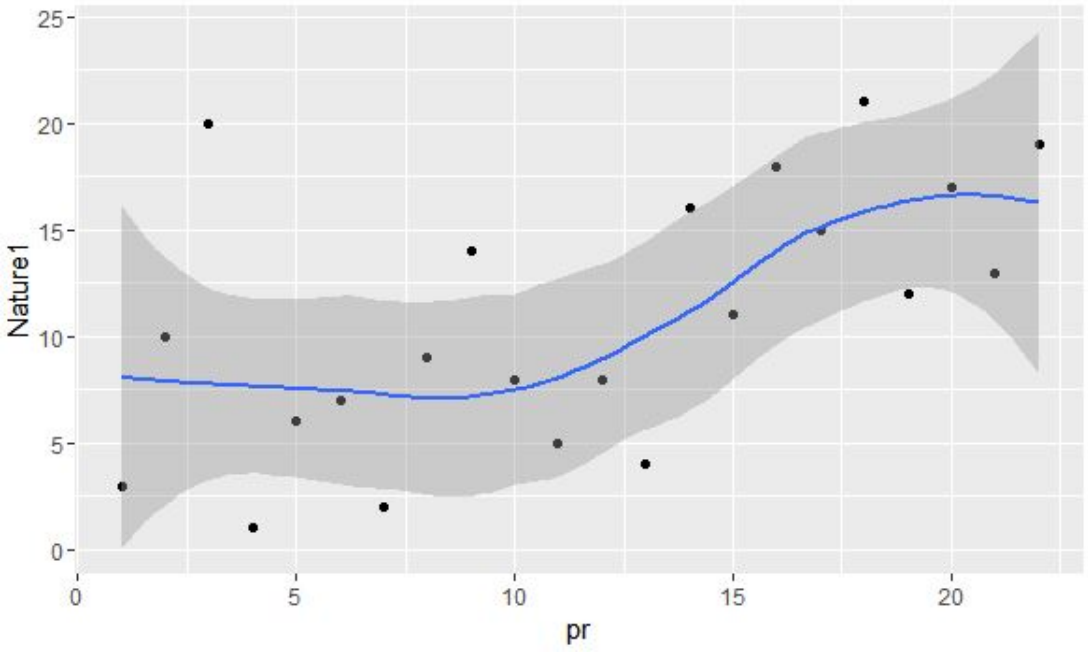


Figure 4: *plot of base model with mindset of people: Liberal or Conservative*

household income z=Nature of the people and A,B,C are the Pearson coefficients denoting the extent to which the factors are correlated to current air and train travel data. This metric "res" is used to rank the cities as a predicted market for hyperloop technology. Higher the metric, better is the candidate for Hyperloop technology.

V. CONCLUSION

The paper describes the importance and relevance of Hyperloop technology as a breakthrough in the travel industry in the coming years. It then identifies features like Distance between cities, Population, Median Household Income and mindset of the people: Liberal or Conservative which can affect the prediction of markets for Hyperloop technology. It then presents an analysis and shows that all these factors have a positive correlation with the base model(current means of transportation metric). However it was found out that some factors have a stronger correlation than others: Median household income has the strongest impact on the prediction followed by mindset of people and population. The Pearson coefficient is used as the parameter for each of the factors to calculate a metric used for ranking the cities. "New York" out-stands as the most favorable candidate for the hyperloop technology followed by Los Angeles and Chicago.

References

[Garber] Garber, Megan (July 13, 2012). "The Real iPod: Elon Musk's Wild Idea for a 'Jetson Tunnel' from S.F. to L.A.".

[SpaceX] Musk, Elon (August 12, 2013). "Hyperloop Alpha" (PDF). SpaceX.

[Elon] Beyond the hype of Hyperloop: An analysis of Elon Musk's proposed transit system". Gizmag.com. August 22, 2013.

[Hawkins] Hawkins, Andrew J. (January 30, 2016). "MIT wins SpaceXâĂŹs Hyperloop competition, and Elon Musk made a cameo". The Verge

1. New York, N.Y.
2. Los Angeles, Calif.
3. Chicago, Ill.
4. Houston, Tex.
5. Philadelphia, Pa.
6. Phoenix, Ariz.
7. Dallas, Tex.
8. San Francisco, Calif.
9. Cleveland, Ohio
10. Indianapolis, Ind.
11. Detroit, Mich.
12. Washington, D.C.
13. Seattle, Wash.
14. Denver, Colo.
15. Boston, Mass.
16. Kansas City, Mo.
17. Minneapolis, Minn.
18. Miami, Fla.
19. New Orleans, La.
20. St. Louis, Mo.
21. Pittsburgh, Pa.
22. Salt Lake City, Utah

Figure 5: *list of cities as a possible market for hyperloop technology*

Correlating Family Survey Data With Birth Health

Jesse Wisniewski

University of Colorado at Boulder
jesse.wisniewski@colorado.edu

Abstract

The health of newborn children has a significant effect on their health and quality of life as they further develop and grow. Finding additional factors that are correlated to the level of health at birth could lead to an increase in understanding of the related influences, which could then lead to the development of practices which improve overall health and well-being of the population as a whole even beyond birth. This paper uses birth health data from the CDC, measured by Apgar score and birth weight, and analyses family survey data from the US Census Bureau to look for such correlations. Chief findings include a significant degree of correlation between this survey data and the birth weight data for 2014, as well as a significant amount of positive spatial autocorrelation in birth weight among counties in the contiguous United States.

I. Introduction

Since the health of a child at birth has a direct influence on their future development, finding correlations between health at birth and other factors might prove useful in working to improve overall health.

Studying data on the health of newborn children from a geo-spatial perspective presents challenges, as privacy concerns lead to a lack of geo-location information tied to the data. The Center for Disease Control's VitalStats system lists birth data for the U.S. by county, but because of those privacy concerns it groups together all counties for each state which have a population less than or equal to 100,000. This means that in that data less than six hundred of the over three thousand U.S. counties have individual listings.

However, this also presents an opportunity to build a model that could be used to predict the data for the other counties, which could be tested by comparing the average against the value given per state for the grouped counties in that state.

II. Data

For this project, data on birth health for 2014 was downloaded in from the CDC's VitalStats system [CDC VitalStats, 2016]. The data included two birth health indicators: Apgar scores and birth weight. The Apgar score system rates the health of newborns using five categories - appearance, pulse, grimace, activity, respiration - each with a value from 0 to 2. The data listed number of births per county per Apgar score of 0-10, and number of births per county in ten ranges of 500g each from 0 - 5000g, and also any above 5000g.

All of this data was averaged for each county to give a single numeric value. For the Apgar scores, this just involved multiplying the birth count for each score by that score, summing those products, and dividing by the total birth count. For the birth weights, the center of each weight range was used as the multiplication factor.

In searching for data that might be correlated with this birth data, census data was first considered since it was easily available at the county level, simplifying direct comparison with the birth data. However, the US Census Bureau also conducts the American Commu-

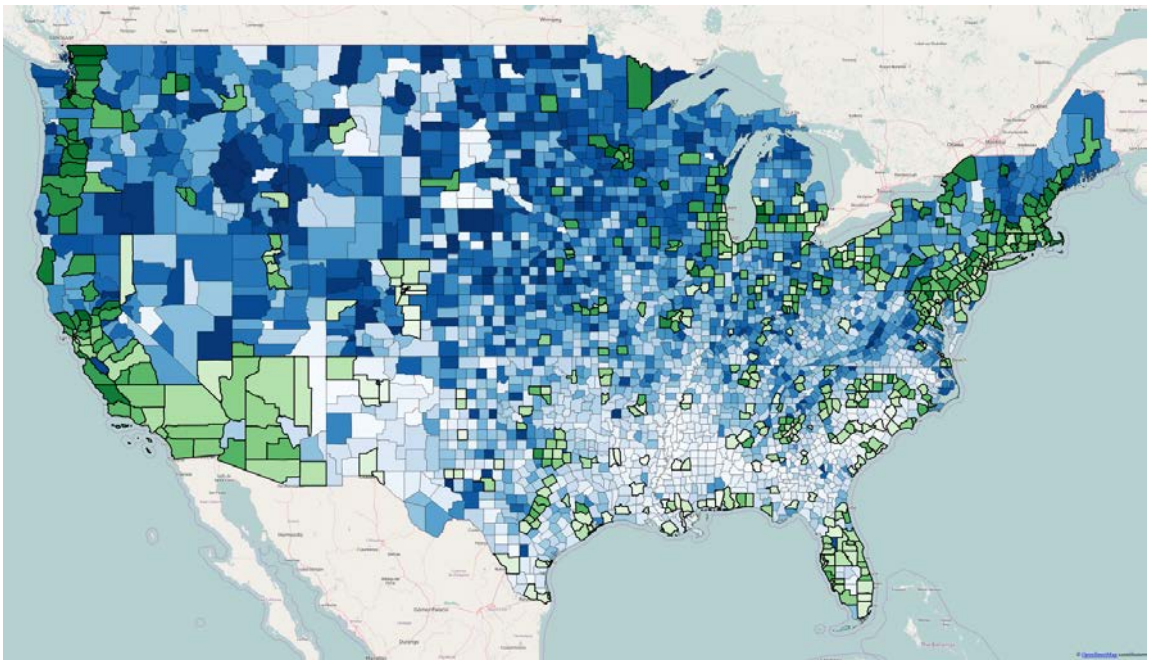


Figure 1: *2014 contiguous U.S. birth health by weight, both from original data (green/outlined) and modeled. Darker means higher values. [TIGER] [OSM]*

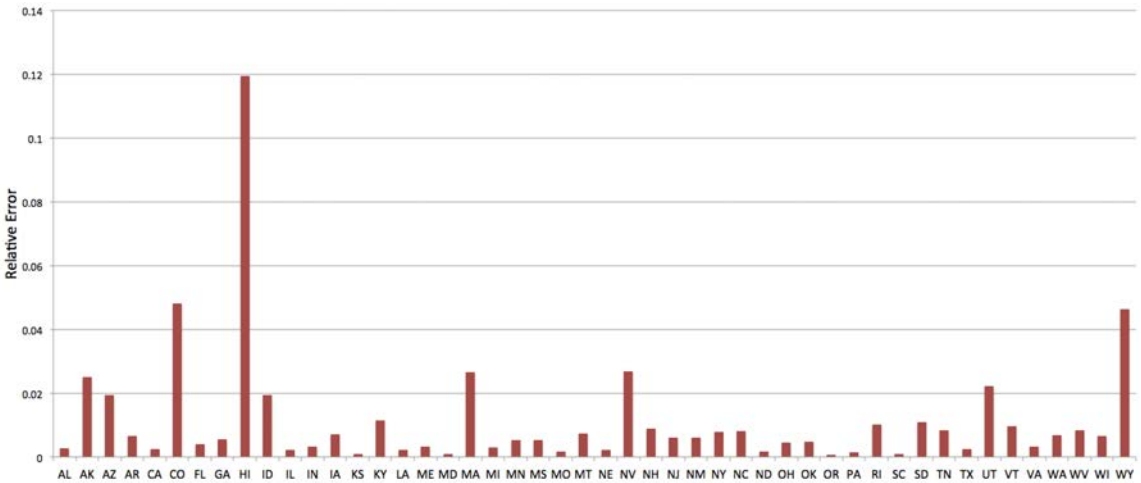


Figure 2: *Relative error, by state, of the estimated averages of remainder counties for each state.*

nity Survey, and that data is also available at the county level. Four tables of ACS data were ultimately chosen, that all concern families:

- Households and Families [S1101]
- Poverty Status Past 12 Months [S1702]
- Employment Characteristics [S2302]
- Financial Characteristics [S2503]

These tables were all available as one-year or five-year estimates. The one-year estimate tables were much less complete than the 5-year, with information for far fewer counties (819 vs 3144), and even for those counties the data was less complete. Therefore, the 5-year estimate tables were chosen for use. The four 5-year estimate tables listed above consisted of 193, 423, 99, and 303 fields of survey data, respectively, for each of 3144 counties.

This data was then merged with the birth health data based on the county, which only used a subset of the survey data as the birth data only contained values for 578 individual counties.

III. Methods

The goal of these analyses was to find the closest possible connections between the birth health data and the family survey data. The primary method chosen was the Pearson product-moment correlation coefficient, so both data sets had to be connected through numeric values. The county was the geographic connection, and so then the data then just had to be reduced to simple numeric values, which is why the birth health data was averaged by county as described above.

The merged data was stripped of any completely non-numeric variables, leaving 477 columns. At this point, correlation coefficients and p-values could be generated for all of the data. Only the correlations for the two variables, the Apgar scores and weight values, were needed, so the information for both of those were each extracted separately from the correlation data. Then a subset of each with p-values less than 0.05 was then taken, leaving 189 variables for the Apgar score and 407 for the weights. Then another subset was taken, for correlation coefficients greater than 0.3, leaving zero variables for the Apgar score and 127 for the weights.

At this point, the Apgar score data was no longer used as it seemed that there were no meaningful correlations between it and the survey data. However, there was some degree of linear correlation - there were 15 survey variables with linear correlation coefficients greater than 0.2 with p-values below 0.05. Since the Apgar score would seem to be a far better indication of birth health than simple birth weight, this research could be improved if other data was found that correlates more meaningfully with the Apgar scores.

In the absence of such data, the weight values became the focus of this analysis. The 127 variables that the correlation analysis had identified were analyzed for similarities using a correlation matrix, and 32 variables were selected as representative of the group. A linear regression analysis was then run on the weight values, using these 32 variables.

The model that this linear regression produced was then used to calculate estimated average birth weights for the additional 2566 individual counties. Additionally, the estimated values for each state were averaged and compared with the averaged for the remainder of each state in the original birth data.

IV. Results

The linear regression generated a linear model with multiple R-squared value of 0.6566. The details of the linear model, including intercept and coefficients for each factor, are detailed in Table 1, sorted by linear correlation coefficient with respect to birth weight (column "CC"), and using the factor descriptions as they are listed in the survey data. The linear model coefficient is listed in column "LMC". Even the largest p-value for these 32 variables was only 7.682E-14, so they are not shown.

Based on the survey data, the model was used to calculate estimated average birth weights for each county, which are displayed in Figure 1 Also, the relative errors for the av-

erage of the estimated average birth weights for the remainder of each state are displayed in Figure 2. These were calculated by comparing the averages of the estimated values with the averages for those areas from the original birth health data. Other than Hawaii with 12%, the greatest relative error was less than 5%.

V. Discussion

I. The Model

The r-squared value of 0.6566 for the linear model seems to indicate that while the model is not nearly perfectly accurate, it could still be meaningful. The fact that the estimated values generated by the model all lie in the range of real values of birth weights (5.36 - 8.45 pounds), and the low relative error in the remainder averages for each state, suggests that the model is indeed accurate.

However, that low relative error may have more to do with the nature of the data, with average birth weights lying in a narrow range. It would be more meaningful to have real data on other individual counties to compare to, possibly even by just excluding a few with a high and low values from the regression and then comparing the estimate from that new model with the real values of those counties.

I think the the remarkably higher relative error for Hawaii is worth noting. Although Hawaii is obviously geographically separate from the rest, the error for Alaska is quite low. Also, aside from Hawaii, the relative error for Colorado and Wyoming is nearly double the remaining states. Apparently there is something about these states which causes the model to not work as well there. It is especially interesting how Colorado and Wyoming have nearly identical error, as they are adjacent states. Perhaps there is some amount of spatial auto-correlation in the relative error of this model.

Looking at the map of estimated values, there is a clear pattern of lower values in the southeastern area of the country, and higher values throughout the central and northeastern areas. The values are more erratic in the western half of the country, and then more consistently higher on the west coast. Depending on the accuracy of the model,it could have importance in revealing these spatial trends.

To quantify the spatial clustering of the data, a Moran's I test was performed on the counties in the contiguous U.S., resulting in a value of 0.4919. This indicates a nearly equal amount of both randomness and order in the estimated data. If that is truly indicative of the nature of birth health, it would limit the possible accuracy of any estimation model.

II. The Factors

Most of the factors included in the model seem to validate traditional family values. For example, family size seems to be important to the model as a positive factor. Income and poverty also involved, as would be expected, as positive and negative factors. Educational attainment and employment factors are also positively included, pointing to the value of both. Two of the most strongly negatively correlated factors - #29 and #30 - point to single mothers as being at a disadvantage, which is also intuitive.

However, the strongest positive correlation (factor #1) has to do with single fathers, which has a less clear connection. Perhaps more single fathers means fewer births, and the nature of the connection has to do with that.

The real importance of this analysis seems to lie in how it indicates that these kinds of family factors are indeed connected with infant health. While that fact would seem to be intuitive at a high level, it may be less considered in practice, where it might easily be forgotten in the business of everyday life. So having a statistical connection could help reinforce the importance that these family factors have in infant health.

Table 1: *Coefficients of the family survey factors in the linear model for birth weight.*

#	Survey Data Factor	CC	LMC
	Intercept		3427.55
1	Families with own children under 18 years, Estimate, Other families - Male householder, no wife present	0.53	1.57135
2	Total, Estimate, WORK STATUS CHARACTERISTICS - Families - 2 or more workers in the past 12 months	0.434	-2.09429
3	Owner occupied housing units, Estimate, HOUSEHOLD INCOME IN THE PAST 12 MONTHS, IN 2014 INFLATION ADJUSTED DOLLARS - , 75 000 to, 99 999	0.418	6.89214
4	Occupied housing units, Estimate, MONTHLY HOUSING COSTS AS A PERCENTAGE OF HOUSEHOLD INCOME IN THE PAST 12 MONTHS - 75 000 or more - 20 to 29 percent	0.351	3.94975
5	Total, Estimate, HOUSING TENURE - Owner occupied housing units	0.337	-0.337445
6	Owner occupied housing units, Estimate, MONTHLY HOUSING COSTS - 1 500 to, 1 999	0.332	-0.342891
7	Total, Estimate, Married couple families - Householder worked part time or part year in the past 12 months - Spouse worked part time or part year in the past 12 months	0.332	6.29061
8	Female householder, no husband present - Percent below poverty level, Estimate, EDUCATIONAL ATTAINMENT OF HOUSEHOLDER - Less than high school graduate	-0.304	-0.188897
9	Occupied housing units, Estimate, MONTHLY HOUSING COSTS - No cash rent	-0.324	-2.07183
10	All families - Percent below poverty level, Estimate, EDUCATIONAL ATTAINMENT OF HOUSEHOLDER - Bachelor s degree or higher	-0.345	-1.30761
11	Total, Estimate, Other families - Female householder, no husband present - Not in labor force	-0.349	0.632702
12	Owner occupied housing units, Estimate, MONTHLY HOUSING COSTS AS A PERCENTAGE OF HOUSEHOLD INCOME IN THE PAST 12 MONTHS - Less than, 20 000 - Less than 20 percent	-0.349	-2.10995
13	Female householder, no husband present, family household, Estimate, AGE OF OWN CHILDREN - Households with own children under 18 years - Under 6 years and 6 to 17 years	-0.351	0.251495
14	Female householder, no husband present - Percent below poverty level, Estimate, EDUCATIONAL ATTAINMENT OF HOUSEHOLDER - High school graduate, includes equivalency	-0.362	0.302561
15	Married couple families - Percent below poverty level, Estimate, Families	-0.37	16.8791

Continued on next page

#	Survey Data Factor	CC	LMC
16	All families - Total, Estimate, INCOME DEFICIT - Mean income deficit for families, dollars	-0.372	-0.00566035
17	All families - Total, Estimate, RACE AND HISPANIC OR LATINO ORIGIN - Families with a householder who is - , One race - Black or African American	-0.372	-0.000153442
18	Female householder, no husband present - Percent below poverty level, Estimate, Families	-0.385	4.51834
19	All families - Percent below poverty level, Estimate, NUMBER OF PEOPLE IN FAMILY - 7 or more people	-0.393	0.0259567
20	Owner occupied housing units, Estimate, HOUSEHOLD INCOME IN THE PAST 12 MONTHS, IN 2014 INFLATION ADJUSTED DOLLARS - , 10 000 to, 14 999	-0.415	-2.82697
21	Married couple families - Percent below poverty level, Estimate, NUMBER OF WORKERS IN FAMILY - No workers	-0.446	0.376787
22	Female householder, no husband present - Percent below poverty level, Estimate, Householder worked - Householder worked full time, year round in the past 12 months	-0.448	-0.960013
23	Occupied housing units, Estimate, MONTHLY HOUSING COSTS AS A PERCENTAGE OF HOUSEHOLD INCOME IN THE PAST 12 MONTHS - Less than, 20 000	-0.452	-0.0446492
24	Female householder, no husband present - Percent below poverty level, Estimate, Householder 65 years and over	-0.458	-0.0717444
25	Owner occupied housing units, Estimate, MONTHLY HOUSING COSTS AS A PERCENTAGE OF HOUSEHOLD INCOME IN THE PAST 12 MONTHS - Zero or negative income	-0.459	-6.56193
26	All families - Total, Estimate, PERCENT IMPUTED - Poverty status for families	-0.462	-0.813742
27	Occupied housing units, Estimate, HOUSEHOLD INCOME IN THE PAST 12 MONTHS, IN 2014 INFLATION ADJUSTED DOLLARS - Less than, 5 000	-0.501	0.637299
28	All families - Percent below poverty level, Estimate, EDUCATIONAL ATTAINMENT OF HOUSEHOLDER - Less than high school graduate	-0.509	-0.376005
29	Female householder, no husband present, family household, Estimate, FAMILIES - Average family size	-0.511	-43.5134
30	Total, Estimate, Other families - Female householder, no husband present	-0.548	1.16379
31	All families - Percent below poverty level, Estimate, Families	-0.564	-16.1921
32	Total, Estimate, WORK STATUS CHARACTERISTICS - Families - 1 worker in the past 12 months	-0.615	-3.11092

Concluded

References

[CDC VitalStats, 2016] Centers for Disease Control and Prevention. National Center for Health Statistics. VitalStats. http://www.cdc.gov/nchs/vitalstats.htm. [April 2016].

[S1101] United States. Census Bureau. American Community Survey, 2014 American Community Survey 5-Year Estimates, Table S1101; generated by Jesse Wisniewski; using American FactFinder; <http://factfinder2.census.gov>; (25 April 2016).

[S1702] United States. Census Bureau. American Community Survey, 2014 American Community Survey 5-Year Estimates, Table S1702; generated by Jesse Wisniewski; using American FactFinder; <http://factfinder2.census.gov>; (25 April 2016).

[S2302] United States. Census Bureau. American Community Survey, 2014 American Community Survey 5-Year Estimates, Table S2302; generated by Jesse Wisniewski; using American FactFinder; <http://factfinder2.census.gov>; (25 April 2016).

[S2503] United States. Census Bureau. American Community Survey, 2014 American Community Survey 5-Year Estimates, Table S2503; generated by Jesse Wisniewski; using American FactFinder; <http://factfinder2.census.gov>; (25 April 2016).

[TIGER] United States. Census Bureau. TIGER Products; Cartographic Boundary Shapefiles - Counties; <https://www.census.gov/geo/maps-data/data/cbf/cbf_counties.html>

[OSM] OpenStreetMap base map, ©OpenStreetMap contributors <http://www.openstreetmap.org/copyright>

Interactive Visualization of Large Geospatial Point Data

Jonathan Blackstock
University of Colorado, Boulder
jobl5669@colorado.edu
26 April 2016

Interactive Visualization of Large Geospatial Point Data

Abstract

Due to the increasing size of geospatial dataset, it is becoming much more difficult to visualize the data using traditional web mapping systems. These are generally limited in the number of data points or the data points are clustered together, which abstracts away the detail of a single point. This paper explores the use of vector tiles in an attempt to create a detailed, fast, and interactive map of Colorado voter registration data. After many attempts to achieve this, we were able to generate a map of nearly 225,000 points with the ability to filter the data based on characteristics such as the voters age or party affiliation.

I. Introduction

As the volume of geospatial data has grown in the last decade, visualizing complete datasets on a map has become increasingly difficult, driving the need for advanced and interactive visualization methods. More traditional mapping systems like Google Maps[1] or Google Earth[2] are limited in the number of points they are able to display and often become very sluggish as the data size increases beyond 10,000 points, leading to a deteriorated user experience. According to Google's documentation, up to 100,000 points can be loaded into Google Maps using Fusion Tables, although this was not tested in this study since it was still not large enough. Previous research has been performed in geospatial data visualization, but only considered datasets with up to 5,000 data points[5]. For this study, data was both clustered by distance based on zoom level and displayed from pre-generated vector tiles. This allowed the data to remain interactive, both by clicking on individual data points as well as filtering the data in real time. The tool that appeared best suited for the task was Mapbox[3].

II. Data

For this project, Colorado voter registration data from 1 February 2016 was used to test the limits of the visualization. The data set was available in 14 compressed files containing CSVs with about 3.6 million total records. The data is all public record, and was re-hosted for free at www.coloradovoters.info. The dataset is updated monthly. The data did not contain any coordinates, so the addresses had to be geocoded. Apartment numbers were removed from the addresses, leaving about 1.6 million unique addresses. This also took into account multiple voters registered at the same address. MapQuest's API[4] was selected to perform the geocoding due to its relatively fast responses, ability to batch process requests, and it's apparent lack of quotas. However, due to the sheer volume of addresses and the limited time allowed, a random sample of about 225,000 addresses were geocoded and used in this project. The results were appended to the CSV files using Python. Irrelevant columns were removed and the CSV files were converted to GeoJSON via a VRT file using GDAL's ogr2ogr command. Finally, these GeoJSON files were converted

into vector tiles using Tippecanoe.

III. Methods

Mapbox was selected to display the data and form the basis of the interactive webpage due to it's use of vector tiles and extensive API. The voter registration vector tiles were loaded into Mapbox Studio as a new tileset. Figure 1 shows the vector layer along with all of the properties associated with each data point, such as name, voter identification number, address, city, state, zip code, phone number, birth year, party affiliation, and voter status. The map was created entirely with JavaScript in the client's browser using the Mapbox GL JS API.

Initially, the data was clustered before being displayed, as shown in Figure 2. The map was very responsive but only allowed the user to click on a single data point once the map was zoomed in sufficiently far for the data to be un-clustered. Additionally, the Mapbox GL JS API did not provide a way to implement dynamic filtering of the data based on each point's properties. Due to these limitations, clustering was abandoned in favor of displaying every data point outright. Given that this map layer was already a vector tile, this was a feasible solution. The map would automatically display more data points as the user zoomed in. Filtering input was provided by client-side JavaScript/jQuery which dynamically generated Mapbox filters based on the properties of each data point. The code was all written in a single, stand-alone HTML page.

IV. Results

Using Mapbox Studio and the Mapbox GL JS API, we were able to successfully display and dynamically filter nearly 225,000 interactive data points at once. The filter (Figure 3) was built from simple input fields and is extremely responsive despite the large data set. In the case of the birth year range selector for example, the data would filter as the slider was still moving.

Pop-ups were created to provide additional information to the user including name, address, phone number, party affiliation, and voter status as shown in Figure 4. These too responded immediately on click.

It is also worth noting that the process of geocoding the addresses through the MapQuest geocoding API did have some noticeable errors. For example, seven points in Buena Vista, CO geocoded to the middle of Kansas. Given that the purpose of this project was to push the ability to visualize a large amount of point data, the geocoding errors had no effect on the results.

Figure 5 shows a wider view of the greater Denver-Boulder-Fort Collins area broken into three sub-images. The first is filtered to show only voters registered as Democrats, the second as Republicans, and the third as Unaffiliated. Although this is a subset of the full data set, the map provides a very fast qualitative glimpse of how voters of various affiliations are distributed throughout the state, while simultaneously providing information down to a specific voter.

V. Conclusion and Future Work

By leveraging vector tiles and the Mapbox API, we were able to successfully create a filterable, interactive map of nearly 225,000 points that could both show large scale distributions as well as information about a specific data point. While we used voter registration data for this particular project, the applications for mapping large volumes of point data go far beyond this.

Due to the limited amount of time to conduct this work, we were not able to accomplish everything that we set out to do. Later work will include increasing the number of data points displayed at once, adding the ability to search the data by name, and revisit clustering the data with the filters present. Additionally, we would like to provide the ability for the user to select if the data is viewed as clusters, a heatmap, or individual data points.

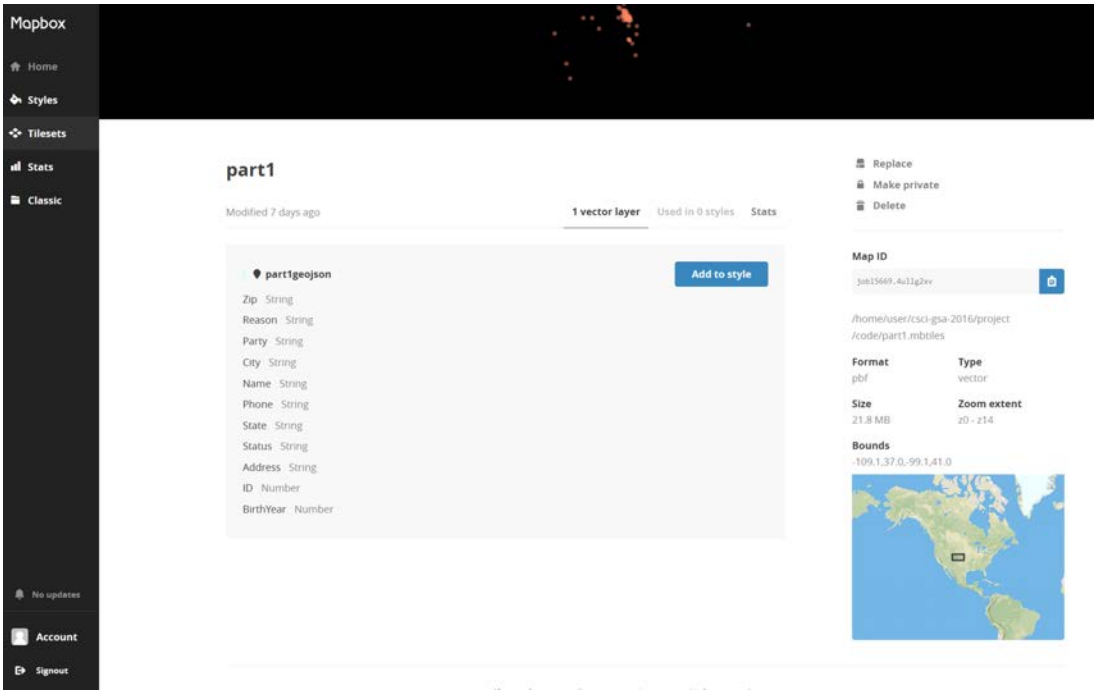


Figure 1: *Mapbox studio view of the vector layer containing a subset of the voter registration data*

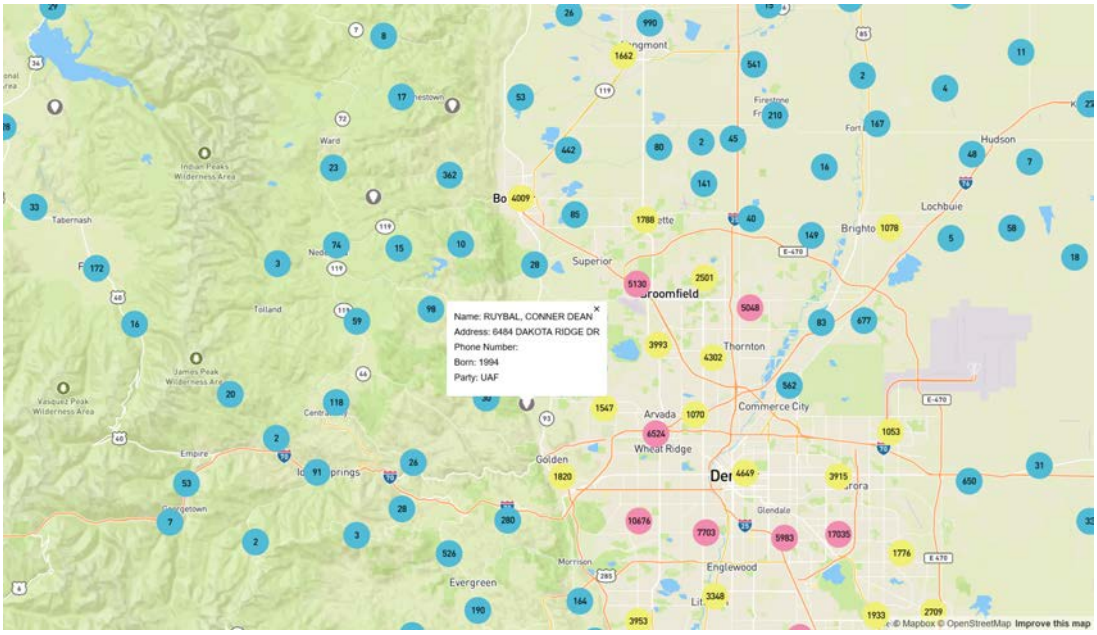


Figure 2: *Clustered Data with Pop-Up*

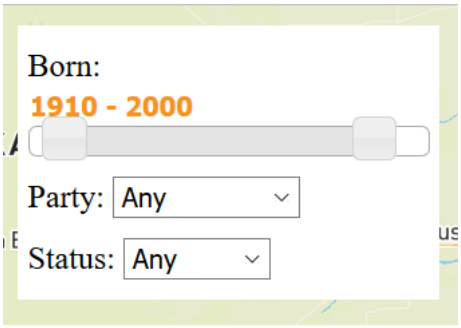


Figure 3: *Available filters are the year the voter was born, the voter's party affiliation (Republican, Democrat, or Unaffiliated), and their status (Active or Inactive)*

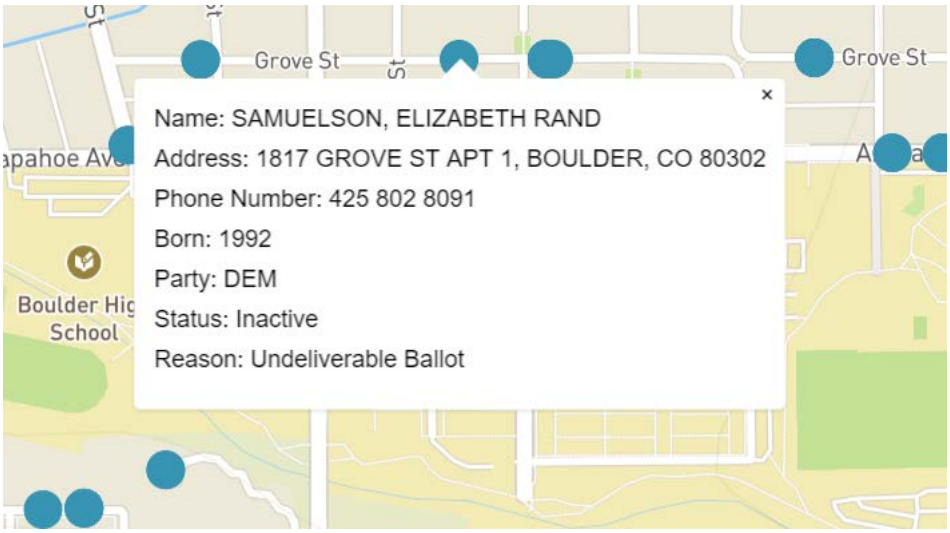


Figure 4: *Pop-Up presented to the user when they click on a point. If the status is 'Inactive' it also includes the reason why they are inactive*

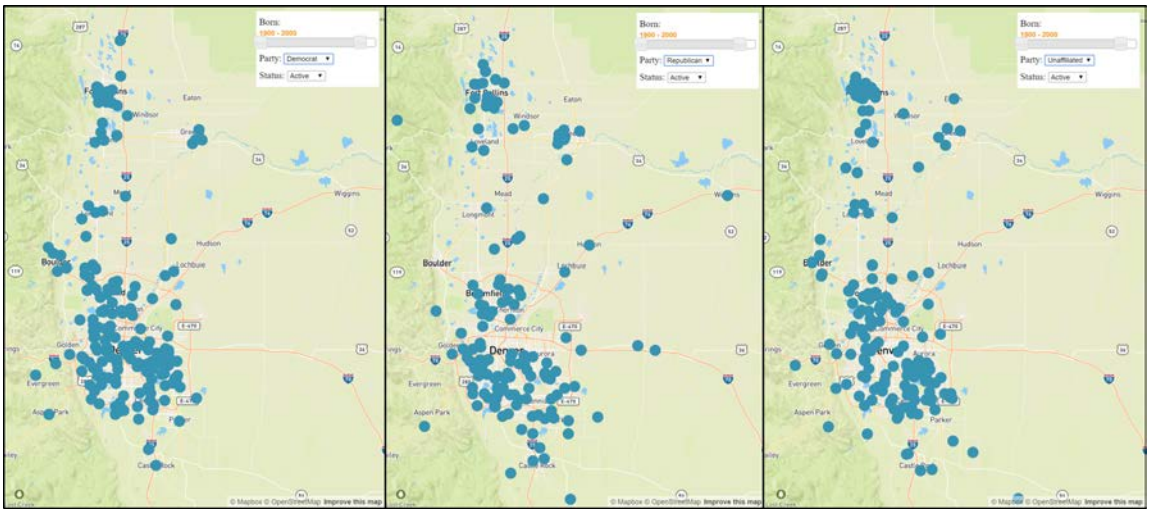


Figure 5: *Comparative map showing the locations of voters identified as Republicans, Democrats, and Unaffiliated*

References

[1] https://www.google.com/maps

[2] https://www.google.com/earth/

[3] https://www.mapbox.com/

[4] https://developer.mapquest.com/products/geocoding

[5] Delort, Jean-Yves. "Vizualizing large spatial datasets in interactive maps." Advanced Geographic Information Systems, Applications, and Services (GEOPROCESSING), 2010 Second International Conference on. IEEE, 2010.

Crime Trend Analysis

Keerthi Chikalbettu Pai
University of Colorado, Boulder
Keerthi.ChikalbettuPai@colorado.edu

Abstract

In studying crime, most research is concentrated on demographics such as age, sex, race, with few papers on the influence of physical environment like weather, movies, hate speech, natural disasters, on criminal behavior. This paper examines the influence of both socio-demographic variables and physical environment on criminal behavior in various cities in USA. The objective is to answer some of the questions with respect to crime rates such as, how do crimes vary with weather? Is there any pattern across various cities? Is there an increase in crime rate when a hate speech is delivered? The study finds that most types of crime vary with weather attributes such as temperature and dew point, this pattern is not found in certain types of crime such as auto-theft and homicides. Furthermore, based on the demographics of a city, high profile events such as the Ferguson shooting has an impact on crime. The results from the study is used to help map safe-neighborhoods in various cities.

I. Introduction

In this paper, we analyze the evolution of crime over the years, and examine the impact of external factors that have affected crime across various cities in the last fifteen years. We map how crime varies with weather attributes such as temperature, dew points, and events such as a terrorist attack or a natural calamity influenced crime. The paper helps us understand with greater certainty the degree to which weather influences crime, and how influential an event is. The analyzed results are used to map crime hot spots and provides guidance in picking safe neighborhoods. The motivation behind this project is to find if there is any trigger that has resulted in the increase or decrease in crime and how crime has evolved in neighborhood across various cities. The data is analyzed using R and represented using R web application framework shiny[1].

There are many existing studies analyzing how the crime trends have been influenced, individually by each of the external factors such as natural calamities, movies, weather, National Football League[2] and economy. Zahran, Shelley, Peek and Brody[3] found that natural calamities in Florida tend to increase the number of domestic violence crimes in the state, but decreases the rates of property and violent crimes. Dahl and Vigna[4] conducted laboratory experiments in psychology to identify the relationship between movie violence and violent crime rates and observed, surprisingly, that the rate of violent crimes decreased on days when there were large number of viewers for violent movies. Zanten[5] has compared crime reports available from 2001 with weather data for each day and found a pattern in the crime trends in accordance with temperature. This comparison was done for the city of Chicago. A report prepared by United Nations Office on Drugs and Crime[6] details the crime types in twelve different countries which are the most affected by economic factors. The report also lists the economic indicators which predict the change in these crime rates. Andrienko[7] studied the change in crime rates with the economic changes in Russia. He observed that the number of violent crimes increased with the increase in unemployment rate, but with a stable economy, there were more property crimes than violent ones. Our study analyses how movies, weather, and other events such as football or acts of terror have influenced crime in major cities in the United States.

II. Datasets

The crime data used in this research was collected from the US City Open Data Census[8], a crowdsourced publicly available database for various municipalities. It is maintained by Code for America, Sunlight Foundation, and Open Knowledge Foundation staff members, Code for America Brigade, and the Open Government Data working group with contributions from many members of the wider community. The weather data was provided by the Weather Underground historic weather database[9] and the Wiki list of Events[10] provided the data for various events and movies.

Of the available datasets, based on how reliable and complete the datasets were, containing data for at least five years and ensuring the cities were spread across the US, the following are used in this paper:

- Atlanta Police Department Open Data, year includes 2009 to 2016[11].
- Baltimore Police Department Open Data, year includes 2011 to 2016[12].

- Baton Rouge Crime Incidents, year includes 2011 to 2016[13].
- City of Chicago – Crimes Data Portal, year includes 2001 to 2016[14].
- City and County of Denver Crime, year includes 2010 to 2015[15].
- City of Philadelphia Crime Incidents, years include 2006 to 2016[17]
- San Francisco Police Department Incidents, year includes 2003 to 2016[16].

There is no standard way to report crime, and one of the first steps was to make the various datasets uniform. This required going through the datasets, identifying the various types of crimes, selecting a set of crimes to be worked in this project and categorizing the crimes in the datasets to the set considered for the project. The set of crimes considered for this project include assault, larceny, sex offense, disturbing the peace, narcotics, auto theft, weapons, trespass, robbery, burglary, kidnapping, fraud, homicide, arson, gambling, human trafficking, and domestic violence.

III. Methods

The first step after the data pre-processing was looking at how crime has changed over the years and over the months for a given year in various cities. This would give a general pattern as to how crime has evolved in a city over the years. Second step was to look at the crime types individually across cities and how they evolve with time and how a neighborhood crime type changes. This information can be used to created heat maps which helped identify crime hotspots in the city and the neighborhood with the highest crime. The third step was to identify and reason out why certain patterns occurred or if there are any anomalies in the pattern. Further, on mapping certain events like the Ferguson shooting, the Cold Wave that ripped Atlanta in 2014, the 9/11 attack, we were able to check if there are any external factors that affect crime. We expect the weather and the shooting incident to impact crime negatively. The last step was to build an interactive application using shiny which can show the above results and can be played around with.

IV. Results

Few interesting findings so far are:

1. Crime in general has decreased over the years in most cities with the exception of Denver. The following plots help understand this.

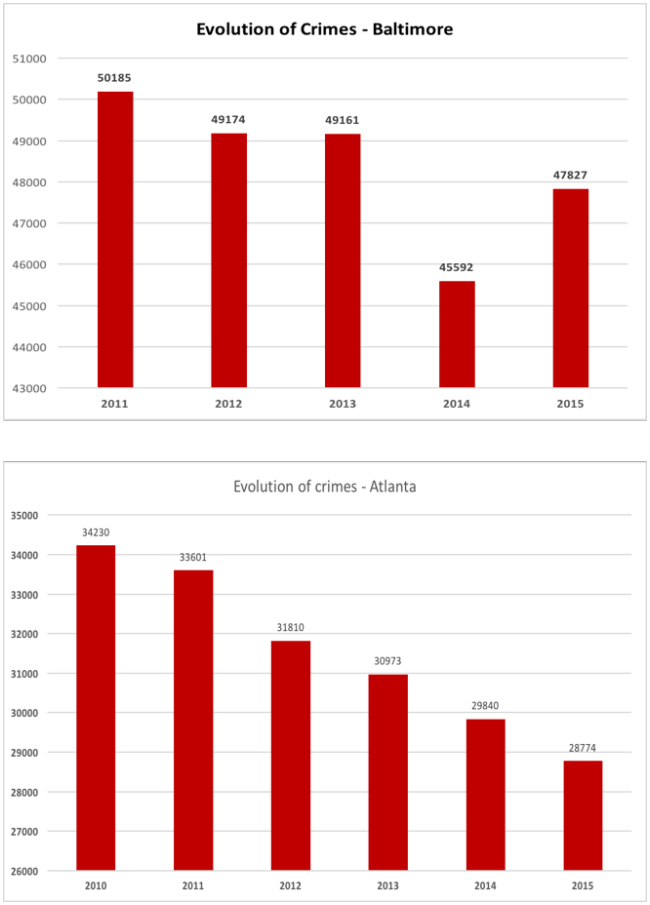


Figure 1: *Crime Over the Years in Baltimore and Atlanta*

2. The interesting factor that emerged while mapping crimes across cities is the dip in crime activity in the month of February as seen in *Fig*.2, *Fig*.3, *Fig*.4 and *Fig*.5.

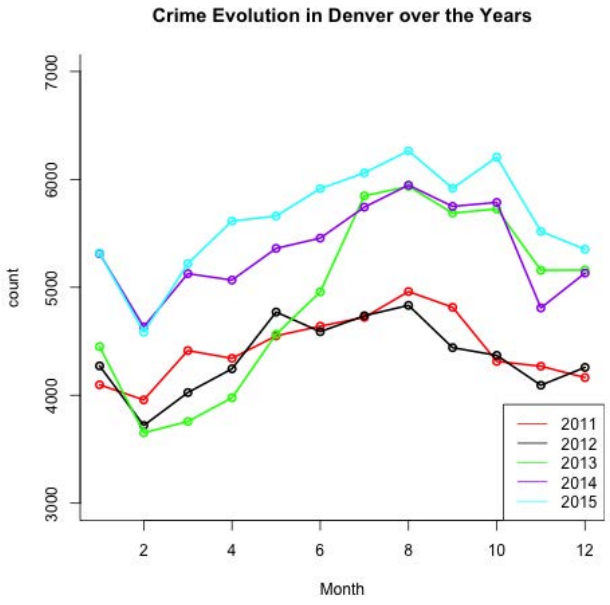


Figure 2: *Crime Over the years in Denver*

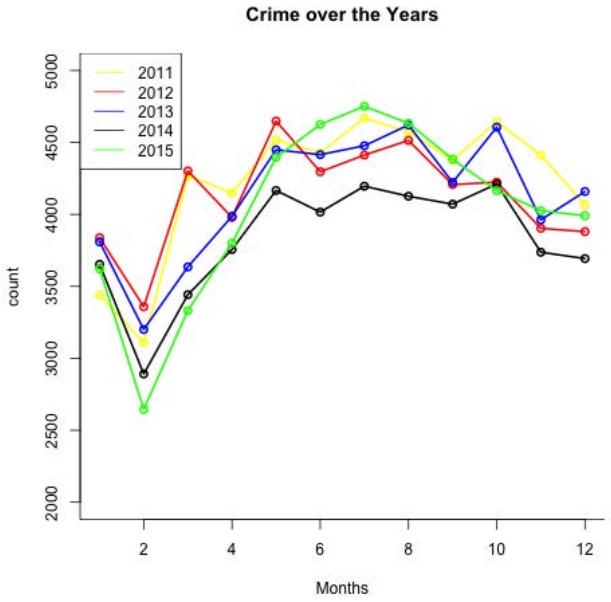


Figure 3: *Crime Over the years in Baltimore*

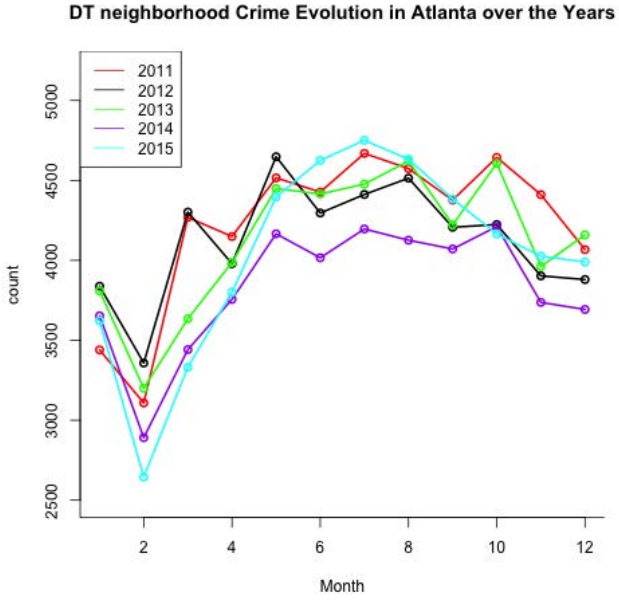


Figure 4: *Crime Over the years in Atlanta*

3. What is interesting about the Denver crime rise is the big jump in Narcotics related crime. Knowing that in 2014 Colorado legalized marijuana the basic assumption is that there should have been a decrease in drug-related crime. The data states otherwise. This may be due a very high tax rates in Colorado, and it might be cheaper to buy drugs illegally. What is also interesting is there is a steep rise in drug related activity after the 4/20 event in 2013 which lead to the death of two people in a gang related shooting. On further analyzing this anomaly, we found that Denver Police started reporting crime statistics in a different way since 2013[18], which caused the sudden rise as shown in *Fig*.6

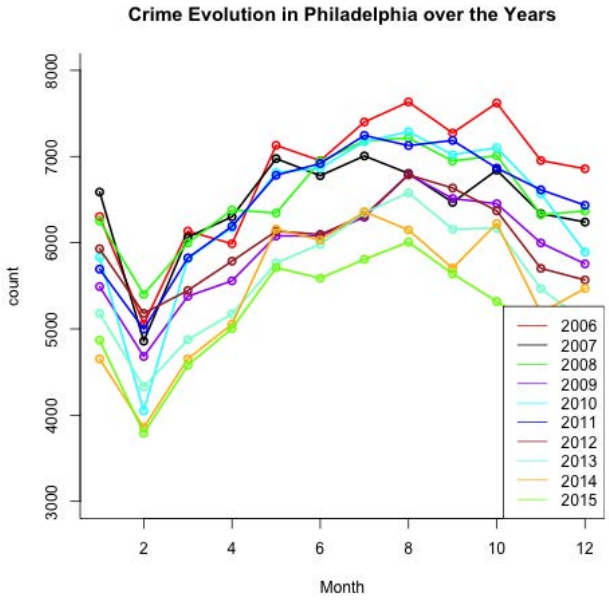


Figure 5: *Crime Over the years in Philadelphia*

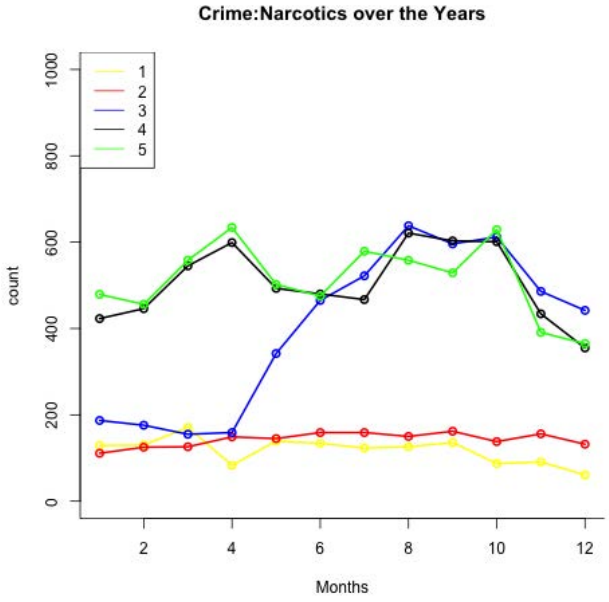


Figure 6: *Narcotics related Crime in Denver from 2011-2015*

4. Further analyzing the crime in the month of February, a similar but not a strong pattern emerged. The crime decreases on two occasions; First, during a Superbowl event, and second around Valentines day. Other than February being the coldest month, the assumption for the decrease during Superbowl event maybe because a lot of people like to gamble and travel to places where gambling is legal

in the Unites States. *Fig*.7 and *Fig*.8 further highlights this factor.

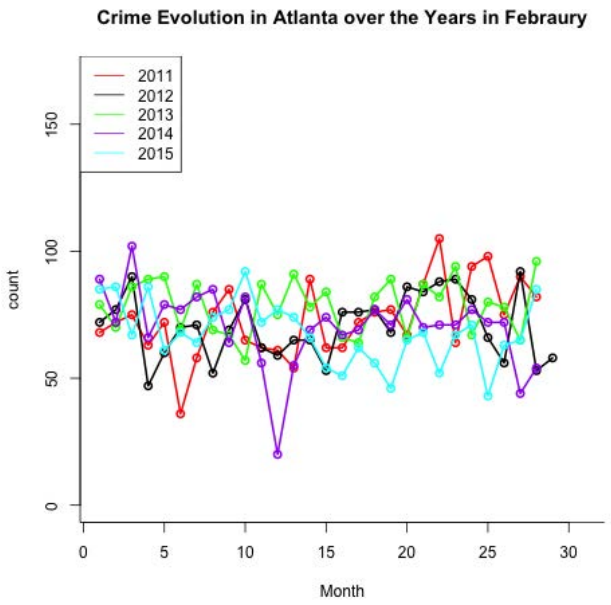


Figure 7: *Crime in the month of February in Atlanta*

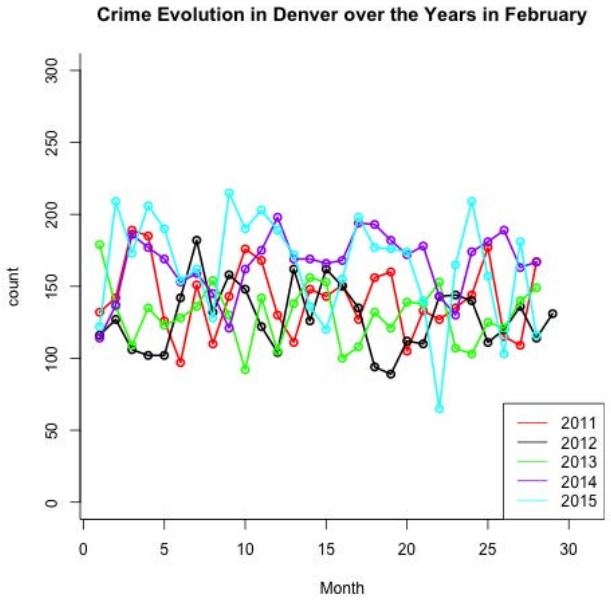


Figure 8: *Crime in the month of February in Denver*

5. Another finding from *Fig*.7 is how crime varies with weather conditions. The sudden drop in crime in 2014 on $13^{th} February$ is due to a snow storm that hit Georgia between $11 - 13 February$. The same is also highlighted in *Fig*.2, *Fig*.3, *Fig*.4, and *Fig*.5, where in overall crime begins to increase as Summer starts, hits a peak during Fall and reduces during Winter.

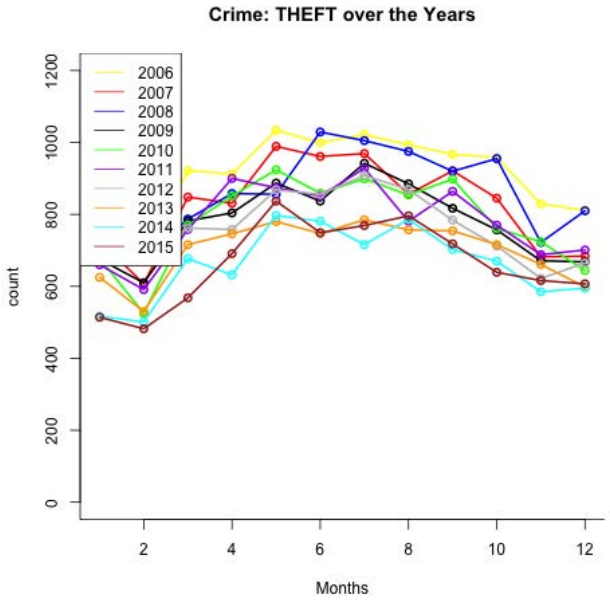


Figure 9: *Crime:Assault over the years in Philadelphia*

6. Next we checked if various types of crime vary with increase as Summer starts, hits a peak during Fall and reduces during Winter. Two interesting patterns emerged in this analysis. Crimes like assault, arson varied as the year progressed, *Fig*.9, and crimes like auto-theft which remained constant throughout the year, *Fig*.10.

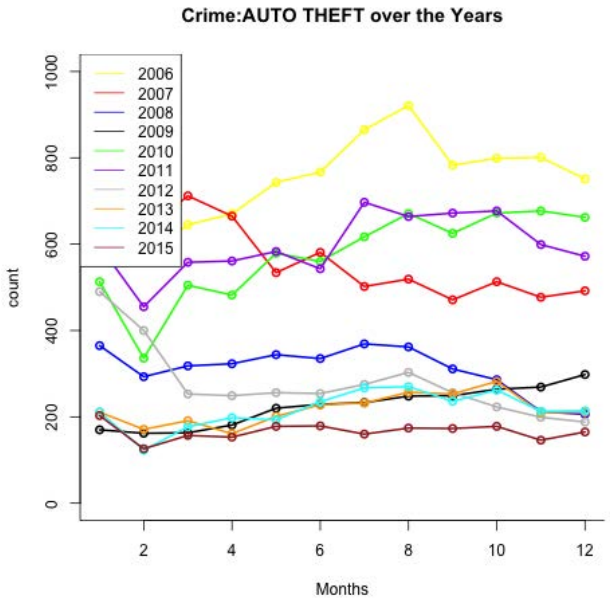


Figure 10: *Crime: Auto Theft over the years in Philadelphia*

7. The correlation between crime and weather attributes such as temperature and dew point had a positive correlation as seen in *Fig*.11 and *Fig*.12.

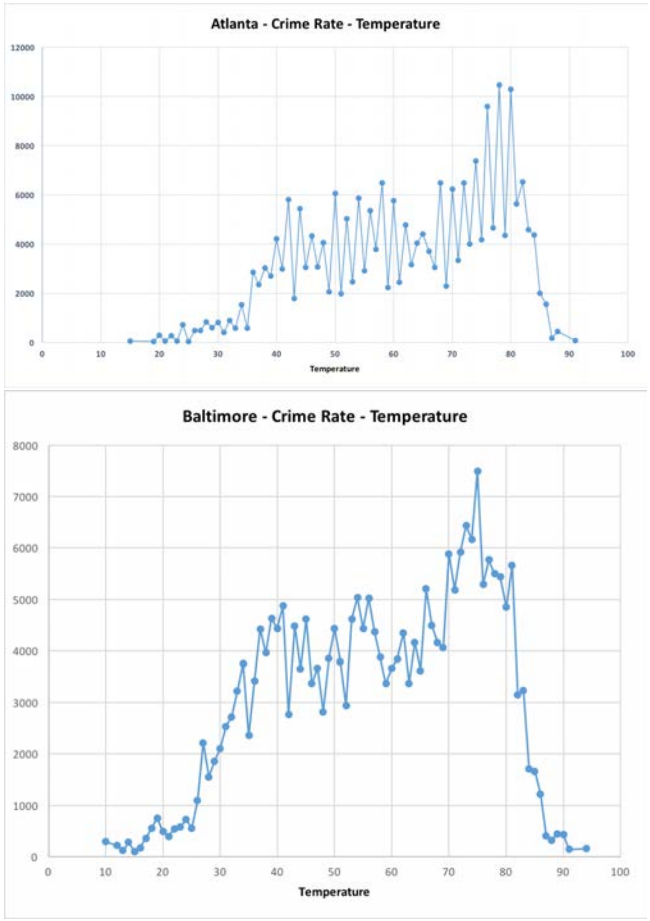


Figure 11: *Crime Rate versus Temperature for Atlanta and Baltimore*

8. Next aim was to check for any relation between the occurrence of certain events and the crimes. To perform this analysis, we picked the following events that could have had a potential effect on the crime numbers: Presidential election in 2012, terrorist attack on the twin towers in 2001, Boston Marathon bombing in 2013, Ferguson shooting incident in 2014, release of violent movies such as Django Unchained and The Expendables and natural calamities like Hurricane Sandy and the North American Cold Storm in February 2014. While we could not see any major effect on crime numbers for most of these events, there was a definite impact of the Ferguson incident on crime numbers in Baltimore, and Atlanta which has predominantly African-American population, as can be seen in *Fig*13 and *Fig*.14. *Fig*.13 and *Fig*.14 shows the plot of crime numbers in August over the years in Atlanta and the crime numbers in August and September 2014 for Baltimore. We can see a marked increase in the number of crimes on August 9, 2014 followed by a steep decline. This could be due to the protest immediately following the shooting of Michael Brown in Ferguson, Missouri.

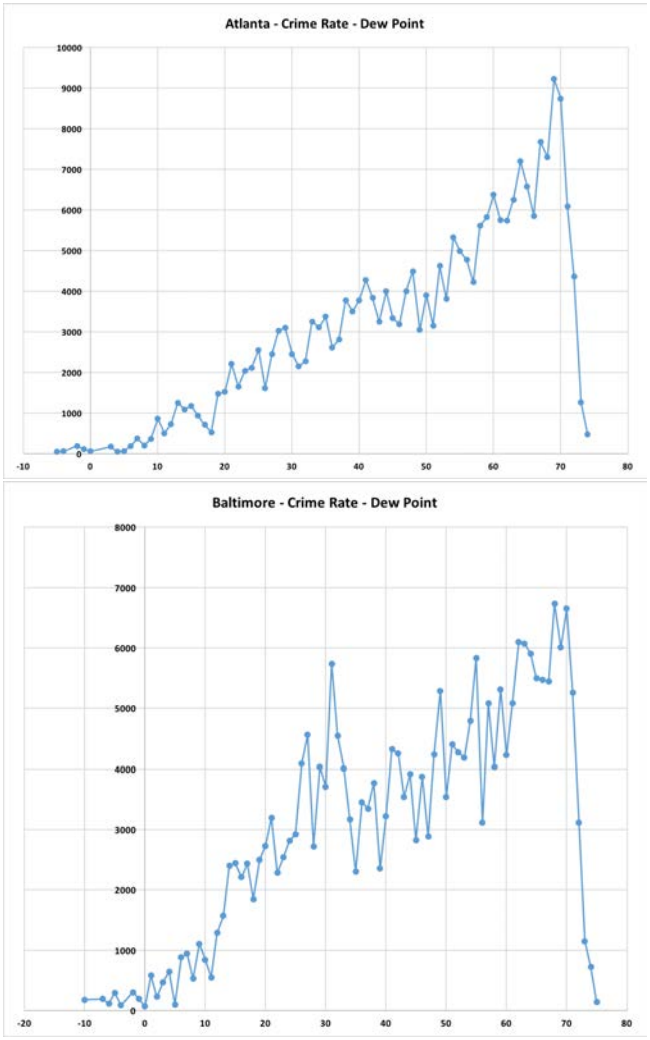


Figure 12: *Crime Rate versus Dew Point for Atlanta and Baltimore*

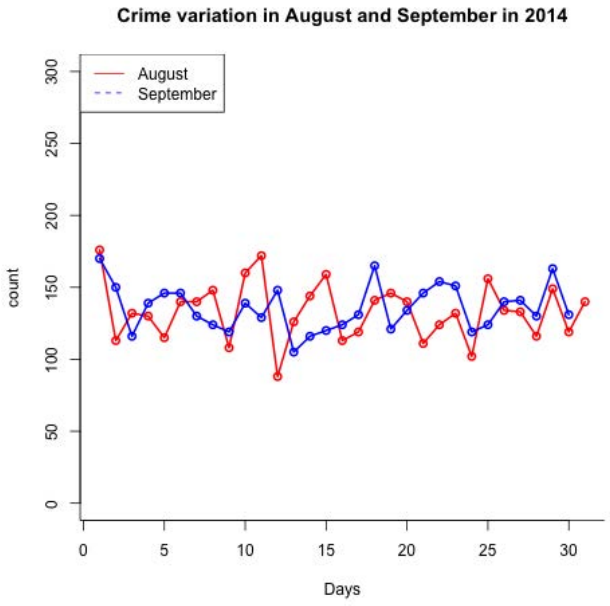


Figure 13: *How crime varied in the months of August over the years in Atlanta*

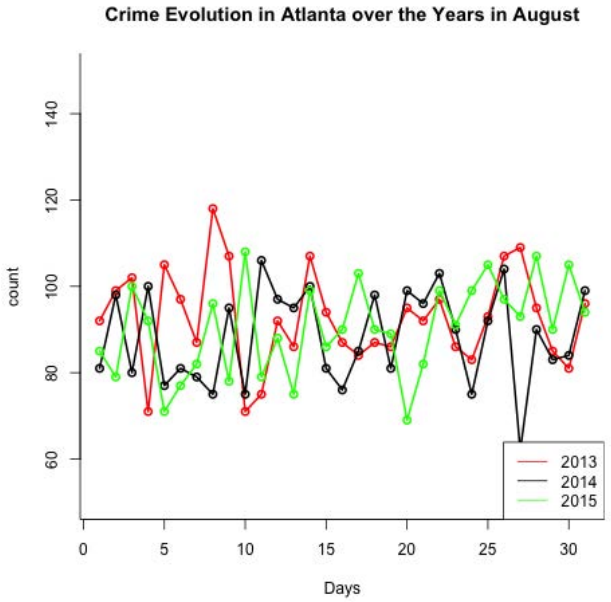


Figure 14: *How crime varies in the months of August in Atlanta and Baltimore*

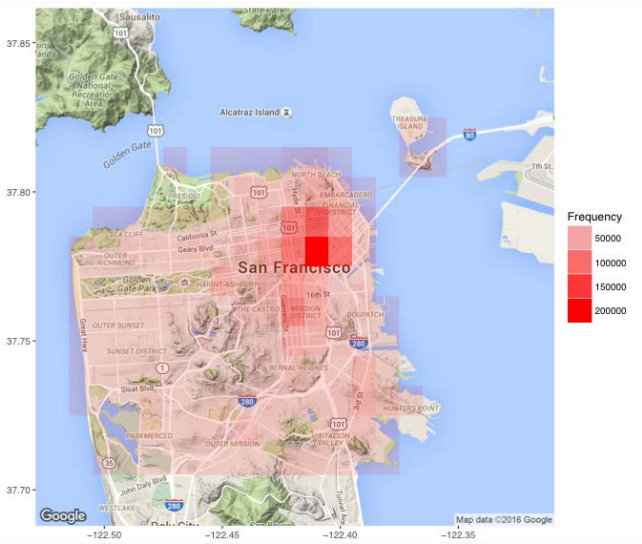


Figure 15: *Heat Map for the city of San Francisco*

9. Finally, we analyzed the distribution of crime in different parts of each city using the library ggmaps in R[19]. This analysis helps identify the safer neighborhoods of each city. The heat maps in *Fig*15 shows the overall spread of crime in San Francisco. Interesting observation here is that the hot spot is in the financial street rather than in the Downtown area as we had expected.

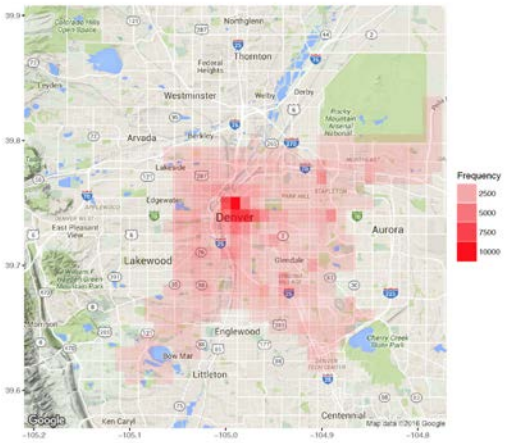


Figure 16: *Heat Map for Denver*

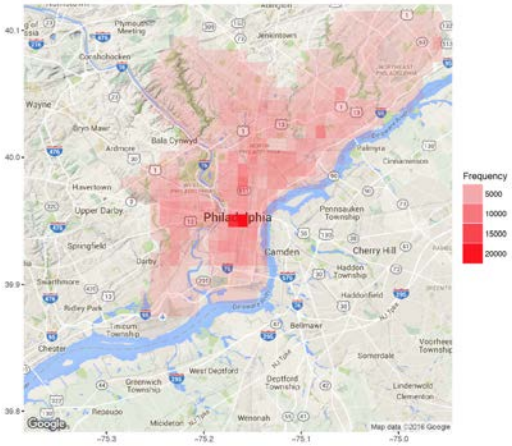
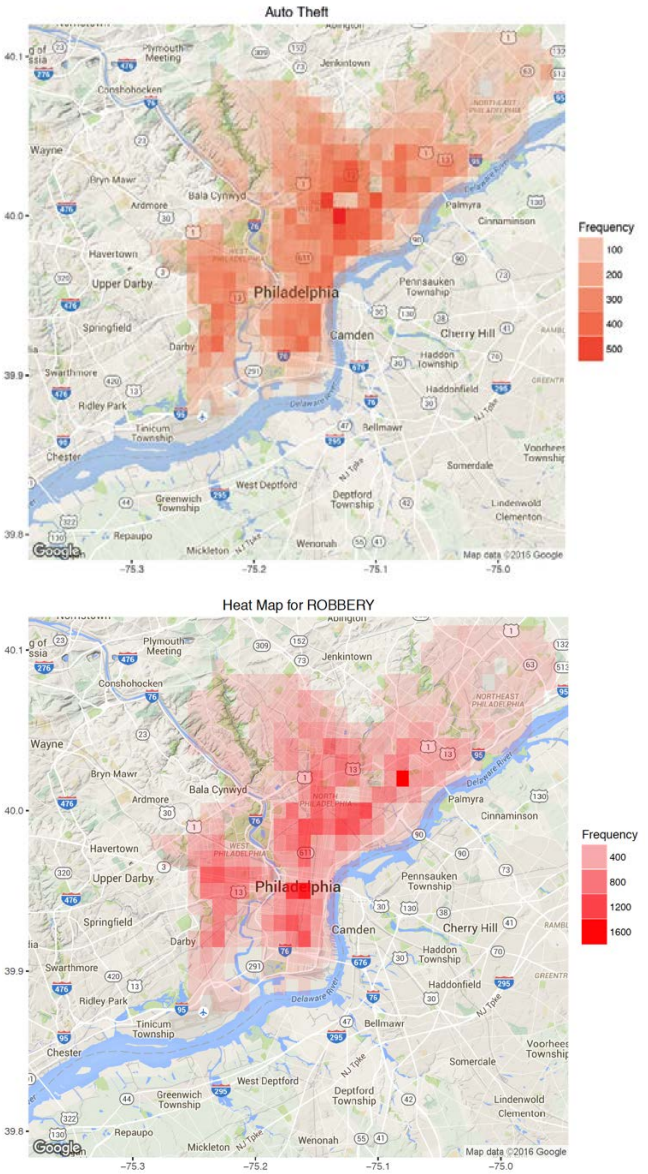


Figure 17: *Heat Map for Philadelphia*

10. We further analyzed the spread of each type of crime in the various cities. This can be seen in *Fig*.18 which shows the overall spread of crime in Philadelphia and the spread of auto thefts, robbery and assaults in Philadelphia.

Figure 18: *Heat Maps for Auto-Theft, Robbery and Assault in Philadelphia*



V. Future Work

We can extend our analysis on crime further in the following direction.

1. By considering historic crime datasets extending up to (say) 40-50 years for various cities, we might find more compelling trends in crime having higher correlations.

2. We considered few events and movies, based on the selected cities and duration of the data available, such as the Boston marathon bombing in 2013, Ferguson shooting in 2014, Breaking Bad, Blood Diamond, and so on. We can extend this analysis further with a historic crime dataset and find better correlation between these events/movies with crime.

3. We can also consider the impact of economy on crime, and the variation of crime with the demographics of the city.

4. In this project we concentrated on the main crime types like robbery, larceny. We can future extend this work by analyzing the trends in sub-crime types like robbery-pedestrian, robbery-commercial, robbery-residence, larceny-vehicle, larceny-nonvehicle.

References

[1] `http://shiny.rstudio.com`

[2] `http://espn.go.com/chalk/story/_/id/14742837/record-132-million-bet-super-bowl-50-nevada-sports`

[3] Zahran, Sammy, Tara O?Connor Shelley, Lori Peek, and Samuel D. Brody. "Natural Disasters and Social Order: Modeling Crime Outcomes in Florida." International Journal of Mass Emergencies and Disasters. N.p., n.d. Web. 25 Feb. 2016.

[4] Dahl, Gordon, and Stefano DellaVigna. "Does Movie Violence Increase Violent Crime?" (n.d.): n. pag. Web. 25 Feb. 2016.

[5] Zanten, Eric Van. "Crime vs. Temperature." Crime vs. The Weather. N.p., n.d. Web. 25 Feb. 2016

[6] Malby, Steven, and Philip Davis. "Monitoring the IMPACT OF ECONOMIC CRISIS ON CRIME." UNITED NATIONS OFFICE ON DRUGS AND CRIME (n.d.): n. pag. Web. 25 Feb. 2016.

[7] Andrienko, Yury. "Explaining Crime Growth in Russia during Transition: Economic and Criminometric Approach*." Policy Documentation Center (n.d.): n. pag. Web. 25 Feb. 2016.

[8] http://us-city.census.okfn.org/dataset/crime-stats

[9] https://www.wunderground.com/history/

[10] https://en.wikipedia.org/wiki/Category:Lists_of_events

[11] http://opendata.atlantapd.org/Crimedata/Default.aspx

[12] https://www.baltimorepolice.org/bpd-open-data

[13] https://data.brla.gov/Public-Safety/Baton-Rouge-Crime-Incidents/fabb-cnnu

[14] https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2

[15] http://data.denvergov.org/dataset/city-and-county-of-denver-crime

[16] https://data.sfgov.org/Public-Safety/SFPD-Incidents-from-1-January-2003/tmnf-yvry

[17] https://www.opendataphilly.org/dataset/crime-incidents

[18] http://www.westword.com/news/marijuana-study-is-rise-in-denver-crime-linked-to-

[19] https://journal.r-project.org/archive/2013-1/kahle-wickham.pdf

[20] S. Sivaranjani and S. Sivakumari, Mitigating Serial Hot Spots on Crime Data using Interpolation Method and Graph Measures in International Journal of Computer Applications September 2015.

[22] Hsinchun Chen, Wingyan Chung, Yi Qin, Michael Chau, Jennifer Jie Xu, Gang Wang, Rong Zheng, Homa Atabakhsh, ?Crime Data Mining: An Overview and Case Studies?, AI Lab, University of Arizona, proceedings National Conference on Digital Government Research, 2003, available at: http://ai.bpa.arizona.edu/

Visual Representations of Disaster in the 2015 Nepal Earthquake

Melissa Bica

University of Colorado Boulder
melissa.bica@colorado.edu

Abstract

In April 2015, Nepal was struck by a massive earthquake and experienced major damage and loss of lives. This event gained much media attention from locally affected Nepalis and the global public alike. In this paper, we investigate how this disaster event is represented visually via the images shared on Twitter. We are interested in what images get shared the most as well as the users who share them. We conduct qualitative analysis on images collected from tweets as well as a statistical analysis of the tweet activity as compared to the actual recorded damage. Results show distinct differences in thematic content over the course of the post-disaster phase and by different types of users, as well as positive correlations between tweet activity and damage. We find that in conducting analysis of social media data for a disaster event such as this, it is important to focus specifically on the content of locally affected users, as it is often overshadowed by messages from the concerned yet well-meaning global public.

I. Introduction

Crisis informatics is the study of how people use technology and social media during disaster events to gather and share information. Social media provides researchers the opportunity to learn about many different human behaviors and decision-making processes during disasters, such as evacuation patterns, information sharing and personal reporting [2], and providing aid as digital volunteers [13]. Twitter is a primary social computing platform used in this domain and provides rich metadata for tweets, including geospatial information, multimedia, and user details, that can be utilized in analysis.

In this paper, we examine images shared via Twitter in the aftermath of the 2015 Nepal earthquake. We are interested in learning how the event is represented to others:

- How does the story that the images portray compare to those that unfolded in the aftermath of the event?
- Are the images representative of the people who were affected?
- Do images portraying destruction or calls for help correspond to the areas that were in most need of relief?

Additionally, we are interested in the patterns and trends of image diffusion.

- How do temporal and spatial proximity to the event affect sharing of images?
- Why do certain images during the post-disaster phase go viral while others are relatively unnoticed?

We believe that a combination of attributes of the image content, tweet content, and the sharing user play a role in what images get shared the most.

This work focuses on geospatial analysis as a means to address these questions. We use geospatial metadata from the tweets as well as geospatial data from various organizations that reflect the damage in Nepal after the earthquakes. This data can identify users who were local to the earthquake and directly affected and users who were part of the global public and not affected. By mapping the tweets and their attached imagery together with the damage assessment data, it is possible to study the

relationship between the imagery shared on Twitter, the users sharing the images, and the actual damage of the event, and evaluate how representative the imagery is of the event.

Imagery and geospatial data have been studied together previously to address various research questions that motivate this work. Several tools have been implemented to allow users to visualize images from geotagged tweets on maps in order to learn about news and current events happening around the world [7, 14]. In [1], the authors combine methods in image processing, text processing, sentiment analysis, and data integration to infer societal happiness based on geotagged images shared on Twitter. Kawakubo and Yanai extended a well-known algorithm for ranking images based on similarity to also consider images' geo-locations [8]. Geospatial metadata of photos has been utilized for automation of tasks such as generating tourist maps and detecting events from distinct cameras [5, 6]. Although not specific to imagery, the relationship between general Twitter activity and economic damage around a particular disaster event is analyzed in [10].

On April 25, 2015, a massive 7.8 magnitude earthquake struck in Gorkha and devastated Nepal. Many large aftershocks continued over the following weeks, with the largest on May 12 near Mt. Everest (see Figure 1). This was one of the worst natural disasters ever to strike Nepal, having killed over 8,000 people and destroyed over 500,000 homes [11].

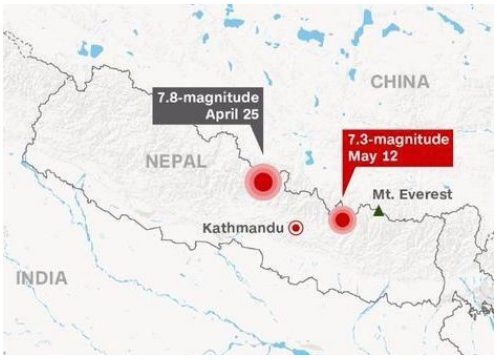


Figure 1: *Epicenters of earthquake on April 25 and major aftershock on May 12, 2015.*

We chose to analyze the Twitter activity surrounding this event based on its severity and wide global audience. People tweeted in different languages, from different locations, with multimedia, and across many months, providing us with a huge amount of rich data to explore.

II. Data

Data from Twitter was collected through a software infrastructure our research group has set up using a Cassandra cluster for 24/7, high-volume Twitter data collection [3]. We started a collection on April 25, 2015 soon after the earthquake hit using the Streaming API with over 100 keywords, including generic terms such as `#nepalearthquake` and `#everest2015` as well as many specific affected location names. The list of keywords changed over the following weeks with new terms being added and some terms that were too generic and produced a lot of noise (e.g. `damage`) being removed, resulting in 151 keywords for data collection. The 24M+ tweets were downloaded in JSON format then loaded into MongoDB. The details of the Twitter dataset are presented in Table 1.

Table 1: *Overview of the dataset*

Start date	04/25/2015
End date	02/15/2016
No. of tweets	24,045,885
No. of unique users	5,228,402
No. of original tweets	11,932,871 (49.6%)
No. of retweets	12,113,014 (50.4%)
No. of tweets with images	6,506,081 (27.1%)
No. of retweets with images	4,873,338 (20.3%)
No. of geotagged tweets	96,368 (0.4%)
No. of tweets geotagged in Nepal with images	2,392 (0.01%)

Although we did not consider all 24M tweets in this analysis, we are studying a relatively rare phenomenon—tweets geotagged

in Nepal containing images after a specific event—and thus we needed to start with a large amount of data in order to ensure a reasonably-sized final sample.

Given the long period of time our dataset covers after the earthquake (nearly 300 days), we decided to focus on only a subset of the data based on date. This is necessary because the data collection has been running continuously since the earthquake hit until present day, and many of the keywords and hashtags that were initially used in relation to the earthquake have likely been repurposed by now. Additionally, daily tweet volume decreased substantially within days after each earthquake. Figure 2 shows the number of tweets collected per day starting on April 25, 2015, the day the earthquake hit, through the end of May. The first spike is on April 26, likely due to the fact that the data collection did not start until after the earthquake hit midday. The second spike on May 12 corresponds to the major aftershock that occurred that day between the districts Dolakha and Sindhupalchowk (see Figure 1).

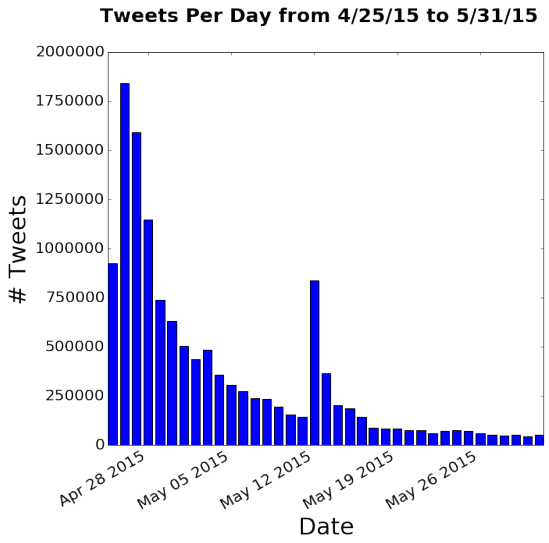


Figure 2: *Number of tweets per day.*

We used the date range represented in this figure as the bounds for our project, assuming based on the numbers that the most relevant tweets to the earthquake were produced during this period of just over one month. In total, there are 12,914,430 tweets during this span, or 53% of the full Twitter dataset. Additionally, we have chosen to split this date range into two ranges, *before-second-earthquake* (4/25/2015 - 5/11/2015) with 10197520 tweets and *after-second-earthquake* (5/12/2015 - 5/31/2015) with 2,716,910 tweets. This allows us to see differences in people's information needs and communication patterns as reflected on Twitter after a primary disaster and after a second, closely-located disaster event which added to the effects of the first in many places and caused new damage and disruption in others.

In addition to Twitter data, we also use damage assessment data from the National Geospatial-Intelligence Agency (NGA), specifically the NGA Nepal Earthquake Open Data Search application.[1] This data is derived from satellite imagery taken from April 26 - May 13, 2015, which covers the immediate aftermath of both major earthquakes that occurred. Each data point has geographic coordinates, the recorded date, and classification as *Affected*, *Minor*, *Major*, or *Destroyed*. Finally, we use population data from the 2011 Nepal census, taken at the district level.[2]

III. Methods

In order to answer the broad question of how disasters are represented on Twitter via shared images, we chose four analytical categories by which to study the data: (1) all tweets with images, (2) all tweets with images in Nepali, (3) all tweets with images not in Nepali, and (4) all tweets with images geotagged in Nepal. We found all the retweets of tweets in each of these categories for both date ranges. The counts of tweets containing images for each dataset are given in Table 2. We are interested in how im-

[1]Data from Latest NGA Damage Assessment Points, `http://nepal.nga.opendata.arcgis.com/datasets/5dd6783700850409080c2c8f2ed019b55_0`

[2]Data from District/VDC wise population of Nepal (CBS 2011), `http://umeshg.com.np/district-wise-population-of-nepal-cbs-2011/`

ages from local, directly affected people diffuse through the network differently from other images, as well as why certain images in these categories gain popularity while others do not.

Table 2: *Number of tweets (including retweets) with images per analytic category before and after the second earthquake.*

Analytic Category	Before Second Earthquake	After Second Earthquake
All	3,047,563	729,563
Nepali	25,513	15,862
Geotagged	3,626	2,244

We first found the top retweeted images in each of these analytic categories for each time range (before and after the second earthquake), resulting in 8 image sets total. We created queries to find all retweets that fit the conditions of each category, then calculated aggregate values to count the occurrences of the retweets' `retweeted_status.id` (i.e., the tweet id of the originating tweet). Since this dataset comes from the Streaming API and does not include the entirety of Twitter data, it is not guaranteed that every original tweet for which there is a retweet exists in the dataset. So, once we obtained the original tweet ids with the highest counts in each category, we obtained the original tweet metadata by querying for a retweet of that tweet, since all retweets contain the full metadata for their originating tweet in the `retweeted_status` field.

A preliminary exploration of the top retweeted images in each category led us to find that categories (1) and (3) are overlapping for the most retweeted images, as the non-Nepali tweets have much higher retweet counts than Nepali tweets overall. Therefore, we treat these two categories as the same.

For the three remaining analytic categories, we coded the top 100 retweeted images for denotative content. The image coding scheme, explained in Table 3, includes the categories: people suffering, damage, branded groups, non-photographic information, celebrities, relief effort work, and concern and prayer.

Part of the representation of disaster on Twitter also stems from the relationship between where users tweet from and where the damage occurred. We map the damage per district based on the number of recorded damage points in that district at any damage level (Affected, Minor, Major, or Destroyed) (Figure 3). We map the tweets per capita using the 5,870 tweets with images that are geotagged in Nepal (this includes both date ranges), normalized by district-level population data (Figure 4). Figure 5 shows the distribution of NGA damage assessment points in the affected districts.

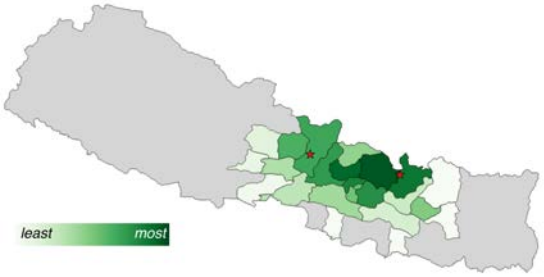


Figure 3: *Damage per district based on the number of recorded damage points at any damage level (Affected, Minor, Major, or Destroyed) per district. The red stars on this and all following maps represent the epicenters of the 2 earthquakes.*

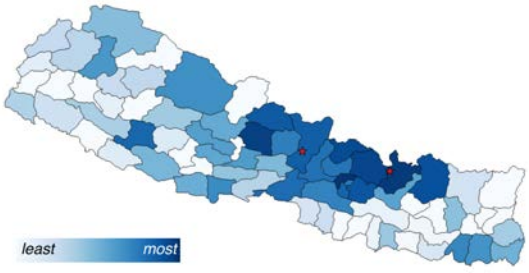


Figure 4: *Tweets per capita based on tweets with images that are geotagged in Nepal and district-level population data.*

Table 3: *Image coding scheme.*

Code	Description
people suffering	Images of people who appear to be suffering, injured, or dead.
structural and environmental damage	Images of damage to buildings and infrastructure, or of natural disaster itself (e.g. landslide).
branded groups	Images advertising or containing logos for organizations/companies.
non-photographic information	Images of text, documents, maps, infographics, political cartoons, or screenshots of social media/news. Not photographs.
concern and prayer	Images suggesting prayers & wellwishes, e.g. "Pray for Nepal."
people who are celebrities	Images featuring celebrities.
relief effort work	Images of relief efforts, e.g. supplies being distributed.
other	Images apparently associated to the disaster (via data collection) not covered by previous categories, such as missing people or landscapes.
N/A	Images that have been either removed from Twitter or posted by users whose accounts have been deleted, changed, or suspended.

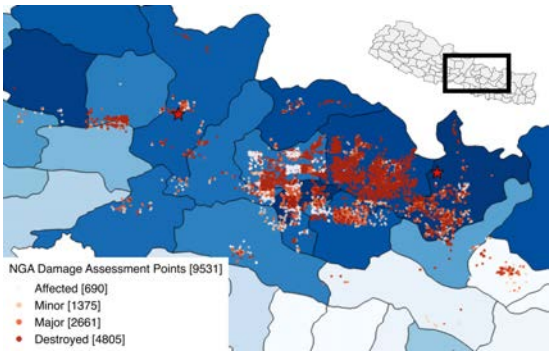


Figure 5: *NGA damage assessment points overlayed on map depicting tweets (geotagged with images) per capita. Numbers in brackets represent total number of points in each category.*

Using these data, we follow the methods of Kryvasheyeu et al. to calculate correlations between normalized tweet activity per capita and damage [10]. In this study, the authors used data from Twitter, FEMA, and insurance claims during Hurricane Sandy to measure rank correlation coefficients for various geospatial areas in New Jersey and New York. They plotted log-log graphs of per-capita Twitter activity vs. per-capita damage for both states and found positive correlations in each. Though this paper aimed to be predict damage geospatially based on Twitter activity, our goal is to show whether a relationship exists between damage and a specific type of Twitter activity to determine whether this activity is representative of the event. We expect to find positive correlations in our data as well, which would imply that the geospatial distribution of tweets is representative of where earthquake damage occurred.

IV. Results

I. Descriptive

Qualitative analysis of the images uncovered the differences in thematic image content shared in both time ranges as well as in the various analytic categories. The top thematic image categories for each set of images based on coding are shown in Table 4. We found that images of people who appear suffering, injured, or dead were retweeted a great deal in the set of all tweets with images both before and after the second earthquake, whereas they were not nearly as common in either the Nepali or geotagged sets. Images of celebrities were very common after the second earthquake in the set of all images but not in any other dataset nor before the earthquake. Finally, in both time ranges for both the Nepali and geotagged sets, we found that images of relief effort work were among the most retweeted. Images in the non-photographic information category were also common in these datasets, assumed to be mainly comprised of users directly affected by the earthquake or at least living in Nepal. Many images coded as non-photographic information seem as though they would be useful only or primarily to locally affected people, e.g. images explaining how to assess earthquake damage to one's house. This prevalance of locally relevant and, in many cases, actionable, information in images shared by users local to the earthquake is consistent with findings in [9].

Table 4: *Most retweeted image categories*

Analytic Category	Before Second Earthquake	After Second Earthquake
All	people, damage	celebrities, people
Nepali	relief, non-photographic	damage, relief
Geotagged	people, relief	non-photographic, relief

The top retweeted image is of a woman on the ground holding her child and looking dejectedly into the camera (see Figure 6). In addition to the strong emotions evoked from the subjects of the image, the tens of thousands of retweets of this image were likely also due to the fact that it is a publicity image for the charity organization Save the Children and

was shared by internationally-known soccer star Cristiano Ronaldo, a global ambassador for the organization. Moreover, a rumor circulated widely soon after the initial earthquake that Ronaldo donated $8 million to Save the Children to support earthquake victims (this was later found to be false) [4].

The second most retweeted image is of a young boy holding (presumably) his younger sister (see Figure 6). Most of the retweets of this image say that it is a "young boy protecting his sister in Nepal." In reality, this image is not related to the earthquake at all—on 5/2/15, the photographer tweeted the image with the text: "This is my photo about two Vietnamese Hmong ethnic children taken in 2007 in Ha Giang province, it's not about Nepal." The image was resurfaced after the earthquake and made to appear to depict two lonely, scared children [12].

Figure 6: *Most retweeted images in full Twitter dataset. Left: Publicity image for Save the Children originally tweeted by Cristiano Ronaldo. Right: Photo of brother and sister from Vietnam in 2007 rumored to be taken after the Nepal earthquake.*

II. Statistical

We illustrate the correlations between damage and (geotagged image) tweet activity in Figure (7). Results are shown for cumulative damage (all levels of damage combined) as well as for each level of damage individually. All five plots show a positive correlation between damage and geotagged image tweet activity in Nepal. Additionally, we calculated rank correlation coefficients—Kendall's τ, Spearman's ρ, and Pearson's ρ—between damage and geotagged image tweet activity for each level of damage as well as cumulative. Results for cumulative damage are shown in Table 5. All correlations are statistically significant with p-values < 0.05, which indicates that the correlation between damage and geotagged image tweet activity in Nepal is positive.

Table 5: *Rank correlation coefficients for all damage levels compared to tweet activity.*

	Statistic	P-value
All Damage		
Kendall's τ	0.47	2.87×10^{-7}
Spearman's ρ	0.58	5.99×10^{-8}
Pearson's ρ	0.54	7.32×10^{-7}
Destroyed		
Kendall's τ	0.46	6.96×10^{-7}
Spearman's ρ	0.57	1.02×10^{-7}
Pearson's ρ	0.48	1.63×10^{-5}
Major		
Kendall's τ	0.47	3.47×10^{-7}
Spearman's ρ	0.58	6.72×10^{-8}
Pearson's ρ	0.60	1.07×10^{-8}
Minor		
Kendall's τ	0.47	2.54×10^{-7}
Spearman's ρ	0.58	5.10×10^{-8}
Pearson's ρ	0.60	1.81×10^{-8}
Affected		
Kendall's τ	0.49	1.50×10^{-7}
Spearman's ρ	0.60	1.05×10^{-8}
Pearson's ρ	0.28	1.42×10^{-2}

The linear-fit regression lines on each plot make clear which districts are outliers. Those districts that fall above the line experienced more damage than was represented by the number of images tweeted by people there—i.e., they were *underrepresented* on Twitter. Conversely, those districts that fall below the line were *overrepresented* on Twitter via images as

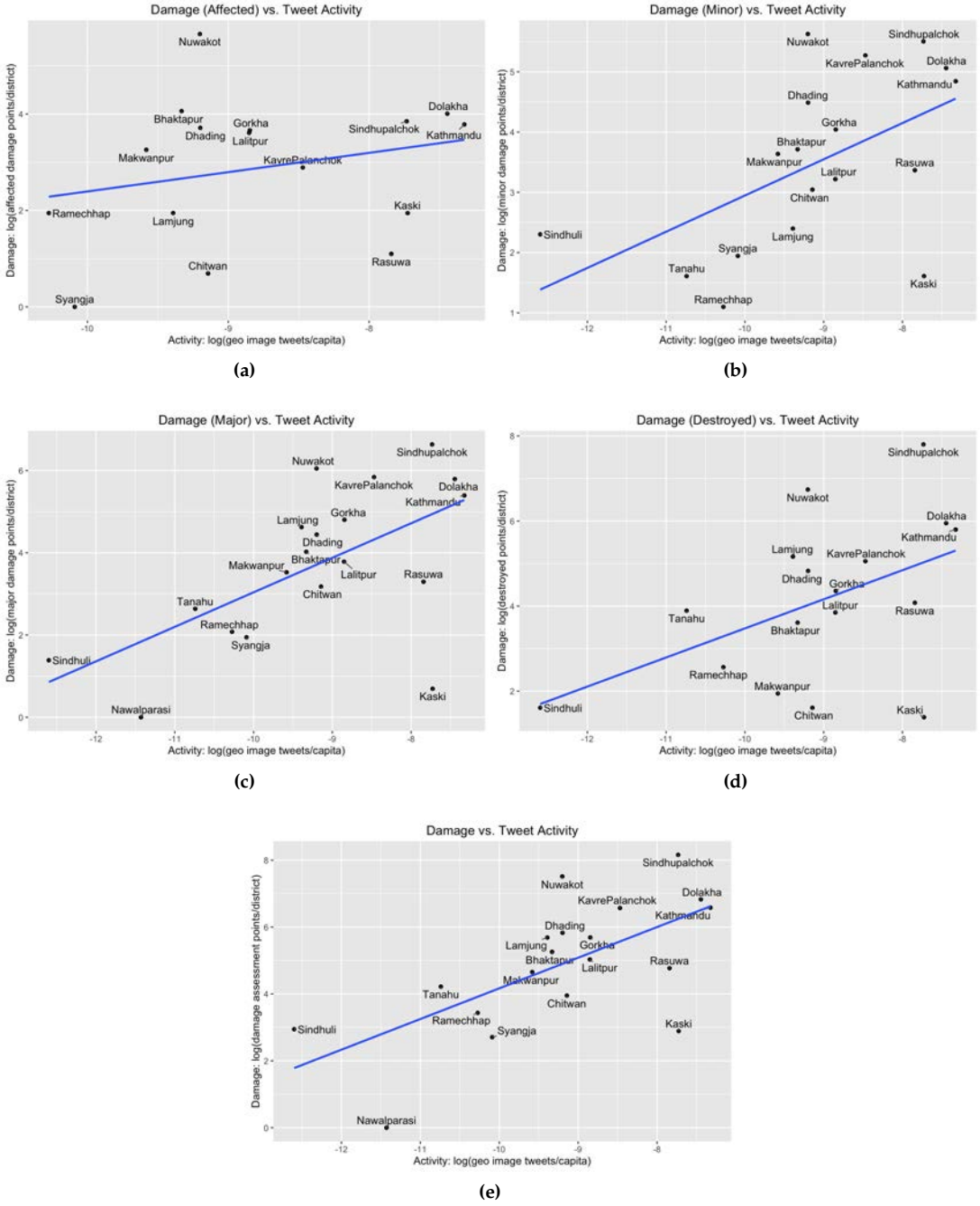


Figure 7: *Log-log plots of tweet activity versus damage at different levels of severity: (a) Affected, (b) Minor, (c) Major, (d) Destroyed, (e) All/cumulative. Tweet activity refers to all tweets in the study date range (4/25 - 5/12) containing images and geotagged in Nepal. Damage is measured in number of NGA damage assessment points per district. The best-fit linear regression lines are included, with all showing a positive correlation between tweet activity and damage.*

compared to the amount of damage that occurred. Interestingly, Kathmandu District, the most densely populated district in Nepal containing the capital city Kathmandu, falls on or very close to the line in all plots. Its location in the plot area shows that it had both a high amount of damage (though not the highest) and a high amount of tweet activity. We had expected Kathmandu to be very overrepresented on Twitter considering it has by far the most people of all the districts. Future qualitative analysis on the images shared from users in the outlier districts will be conducted in order to determine why there is a difference in representation for these areas.

The log-log plots show a positive correlation between damage and tweet activity for all levels of damage based on the positive slope of each linear regression line. In particular, the first four plots (for All Damage, Destroyed, Major, and Minor vs. tweet activity) have similar slopes, whereas the plot for minor damage vs. tweet activity has a distinctly slighter slope, indicating a less positive correlation for this lowest level of damage. This could be due to the much lower amount of damage assessment points at this level in the data—only 680 points are classified as Affected, whereas Minor, Major, and Destroyed have 1375, 2661, and 4805 points, respectively. The lower amount of data for the Affected level presumably leads to less accurate findings concerning the relation between damage and tweet activity.

V. Discussion

From the geospatial analysis of image tweets and damage, we see that some districts are overrepresented and others are underrepresented on Twitter as compared to the amount of damage experienced there. We also find instances of unequal representation of image content, particularly between all tweets and tweets either in Nepali or geotagged in Nepal. Images of people suffering and of celebrities appear the most in the all tweets dataset, whereas images of non-photographic information, such as damage assessment advice and social media messages, were retweeted frequently among local users. Given this, we can speculate that locality to the earthquake does not necessarily increase diffusion, or retweeting, of images. Future work could explore whether this is true of other disaster events, and investigate the underrepresented districts to determine why locally-sourced images are not as popular on Twitter as images from the global public.

An interesting finding from the image analysis is the presence of misinformation among images shared by the global public surrounding this event. This is manifest in the top retweeted images shared in our dataset: the photos of the woman and child and of the brother and sister. The first was shared by a well-known soccer star who was falsely thought to have donated millions of dollars to the organization for which he was promoting in the image. Moreover, photos of him attached to tweets about this false donation appeared in nearly every dataset. The second was taken in a different country and a different year, although many Twitterers shared it along with a statement about the young boy protecting his sister after the earthquake. Both these images involve some sort of misinformation, whether a rumor about the original user who shared the image, or a blatant misuse of an image to tell an untrue story. The fact that they were retweeted so much suggests that people feel connected to an event by sharing stories told through sentimentally powerful images; this is a theme worth exploring further.

VI. Conclusion

We have investigated the representation of the Nepal earthquakes via images shared on Twitter. Qualitative analysis of the images showed that the types of images people share differ based on both temporal proximity to the initial and second earthquakes and whether the images are geotagged in Nepal, in Nepali, or from the full dataset. Images related to rumors or misinformation are more common amongst the set of all tweets with images, suggesting the significance of sentiment over other attributes

in images shared by the global public. Images from tweets in Nepali or geotagged in Nepal tend to depict more locally relevant, actionable content. Geospatial analysis of geotagged images in Nepal showed positive correlations of tweet activity with damage. By studying the visual representations of disaster on social media, we can learn about the changing needs and concerns of people, especially those who are most vulnerable, and how to address them more effectively in future disaster events.

A limitation of this study is the lack of data on baseline usage of Twitter by people in Nepal. Though we calculate tweets per capita in looking at the relationship between geotagged image tweets and damage, a more effective metric would be to determine tweets per Twitter user in Nepal and each district. This alternative normalization was implemented by Kryvasheyeu et al. with no change in correlation, however their study was on Twitter users in New York and New Jersey, places with a much higher proportion of Twitter users per capita than Nepal [10]. Thus, we would expect a change in our results using this type of normalization.

Future work will include conducting qualitative analysis of images by district, focusing especially on districts that we found to be either over- or underrepresented based on tweet activity and damage. Additionally, we would like to look at the geospatial data at a more fine-grained level by using Village Development Committees (VDCs) as a new level of analysis. For comparison, there are 3157 VDCs and 75 districts (our level of analysis in this study) in Nepal.

References

[1] Saeed Abdullah, Elizabeth L. Murnane, Jean M.R. Costa, and Tanzeem Choudhury. Collective Smile: Measuring Societal Happiness from Geolocated Images. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '15*, pages 361–374, New York, New York, USA, February 2015. ACM Press.

[2] Jennings Anderson, Marina Kogan, Melissa Bica, Leysia Palen, Kenneth Anderson, Kevin Stowe, Rebecca Morss, Julie Demuth, Heather Lazrus, Olga Wilhelmi, and Jennifer Henderson. Far far away in Far Rockaway: Responses to risks and impacts during Hurricane Sandy through first-person social media narratives. In *Proceedings of the 13th International ISCRAM Conference*, 2016.

[3] Kenneth M Anderson and Aaron Schram. Design and implementation of a data analytics infrastructure in support of crisis informatics research (NIER track). In *Proceedings of the 33rd International Conference on Software Engineering*, pages 844–847. ACM, 2011.

[4] Des Bieler. Cristiano Ronaldo donated nearly $8 million for Nepal earthquake relief. `https://www.washingtonpost.com/news/early-lead/wp/2015/05/14/cristiano-ronaldo-did-not-actually-donate-8-million-to-nepal-earthquake-victims/`, May 2015. Accessed online: 2016-05-30.

[5] Wei-Chao Chen, Agathe Battestini, Natasha Gelfand, and Vidya Setlur. Visual summaries of popular landmarks from community photo collections. In *Proceedings of the seventeen ACM international conference on Multimedia - MM '09*, page 789, New York, New York, USA, October 2009. ACM Press.

[6] Hugo Feitosa de Figueirêdo, João Paulo Ribeiro da Silva, Daniel Farias Batista Leite, and Cláudio de Souza Baptista. Detection of photos from the same event captured by distinct cameras. In *Proceedings of the 18th Brazilian symposium on Multimedia and the web - WebMedia '12*, page 51, New York, New York, USA, October 2012. ACM Press.

[7] Brendan C. Fruin, Hanan Samet, and Jagan Sankaranarayanan. TweetPhoto: photos from news tweets. In *Proceedings of the 20th International Conference on Advances in*

Geographic Information Systems - SIGSPATIAL '12, page 582, New York, New York, USA, November 2012. ACM Press.

[8] Hidetoshi Kawakubo and Keiji Yanai. GeoVisualRank: A Ranking Method of Geotagged Images Considering Visual Similarity and Geo-location Proximity. In *Proceedings of the 20th international conference companion on World wide web - WWW '11*, page 69, New York, New York, USA, March 2011. ACM Press.

[9] Marina Kogan, Leysia Palen, and Kenneth M. Anderson. Think Local, Retweet Global: Retweeting by the Geographically-Vulnerable during Hurricane Sandy. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '15*, pages 981–993, Vancouver, Feb 2015. ACM Press.

[10] Yury Kryvasheyeu, Haohui Chen, Nick Obradovich, Esteban Moro, Pascal Van Hentenryck, James Fowler, and Manuel Cebrian. Rapid assessment of disaster damage using social media activity. *Science Advances*, 2(3), March 2016.

[11] Government of Nepal. *Nepal Earthquake 2015: Post Disaster Needs Assessment Executive Summary*. 2015.

[12] Nga Pham. Haunting 'Nepal quake victims' photo from Vietnam. `http://www.bbc.com/news/world-asia-32579598`, May 2015. Accessed online: 2016-05-30.

[13] Kate Starbird and Leysia Palen. "voluntweeters": Self-organizing by digital volunteers in times of crisis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 1071–1080, New York, NY, USA, 2011. ACM.

[14] Keiji Yanai. World Seer : A Realtime Geo-Tweet Photo Mapping System. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval - ICMR '12*, page 1, New York, New York, USA, June 2012. ACM Press.

Gut Diversity and Health in the Western Culture

Michelle D. Bray*

University of Colorado at Boulder
michelle.bray@colorado.edu

Abstract

Our microbiomes are an integral part of our genetic makeup and outnumber our own cells at an astonishing ratio of ten to one. The more we learn about our resident bacteria, viruses, fungi, etc., the more we learn about our own bodies. This study aims to explore any potential health implications of diversity– or lack, thereof– in our gut flora. In particular, records of respiratory illnesses such as allergies and asthma will be compared against cesarean section births in the United States. Data sets are taken from both American Gut Project samples and Center for Disease Control (CDC) publicly available records.

I. Introduction

Initially, one may be inclined to view human beings and our bacteria as separate, autonomous organisms. Recent advancements in genetics, however, prove that this is far from the truth. In current understanding, the human body is made up of about ten times more microbial cells than human cells. Many of these bacterial inhabitants are obligatory symbionts, which means that we depend on one another for survival.

In his article "Some of My Best Friends Are Germs" (New York Times Magazine, 5/15/2013), Michael Pollan addresses the importance of gut diversity. He builds the argument that being overly sanitary is a detriment to our microbiome, and ultimately our health. Pollan's article ties to relevant microbiome research being done here at CU, namely The American Gut Project, and serves as inspiration for this study.

II. Data

Data used for this study is compiled from multiple sources. Gut diversity and flora information is provided by The American Gut Project, an effort that analyzes and records fecal samples that are voluntarily sent in from participants across the United States. Information regarding chronic asthma is derived from the CDC's Division of Population Health.

The OTU table obatined from the American Gut Github repository contains relevant OTU (operational taxonomic unit) data and metadata. Each row in this table corresponds to a unique participant, and thus, fecal sample. Each column respresents a particular taxanomic identifier, and the values equate to occurances of each microbe in that sample. This data is parsed with a Python library called 'biom' to create a single .csv with the number of nonzero taxa for each sample–a simple representation of diversity– for further analysis in R.

Fecal sample data is distributed across the United States, with the highest density of samples in the north east. The west coast also has a relatively high density, but the east coast appears to have more samples. Individual states that stand out, apart from those in the New England area, are California, Washington, Colorado, Illinois, and Florida. However, smaller states in the north east are closely packed, and data thus appears more dense. If evaluating data on a state by state basis, this is important to keep in mind.

*A thank you or further information

Health related data sets are obtained as .csv (comma separated values) files, and manipulated in R to produce numerical latitude and longitude columns, excluding empty or NA values.

III. Methods

Because this study aims to seek out correlations between gut diversity and health (particularly cesarean birth, asthma, diabetes, and allergies), data is analyzed both statistically and visually. Fecal samples are first analyzed to determine general gut diveristy. Health records (cesarean birth, asthma, diabetes, and seasonal allergies) are also analyzed individually to get a feel for data distribution across the United States. Once this is done, spatial distribution of each health category is directly compared to that of gut diversity. Finally, all health implications are simultaneously compared to gut diversity in an attempt to find overall trends.

For statistical analysis, each health category is directly compared to gut diversity and graphed as a boxplot. Using an R library called 'ggplot,' health data is also plotted against a map of the United States using longitude and latitude from the .csv files as X and Y coordinates respectively. In the case of data extracted from American Gut Project records, the color of each geometric point corresponds to the the level of diversity for that sample. Health data taken from the CDC (saved as cdc_asthma.csv and cdc_diabetes.csv once cleaned up using a custom R script) geometric points are plotted with varying color and size according to the data value, or prevalence recorded at each location.

In terms of clustering diversity data, a nonmetric multidimensioinal scaling (NMDS) test is utilized. This approach collapses data across multiple dimensions into just a few, so that they can be visualized/interpreted. Visualizing clusters, and their relative locations gives insight into spatial correlations between gut diversity and chronic health implications.

IV. Results

According to the p-values generated by the Moran I test, the diversity value for fecal samples is statistically insignificant. This implies that we cannot reject the null hypothesis that there is zero spatial correlation present in fecal diversity values.

Cesarean Birth: According to the generated boxplot, those who were born through cesarean means, as opposed to natural, vaginal means, have an average of 400 unique gut flora. On a scale of 0-800 different types of gut bacteria, this value is right in the middle. See Figures 1 and 2.

Asthma: Those with asthma show an average of slightly below 400 different gut flora. See Figures 3 and 4.

Diabetes: Type II diabetes is most common among those whose fecal samples show an average of around 300 unique gut flora. See Figures 5 and 6.

Allergies: Individuals with seasonal allergies also have an average of about 400 different gut bacteria strains. However, there are more outliers with a higher number of unique flora, ranging anywhere from 900 to 1700. See Figures 7 and 8.

V. Conclusion

In conclusion, the results do not indicate that there is any direct correlation between lack of overall gut diversity and incidents of chronic illnesses (asthma, diabetes, and seasonal allergies, in particular) in the United States.

Paritally incomplete datasets could have some bearing on the discussed results. Participants of the American Gut Project must volunteer and pay to have their fecal sample analyzed. This restricts the range of samples collected. Furthermore, samples that do not have a latitude, longitude, or the desired field in question must be discarded.

References

[Pollan, 2013] Michael Pollan. (May 15, 2013). Some of My Best Friends Are Germs, *The New York Times Magazine*.

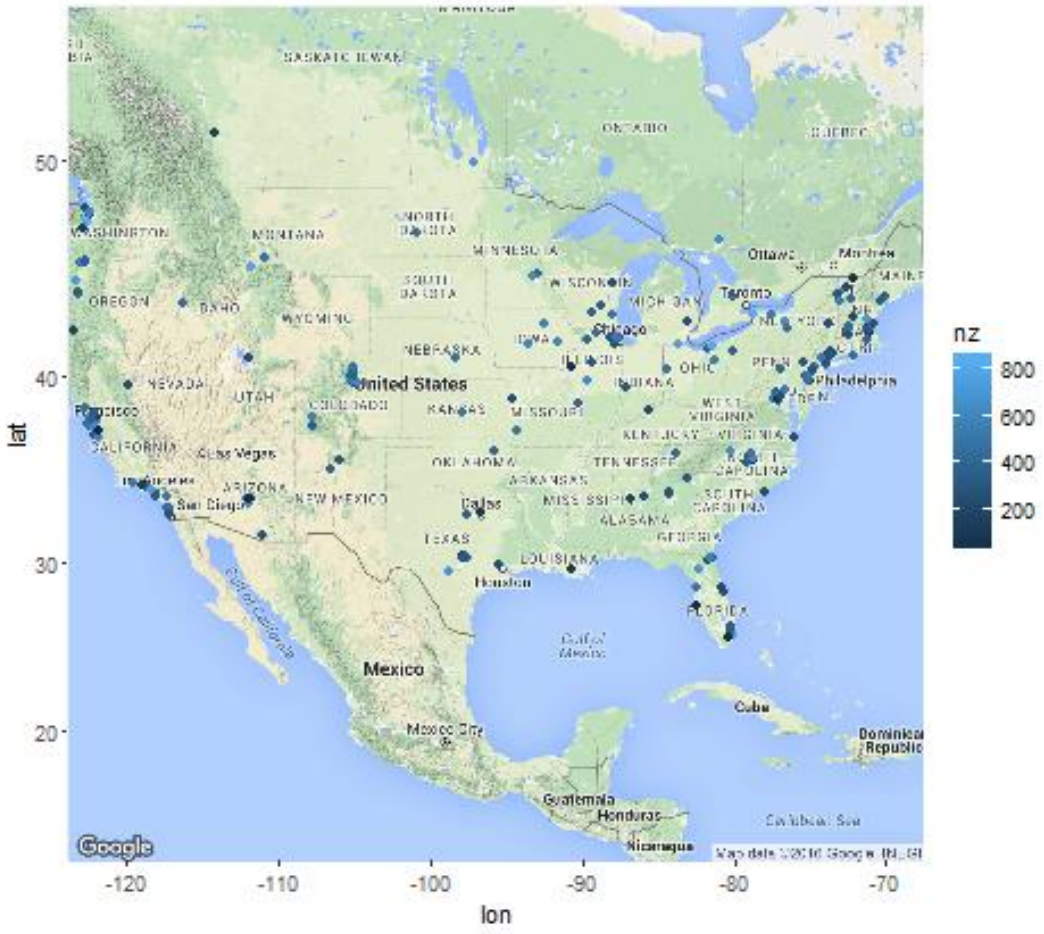


Figure 1: *Cesarean Births Map*

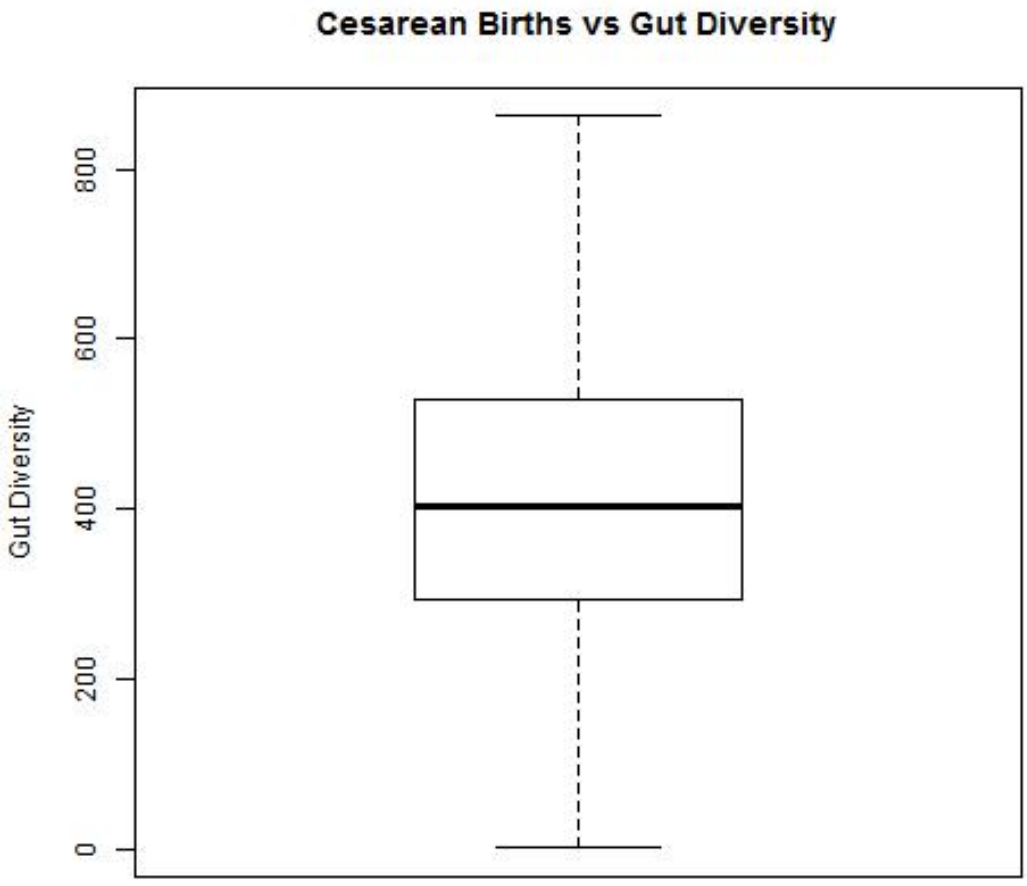


Figure 2: *Cesarean Births Boxplot*

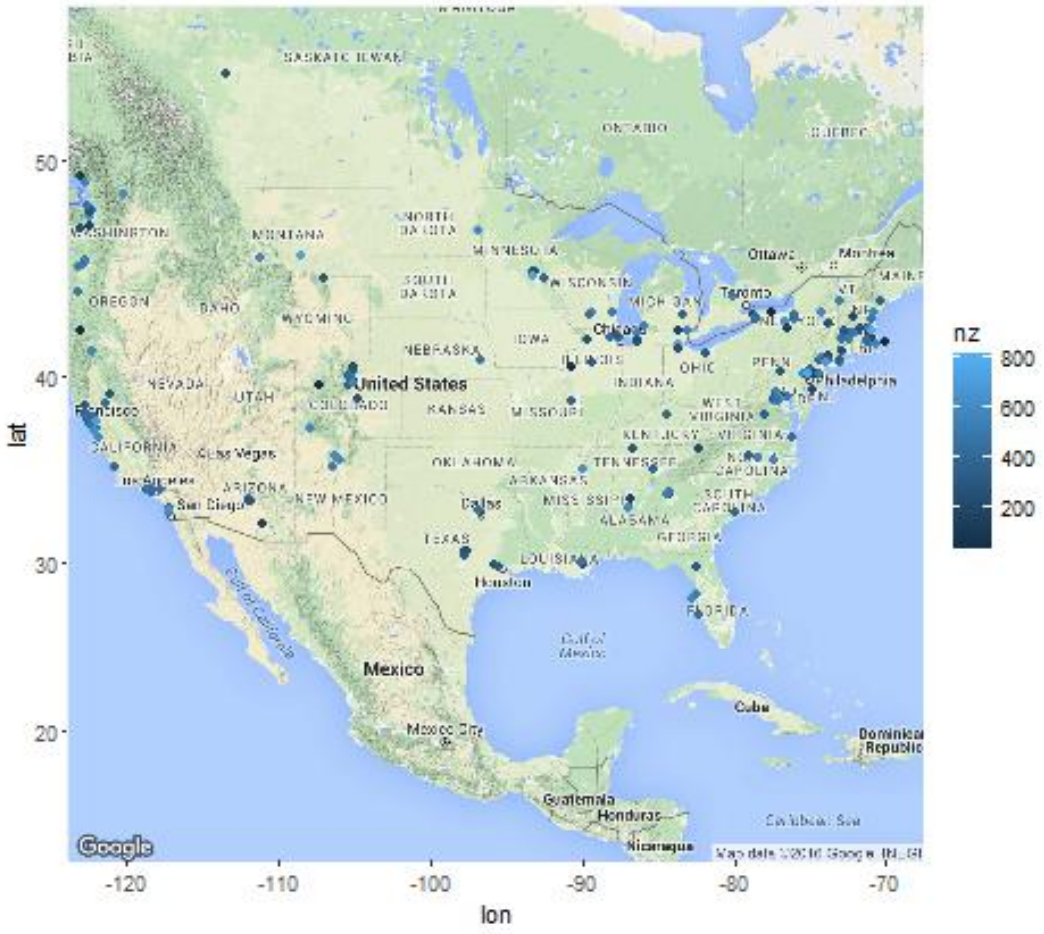


Figure 3: *Asthma Map*

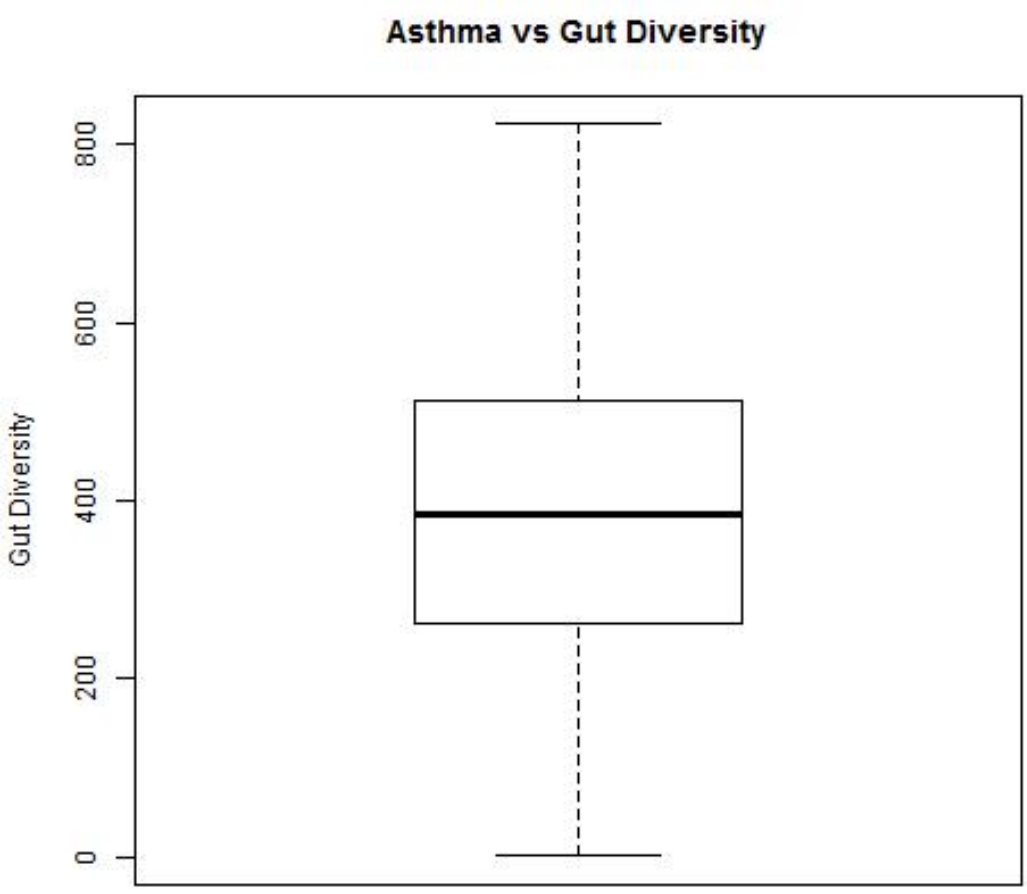


Figure 4: *Asthma Boxplot*

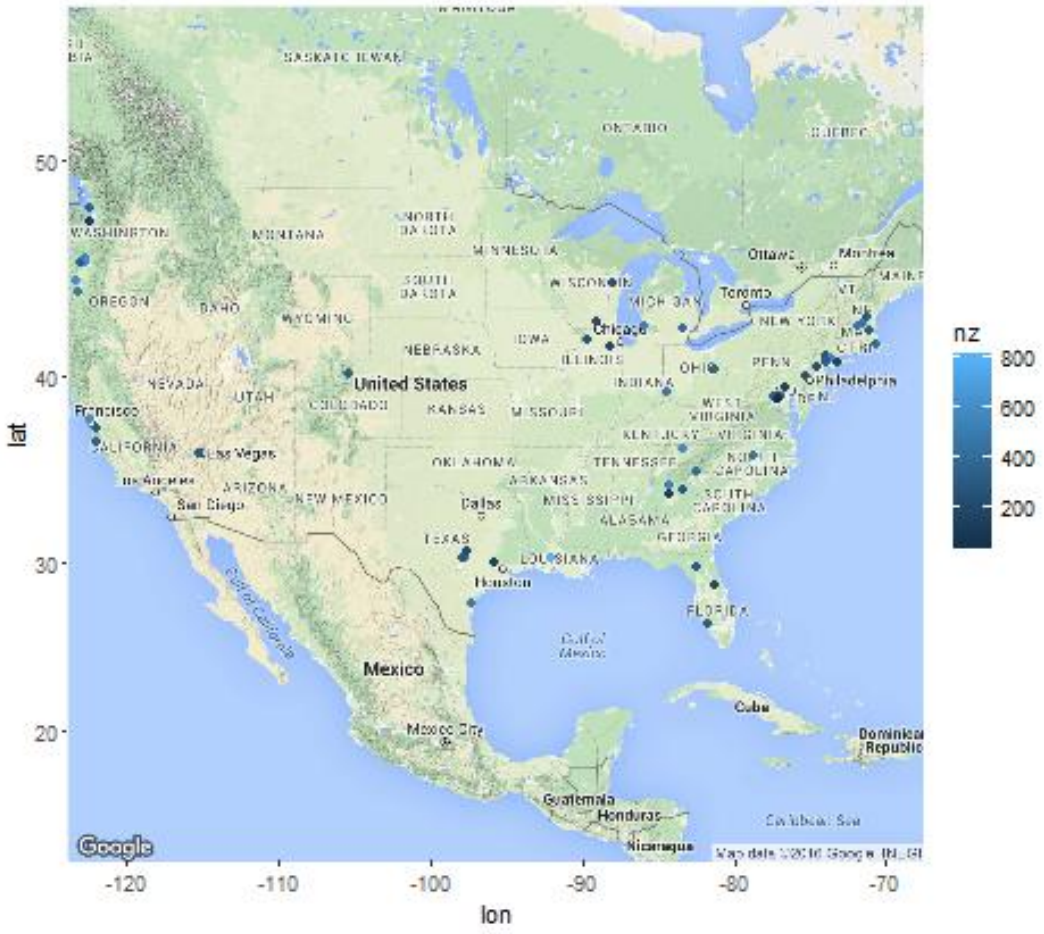


Figure 5: *Type II Diabetes Map*

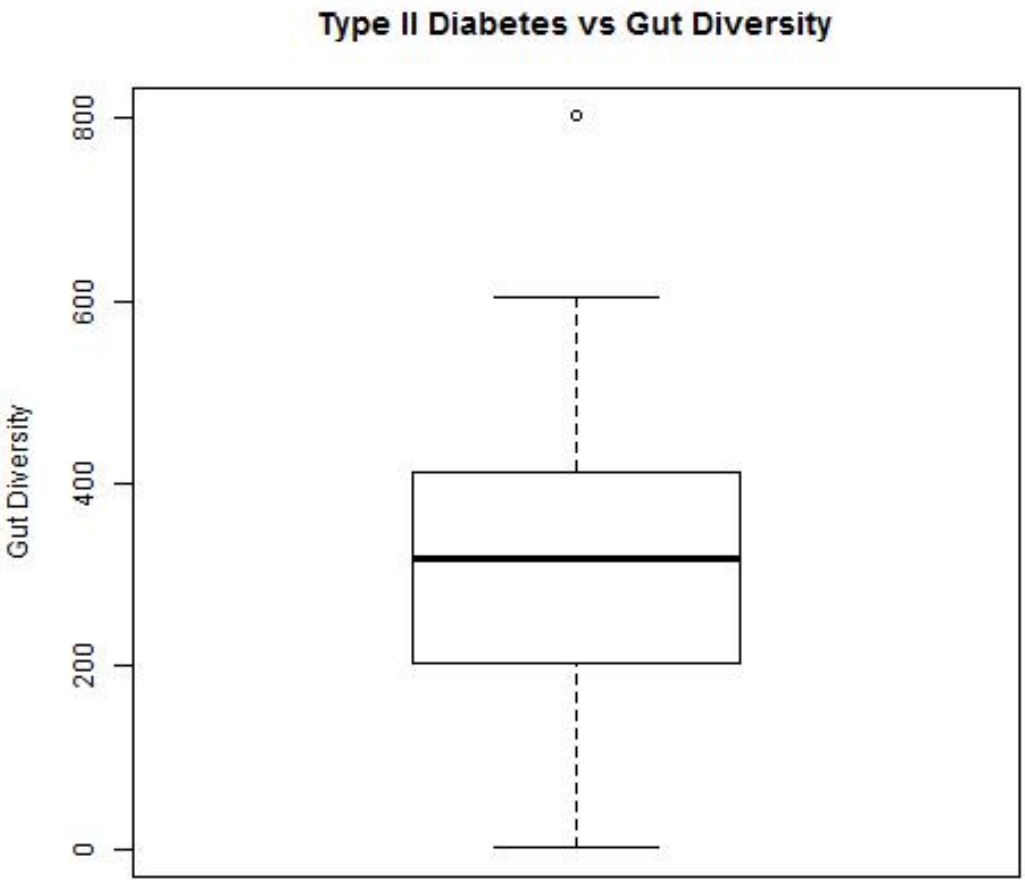


Figure 6: *Type II Diabetes Boxplot*

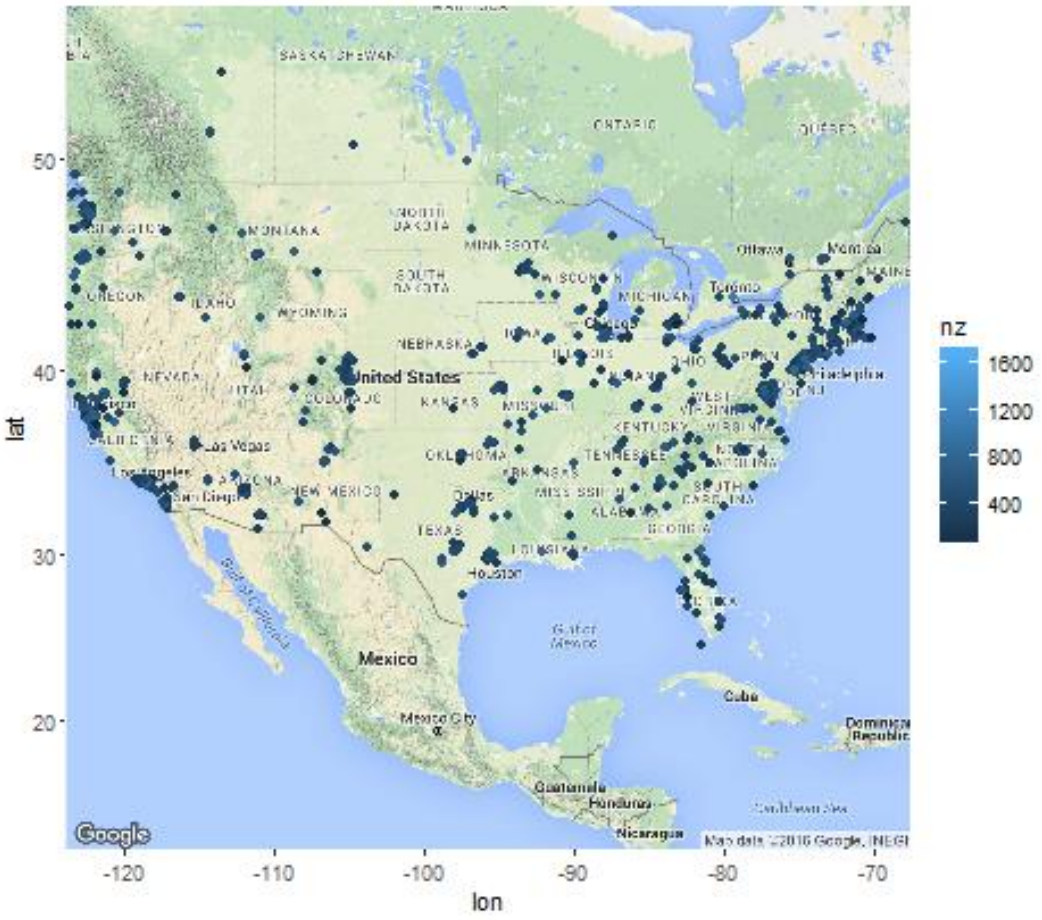


Figure 7: *Seasonal Allergies Map*

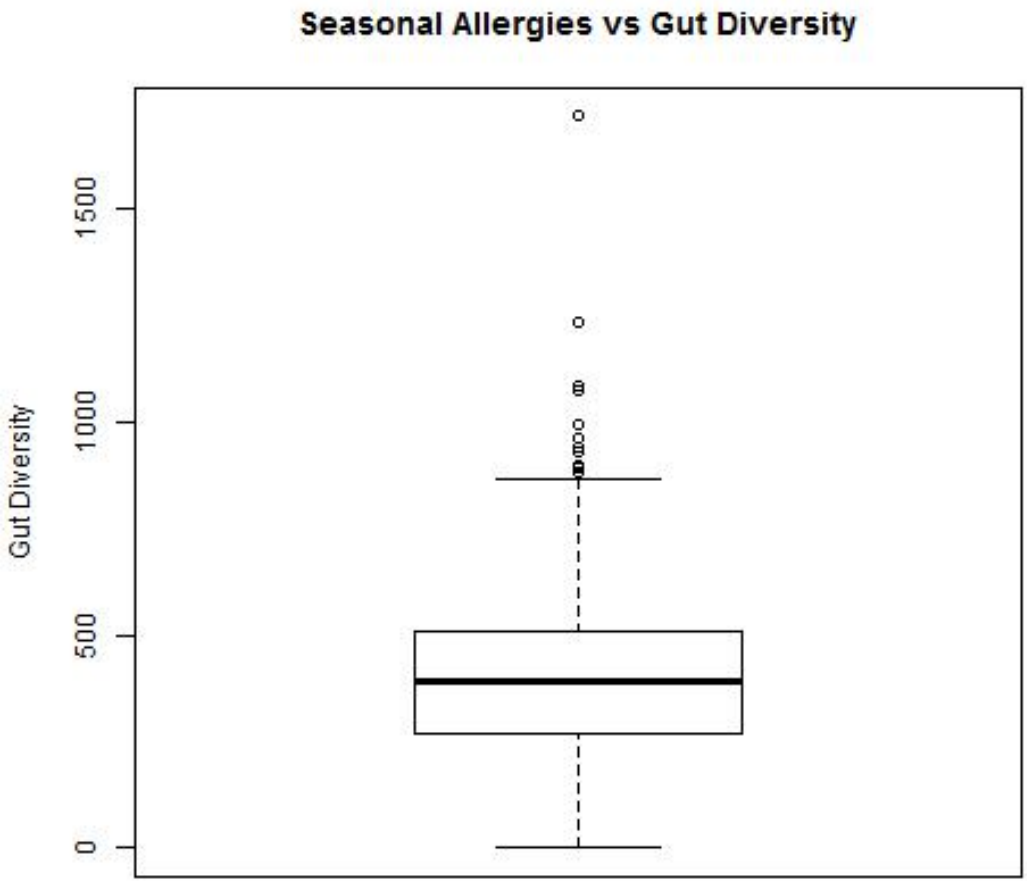


Figure 8: *Seasonal Allergies Boxplot*

A Geospatial Analysis of Light Pollution Density

Mikhail Chowdhury
University of Colorado, Boulder
mikhail.chowdhury@colorado.edu

Abstract

Man-made light pollution has an impact on natural life ranging from biological systems such as plants, animals, and ecosystems, to energy consumption, and impacts the astronomical researcher, to the casual night sky observer. This paper analyzes light pollution in densely populated areas across the globe using night-sky brightness data contributed by citizen scientists during 2010-2015 from the Globe at Night project. We use naked-eye limiting magnitudes (NELMs) as a metric of light pollution to determine the prevalence of light pollution and its relationship to socioeconomic, health, geographic, and environmental factors. Specifically, this paper utilizes World Development Indicator data provided by the World Bank, to study the influence of energy consumption, elevation, life expectancy, population density, and gross domestic product (GDP) have on the light pollution of an area. Furthermore, this paper anticipates to be a catalyst to further research of factors that contribute to significant light pollution in a region.

I. Introduction

Over 100 years ago, if planet earth at nighttime was photographed from space it would look very different from today. Most notably, there would be almost no light pollution to observe. Light pollution is the inappropriate or excessive use of artificial light that can adversely affect our environment, our health, and our energy consumption. Currently, researchers do not understand the direct health implications of light pollution, but what they do comprehend is that light pollution disrupts the 24-hour day/night cycle known as the circadian clock which exists in almost all organisms, including humans. A disruption of this cycle can cause irregularities in biological processes such as cell regulation and hormone production, which can lead to health problems such as insomnia, cancer, depression, and cardiovascular disease [1]. Less seriously, light pollution is also a hindrance to astronomy because it impairs the number of observable stars and objects in our view of the night sky. As a result of light pollution, the night sky over cities can be hundreds of times brighter than a naturally lit starry sky, and can even impact places hundreds of kilometers away from urban development. Today, more than $1/5^{\text{th}}$ of the world's population have lost naked-eye visibility of the Milky Way at night [2].

The impact of light pollution across our globe has inspired the citizen-science project Globe at Night, which tracks night sky brightness where contributors submit their data and observations to a database through a smartphone app that measures light pollution. Given this data, this study examines what socioeconomic, geographic, and population density factors cause a city or location to have more light pollution.

II. Data

The light pollution data used was taken from the Globe at Night Project during the years 2010-2015. All data was recorded by citizen sci-

Year	# of Observations	# of Countries Participated
2015	23,050	104
2014	20,911	103
2013	16,342	89
2012	16,850	92
2011	14,249	115
2010	17,805	101
Total:	**92,865**	**–**

Table 1: *Globe at Night observation data summary.*

entists reporting light pollution observations through the Globe at Night website or smartphone app. To submit an observation, the user begins by matching the appearance of a constellation corresponding to the user's location listed on the Globe at Night website (the smartphone app automatically selects and displays the correct constellation). The user then matches their observation with one of seven constellation magnitude charts and records the amount of cloud cover. Finally, the user records the date, time, and location (latitude / longitude) of their observation (note: the smartphone app automatically records this information).

The Globe at Night project has been collecting user-reported light pollution data since 2006. For the purposes of this study, we decide to focus on more current light pollution observations reported between 2010 and 2015. Each year, the project's organizers provide a map of light pollution levels worldwide. The Globe at Night Project is a program of the National Optical Astronomy Observatory backed by the National Science Foundation. The data sets can be accessed at `http://www.globeatnight.org/maps.php`.

III. Methods

Night sky brightness is measured from the ground using two different quantitative methods. One method is using a Sky Quality Meter (SQM) device that measures night sky brightness in magnitudes per square arcsecond. Another different method is using naked-eye limiting magnitude (NELM) as a metric of sky brightness, where the observer compares the visibility of a constellation to a picture. This method is based on a 1-7 scale where larger magnitudes represent greater visibility see Figure 1.

The Globe at Night Project accepts both measurements. While the NELM metric is less reliable because it is influenced by subjectivity, it is more prevalent in the data set. Only 22.06% of observers reported SQM readings between 2010-2015, whereas 86.93% of observations included NELM measurements. The SQM readings can be converted to NELM measurements with the equation in below, where α represents the SQM reading in magnitudes per arcsecond [3].

$$NELM = 7.93 - 5 * log(10^{4.316-(\alpha/5)}$$

This relationship equation between limiting magnitudes for naked eye-visibility (originally developed for comparing naked-eye visibility to that of a telescope) and night sky brightness was developed by Bradley Schaefer in 1990.

To commence our initial data analysis, we combined each data set through the years 2010-2015 into a single comma separated value file and converted all singly reported SQM readings to NELM to provide a common metric to compare light pollution. The 2010-2015 timeframe was chosen because it provided a reasonably recent data set that had enough observation data (approximately 113,459 observations) spread out

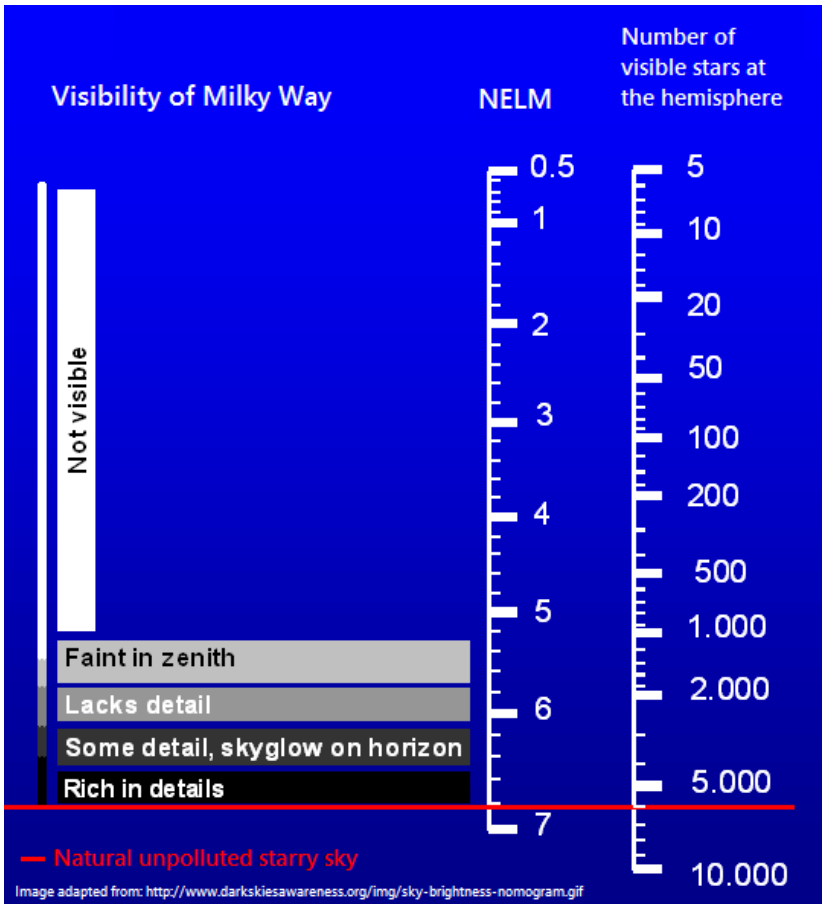


Figure 1: *Naked-Eye Limiting Magnitude (NELM) visibility comparison.*

across most of the world (see Figure 2). We then acquired the *World Development Indicators* data provided by the World Bank (`http://data.worldbank.org/data-catalog/world-development-indicators`), and utilized a database query program to match this data by location to the Globe at Night observation data. This allowed us to examine the relationship between socioeconomic, health, and environmental factors to the night sky brightness.

IV. Results

In combining the Globe at Night Project observation data with the World Development Indicators data provided by the World Bank, we examined the relationship of night sky light pollution with a number of socioeconomic, health, energy consumption, and geographic factors. The most notable relationships are shown in Figure 2 - Figure 7.

Figure 2 shows a geospatial mapping of the maximum reported light pollution in each country. From this visual we were able to recognize that most countries have a minimum of one reported observation. Additionally, the countries are color coded based on the maximum light pollution reported in that country. Although this representation gives an overview of the light pollution, it's usefulness is limited due to the variability in light pollution within a country's territory. **Figure 3** is a United States and Canada focused map, that shows a much

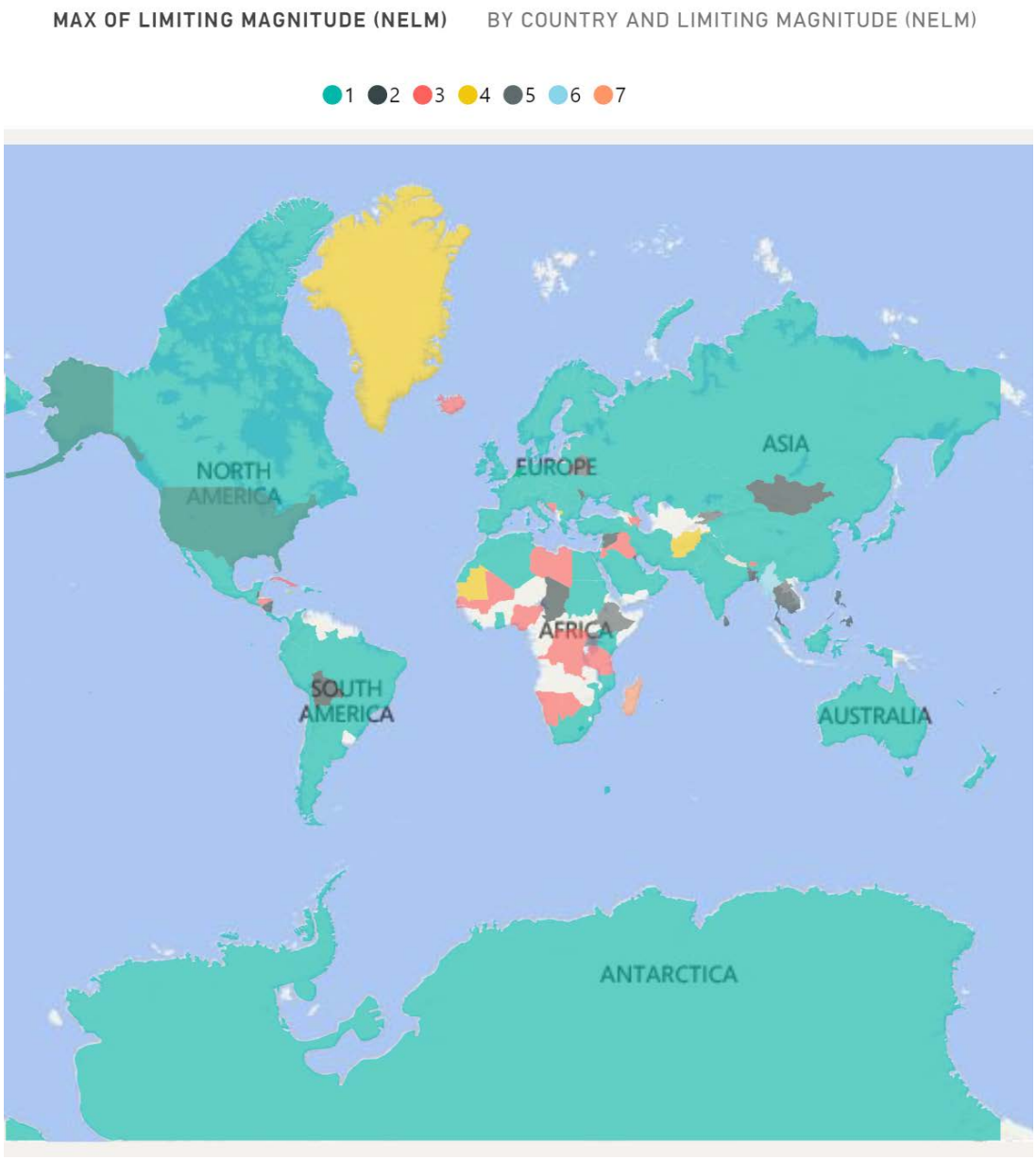


Figure 2: *Maximum Naked-Eye Limiting Magnitude (NELM) observed by country*

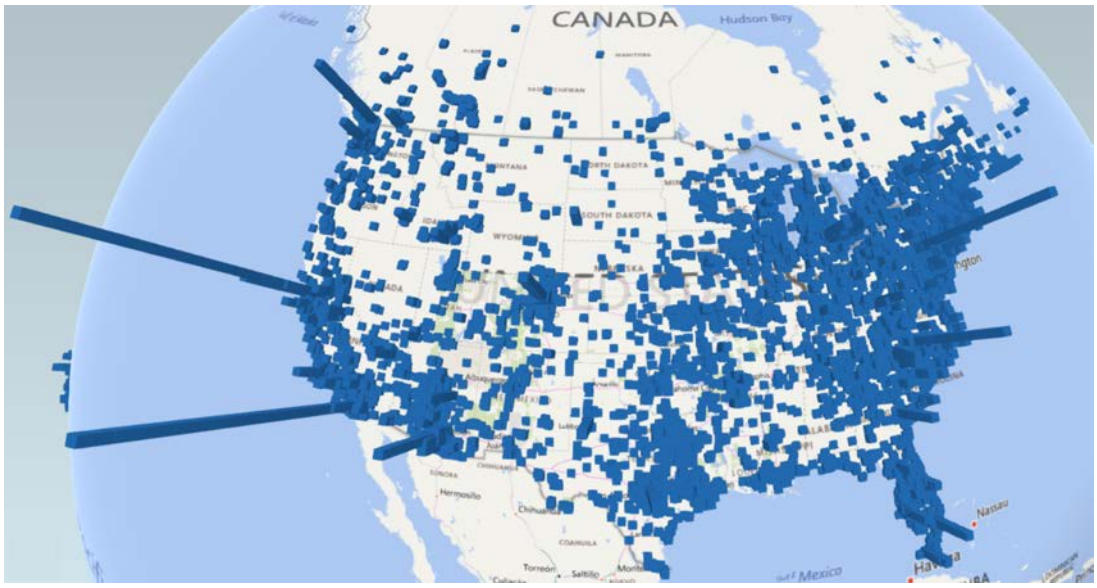


Figure 3: *Map of NELM Observations in the United States and Canada.* ***Note:*** *Longer bars represent greater density of light pollution*

more detailed spread of night sky light pollution, where longer bars indicate greater light pollution. In this representation, it is easily noticeable that cities with high population density, such as New York, San Francisco, Los Angeles, Vancouver, and Denver suffer from the high density of light pollution.

Once we gained a better understanding about the spread of our data, we wanted to see if geographic factors, such as elevation, had an influence on light pollution. In **Figure 4**, a plot showing the average elevation against NELM shows an exponential relationship between the two variables. As expected, as altitude increases, night sky light pollution is diminished due to the decrease in the amount of air particles that scatter natural light from stars and the moon [4]. Confirming this relationship also helped us gain more confidence in the accuracy of the Globe at Night observation data, as all observations are automatically marked with elevation data based off of GPS coordinates. Additionally, this may help to explain why (as seen in **Figure 3**) Denver, which is at significantly higher elevation than a city like Seattle, has less light pollution, despite both cities having similar population densities.

Note: *For Figures 5-7 we match the factors of life expectancy, electric power consumption per capita, and GDP per capita of a country, to the country the NELM observation was taken.*

Figure 5 shows our attempt to determine if to light pollution density relates to health. We decided to choose life expectancy as our factor to test this theory. As interpreted from the box plot, there does not seem be any direct correlation between life expectancy and NELM. However, this does not rule out the possibility that there is a relationship between light pollution density and life expectancy. Many outside variables such as poverty level, prevalence of disease, access to healthcare, and a plethora of other factors most likely have a greater influence on life expectancy than light pollution. In our research, we could not find an easy way to control for these variables in our dataset, and thus could not draw any conclusions from this relationship.

Figure 6 shows the average electric power consumption per capita for the observation lo-

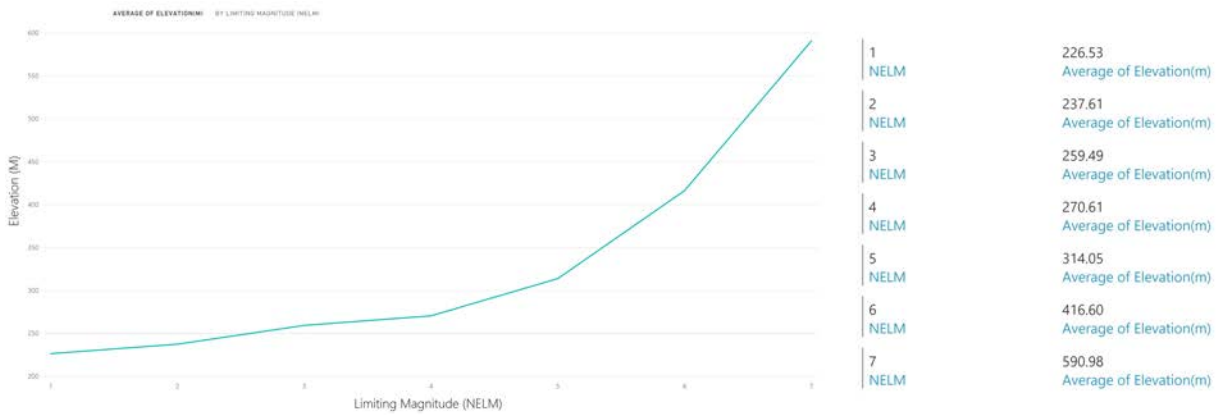


Figure 4: *Plot of Average Elevation of Observations by Naked-Eye Limiting Magnitude (NELM)*

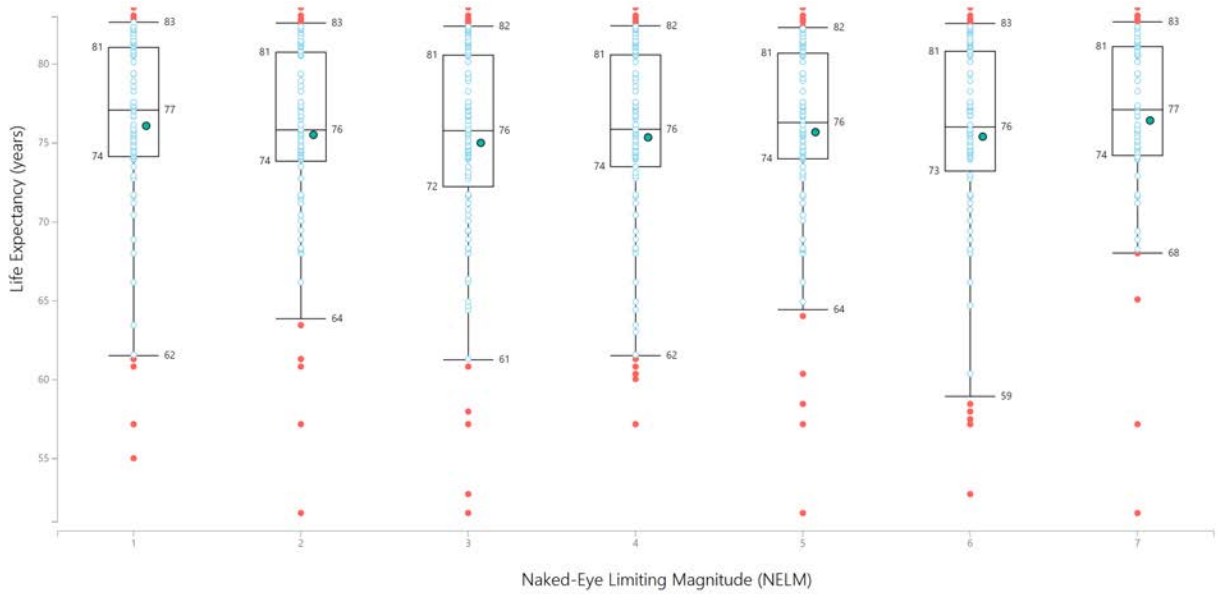


Figure 5: *Box Plot of Life Expectancy (years) by Naked-Eye Limiting Magnitude (NELM)*

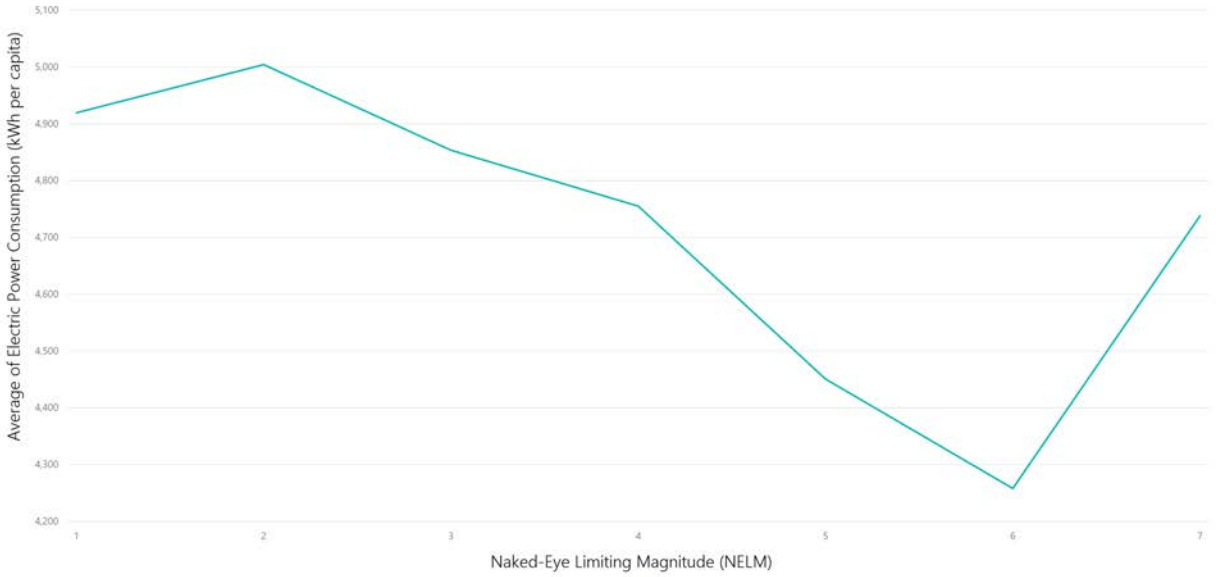


Figure 6: *Plot of Average Electric Power Consumption (kWh per capita) by Naked-Eye Limiting Magnitude (NELM)*

cation of each factor of NELM. This plot shows that there is a strong association of greater light pollution density with energy consumption. Intuitively this is expected, because places with higher light pollution must consume more energy to power light sources that are causing greater levels of light pollution. The spike for energy consumption at a NELM factor of 7 is most likely attributed to locations that have significantly higher daytime energy consumption than at night.

Figure 7 shows a box plot of gross domestic product (GDP) per capita for the observation location of each factor of NELM. GDP is a metric of the economic status of a country. In observing the GDP of countries where NELM observations were made, we were trying to determine if economically successful nations tended to have more light pollution. In our box plot it is apparent that locations with an NELM metric of 1 (highest amounts of light pollution) tend to have higher GDP. Overall, there does not seem to be a strong relationship between GDP and light pollution. As a result, we conclude that GDP has a minor influence on night sky light pollution.

V. Limitations

The greatest source of uncertainty in our conclusions and data sets was that fact that the majority of the Globe at Night data was based on subjective observations to determine the NELM light pollution metric. The best approach to determine light pollution of an area is to observe night satellite imagery of earth and utilize computer vision algorithms to provide a consistent light pollution metric across different locations. However, at this moment we did not have the time or the resources to pursue this option. An additional limitation of the Globe at Night data set, was the variability of the amount of observation data across different geographic locations. Some areas have a large amount of observations data, while other geographic locations were represented by only a few observations. Despite these limitations, the Globe at Night data was subject to a statistical analysis performed by the *American Association of Variable Star Observers* that deemed the observation

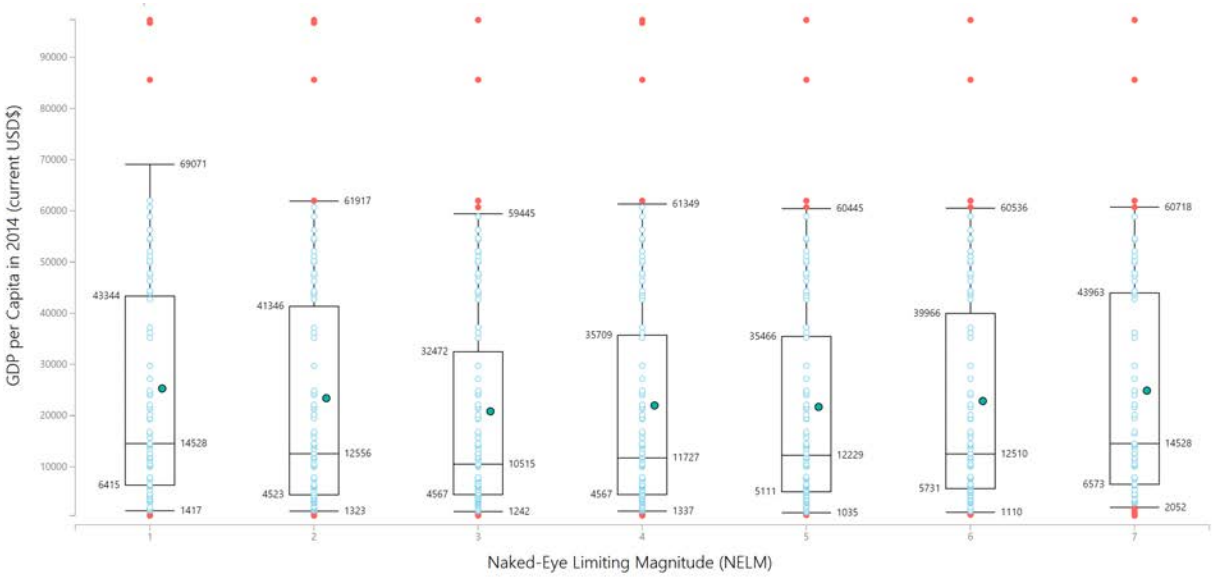


Figure 7: *Box Plot of Gross Domestic Product (GDP) in current USD by Naked-Eye Limiting Magnitude (NELM)*

data significantly reliable for a citizen-science project, however noted that more sophisticated methods of filtering data were needed to determine the success of the data on a local or regional scale [3].

VI. Conclusion

Man-made light pollution has an impact on natural life ranging from biological systems such as plants, animals, and ecosystems, to energy consumption, and impacts the astronomical researcher, to the casual night sky observer. In our research we were able to examine the socioeconomic, health, and geographic factors that contribute to light pollution using data from *World Development Indicators* compiled by the World Bank, and citizen-science driven light pollution observation data from the Globe at Night project. In examining these factors, we determined that geographic locations with higher altitude tend to have greater visibility of the night sky as a result of less natural light pollution. We determined that the geographic locations that have higher energy consumption and higher population density tend to endure more significant light pollution. Lastly, we conducted a surface-level analysis of the relationship between light pollution density of a geographic location and the life expectancy and economic success (GDP) of that particular location. While our results were inconclusive, we still believe that a relationship exists in these factors and hope that further research finds closure on this relationship. Furthermore, we hope that our initial research inspires others to continue to examine additional factors that contribute to significant light pollution prevalence in a region.

References

[1] Chepesiuk, R. (2009). Missing the Dark: Health Effects of Light Pollution. *Environmental Health Perspectives*, 117(1), A20–A27.

[2] Cinzano, P., Falchi, F. and Elvidge, C.D. (2001), The first World Atlas of the artificial night sky brightness. *Monthly Notices of the Royal Astronomical Society*, 328:689–707.

[3] Birriel, J. J., Walker, C. E., and Thornsberry, C. R. (2014), Analysis of Seven Years of

Globe at Night Data. *J. Amer. Assoc. Var. Star Obs.*, 42:219–228.

[4] Dinkel, K., Klein, V., Schuette, S., Truesdale, N., and Zizzi, A. (2010), A Model of Sky Brightness in the Stratosphere.

Predicting and mapping the Northern Lights

Monal Narasimhamurthy

University of Colorado Boulder
monal.narasimhamurthy@colorado.edu

Abstract

Northern Lights (also known as Aurora Borealis) are natural electromagnetic phenomena characterized by the appearance of colored lights in the night sky, usually near the higher latitudes. This paper explores machine learning techniques to predict the Northern Lights. Northern Lights are created when charged solar particles hit the upper layers of the Earth's atmosphere. This geomagnetic activity is calculated in terms of two indices - 24 hour A index, 3-hourly K indices. The National Oceanic and Atmospheric Administration (NOAA) maintains a daily record of the estimated planetary A indices and K indices obtained from a network of magnetometers spread across the western hemisphere. Using a historical dataset of recorded indices between the years 1997 and 2015, a supervised machine learning model is trained to predict the intensity of auroral activity on a given day. An accuracy of 76.16% was achieved by the model.

I. Introduction

Current techniques for predicting the northern lights are based on atmospheric models that make use of solar wind measurements at L1 liberation point(the point between the Sun and the Earth where their gravities are cancelled). The research in the field is still at a very early stage, since space data is hard to collect. NOAA's prediction model uses data from just one satellite, as of now [1]. This paper looks into the novel approach of using machine learning techniques for predicting the northern lights.

Historical data of when the auroras were observed from the Earth's surface is not documented anywhere. Efforts are being made in this direction. The Aurorasaurus project collects citizen reported data via an android application, and uses this to complement their prediction methods [2]. However, this data is proprietary.

In this work, instead of predicting the auroral activity, we predict the K-index. The K-index measures disturbances in the horizontal component of the Earth's magnetic field. It is a number between 0 and 9, with It is derived from the maximum fluctuations of horizontal components observed on a magnetometer during a three-hour interval.Historical data for the estimated planetary K-indices is maintained by NOAA [3].

The geomagnetic disturbances correlate directly with the occurrence of the aurora borealis. Hence, K-index can be used as a proxy for whether an aurora occurred or not.

For instance, a K-index of 0 would correspond to a very weak or non-existent aurora and a K-index of 9 for a major geomagnetic storm with auroras likely in France and even Northern Spain.

The features that most likely influence the K-index are

- Solar Cycle - The solar cycle is the nearly periodic 11-year change in the Sun's activity. The levels of solar radiation and

ejection of solar material changes during the cycle resulting in heightened auroral activity. The solar cycle is measured by scientists through

  - Average number of sunspots
  - Solar radio flux at 10.7 cm

- Time at which k-index is measured
  - Day - Historically, the intensity of the auroras have been observed to increase every 27 days
  - Hour - Auroras are believed to be brightest just before midnight
  - Month - Statistically, the equinox months of September and March are best for aurora activity
- Solar Wind Intensity - The solar wind is a stream of charged particles released from the Sun's atmosphere. Stronger the solar wind, higher the auroral activity

II. Data

National Oceanic and Atmospheric Administration(NOAA) maintains historical data [3] for

- Daily Particle Data
- Daily Solar Data
- Daily Geomagnetic Data

between the years 1994-2016.

The daily geomagnetic data contains Fredericksburg, College, and Estimated Planetary A and K Indices. The daily 24-hour A index and eight 3-hourly K indices from the Fredericksburg (middle-latitude) and College (high-latitude) stations monitoring Earth's magnetic field. The estimated planetary 24 hour A index and eight 3-hourly K indices are derived in real time from a network of western hemisphere ground-based magnetometers. Missing indices are shown as -1.

The Royal Observatory of Belgium, Brussels maintains the Sunspot Index and Long-term Solar Observations(SILSO) repository [4], which records the daily sunspot SWO and RI indices(smoothed), the solar radio flux at 10.7cm and observed daily planetary A-index. A-index is another metric to measure Earth's geomagnetic activity. The data is available for the years 1818-2016.

The solar wind intensity is collected by NASA Advanced Composition Explores(ACE) satellite and the data is maintained by Space Weather Prediction Centre(SWPC), NOAA [5]. ACE measures four, hourly metrics - MAG (Magnetometer) , SWEPAM (Solar Wind Electron Proton Alpha Monitor), EPAM (Electron Proton and Alpha Monitor), SIS(Solar Isotope Spectrometer). The data is recorded between the years 2001-2016.

III. Methods

I. Feature Engineering

Standard correlation tests like Pearson, Spearman etc cannot be applied between the feature variable and k-index variable since we expect the correlation between the variables to be non-linear and non-monotonic.
Instead, it is best to visualize the data to observe if a feature has any correlation with the k-index. If it does not, then the feature is not included in the prediction model.

K-index is measured in a quasi-logarithmic scale. Before doing any calculations, we need to convert it back to the normal scale of real numbers. When converted the numbers in the new scale is called the A-index.

Time at which A-index is measured

Day - Bartels' Rotation Number is a serial count that numbers the apparent rotations of the sun as viewed from Earth, and is used to track certain recurring or shifting patterns of solar activity. Each rotation has a length of

exactly 27 days. A cyclic pattern was observed when the Bartels' Rotation Number was plotted against A-indices for the years 1997-2014 [Figure 1]. The first and the last day in the cycle have the highest values of A-index. The mean A-index for days grouped based on the Bartel's Rotation Number tends to follow a semi-sinusoidal curve.

Month - The equinoctial months of March-April, September-November see a higher lever of auroral activity. [Figure 2]

Hour - The hour of the day has no correlation with the auroral activity. Hence, the hour was not included as a feature in the model.

Solar Cycle

Average number of sunspots - The chance of auroral activity decreases with the increase in the average number of sunspots. [Figure 3]

Solar radio flux at 10.7 cm - Solar radio flux is directly correlated with auroral activity as well. [Figure 4]

Solar Wind Intensity - A strong correlation was found between solar wind speeds and measured geomagnetic activity. Higher the wind speeds, stronger the auroral activity.

II. Prediction

Geomagnetic activity as a proxy for auroral activity

The geomagnetic activity is related to aurora activity as follows,

A-index	Auroral Activity
0 - 7	Quiet
8 - 15	Unsettled
16 - 29	Active
30 - 49	Minor storm
50 - 99	Major storm
100 - 400	Severe storm

Hence, A-index was used as a proxy for auroras to label the data for prediction. An A-index greater than 15 was labelled with 1 indicating the presence of auroral activity. If the A-index was below 15, then absence of auroral activity was assumed.

Classification

A Linear Discriminant Analysis (LDA) algorithm was used to classify the data. The model is used to find a linear combination of features that characterizes or separates two or more classes.

The features in a LDA classifier can be of any type - nominal, ordinal, ratio or interval, which is apt for this use-case. If there are more than 3 features in the model, then the classes gets separated by a hyper-plane.

The classification rule is to an object to a group with the highest conditional probability (Bayes Rule). Thus,

If C = Set of all classes,

$$P[c'|X] > P[c|X], \forall c \in C, c \neq c'$$

then, c' is the assigned class for the object.

The data from the years 1997-2014 was used as the training set. The data from the year 2015 was used as the testing set.

IV. RESULTS

The prediction model gave an accuracy of 76.16% on the test dataset from the year 2015.

Historically, there has been no documentation of whether an aurora occurred on not a given day. This model can be used to fill that void. Knowing when an aurora had occurred is crucial to predicting future auroras.

Figure 1

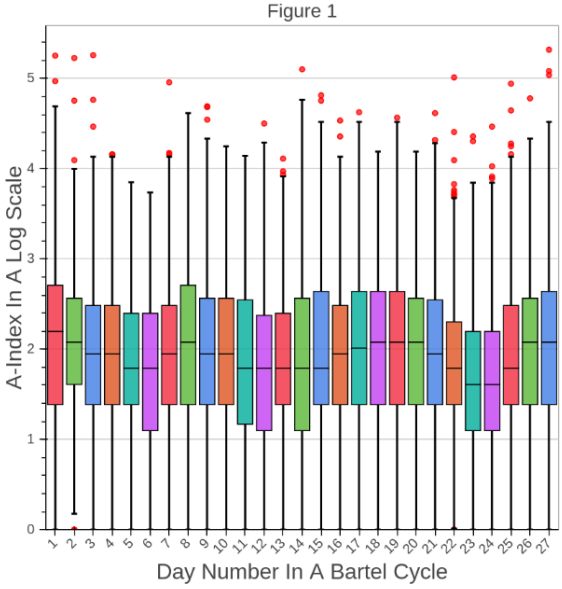


Figure 2

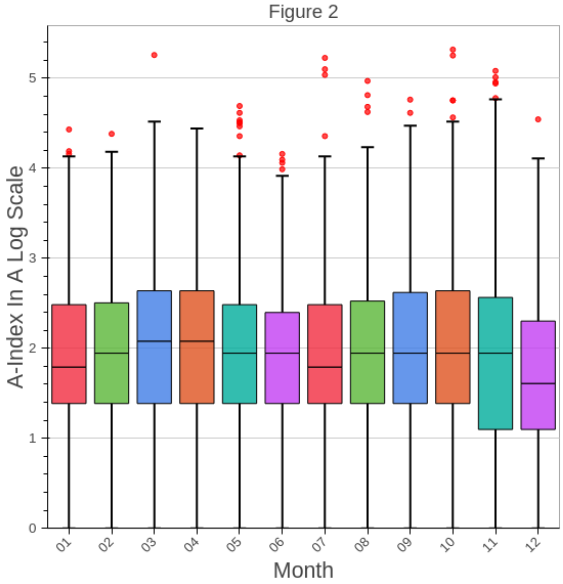


Figure 3

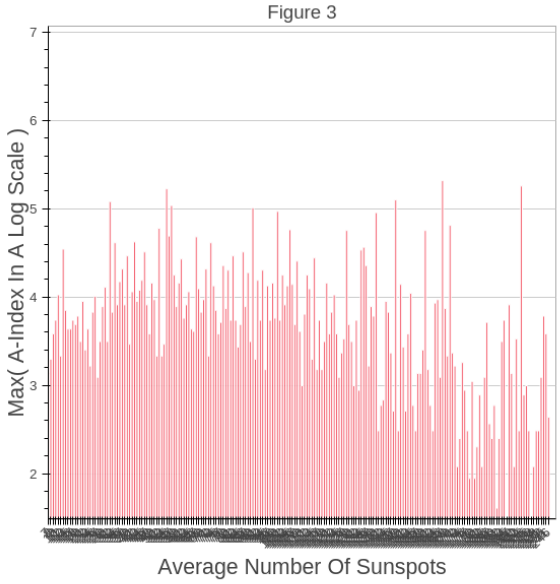
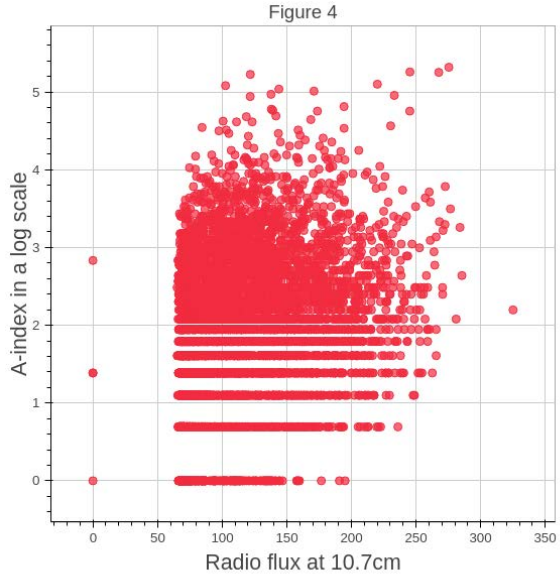


Figure 4



V. Discussion

The model can be used to fill in the gaps in historical auroral observation data.
It can further be used to predict auroral activity in popular tourist spots, which is a booming industry.

References

[1] NOAA Ovation Model http://www.swpc.noaa.gov/products/aurora-30-minute-forecast

[2] The Aurorasaurus Project http://www.aurorasaurus.org/

[3] NOAA geomagnetic data ftp://ftp.ngdc.noaa.gov/STP/GEOMAGNETIC_DATA/INDICES/KP_AP/

[4] SILSO sunspot data http://www.sidc.be/silso/datafiles

[5] ACE solar wind data http://www.swpc.noaa.gov/ace

[6] Scikit http://scikit-learn.org/stable/

An Analysis of Diet Trends and Potential Effects on American Health

Nika Shafranov

University of Colorado
nika.shafranov@colorado.edu

Abstract

This paper aims to look at the correlation between American obesity rates over the years 2004-2012 and three trending diets over that period. In order to determine whether or not the popularity of these diets is significant to the changes in a state or county's obesity rates, we examine possible factors that may bias the results of our observation, taking into account influences such as income, ethnicity, age, behavior, environment and more. Using Google search data from these 9 years alongside obesity statistics and leisure-time activity data obtained from the Center for Disease Control and Prevention (CDC), this paper examines how the obesity rates of areas in which these diets are most popular compare to their surroundings as interest in each diet ebbs and flows. Focusing primarily on changes in physical activity and diet searches of Americans aged 18 or over, we examine whether corresponding changes in the prevalence of obesity across the U.S. affect the changing obesity rates across the nation.

I. Introduction

In today's world of instantly transferred information and shared ideas, the spread of trends across the United States has a significant impact on how people behave. Diet trends are particularly important due to their role in American health; their sudden, rapid expansion from impacting only a few small communities to determining how a large portion of the population chooses to nourish themselves may play a massive role in American culture and overall population health.

In recent years, with the invention of simple large-scale data analysis tools such Google Trends, investigating the spread of ideas across the globe has become incredibly easy. Using this API in conjunction with CDC data on American obesity rates, it is now possible to see how rapidly diet trends can gain popularity, e.g., originating in a small town in Colorado and growing influence until it becomes the most "Googled" diet in the country.

Even so, obesity is on the rise; the percentage of adults citizens with a healthy weight in the U.S. has decreased from 41.6% (1988) to 29.6% (2012), alongside a 12.2% increase in obesity prevalence. [9]

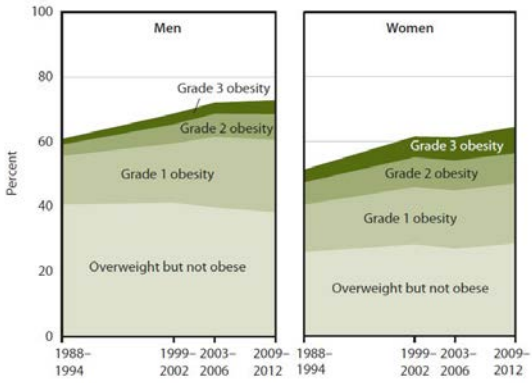


Figure 1: *Rates of overweight and obesity among adults aged 20 and over, 1988-2012*

This is a major health concern, seeing as obesity is responsible for a number of serious diseases. While many Americans may display interest in trying out new diets in pursuit of a healthier lifestyle, it is unlikely that every fad diet sweeping through town will be permanently adopted. In fact, a study performed in 2012 by the NPD Group demonstrated that not only has the percentage of dieting adults dropped to 20%, an 11% decrease since 1992,

but that the number of people reporting following a diet for more than 6 months has slowly begun to drop as well. [1]

One of the primary goals of this project will be to balance the rates of physical activity in an area against the corresponding interest in trending diets. A study published in 2009 examining the associate between insufficient physical activity and the prevalence of obesity demonstrated that after adjusting for age, gender, race and median household income, insufficient physical activity was still a significant predictor of the state's prevalence of obesity. [5] Performed by David Brock, this analysis revealed a strong correlation between rising obesity rates in areas of minimal physical activity. Consequently, this paper will be focusing largely on working around the degree of sedentary lifestyles of various counties, as well as factors such as median income and poverty rates in the area, in order to most accurately examine whether diet trends are significant to obesity prevalence.

II. Data

I. Center for Disease Control and Prevention

Two sets of data were downloaded from the Center for Disease Control & Prevention (CDC). Each came from the CDC website in an Office Open XML Spreadsheet (.xlsx) format and had to be converted to a more widely usable .csv formatting in order to later analyze it using RStudio.

I.1 Obesity Prevalence

A set of county-level obesity rates across the U.S for years 2004 - 2012. There are three introductory categories, comprised of State, FIPS Codes, and County, as well as six categories for each year of data: number, which corresponds to the number of obese residents in each county, and age-adjusted number, which was not used for our purposes. Following these are the lower confidence limit and upper confidence limit for both the number and the age-adjusted number of obese residents.

I.2 Leisure Time Physical Inactivity

A set of county-level obesity rates across the U.S for years 2004 - 2012. There are six categories for each year of data: number, which corresponds to the number of obese residents in each county, and age-adjusted number, which was not used for our purposes. Following these are the lower confidence limit and upper confidence limit for both the number and the age-adjusted number of obese residents.

II. Google Trends

The data on which diets were trending in locations across the U.S. was obtained through Google Trends, a public web facility that provides users with detailed information on how often a particular query is made through the search engine. With this API we were able to select and monitor the ebb and flow of these diets' effects across the states prior to even downloading the raw data, allowing for careful consideration of each diet before choosing the subjects of the study.

Google Trends reports searches on a weighted scale, wherein a score of 100 corresponds to the most searched term in the category. By searching for queries containing the word 'diet,' a list of highest trending diets was provided, including "south beach diet," "hCG," "Atkins" and more. While the hCG diet was among the most popular searches, it was decided best not to include it in the analysis due to the FDA's declaration that the diet was "fraudulent," "dangerous," and "illegal." [7] In order to stay within the scope of the CDC data on obesity rates (measured from 2004-2012) we ended up choosing 3 diets to focus on: the south beach diet, which peaked in 2004 then dramatically lost popularity, the vegan diet, which has linearly gained interest over those years, and the gluten free diet, which has grown exponentially over the course of the past few years.

The query data for each diet was limited to those originating from within the United States, pulling all information from Jan 20014 âĂŞ Dec 2012. Additionally, only searches classified as belonging to the Food & Drink category were viewed in order to primarily focus on those individuals searching for cooking related results or looking to find restaurants in the area adhering to these diet restrictions. A final limitation was placed on the method of searching, allowing only Web Searches to be returned and ignoring all Image, News, Shopping and YouTube searches as those were deemed irrelevant.

When retrieving this data from Google's API, it became necessary to accept some sacrifices in the information extraction and formatting. In order to avoid missing a large amount of data, the search terms "gluten free" and "vegan" were not paired with the word "diet," as this would have resulted in a near 65% loss of total searches relevant to those diets. The South Beach diet proved too difficult to search for without the modifier "diet" due to a variety of searches relating to the location "South Beach Miami" or "South Carolina Beach" or other non-SBD related queries. Thus it became necessary to use the identifying word, "diet," in this selection. To avoid bias that may have come from those merely interested in the diet from a news or entertainment perspective, a Food & Drink limitation was placed on the pool of search history for both the Gluten Free diet and the Vegan diet, allowing only nutrition-based searches to be analyzed. Luckily in the case of the South Beach diet, the results from this limitation proved almost identical to the search "South Beach Diet," therefore it was acceptable to simply use the non-limited data in order to get the maximum amount of searches of this diet.

The raw data from each diet was downloaded separately in a .csv format. Each of the diets came with varying amounts of data, but all three contained:

- Interest over time (score): a popularity score of 0-100 assigned for the query, calculated every week from January 2004 - December 2012.
- Interest over time (rate): a percentage of growth observed to range from -100% - 1000%, calculated every week from January 2004 - December 2012.
- Top subregions for the diet: A list of 51 subregions with the highest rate of queries for the diet.
- Top metros for the diet: A list of 210 or less metro areas with the highest rate of queries for the diet.

An issue that was discovered midway through calculations was that once Google Trends detected a search was growing in popularity at a rate higher than 10x, it would assign that value a peak popularity rating. [1] For any result gaining interest at a rate faster than 9999%, Google's interpreter would register that subject's popularity score as âĂŸ1' and log the remainder of each interest rate in the next available column until the score dropped below this value again. This brought about some incoherent results for the initial analysis due to both a difference in value type (percent & integer) and in size, leading to some frustration before the issue was discovered.

Additionally, the data downloaded from Google Trends seemed arbitrarily assigned for each category; when viewed on their interface (shown in Figure 2), it appeared that during 2004 the South Beach diet was almost twice as popularly searched as the Gluten Free diet.

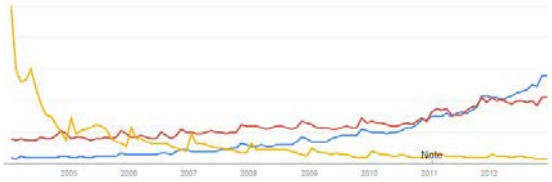


Figure 2: *Comparison of Gluten Free diet (blue), Vegan diet (red), and South Beach diet (yellow) as shown on the Google Trends web plot.*

[1] This may be false, there is no citation. These are merely my assumptions based on the given results.

However upon examining the raw data for popularity growth and interest rates, the results were far from agreeable. Not only did each category's interest rate seem to be calculated using different scales, the data also behaved differently when downloaded in bulk 4 than when pulled individually 3.

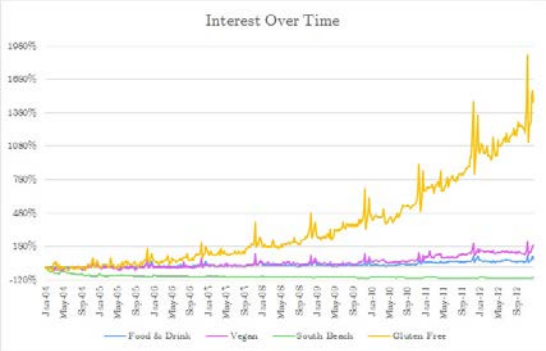


Figure 3: *Comparison of Gluten Free, Vegan, and South Beach diets from January 2004 to December 2012 using data obtained by three individual downloads.*

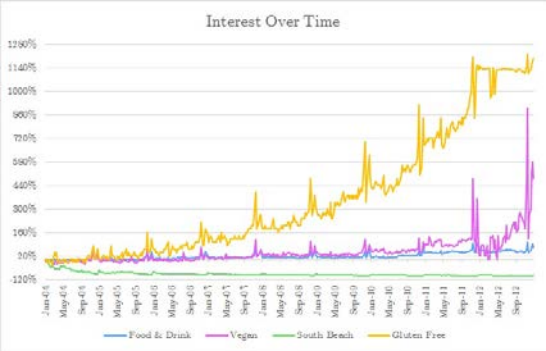


Figure 4: *Comparison of Gluten Free, Vegan, and South Beach diets from January 2004 to December 2012 using data obtained from one simultaneous download of all diets into one .csv file.*

For the purposes of this study, the results pulled from individual searches seemed more reliable than those downloaded in bulk, therefore the data shown in Figure 4 was ignored.

III. Methods

In order to determine which factors to control for in the upcoming calculations, a number of studies were consulted that shed some light on possible causes for changing obesity rates.

A study on the geographic distribution of obesity was released in 2014 by Adam Drewnowski, PhD.; concentrating his evaluation on the citizens of King County, WA, Drewnowski examined the geographic concentration of adult obesity prevalence by census tract in relation to both social and economic factors. [6] This study found that the spatial pattern of obesity was non-random, and revealed that home values and college education were more strongly correlated with obesity than household incomes. Unfortunately due to time constraints this project was not able to examine how this result held out on a larger scope, but it will be an area of interest if we perform further analyses on this subject.

I. Excel

Microsoft Excel proved to be an invaluable tool for data analysis. Due to the high volume of data and the scope of the initial project, it seemed impractical to attempt any manipulation in RStudio before the foundation of the analysis was laid out in a more manageable format.

II. Google Trends

Due to a major oversight on my part, a particularly disappointing last-minute discovery was made, finding that Google's trend data did not contain any records that were both time-stamped and location sensitive left this project sadly short of its original goal to examine the spatial correlation between diet trends and changing obesity rates.

However, despite the data not being offered for download it was still freely displayed in a purely visual map-based format on the web application (shown in Figures 5 and 6). While this prevented any serious calculations and comparisons of the datasets, it did provide a

rough outline of where each diet was gaining popularity at what times.

Figure 5: *Google's geospatial display of the Gluten Free diet's popularity, January 2004. A lack of blue shows only slight interest in this trend during this month*

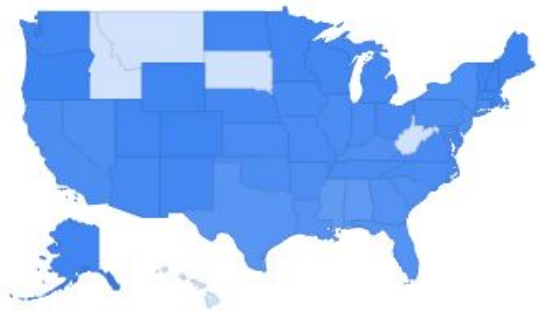


Figure 6: *Google's geospatial display of the Gluten Free diet's popularity, October 2012. The bold blue across most of the states shows a high interest in this trend during this month*

IV. Results

Unfortunately, due to a lack of corresponding datasets the results of the original experiment were inconclusive. With additional time and some practice with QGIS, it may be possible to analyze the movement of these trends using only Google's visual dataset.

[2]determined by the glycemic index

V. Glossary

Body mass index (BMI): An individual's weight in kilograms divided by the square of height in meters. A high BMI can be an indicator of high body fat, but is not diagnostic of body fat percentage or health of an individual. In adults, overweight and obesity ranges are calculated based on their BMI due to the high correlation between the amount of body fat and an individual's index.

- An adult with a BMI between 25 and 29.9 is considered overweight.
- An adult with a BMI of 30 or higher is considered obese. [8]

South Beach Diet: A commercial weight loss diet developed by Arthur Agatson, MD, and promoted in his best-selling 2003 book. It prescribes a balanced, nutrient-dense diet of lean protein, high-fiber, low-glycemic, good carbs [2] and unsaturated fats. [3]

Gluten Free Diet: A diet that excludes the protein gluten, which is found in grains such as wheat, barley and rye. It is primarily used by those with celiac disease, who suffer from inflammation in the small intestines when eating foods containing gluten, as this diet enables them to control their symptoms and prevent disease-related complications. The gluten free diet restricts intake of foods containing barley, rye, wheat, triticale, food additives, food starch, medications and vitamins using gluten as a binding agent, and certain grains contaminated with wheat during growing and processing stages of production. [2]

Vegan Diet: A diet that expands on vegetarianism (removing meat, fish and poultry from the diet), restricting both intake and use of animal products and by-products such as eggs, dairy products, honey and more. A healthy vegan diet includes fruits, vegetables, leafy greens, whole grain products, nuts, seeds and legumes. Sources of protein include lentils, chickpeas and tofu.[4]

References

[1] Dieting is at an all time low. *National Eating Trends*, 2012.

[2] Gluten free diet. *Healthy Lifestyle: Nutrition and healthy eating*, 2014.

[3] South beach diet. *Healthy Lifestyle: Weight loss*, 2014.

[4] What is a vegan? *Veganism in a Nutshell*, 2014.

[5] Luma Akil and H Anwar Ahmad. Effects of socioeconomic factors on obesity rates in four southern states and colorado. *Ethnicity & disease*, 21(1):58, 2011.

[6] Adam Drewnowski, Colin D Rehm, and David Arterburn. The geographic distribution of obesity by census tract among 59 767 insured adults in king county, wa. *International Journal of Obesity*, 38(6):833–839, 2014.

[7] US Food, Drug Administration, et al. Hcg diet products are illegal. *FDA Consumer Health Information/December*, 2011.

[8] Center for Disease Control and Prevention. Disability and obesity. 2016.

[9] National Center for Health Statistics (US et al. Health, united states, 2014: With special feature on adults aged 55–64. 2015.

Predicting transportation modes through GPS logs

Sachin Muralidhara

University of Colorado, Boulder
sachin.muralidhara@colorado.edu

Abstract

This paper presents a method for predicting a person's mode of transportation using a GPS trajectory. Predicting a person's mode of transportation here is based on a supervised machine learning approach. The system considers a user's mean velocity and mean speed as primary features to learn and predict the mode of transportation for an unlabeled GPS log. The paper also studies the different kinds of classification models such as Support Vector Machine, Logistic Regression, Ensemble Methods like Gradient Boosting and Random Forest and their performance in making the prediction. Further, the paper studies the people's movements and establishes hotspots in terms of latitude and longitude based on the GPS travel logs using the k-means clustering technique. The dataset was collected by Microsoft Research Asia and is represented by a sequence of time-stamped points which consists of longitude, latitude, altitude, start-date and end-date. The system developed has an 81.2% accuracy when velocity is used as a feature and 78.91% accuracy when mean speed is used as a feature, with Gradient Boosting outperforming the other classifiers when average velocity is used as a feature and Support Vector Machine outperforming the other classifiers when mean speed is used as a feature. This is a multi-class classification problem, with the modes to predict being walk, run, airplane, car, bus, train and bike.

I. Introduction

Understanding a user's mobility can help find answers to a rich set of questions such as: Do people from a similar geographic area exhibit similar travel patterns? Can knowledge of other people's travel modes help create better a transportation framework in that locality? What mode of transportation is better suited for a particular area? Also, by knowing the locations where people frequently travel, better infrastructure can be created around that location which is advantageous for both the users as well as the establishments that are involved in creating that infrastructure.

The main features that are crafted here in making a prediction are the average velocity and mean speed of a user's travel log. The user's GPS log does not explicitly provide us with speed at which the distance is covered. We make use of the starting time and ending time given in the labeled travel log to calculate the duration of the travel in seconds. The labeled travel log provides us the starting coordinate and the ending coordinate along with the start and times of the journey. To calculate the total distance covered by a user, between a pair of gps coordinates, we make use of the haversine formula to calculate the great circle distance between two points on the Earth.

Prior work has been done in predicting modes of transportation for this particular dataset. [5] Zhen et al. partition the GPS trajectories into segments, based on change points and then extract features from each segment. They also use a density-based clustering algorithm to group the change points of each user and group these change points into clusters. They then build a graph using the generated clusters.

II. Data

The dataset being used is named as the Geolife GPS trajectories[6][7][5] dataset, maintained by

Microsoft Research. The GPS logs were collected for 182 users in a period of over five years. The GPS logs in this dataset is represented by a sequence of time-stamped points. Each time-stamped point has a set of six attributes - latitude, longitude, altitude, number of days passed since 12/30/1989, Date (as a string), Time (as a string). Out of the 182 users, 69 users have transportation mode labels, which denotes the possible transportation modes undertaken by each user. The possible transportation modes are: walk, bike, bus, car, subway, train, airplane, boat, and run. The labels also have a timestamp associated with them, denoting the start and end time for each mode of transport. The date/time have all been converted to GMT.

III. Methods

In this paper, we study the problem of identifying mode of transportation an observation belongs to, as a classification problem and identify the co-ordinates where the user activity is thriving using a clustering technique.

I. Classification

The labels file present in the dataset is used to train our model. The first feature is the average velocity denoted as an object's speed in a particular direction. Given a start time and an end time, the difference between the gps co-ordinates that correspond to the start and end times gives us the displacement over time. This is calculated using the Haversine[4] formula.

Haversine Formula:

$$hav(\tfrac{d}{r}) = hav(\varphi_2 - \varphi_1) + \cos(\varphi_1)\cos(\varphi_2)hav(\lambda_2 - \lambda_1)$$

- hav is the haversine function
- d is the distance between the two points
- r is the radius of the sphere
- φ_1, φ_2 is the latitude of point 1 and latitude of point 2, in radians
- λ_1, λ_2 is the longitude of point 1 and longitude of point 2, in radians

Given a particular start date and end date, the number of seconds elapsed since the epoch time is calculated and the difference between the end time and the start time gives the duration of the journey undertaken in seconds. The average velocity associated with the mode of transport is then calculated as the total displacement over a period of time divided by the difference between the final time and initial time.

$$\text{Average velocity} = \frac{x_f - x_i}{t_f - t_i}$$

where x_f is the final position (final lat/long pair), x_i is the initial position (initial lat/long pair), t_f is the final time and t_i is the initial time, both calculated in seconds.

The second feature being used is the mean speed, which is a measure of the distance travelled in a given period of time. Given a trajectory for each user, the speed between successive GPS points is calculated. The results of the speed between successive points are added together and then divided over the number of points appearing in that trajectory.

This is done for 69 users, who have labels associated with their GPS trajectories. Once the average velocity and mean speed is calculated for each user, the classifier is trained, individually on the average velocity and mean speed. However, there are prominent outliers for each mode of transportation. This might possibly be due to a weak GPS signal, which could disorient the location accuracy.

Once our classification has been done, we observe that an ensemble method like Gradient Boosting[3] performs better than most of the other classifiers.

Gradient Boosting

Usually, in supervised learning, there is a label (output variable y) and a vector of input variables (x). These are connected by a joint probability distribution. The goal is to find

an approximation to a function such that the expected value of a particular loss function is minimized.

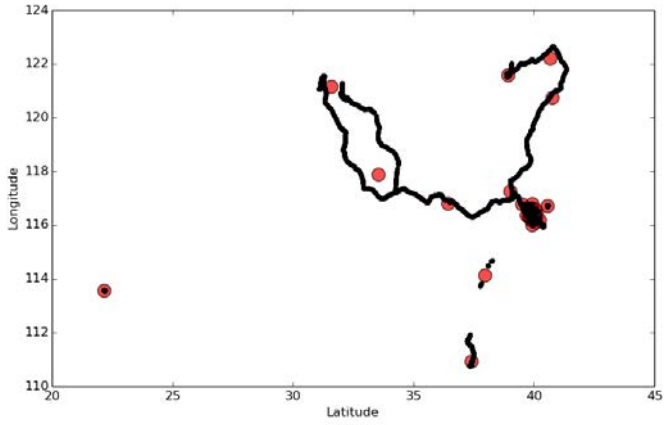
Gradient Boosting[1] is a combination of the Gradient Descent, which tries to find the local minimum for a given function and Boosting, where an additive-model is built in a forward stage-wise fashion. That is, a combination of weak learners are combined to form one strong learner.

In order to average out the biases and to be less susceptible to noise, we use this method. We hypothesize that a combination of models is less susceptible to noisy data than a single model. On observation, we find that there are prominent outliers with respect to the average velocity and mean speed for each mode of transportation. For example, the average velocity for *walk* is in the range 0.2 to 0.4 m/s. However, there are quite a few data points which have an average velocity ≥ 5 m/s. To be less susceptible to noisy data like these and to average out the biases, if there are any, we use Gradient Boosting.

II. Clustering

In order to determine the co-ordinates where the user activity is thriving, we perform cluster analysis. This analysis is done using the k-means clustering algorithm[2]. The aim here is to find the best division of 1,845,936 GPS points, across 10 users, into 100 groups (100 clusters), where each group represents co-ordinates where the user activity is thriving.

The latitude and longitudes are converted into a two dimensional array. Then, we describe variables for clustering, the variables being the number of clusters, the number of iterations to be performed and having a random distribution of the actual data points. We create 100 clusters, since we are trying to find the best division of points and then perform 75 iterations on the coordinates data using the *k*-means clustering algorithm. We first normalize the data by dividing each of the feature (latitude, longitude) by the standard deviation across all the points. We observe that normalizing skews the clustering process and satisfactory results cannot be extracted. As a result, we use the un-normalized data to cluster the coordinates. By doing so, the cluster centroids give us the co-ordinates where the user activity is thriving.



The gray scale darkness in the above plot indicates the number of points that are stacked on top of each other. To a large extent, the gray scale darkness indicates the travel pattern for a set of users over a period of time. The cluster centroids represent the data as a whole. The cluster centroids can be viewed as points where most of the user activity is concentrated.

IV. Results

I. Classification Results

The prediction is done for 5 transportation modes, namely *walk*, *run*, *airplane*, *car*, *bus*, *train* with average velocity as a feature. The results (in terms of accuracy) were obtained with **Hold One Out** cross-validation.

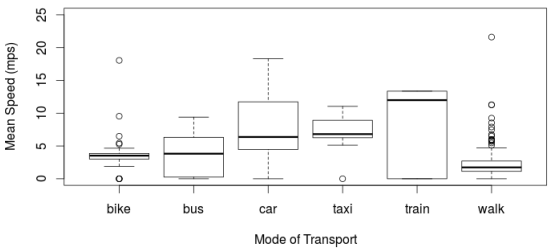
- Logistic Regression with l1 penalty - 80%
- Gradient Boosting - 81.2%
- Support Vector Machine with rbf kernel - 79.41%
- Support Vector Machine with linear kernel - 81.2%
- Random Forest Classifier - 70.58%

By taking mean speed as a feature, the prediction has been done for 6 transportation modes, namely $walk, car, bus, bike, train, taxi$. The results (in terms of accuracy) were obtained with **Hold One Out** cross-validation.

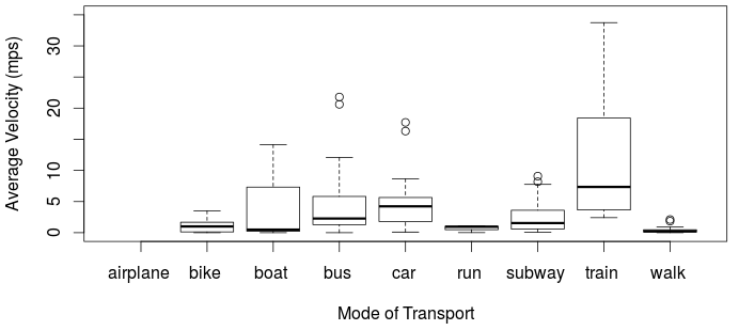
- Logistic Regression with l1 penalty - 75.9%
- Gradient Boosting - 78.12%
- Support Vector Machine with rbf kernel - 78.91%
- Support Vector Machine with linear kernel - 66.26%
- Random Forest Classifier - 69%

II. Visualization with Plots

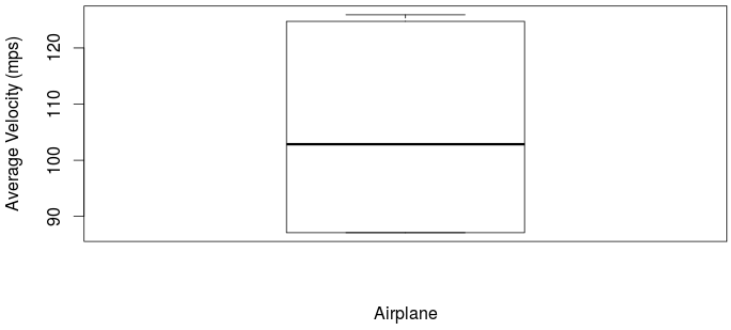
The first plot is that of mode of transport against mean speed. We see the median for cars and taxis are almost identical. However, the surprising find is that the median for bikes and bus is identical as well. A probable reasoning behind this could be that - most of the coordinates come under the populous cities in China, them being Beijing and Shanghai. The congestion in traffic is extremely high in these cities and the traffic is slow moving as well. This might be a good reason as to why the median for bikes and bus is identical. Also, these values being identical can also be the reason as to why some data points get misclassified by our algorithm



The second plot is that of mode of transport against average velocity. From this plot, we can see that the median for bikes and subway is almost identical. The average velocity for walk is about 0.2 m/s and train is about 13 m/s. Airplane has been excluded in this plot since it has very high average velocity.



The third plot is that of mode of transport (airplane) and average velocity. We can come to know that the average velocity for an airplane is greater than 100 m/s.



V. Conclusion and Future Work

We study a big dataset maintained by Microsoft Research[7][6][5] to predict transportation modes using GPS trajectories and also study locations where most of the user activity is concentrated. We find that using average velocity as a feature helps us predict better,

the mode of transportation a user undertakes, given his GPS trajectory. Also, we postulate that performing a cluster analysis on a huge set of GPS points will yield us locations where the user activity is thriving.

In the future, we aim to incorporate these features, along with maximum speed to see if they can help in making better predictions. We also intend to use different clustering techniques, such as DBSCAN and Mean-Shift to compare the results, with that of k-means.

References

[1] Scikit. Machine learning library for python. `http://scikit-learn.org/stable/`.

[2] Scipy. Python. `http://docs.scipy.org/doc/scipy/reference/cluster.vq.html`.

[3] Wikipedia. Gradient boosting. `https://en.wikipedia.org/wiki/Gradient_boosting`.

[4] Wikipedia. Haversine formula. `https://en.wikipedia.org/wiki/Haversine_formula`.

[5] Yu Zheng, Quannan Li, Yukun Chen, Xing Xie, and Wei-Ying Ma. Understanding mobility based on gps data. In *Proceedings of the 10th International Conference on Ubiquitous Computing*, UbiComp '08, pages 312–321, New York, NY, USA, 2008. ACM.

[6] Yu Zheng, Xing Xie, and Wei-Ying Ma. Geolife: A collaborative social networking service among user, location and trajectory. *IEEE Data Eng. Bull.*, 33(2):32–39, 2010.

[7] Yu Zheng, Lizhu Zhang, Xing Xie, and Wei-Ying Ma. Mining interesting locations and travel sequences from gps trajectories. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 791–800, New York, NY, USA, 2009. ACM.

Demographic Access to Wireless Security

Satchel Spencer
University of Colorado Boulder
satchel.spencer@colorado.edu

Abstract

Early Wireless networks were either entirely unsecured or used protocols with serious flaws. Despite this insecurity, wireless local area networks or WLANs have become ubiquitous, simply because of their low cost and convenience, especially in consumer networking. The past 15 years has seen a transition towards more effective and widely implemented wireless security, but many insecure legacy systems are still in use today. As of 2016, in New York City and its surrounding areas, low income demographics as a whole have a considerably more secure wireless landscape. Excluding the richest demographics, Mean Income has a positive correlation with the fraction of insecure networks. The highest income areas appear to have an insignificant correlation with wireless security, though they are still less secure than the poorest areas.

I. Introduction

Wireless networking is a now ubiquitous part of our communications infrastructure. Privacy is one of the key challenges in computer networking. When compared with their wired equivalents, wireless networks are inherently less secure [Singh, 2014]. The first popular wireless security implementation was WEP — Wired Equivalent Privacy. As discovered by Fluhrer, Mantin, and Shamir, WEP was vulnerable to attack due to misuse of an Initialization Vector for the RC4 stream cipher [Stubblefield, 2001]. In the following 15 years there was an industry-led transition to more modern protocols such as WPA 2. This transition period saw more flawed protocols, hesitant adoption and little consumer awareness [Bhagyavati, 2004].

Early on, only newer access points and consumer devices were compatible with WPA, justifying a potential for inequality among socioeconomic groups. Even today, many WEP, WPA and unsecured networks are still in use. As protecting the ever growing flood of personal data will only become more important, understanding how demographics affect the adoption of new technologies will be essential in responding to future security threats.

II. Data

I Sources

The primary data source for this analysis of locations and meta-data of wireless APs is wigle.net [Wigle, 2016]. Between 2001 and 2016 Wigle has amassed database of 250 Million unique APs. Unfortunately, Wigle does not make their dataset publicly available and has numerous mechanisms to discourage large scale scraping. With this in mind, the scale of the analysis was reduced to a square section New York City and surrounding areas (40.830, -74.049 to 40.679, -73.862). NYC is a convenient location due to its high population density, excellent data coverage, and very heterogeneous demographics.

Demographic data for New York was gathered from the American Housing Survey 2013 [AHS, 2013]. The AHS is available at a smallest scale of ZIP Code Tabulation Areas — ZCTAs. There are 124 ZCTAs in and around NYC included in this study.

II Collection

Initially, the scraping of Wigle's database was implemented using a 10 by 10 grid of 2 kilometer square sample locations. Due to Wigle's

restrictions on scraping, only 100 APs could be retrieved at a time. This proved to be an ineffective sampling technique because the results were sorted geographically, yielding a non-uniform spatial sampling distribution. Since the American Housing Survey is available for ZCTAs,the chunks of APs were instead randomly sampled within he confines of each ZCTA. Since the number of samples within each ZCTA is held equal and the sampling does not include all APs within a given area, this approach does not provide an absolute measure of the number of networks. However, it is sufficient to examine the properties of the networks as a group, within a given ZCTA.

III. Methods

I Exclusion

For the purposes of this analysis, insecure networks are those recorded as using no security, WEP, and WPA. Although WPA fixes WEP's key scheduling flaw, it has its own vulnerabilities in PSK mode [MacMichael, 2005]. Only APs implementing WPA2 will considered secure. Approximately 10% of the observed networks had unknown security and were excluded entirely. Unsecured networks with non-unique SSIDs — Service Set Identifiers, and those associated with known vendors of public access points (xfinity, optimumwifi, etc.) were excluded in calculations of the overall securty of a region, since the Wigle metadata cannot distinguish between them and an unsecured network. Wigle also includes an optional *paynet/freenet* flag that helps to identify these public networks.

II Variables

Using the above definition of "secure", the access points were aggregated by ZCTA. Each access point included the following metadata:

- *ssid:* network name
- *unique:* unique, and non-vendor network
- *secure:* uses WPA2
- *lastupdt:* time of original observation

These, along with AHS data were used to compute the variables for each ZCTA, in an arbitrary time frame:

- *insecure:* proportion of insecure networks within the ZCTA
- *unique:* proportion of unique (personal/commerical) networks within the ZCTA
- *income:* mean household income
- *households:* number of households

III Cluster Analysis

Initial analysis of the ZCTAs mean income, AP insecurity and uniqueness, revealed apparently separate groups with vastly different behaviors. In order to adequately analyze these independently, the ZCTAs were clustered based on income, security, household count and uniqueness. K-means clustering from R's cluster package was used [Maechler, 2015, R Core Team, 2015]. All following analysis was done on the resulting clusters individually, or comparatively.

IV. Results

I Clusters

Experimentation with cluster counts and their resulting variance revealed 3 to be the optimal number of clusters. Figure 1 shows the resultant clusters. Though the clusters take into account income, insecurity, households and uniqueness, they divide more or less along income, separating the ZCTAs into low, medium, and high income.

As seen in Figure 2, the low and medium income clusters have markedly different distributions of insecure networks, with low income ZCTAs having a mode of approximately 12% , and those with medium income at 23%. Interestingly, the high income cluster has a similar distribution to that of low income, though there are an insufficient number of ZCTAs in this group to be significant.

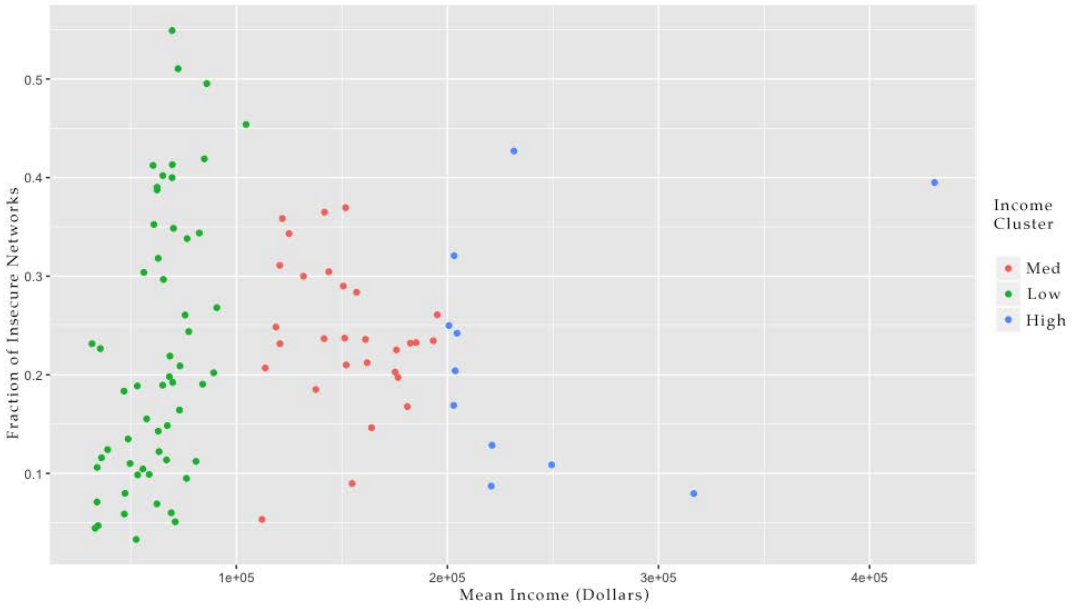


Figure 1: *Resulting clusters by Mean Income and Fraction of insecure networks*

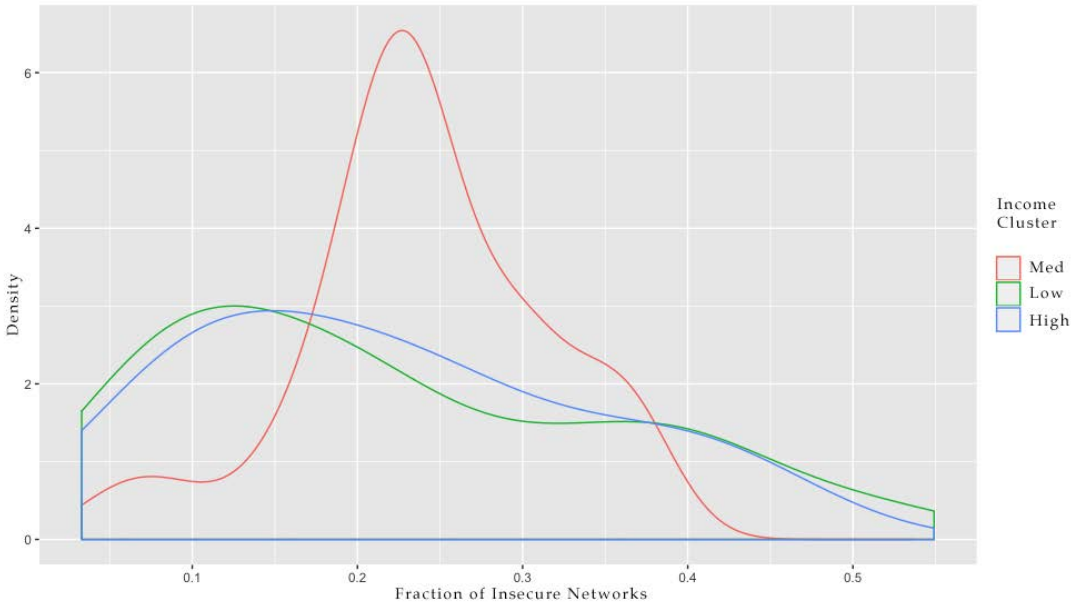


Figure 2: *Distribution of network insecurity in ZCTAs colored by cluster*

II Bivariate Analysis

In the low income cluster, Spearman rank correlation coefficients were positive and significant when comparing ZCTAs' mean income to the fraction of insecure networks. Likewise, linear regression supported this significant, but relatively weak relationship. See Table 1, Figure 3.

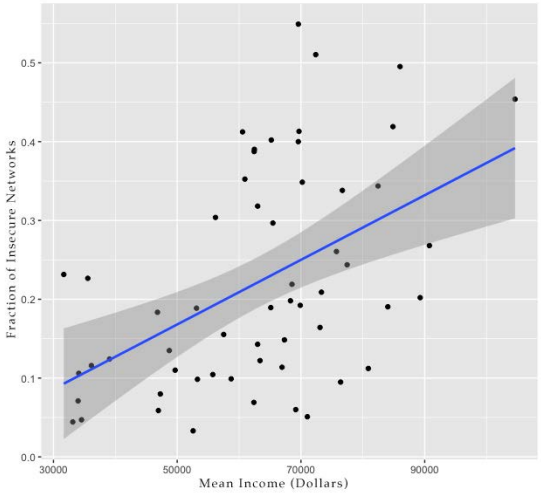


Figure 3: *Mean income vs network insecurity with linear regression, for low income cluster.*

In the medium income cluster, with mean yearly incomes of approximately $100,000 and above, the trends are insignificant, with $p > 0.05$ for both Kendall correlation tests, and linear regression.

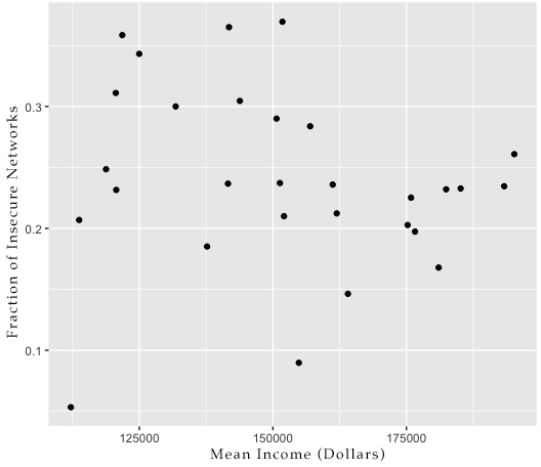


Figure 4: *mean income vs network insecurity, for upper income cluster.*

Table 1: *Analysis on mean income and fraction of insecure networks, in the low-income cluster*

Test	Result	P-value
Kendall	$\tau = 0.485$	$\ll 0.01$
Linear Regression	$r^2 = 0.253$	$\ll 0.01$

V. Visualization

As a supplementary tool for analyzing the Wigle data alongside demographic information, I created a browser-based visualizer that displays the mean income and security of wireless networks from 2004 to 2016 [Spencer, 2016]. The visualizer can filter based on both mean income and observation date. The filters adjust the aggregation of the APs in real time and produce a choropleth, histogram and scatter plot of the desired data.

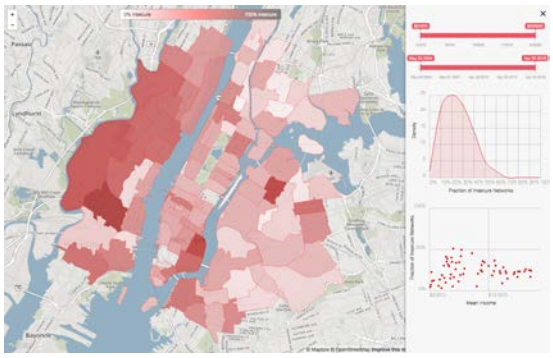


Figure 5: *Browser-based visualizer*

VI. Discussion and Conclusions

It seems clear that the very poorest demographics have fewer insecure networks than any other group. Most likely, this is a result of legacy wireless systems that were economically unavailable in low income areas until the more modern protocols were already implemented. The highest income areas similarly have more secure networks, presumably because those with sufficient income will upgrade their technology as soon as it becomes available.

As a result of Wigle's data collection limitations, the scope of this study was relatively

restricted, spatially and temporally. The vast majority of the observations were recorded after 2011, but cursory analysis has hinted towards interesting trends in the transition between technologies over time as it relates to an area's demography. Given more dense temporal data, a future study could examine the relationship between demographics and the state of wireless security over time.

References

[AHS, 2013] American Housing Survey 2013 *U.S. Census Bureau http://www.census.gov/programs-surveys/ahs.html*

[Bhagyavati, 2004] Wireless Security Techniques: An Overview. *InfoSecCD '04 Proceedings of the 1st Annual Conference on Information Security Curriculum Development*

[MacMichael, 2005] MacMichael, John (2005). Auditing wi-fi protected access (WPA) preshared key mode. *Linux Journal archive. Volume 2005 Issue 137, September 2005, Page 2*

[Maechler, 2015] Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K.(2015). cluster: Cluster Analysis Basics and Extensions. *R package version 2.0.3.*

[Singh, 2014] Singh, Prashant, Mayank Mishra, and P. N. Barwal (2014). Analysis of security issues and their solutions in wireless LAN. *Information Communication and Embedded Systems (ICICES), 2014 International Conference on. IEEE, 2014.*

[Spencer, 2016] Wireless Security Visualizer. Web based visualizer for income and wireless security data in NYC *http://elldev.com/gsa/vis/*

[Stubblefield, 2001] Stubblefield, Adam (2001). Using the Fluhrer, Mantin, and Shamir Attack to Break WEP. *ATT Labs technical report*

[R Core Team, 2015] R Core Team (2015). R: A language and environment for statistical computing. Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

[Wigle, 2016] Wigle.net. Maps and database of 802.11 wireless networks, with statistics, submitted by wardrivers, netstumblers, and net huggers. *https://wigle.net/*

Geoanalysis of the 2016 US Presidential Elections using Twitter Data

Saurabh Sood
University of Colorado Boulder
saurabh.sood@colorado.edu

Abstract

This paper attempts to analyze the 2016 United States Presidential Elections from a geospatial perspective. Analysis is done on a collection of Geotagged tweets. The main purpose of the paper is to find out which agendas of the prospective candidates resonates with voters, which ones may negatively impact a candidates chances. As the collected tweets are geotagged, they are analyzed from a geospatial perspective, and the sentiment for the candidate is plotted on a choropleth map. One of the main observations from the analysis was, that most of the tweets carried negative sentiment, and that people generally rant on twitter, rather than discussing issues related to the election.

I. Introduction

The analysis of the Presidential Elections in the United States is a challenging task, mainly due to the sheer number of factors involved. It is a hard job to statistically predict the outcome of the elections. Sentiment Analysis is one of the techniques which could be used to determine whether a candidate could win the election or not. The candidate's views on various topics, policies, and the sentiment associated with it could go a long way in determine whether a candidate would win or not.

Twitter is a very important social network. What sets it apart from other social networks is the fact that Twitter data is public. It could be viewed by someone who doesn't have a Twitter account. The fact that a Twitter feed could be embedded in other pages is also a significant factor in Twitter being all pervasive in the social media space.

The tweets collected for the purpose of this paper are geotagged, which becomes possible to localize the sentiment in a particular area. For the purpose of this paper, the sentiment on various agendas of the candidates would be gauged based on the origin of the tweet. With this, the popularity of a particular candidate in a specific area/state will be determined. The paper focuses on the main presidential candidates in 2016, namely Hillary Clinton, Bernie Sanders, Donald Trump, and Ted Cruz.

There has been prior work in gauging public sentiment using Twitter. The work of Hao Wang et al[7] deals with large scale Twitter analysis for the 2012 presidential election. This paper attempts to build on that paper, and factor in the geospatial aspect by using geotagged tweets for the analysis.

II. Data

The data for the purpose of the paper is collected from the Streaming API of Twitter. The Streaming API of twitter provides an asynchronous long polling mechanism to retrieve tweets from the Twitter API. The advantage of the long polling is, that it does not hold the connected between the API server and the client. This enables the script to execute code only when tweets are made available. As the US presidential elections carries global interests, a lot of people not from the US could be tweeting about the candidates. To gauge the public sentiment in the US, it makes sense

to restrict the tweets being analyzed to those from the US. For this purpose, the *location* filter was applied in the calls made to the Twitter API, so that only tweets from the USA were retrieved from the API. The *location* filter specifies a bounding box, which will be used to localize the tweets. A tuple of the Southwestern, and Northeast coordinates is needed by the Twitter API. The bounding box used for the purpose of this paper is [(-117.477, 32.496), (-67.2038, 44.6103)].

To retrieve the tweets from the API, the Tweepy[6] library for Python was used. Tweepy provides a consistent access to the Twitter API, and includes OAuth authentication. A Drawback with the Twitter API is that it does not allow for multi-filtering. This means, that if the *location* filter has been applied, then filtering for a particular hashtag, or text does not work. The workaround to this limitation is to filter for particular hashtag's in Python code. Once the geotagged tweets are retrieved, they are filtered for the keywords *cruz*, *clinton*, *trump*, *sanders*, *bernie*. All the tweets pertaining to the presidential candidates are retrieved separately.

Twitter data is very unstructured. People tweet all sorts of stuff, such as links, images, sound clips, videos, and a lot of garbled text. This makes sentiment analysis and topic modelling challenging. In addition to the casual text, a lot of tweets are just retweets. A retweet is a repost of an original post. These also need to be handled properly, before further analysis could be done. For this purpose, before running the analysis, the tweets are preprocessed, so that only relevant text is used. For instance, URL's, images, mentions, usernames, are stripped out of the tweet.

III. Methods

Once the data has been collected and preprocessed, it is ready for analysis. The data consists of 7800 tweets collected for the presidential candidates. As an additional preprocessing step, all the tweets are combined into a single CSV(Comma Separated Values) file, and annotated with a tag, so as to identify that the tweet is for that presidential candidates. This will be required for analyzing sentiment for that presidential candidate.

The system developed can be represented by the following block diagram:

I. Sentiment Analysis

For performing Sentiment Analysis, a Naïve Bayes classifier[9] is trained on a training corpus of 3200 tweets, equally divided among the presidential candidates is prepared. The remaining tweets are training data for the classifier. The training data is manually annotated as being positive or negative.

Naïve Bayes Classifier

The Naïve Bayes classifier is a bag of words classifier, which splits the data based on the word features. It is a conditional probability model. Given a set of k classes, it computes the conditional probability of the given data belonging to all the classes, and returns the class with the highest conditional probability. Formally, this could be expressed by:

$$p(C_k|x_1, x_2...x_n) = \frac{1}{Z} p(C_k) \prod_{i=1}^{n} p(x_i|C_k)$$

where C_k is one of k classes, x_i represent the set of word features, Z is the scaling factor, $p(C_k)$ is the class prior.

For the purpose of sentiment analysis of tweets, the tweet was divided into a set of words, and the conditional probability of it being in the positive and negative classes was calculated. The class it belongs to has the highest conditional probability. It is highly efficient for text classification.

II. Web Application

As the tweets are tagged by location, they can be geocoded to find out the coordinates required for plotting on a map. For geocoding, the Python *geocoder* module [5] was used. Once the tweets are geocoded, the tweets are written to a JSON file, with the following structure:

```
{
    coordinates: [], //Lat-Lng pair
    candidate: '<candidate>', //Clinton, Cruz, Sanders, Trump
    sentiment: '<sentiment>' // POS or NEG
}
```

A web application was created which reads the JSON[8] file, and is used to plot points, representing the origin of the tweets on a map. The Google Maps API[4] was used to create the required tiles. The tweets are then aggregated by the State of Origin, and the percentage of negative tweets, and positive tweets was calculated using a Spreadsheet. The sentiment for the candidates was visualized using Datamaps[2], which is a D3.js[1] library for creating maps.

IV. Results

I. General Observations

While annotating the tweets for training the sentiment classifier, an interesting observation was made. Most of the tweets carried a negative sentiment. This could be attributed to the fact that there is a tendency to rant on a public social media platform. Supporters of a particular presidential candidate often target supporters of a rival candidate. There are frequent flame wars between supporters of rival candidates. This leads to a significantly large number of tweets carrying a negative sentiment.

On analyzing the tweets, another interesting observation was made with respect to the hashtags used. The hashtags are very consistently used by the supporters or detractors of the candidates. Some of the commonly used hashtags are *#ImWithHer*, *#NeverTrump*, *#FeelTheBern*.

II. Sentiment Analysis Results

On the test set of 6800 tweets, the Naïve Bayes algorithm gave an accuracy of 68% with 5 folds cross validation. For the purpose of cross validation, 80% of the training data was used for training, and the remaining 20% of the training data was used as test data.
A major observation that was made in the classification was, that most of the tags were classified as negative. This could be explained by the fact that the majority of the training set comprised of negative tweets.

III. Geovisualization

A web application was created in Flask[3]. Flask is a framework for developing web applications quickly. The application exposes the following endpoints:

- */*
  This endpoint plots a Google Maps marker on a map, signifying the origin of a tweet. The HTML page provides option to select the candidate, and the sentiment, and the markers change corresponding to the candidate and sentiment selected.
- */<sentiment>/<candidate>*
  This endpoint returns a JSON (JavaScript Object Notation) object for all the tweets of the specified candidate, and the specified sentiment
- */trumpchor*
  This endpoint shows a choropleth map of the the sentiment for Donald Trump
- */clintonchor*
  This endpoint shows a choropleth map of the the sentiment for Hillary Clinton
- */cruzchor*
  This endpoint shows a choropleth map of the the sentiment for Ted Cruz
- */sanderschor*
  This endpoint shows a choropleth map of the the sentiment for Bernie Sanders

The following Google Map showing all the tweets bearing Positive Sentiment for Hillary Clinton:

The following Google Map showing all the tweets bearing Negative Sentiment for Hillary Clinton:

The following Google Map showing all the tweets bearing Positive Sentiment for Bernie Sanders:

The following Google Map showing all the tweets bearing Negative Sentiment for Bernie Sanders:

The following Google Map showing all the tweets bearing Positive Sentiment for Ted Cruz:

The following Google Map showing all the tweets bearing Negative Sentiment for Ted Cruz:

The following Google Map showing all the tweets bearing Positive Sentiment for Donald Trump:

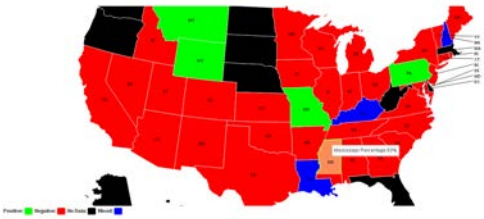
The following Google Map showing all the tweets bearing Negative Sentiment for Donald Trump:

Statewise Sentiment

Once the tweets are geotagged, and a sentiment is attached to each tweet, they are grouped by the state of origin. Using a spreadsheet, the percentage of positive, and negative tweets is calculated for each candidate. This information is then used to plot a choropleth map, which shows the sentiment of each candidate in a

particular state.

The following Choropleth map shows the statewise sentiment for Hillary Clinton:



The following Choropleth map shows the statewise sentiment for Bernie Sanders:

The following Choropleth map shows the statewise sentiment for Ted Cruz:

The following Choropleth map shows the statewise sentiment for Donald Trump:

Legend

- Red: Negative (more than 50% negative sentiment)
- Green: Positive (less than 50% negative sentiment)
- Black: No Data
- Blue: Mixed Sentiment (50% negative sentiment)

The choropleth maps throw some interesting observations:

- Most of the states carry a negative sentiment for most of the candidates. This is to be expected, as the number of negative tweets far outweigh the number of positive tweets.
- The sentiment of Bernie Sanders is negative in most of the states. This is a little unexpected, as the number of positive tweets for Sanders is the highest for any candidate. The number of states for which Sanders has a positive sentiment is the least in comparison with all the other candidates.

IV. Sentiment on various topics

While training the data, the sentiment for various topics for the particular candidate was observed and recorded, based on the content of the tweets.

Hillary Clinton

- Negative on lying
- Negative on the crime bill passed during Bill Clinton's presidency
- Negative on the Email leaks
- Negative on fracking
- Negative on the Benghazi attack
- Negative on corruption, and election funding
- Negative on Surrogate Voting fraud
- Negative on the Panama Papers

- Mixed Sentiment on Black Lives Matter
- Positive on Sexism
- Positive on Gun Laws
- Positive on Planned Parenthood

Bernie Sanders

- Negative on supporting the crime bill passed during Bill Clinton's presidency
- Negative on Gun Laws
- Negative on frequent attacks on Clinton

Ted Cruz

- Negative on 14 year citizenship requirement
- Negative on stealing delegates in Colorado

Donald Trump

- Positive over losing Colorado delegates
- Positive on being non establishment
- Negative of not being informed about political process
- Negative on racism
- Negative on Lying
- Negative on Corruption
- Negative on running a negative campaign

V. Conclusions

The analysis of the US Presidential elections resulted in some interesting observations. A claim could be made that Twitter data is unreliable in predicting the outcome of an election, as most of the sentiment is negative. Also, people tend to rant on Twitter, leading to flame wars with supporters of rival candidates. However, Twitter is very effective in gauging the sentiment on various topics.

For the purpose of this paper, the sentiment on various topics was gauged by manually looking at the tweets, which were used for training the data. This could be done automatically by using a Topic Modeling algorithm, and then running sentiment analysis on the result. This is a good candidate for future work.

At this point, the paper doesn't deal with real time Twitter data. In the future, a system could be developed which would gather tweets in real time, perform the classification, recompute the sentiment for the state, and update the choropleth maps.

References

[1] D3.js. D3.js - data-driven documents. `https://d3js.org/`.

[2] DataMaps. Datamaps library for mapping. `http://datamaps.github.io/`.

[3] Flask. Flask microframework for web applications in python. `http://flask.pocoo.org/`.

[4] Google. Google maps javascript api. `https://developers.google.com/maps/documentation/javascript/`.

[5] PyPi. Python geocoding library. `https://pypi.python.org/pypi/geocoder`.

[6] Tweepy. Tweepy library for python. `http://www.tweepy.org/`.

[7] Hao Wang, Dogan Can, Abe Kazemzadeh, François Bar, and Shrikanth Narayanan. A system for real-time twitter sentiment analysis of 2012 u.s. presidential election cycle. In *Proceedings of the ACL 2012 System Demonstrations*, ACL '12, pages 115–120, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[8] Wikipedia. Javascript object notation. `https://en.wikipedia.org/wiki/JSON`.

[9] Wikipedia. Naive bayes classifier. `https://en.wikipedia.org/wiki/Naive_Bayes_classifier`.