
Authors

Mohammad Alasmary, Landon Bedell, Tom Erickson, Joshua Ferge, Taylor Graham, Amir Kashpazha, Hannah Keller, Hui Soon Kim, David Levin, John Raesly, Forrest Tagg Ridler, Stephen Thoma, and Caleb Phillips

Projects in Geospatial Data Analysis: Spring 2015

Editor:

Caleb Phillips

Authors:

*Mohammad Alasmary**

Landon Bedell

Tom Erickson

Joshua Ferge

Taylor Graham

*Amir Kashpazha**

Hannah Keller

*Hui Soon Kim**

David Levin

John Raesly

Forrest Tagg Ridler

Stephen Thoma

University of Colorado

Department of Computer Science

Boulder, CO, 80309-0430 USA

August 25, 2015

* 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 6 weeks of class to work on a project that applies skills and concepts from the class to a problem of their choosing. Aside from some readings, lectures and class discussion, students were allowed to work on their projects exclusively during that time. 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. Social media data mining (e.g., Twitter and Yelp) was a popular topic due to the ease of access to large data sets or APIs with high quality geospatial data. Some students used these data sources directly, while others developed systems for mining, filtering, and storing vast amounts of real-time data from public streams. Projects involving open environmental data sources were also popular. Students who took that route made use of open, public GIS data from federal and environmental agencies, as well as research data sources. Several students pivoted their projects midway due to limitations in data access, particularly a small number who found themselves mired in bureaucratic delays to obtain access to 'semi-open' data. One student designed their own sampling design and performed their own data collection in the field. Analysis approaches are similarly varied: geovisualization, geostatistical modeling, multiple regression, graph analysis, etc.

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

Table of Contents

Spatial trends in online ratings

Mohammad Alasmary

11 pages

The Characteristics of America's Most Musical Cities

Landon Bedell

8 pages

Google's Expansion and The Potential Impact on Boulder's Housing Market

Tom Erickson

8 pages

Examining the relationship between economies of agglomeration and business performance

Joshua Ferge

9 pages

Geospatial Clustering and Classifying of Twitter Data

Taylor Graham

6 pages

Creating Distributed Temperature, Precipitation, Solar Radiation, and Humidity Maps for Boulder, CO

Amir Kashipazha

12 pages

A geospatial analysis of climate change and its effect on the American Pika's habitat in the Great Basin

Hannah Keller

6 pages

Identifying landslide prone areas of Colorado

Hui Soon Kim

15 pages

Boulder Neighborhood Price Index

David Levin

7 pages

Boulder Food Impacts Pre-flood and Post-flood - 2013

John Raesly

5 pages

A Geospatial Analysis of Oppressive Language On Twitter

Forrest Tagg Ridler

10 pages

Evaluating Potential Correlation Between Water Deficit and Mountain Pine Beetle Infestation in Whitebark Pine

Stephen Thoma

6 pages

Spatial Trends in Online Ratings

MOHAMMAD ALASMARY*

University of Colorado Boulder
Mohammad.Alasmary@Colorado.edu

CALEB PHILLIPS†

University of Colorado Boulder
Caleb.Phillips@Colorado.edu

Abstract

Many people use online rating websites as an assistance tool in decision-making, and these websites popularity has increased with adding such information to current electronic geographic maps. In fact, these sources have different types of information that might reveal a significant fact or trend about business. Yelp, is one source providing rating for many places including restaurants. This study investigates the association between the ratings of yelpers (Yelp users) representing the quality of restaurants and the ancestry percentage of people living in the surrounding regions. It examines the hypothesis that a restaurant has an excellent rating because the existence of peoples' ancestry in the surrounding area. We hope to understand whether a restaurant has an excellent rating because its cuisine matches the majority of community people ancestry. Yelp data will be used on state-level and ZCTA-level to check this association. The first part of the study tries to classify Yelp's restaurants around the ancestry information of people who live in the same area. Then, the yelpers' ratings of the resulted classifications will be statistically tested with ancestry percentage.

I. INTRODUCTION

Ancestry (people origins) plays a vital role in food and their followed cuisine in restaurants. For historical, geographical and economical reasons, people from different countries have different dishes. Asian dishes, for instance, have less red meat than Western dishes because it is scarce and costly. Chinese, in particular, adapt this and other changes to their dishes to face the growing population problem [6]. These differences sometimes form a direct influence on different aspects. From medical side, Kumar [2] reported that peanut sensitization was higher in African ancestry than other tested ancestries.

Yelp, Google places, TripAdvisor, Yahoo local and Judy's Book are examples of consumer rating websites that have become more prevalent and part of everyday decision-making. The content of these websites (the thing being rated) might differ from one to another, except you can say all of them start rating everything in-

cluding places, services and products. According to Yelp website, it has over 71 million local reviews and about 30 countries and of course their different cuisines.

In this study we are looking into the relationship between the quality of restaurants representing different people cuisines and ancestries percentages to their geographic locations. For clarifications, we defined the word "Ancestry" as people from different countries and origins not race. Therefore, we looked at Ethiopian restaurants not African or black as race. We also define the term "pure ancestry" restaurants as restaurants that serve the countries' people dishes. We use two data sets: Yelp data (for all information about restaurants) and U.S. Census Bureau, 2009-2013 (B04003/B04006-TOTAL ANCESTRY REPORTED on ZCTA and states) for peoples' ancestry.

Assessing the ground-truth of food quality is difficult. The only measure we used is the number of restaurants' stars. The diffi-

*

†

culties come from the fact that Yelp conspicuously rounds average ratings for restaurants [5]. Despite Yelp’s filter for fraudulent reviews, assuming that people will tell the truth, still they might give rating considering other factors, such as customer service and the price. Therefore, we took the number of stars as it is without looking into customers review. For ancestry data set, clear understanding of ancestry contrast is difficult in US populations, because self-identified race is imprecise [2] and many ancestry were missing. For more and other limitations see future work section.

I. Related Work

There have been a number of studies about restaurants, race and ancestry, but from financial aspects, such as [3] that uses semantic scales to weigh the relationship between tip and customer’s ratings of service since [4] proved how races constitute a difference in tipping. Luca in [5] investigated the impact of Yelp users reviews on restaurants. He found that independent restaurants “not chain restaurant” registered from 5% to 9% revenue with one star increase. Furthermore, the study showed that the size of rating information, the reviewer’s status, and the number of reviews are significant contribution factors of people’s behaviors change and their responses.

To our knowledge we could not find any notion similar studies. Therefore, we consider this study as a first step that might reflect useful information towards deep relationship with restaurant’s quality.

II. METHODS

I. Data

We used two data sources: Yelp academic data set for restaurants and their categories, and Census Bureau ancestry population. We extracted only the restaurants business that are in the US. We kept only what has matched between the two data sets. Ancestry data has Egyptian, Iraqi, Jordanian, Lebanese, Moroccan, Palestinian, and Syrian as separate

columns and their total. So, we checked Yelp data restaurants if they have the same classification, then we took only the total number of them in ancestry data (provided) and restaurants (calculated). We could not do the same for the rest of ancestries due to several problems, such as the differences in the term used and their real meaning or the geography boundaries. For example, West India and India, and Modern Europe and Europe. We also excluded American restaurants since we were not planning to analyze any other country.

Based on the rest of ZCTA we calculated the percentage of every ancestry and margin of errors instead of numbers. Out of 33120 ZCTA, only 370 (1.1%) were 0 and therefore omitted, the rest may have one or more reported ancestry after classification.

II. Tools

We built some scripts to manipulate the data in python, such as the one in fig.1. We also used QGIS for geospatial analysis looking for nearest neighbors ZCTA for a restaurant’s ZIP code, see Figure.1 for the main function aggregate.

III. Experiment setup

After we filtered the U.S. restaurants per state using aggregate function, we got the following number of restaurants (see table.2).

Out of all restaurant categories we ended up with 16 different category of restaurants representing 16 different countries see Table.1.

```

function aggregate (filtered ancestries, filtered restaurants) returns final_data
  for each ancestry do:
    for each restaurant do:
      if ancestry_name ∈ restaueant_categories and
ancestry_ZCTA = restaurant_ZIP then:
        info ← extract_required_info(ancestry, restauarnt)
        Add_final_data(info)

```

Figure 1: This function aggregates the data, that has filtered manually, together by using ZIP code and category for restaurants and ZCTA for ancestry data. Note that ancestry *name_uses* ∈ because restaurants were classified by more than one category that affected the results and made us filter them. The key in the *final_data* is ZCTA and ancestry together.

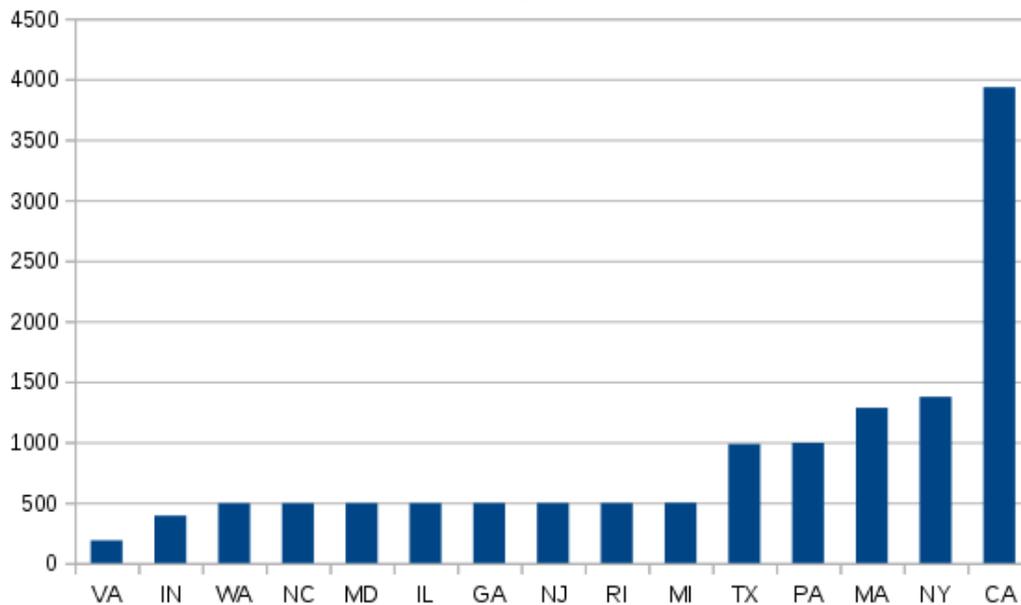


Figure 2: Number of restaurants per state, total:13118. Unmentioned states contain 0 number of restaurants. Note, we have not filtered the results yet to match the ancestry, see Table.1

Table 1: Number of restaurants based on 16 different cuisines. This classification represents about 3.3 % of total number of US restaurants in Yelp academic data set reported in fig.2.

Category	Number of restaurant
Turkish	12
Arabian	10
Brazilian	6
Belgian	3
French	50
Russian	3
Scandinavian	2
Greek	66
Italian	244
Portuguese	3
Ethiopian	13
British	4
Basque	5
Polish	2
Afghan	4
German	8

We needed to know that the used ZCTA is a valid metric, means there are no huge interferences between restaurants' locations and missing restaurants belonging to the same ancestry. We firstly assumed that once a restaurant's ZIP matches a ZCTA then the restaurant belongs to the same area with ancestry percentages. During the process we counted the unmatched restaurants, too. The results are summarized in Table.2

The perfect analysis should be done on three levels: country's level, state's level, and ZCTA level. Country's level analysis means comparing the quality of American restaurants, for example, in America and other countries. This type of analysis need large scale data and ancestry information for each country. State's level means comparing ,for example, Italian restaurants quality in more than one state with ancestry (Italian) percentage in each state. The last level is ZCTA level which is the smallest unit ancestry information could use as a key.

To encounter time and data scarcity issues, we decided to start with a big association analysis process and keep narrowing it based on

the results to get the clearest association so we could confirm or reject our hypothesis. Firstly, this available data does not let us go for country's level analysis because we only have U.S. restaurants. Secondly, For ZCTA-level, we performed the aforementioned aggregation process that lead to lose a vast part of the data, but it was unavoidable.

It is noteworthy that sometimes the number of restaurants was a little bit higher for some ancestry, but the script excluded it due to the high percentage of margin error. we were not interested in any ancestry of ZCTA area that has margin error higher than 10%.

According to [9] no more than 27% of people tend to drive more than 5km for restaurants purpose. We, therefore, for ZCTA-level analysis were willing to add the percentages of people ancestry from any ZCTA that interfere with restaurant's circle as long as it makes the relationship clearer and stronger, see Figure.3.

We started looking for all restaurants, hoping to find a global trend, then we came to the most three frequent cuisines ancestry : Italian, Greek, and French, respectively. Finally we looked deeper for the highest cuisine ancestry, Italian.

III. RESULTS

A quick vision at figure.4 shows a weak negative weak association between the quality of restaurants and the percentage of Ancestry. Applying Spearman's rank correlation test since the data were not normally distributed , $p=.03$ & $r=-0.19$, indicated that quality of restaurants was greater for lower ZCTA ancestry percentages. Since $p<.05$ we tend to reject the null hypothesis and consider this association. From the figure.4 we can say that for all ancestry the quality of restaurants that serve one ancestry cuisine goes down with the increase of ancestry percentage in the same ZCTA. We call this the global trend. French, Greek and Italian restaurants registered the largest number of places in our data. So, we took a look at them one by one.

French Restaurants, see figure 5, tend to

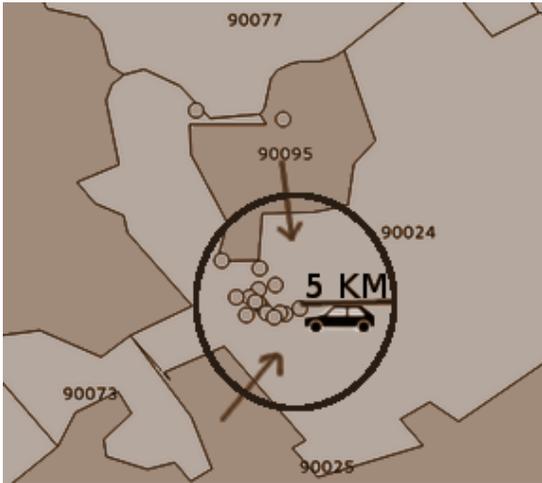


Figure 3: F

or nearest ZCTA neighbors, the quality of the restaurant at the center of the circle will be linked to the ancestry percentages of ZCTA 90024,90025,and 90095 since part of their areas are inside the restaurant's circle driving distance.

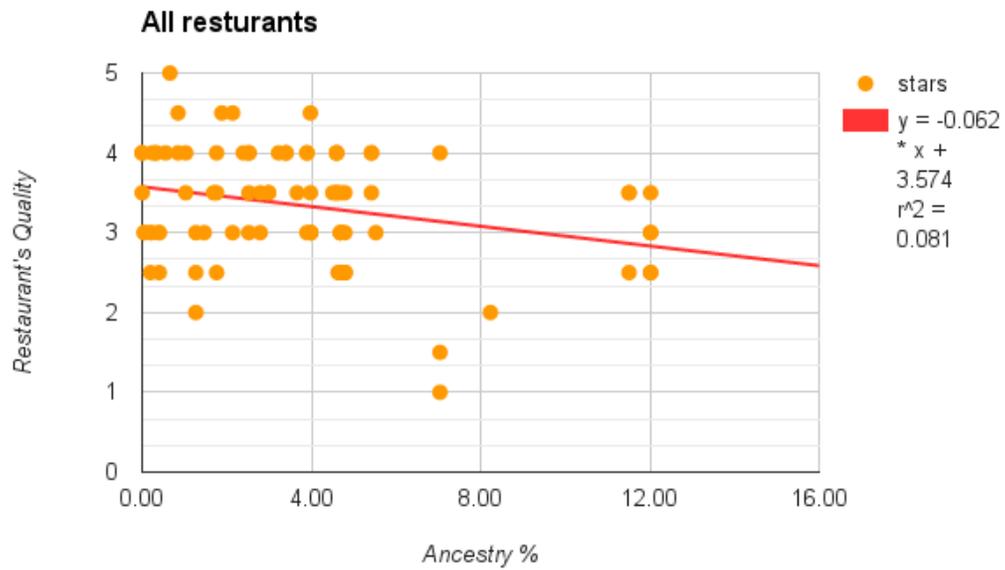


Figure 4: An overview of what is it look like for all restaurants vs their ancestries (ZCTA-level used, no nearest neighbors).

Table 2: Number of cuisines per ancestry for every state. Missing column represents number of unclassified restaurants and their percentages because of the differences between ZIP and ZCTA. According to the previous table, six different cuisines do not have problems at all.

State	Missing	VA	WA	CA	IL	GA	IN	MD	TX	NC	MI	NY	PA	Total
Afghan	0	0	1	1	0	0	0	1	0	0	0	1	0	4
Arabian	2	0	2	1	0	0	0	0	2	0	0	0	0	4
Basque	1	0	0	1	1	0	0	0	2	0	0	0	0	4
Belgian	2	0	1	0	0	0	0	0	0	0	0	0	0	1
Brazilian	0	0	0	4	0	1	0	0	1	0	0	0	0	6
British	1	0	0	2	0	0	0	0	0	0	0	0	0	2
French	8	0	0	4	2	1	0	0	6	4	4	8	7	36
German	0	0	2	0	1	0	0	0	0	0	1	3	0	7
Greek	6	3	5	19	5	3	3	3	8	1	1	6	2	59
Italian	18	5	5	50	11	10	3	10	12	11	6	45	22	190
Polish	1	0	0	0	0	0	0	0	0	0	0	1	0	1
Portuguese	2	0	0	1	0	0	0	0	0	0	0	0	0	1
Russian	0	0	0	2	0	0	0	0	0	0	1	0	0	3
Scandinavian	0	0	0	0	0	0	0	1	0	0	0	0	0	1
Ethiopian	1	1	0	1	0	0	0	1	0	0	1	5	1	10
Turkish	0	0	0	7	1	0	0	0	3	1	0	0	5	12

follow the general trend, if we use all french restaurants, restaurants that are pure French and the others that have additional cuisines, such as American dishes. In fact, we can barely see the trend because it is very weak association but we cannot infer that the French restaurants is the opposite. However, pure French restaurants reflect a weak positive association. We have only 36 French restaurants and about 10 that categorized as pure, serve only french dishes, so it is difficult to study why they have different relationships than the global trend.

Greek Restaurants at figure 6 show a different kind of associations. We can barely see non-linear relationship between the quality of restaurants and the percentage of Greek. This association has a positive trend before the percentage of Greek people reaches 1%. Since this kind of relationships might imply other factors and due to data scarcity, only 59 Greek restaurants and about 11 categorized as pure. California has the maximum number of them, 19 Greek restaurants and only 4 of them are pure, we cannot investigate further what is it with Greek restaurants.

All Italian restaurants, on the right of figure 7, have very weak association. However, pure Italian restaurants (figures:7) consistently follow the general trend that appears clear on the left chart of figure 7 when we fixed the number of additional cuisines to zero.

Since the number of Italian restaurants is the largest, 190 Italian restaurants and 90 are pure, we had to look deeper at this ancestry and its restaurants. Figure. 8 shows a weak positive association between the percentage of Italian and mixed Italian restaurants, that only serves additional dishes to the Italian dishes.

Using nearest neighbor ZCTA, we narrowed it to only California (CA) Italian restaurants due to the time and because CA has the largest number of restaurants. The resulted association did not give any different kind of relationships from ZCTA results but weaker. Also, it did not reflect any strong change in the predictions from their results. Therefore, we did not apply it on any other ancestry.

Moving to state_level, the concept of nearest neighbor is equal to zero, means we assumed that no one would go to another state

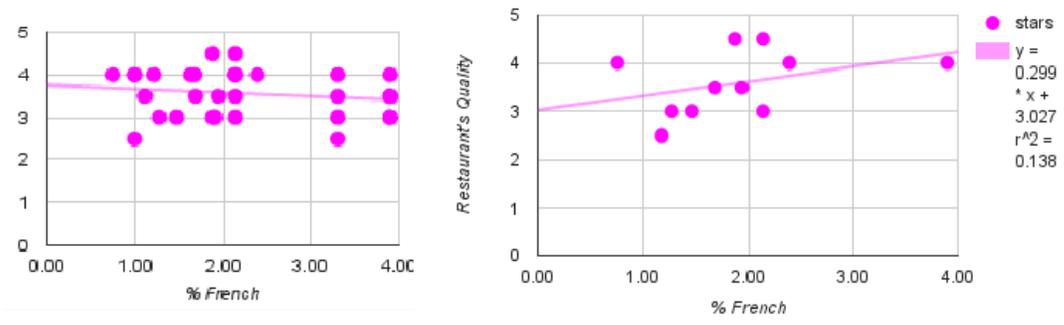


Figure 5: All French restaurants, on the left, follow the global trend, whereas pure french restaurants have the opposite

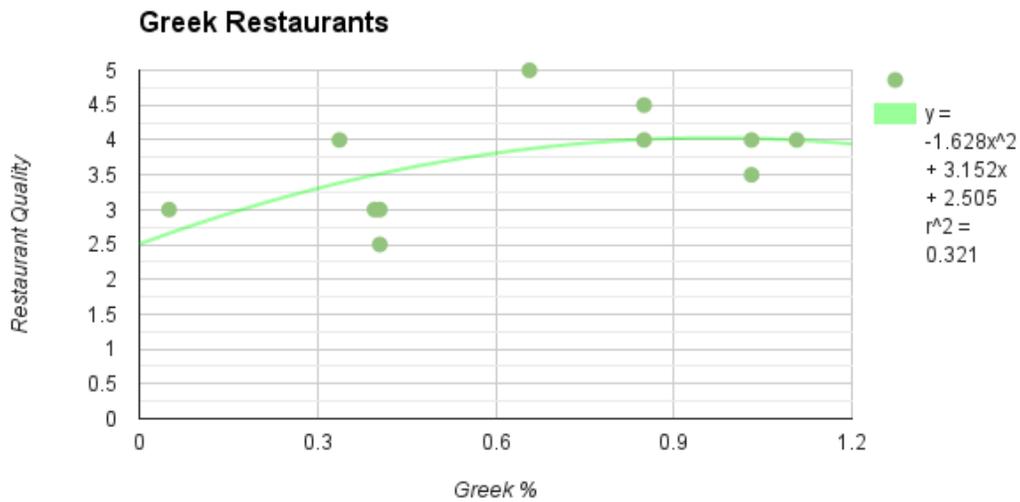


Figure 6: All Greek restaurants draw this non-linear relationship that reveals the role of other factors

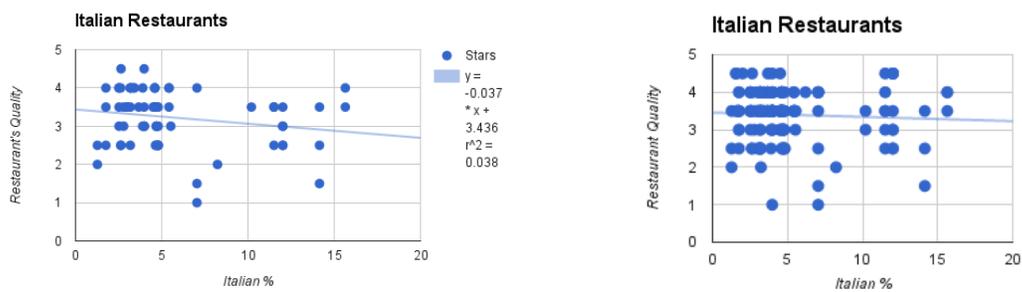


Figure 7: The global trend appears clearer with pure Italian restaurants on the left

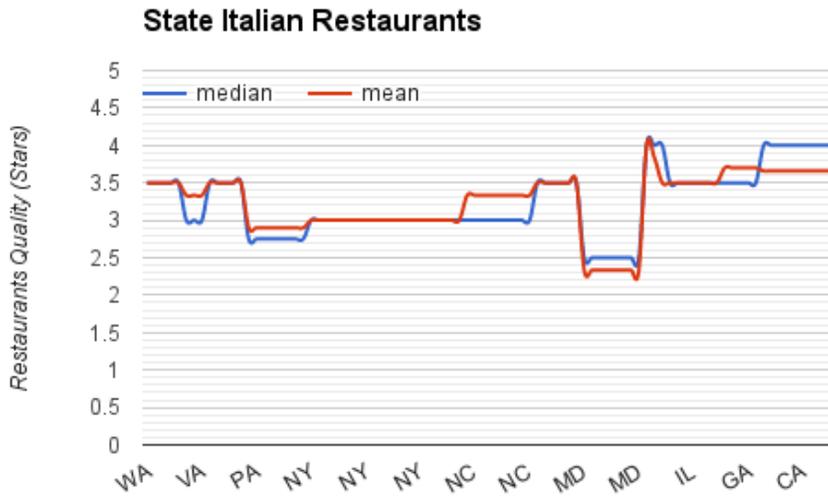


Figure 9: The difference between mean and median for all Italian restaurants quality (ordinal data)

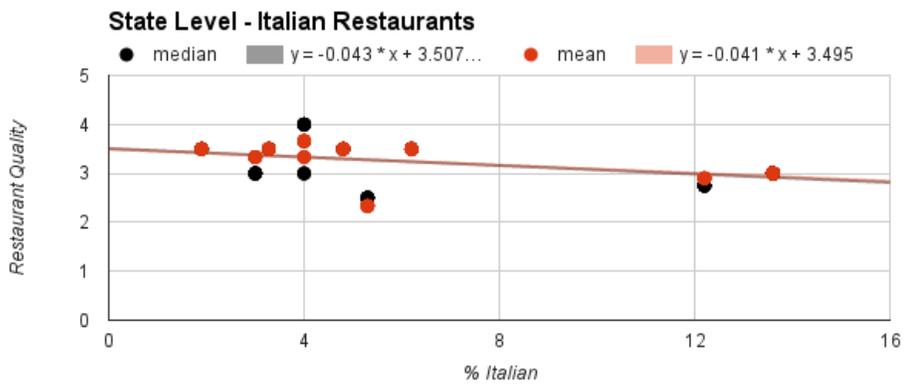


Figure 10: Both mean and median follow the global trend. Mean is over median

the number of stars, can be described as ordinal data. to avoid the long debate of statisticians whether we could apply the mean or not, we used both the mean and the median. since there is a difference between mean and median results, we had to consider both the results. According to the data, California state has the largest number of restaurants, so we thought this bias might affect the results, but graph.9 proved that Maryland, which has fewer number of restaurants, registered higher results although New York and North Carolina could not.

We believe that the idea of nearest ZCTA for the restaurant is more accurate and reasonable. The probable reason for the weaker result was data limitation. Most of the Italian restaurants locations in CA are close to each other, so their ZCTA, and some of them in the same ZCTA. This closeness needs accurate information about ancestry.

Although Italian restaurants that serve additional dishes beside Italian dishes show the opposite of our global trend, we did not report that because Italian state-level test did not show the same relationship, but followed the global trend. Also, our primary goal is to find any association in the data. Investigating this kind of relationship needs more data and can

be performed as a second step.

V. CONCLUSION

This study reflects when people ancestry below 15% there is a simple dropping of restaurant's quality that increases with ancestry percentage. This relationship works better with pure categorized restaurants, serve only one cuisine. Italian restaurants were tested further and proved following this simple trend on state-level and ZCTA level.

I. Future work and limitations

Ancestry data was not the optimum way to collect census information. It lead to exclude Indian and Chinese restaurants representing large numbers since the ancestry data does not have any related statistics. Moreover, the definition of ancestry could be changed to match more people and includes partially race. The filter of restaurants could be improved by adding any common dishes to both ancestries and leveraging them. For a greater Yelp dataset, the study might be improved by looking at country's level, American restaurants in the U.S. and other country, beside ZIP or state level. Additionally, other ancestries than Italian could be tested against this trend.

REFERENCES

- [1] U.S. Census Bureau; 2009-2013 TOTAL ANCESTRY REPORTED; Tables B04003; generated by Mohammad Alasmary; using American FactFinder; <<http://factfinder2.census.gov>>; (6 April 2015).
- [2] Kumar, R., Tsai, H., Hong, X., Liu, X., Wang, G., & Pearson, C. et al. (2011). Race, Ancestry, and Development of Food-Allergen Sensitization in Early Childhood. *PEDIATRICS*, 128(4), e821-e829. doi:10.1542/peds.2011-0691
- [3] Lynn, M. (2003). Restaurant tips and service quality: A weak relationship or just weak measurement? [Electronic version]. Retrieved [insert date], from Cornell University, SHA School site: <http://scholarship.sha.cornell.edu/articles/132>
- [4] Lynn, M. (2011). Race Differences in Tipping: Testing the Role of Norm Familiarity. *Cornell Hospitality Quarterly*, 52(1), 73-80. doi:10.1177/1938965510389297
- [5] Luca, M. (2011). Reviews, reputation, and revenue: The case of Yelp. com. Com (September 16, 2011). Harvard Business School NOM Unit Working Paper, (12-016).

-
- [6] <http://www.foodbycountry.com/>
- [7] People.hbs.edu,(2015).Retrieved 20 April 2015, from [http://people.hbs.edu/mluca/ FakeIt-TillYouMakeI](http://people.hbs.edu/mluca/FakeIt-TillYouMakeI)
- [8] Yelp Academic Dataset. <http://www.yelp.com/academicdataset>.
- [9] Iacono, M., K.J. and A.M. El-Geneidy (2008), Access to destinations: How close is close enough? Estimating accurate distance decay functions for multiple modes and different purposes (Minneapolis: Center for Transportation Reseach, University of Minnesota).

The Characteristics of America's Most Musical Cities

LONDON BEDELL*

University of Colorado
london.bedell@colorado.edu

Abstract

The Popular Music Artist holds an influential position in modern American society. To better understand the origins of these figures, this study attempts to find common features among the cities from which America's top music artists came. Using a data set derived from the Echo Nest API, which contains the city-of-origin of 50,000 of the top Music Artists in the United States, and matching it up with 2010 census data, this study attempts to find some of the key characteristics of the cities that produce the most musical artists in the U.S.. The results suggested that demographic and spatial characteristics of a city have the number of artists it gives rise to.

I. INTRODUCTION

IN modern American culture, the popular music artist represents more than just a musician. They stand among the top celebrities, and receive the same excessive media attention that comes along with it. Their art extends beyond their music to music videos, clothing lines, merchandise, and more. The introduction of social media platforms like Facebook, Twitter and Soundcloud has enabled artists to connect with their fans on a more intimate level, making the personality and character of the artist ever more essential in drawing in followers.

The medium of music itself has also evolved in the past decade, shifting away from permanent ownership with the rise of streaming applications and websites such as Youtube, Spotify, and Soundcloud. Recently, Spotify was reported to have surpassed the entire U.S. record industry in market value. Popularity is become a measure of play count rather than album sales.

This shift has also enabled the development of data analysis in the music industry. Companies like Nielsen and Echo Nest can now monitor play counts, blog posts, news articles, event rosters and more, to develop a better under-

standing of trends within the music industry[2]. There is a high demand for these types of services as trends can now rapidly evolve with the help of the internet, and the consumers now expect relevant and personalized music suggestions.

Echo Nest has become an industry leader in music data. With customers such as Spotify, Twitter, and MTV, their methods of data compilation and interpretation have proven effective in satisfying consumers' expectations [1]. They also provide some of their services to developers for free, with their Echo Nest API, which has lead to some impressive implementations in the field of Geo-spatial data analysis. Paul Lamere, the Director of Developer Platform at Echo Nest has a blog dedicated to technology used in online music discovery. He has used echo nest to map the most popular artists for each state in the United States. He has also shown how listener demographics such as age and gender can effect music preferences.

For this study, data was taken from one of Lamere's posts that retrieves the hometowns of America's top music artists to find out which U.S. cities are the most musically endowed[3]. In attempt to look deeper into this data, this study attempts to uncover what qualities con-

*A thank you or further information

tribute a cities ranking. By analyzing a cities per capita popular artist count with census data and spatial mappings, I try to isolate the main characteristics of America's most musical cities.

II. DATA

The data used in this study was collected from two different sources. The Artist data was retrieved from Paul Lamere's post on his Music Machinery blog that found the most musical cities in the U.S.. This was done by first looking up the top 50,000 Artists in the U.S. using Echo Nest's 'rank_hottness' classifier that ranks Artist's popularity based on properties like play counts and blog posts, and then finding the hometowns for those artists using the 'location' classifier. Only cities with populations greater than 5,000 were included in the post. The data was converted from its original HTML format into CSV to be processed using R. The ZCTAs (zip code tabulation area) corresponding to each city were found using a data set retrieved from United States Zip Codes.org [4].

The demographic data was retrieved via the US census TIGER data system. Data sets for education, economics, and general census demographics for 2010/2011 in CSV format were used.

III. METHODS

In order to identify which population demographics are related to the musicality of a city, a range of statistics were pulled from the census data, and compared to the artist count, represented by the number of artists from the city per 1000 inhabitants. The features tested were:

- Mean Age
- Percent Male
- Ethnic Diversity
- Median Household Income
- Percent of Population with an Occupation in Business, Management, Science, or Arts.
- Percent Self-Employed

Data for all of these features was retrieved from the census data using R by extracting the elements whose GEO.id2 matched each relevant zip code. The values for zip codes contained within each city were averaged using R's mean() function. The only attribute which required further computation was Ethnic Diversity, which represents the likelihood of two individuals being from a different ethnicity when chosen from the population at random. This is not a feature measured by the census bureau, so it was manually computed. First, figures for the percentage of the population that represent each distinct ethnicity listed in the census data, were extracted in the same manner discussed above. Each of these values was then squared to find the likelihood of choosing two people of that ethnicity at random. The values were then added together and subtracted from 1 to find the final figure for Ethnic Diversity.

The attributes selected were used for various reasons. The mean age and percentage of male inhabitants were chosen because gender and age have been shown to have an effect on music preference [5]. Ethnic diversity was chosen, to see if the integration of different cultures might have an effect on a cities musicality. Income was looked at to evaluate whether the wealth of the population increased or decreased the likelihood of producing a popular music artist. The occupation and self-employed categories were chosen because music artists fall into these categories.

To see if these attributes correlate with the musicality of the city, each was plotted with respect to the artist count and log 10 artist count using ggplot. Kendall tests were also used to detect correlations.

In addition to looking for statistical trends in the census data, the Artist data was also mapped to identify trends spatially. Using ggmap, several plots were made of the data in various regions of the U.S.. To display the information as accurately as possible with the given data set, the coordinates of one zip code for each city was used for the plot points. To represent the relative density of popular artists per capita, the artists per 1000 inhabitants fig-

ure was first multiplied by 100 to get rid of any decimal values. Each row was then duplicated as many times as its modified value, so if Boulder,CO had .35 popular music artists per 1000 people, then its row, and in turn its plot point, would be repeated 35 times. This data set was then processed using ggplot's 2d density function to plot a density map layer with respect to the number of artists per-capita for each location.

IV. RESULTS

I. Census data

Neither the comparisons with the census data nor the mappings of the artist data revealed any significant trends. When comparing the the artist count data with the selected population attributes nearly all of the resulting plots and correlation tests suggest that change in the given variable from city to city has no effect on the number of artist's from that city.

Bellow, figure 1 and figure 2 compare the average age and percentage of male inhabitants, respectively. A log 10 distribution was used for the number of popular artists per 1000 people. Log 10 was used for many of the plots to spread the high density of data points with an artist count between .10 and .01, and to check for logarithmic trends. Both figures 1 and 2 show a normal distribution of data, centered around a mean value. The average age is around 35 for most cities, with deviations of increasingly smaller numbers on either end of the mean. The distribution for the percentage male is very similar, with most point around 49%. Neither distribution shows any increase or decrease in the number of inhabitants as the artist count increases.

Figure 3 shows the distribution for the calculated ethnic diversity, with respect to the normalized artist count. The attribute has greater variance among the plotted points than in the previously mentioned figures, but the data still appears to have no correlation to the number of popular artists from a city. Although the lower values seem to have a slight negative shift with

the increase in artist count, it is a negligible trend, particularly with such a large horizontal spread in the data.

In Figure 4, the plot of median household income has been left on a non-logarithmic scale, to show the skewed distribution of the data. The bulk of the points is positioned at a low income, with an almost entirely upward variance. While this data may indicate how income is distributed in the U.S., it does not suggest any correlation between income and artist count.

The plot for the percentage of people with a job in Buisness, Management, science or the arts is shown in Figure 5. Out of all the tested variables, this one comes the closest to showing some correlation with the artist count. The plot appears to have a slight upward trend, with the number of popular artists per 1000 increasing as the percentage of people with the occupation type increases. Using ggplot, a line was fitted to the mean distribution, and shows that the trend does exist. A Kendall correlation test on the variable gives a p-value of less than $2.2 * 10^{-16}$ suggesting that it is very unlikely the trend is due to chance. The correlation coefficient however is 0.074, meaning the correlation is weak, so the increase in the attribute only has a subtle effect on the artist count. Typically, trends with a correlation coefficient < 0.3 are neglected.

The data plot for the percentage of people self employed also seems to show a slight upward trend, as seen in Figure 6. However, the results of the Kendall test suggest that there is a high likelihood this trend is due to chance, giving a p-value equal to 0.44. This, along with a correlation coefficient of 0.007, indicates that the trend should be neglected.

II. Spacial Data

At first glance, the plot of the artist data over a map of the United States, shown in Figure 7, appears to exhibit some spatial logical groupings in the distribution of cities with popular artists. The map shows the points of every city, with the point size corresponding to that city's artist count. The colored layer represents the

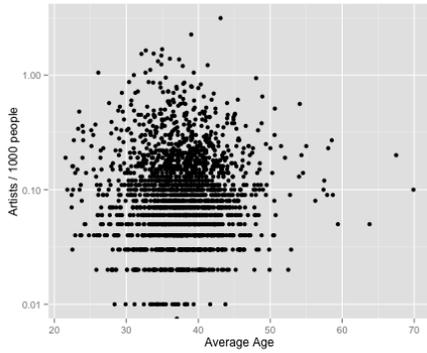


Figure 1: The mean age for each city vs. the artist count with a Log 10 distribution

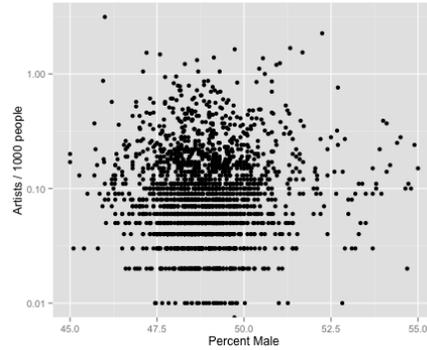


Figure 2: The percentage of male inhabitants vs. the logarithmic artist count.

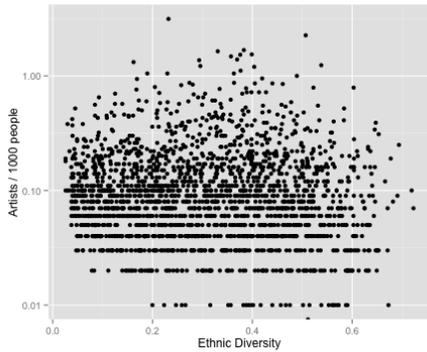


Figure 3: Ethnic Diversity variable vs. Log 10 artist count.

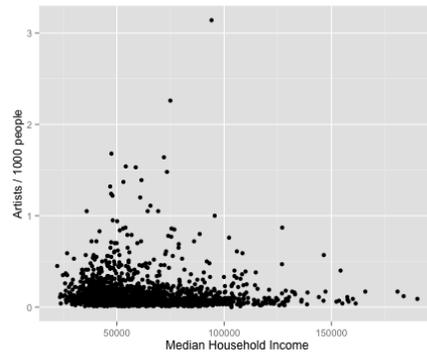


Figure 4: The median household income for each city with respect to the artist count (non Log 10).

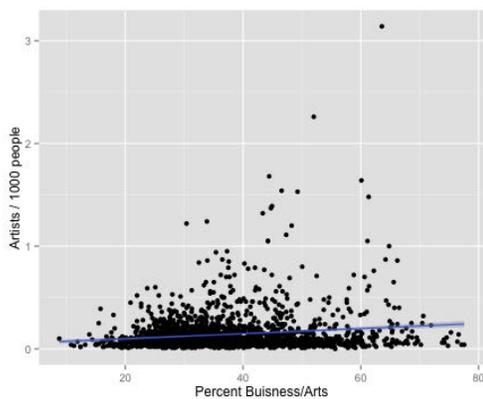


Figure 5: Percent of population with occupation in Business, Management, science or Art vs. Artist count.

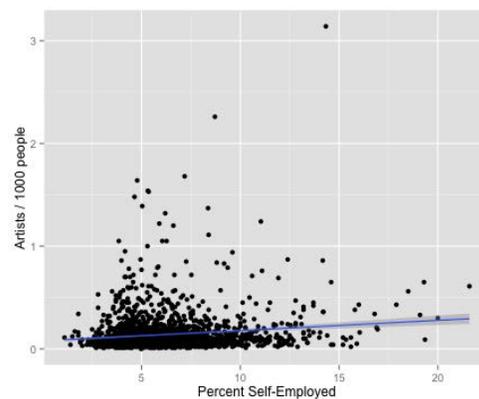


Figure 6: Percentage of population that are self-employed with respect to the number of popular music artists per 1000 people

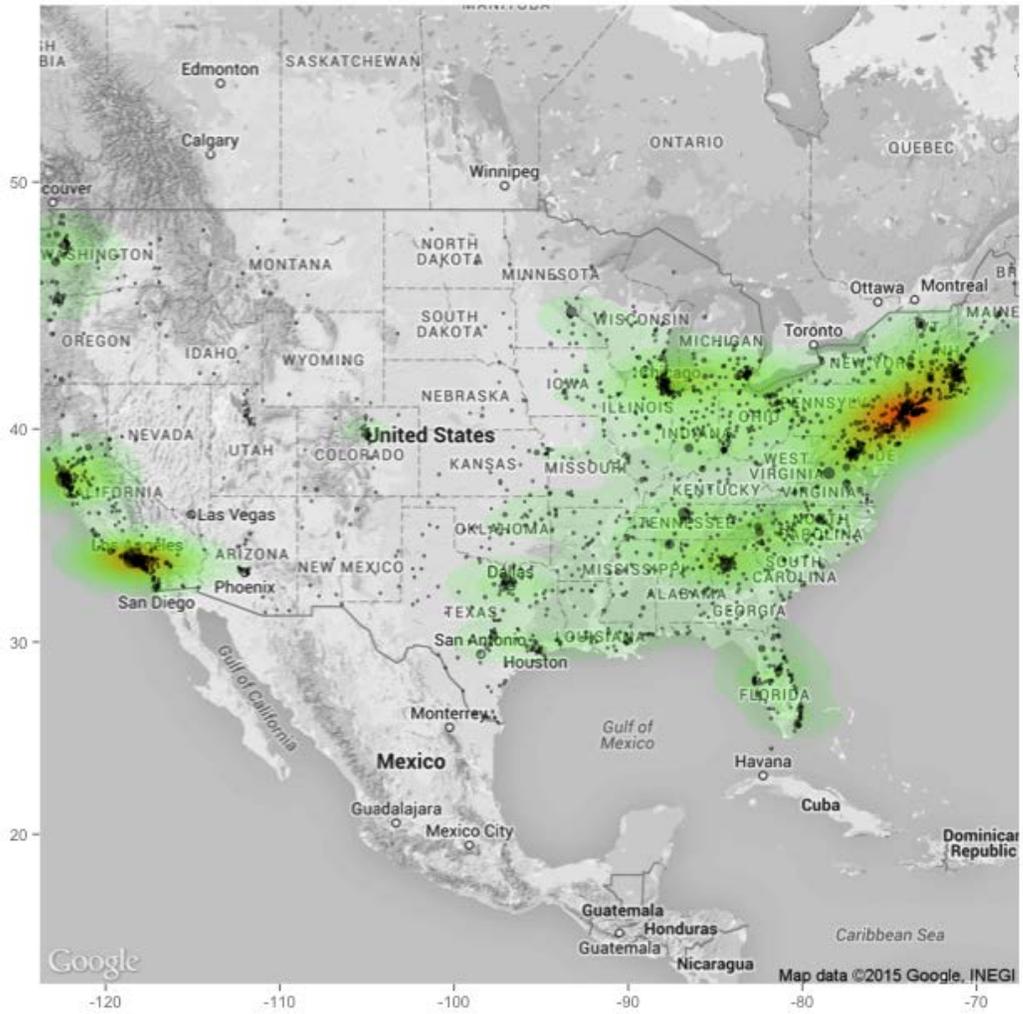


Figure 7: [6]

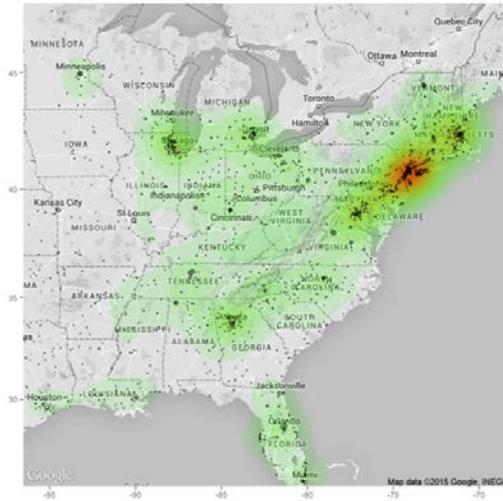


Figure 8: [6]

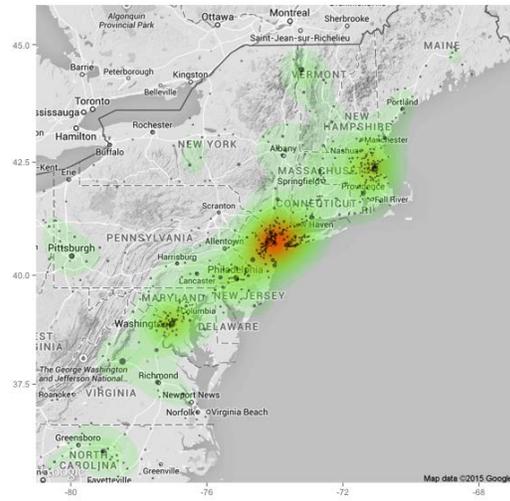


Figure 9: [6]

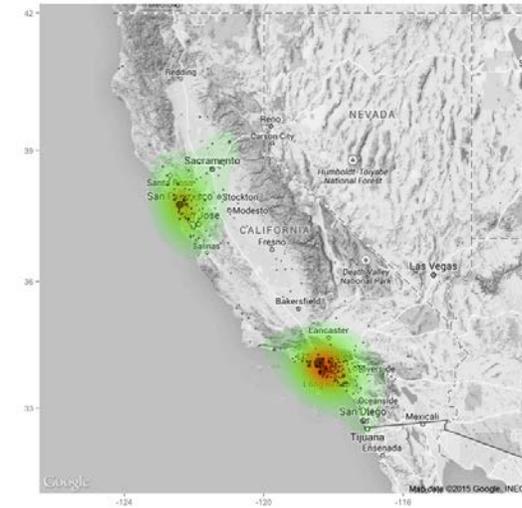


Figure 10: [6]

density of the artist count, with more opaque green corresponding to a low density, and less opaque red corresponding to a high density.

In Figure 7 the east side of the U.S. seems to have a higher artist density in general, with a light green layer covering almost the entire area, and with redder spots circling major metropolitan areas. The density is particularly high on the upper east coast around New York City. More detailed maps of this eastern region and the high density area around NYC are shown in Figures 8 and 9 respectively. By taking a closer look at the eastern region we see that there are dense cluster of cities around the major cities like Chicago, Atlanta, and New York. In Figure 9, New York City is surrounded by a dense cluster of cities surrounded by a red glow.

The California also has areas with a high artist density, namely Los Angeles and San Francisco. As seen in Figure 10, these two metropolises each have a high artist density, but are completely isolated from one another, unlike on the east coast where high density levels string from one major cluster to the next.

Referring back to Figure 7, with the density plot of the entire country, it can be seen that other than the high density areas in the east and on the west coast, much of the country has a very low artist density. Other than a slight spike in density near Denver, the large span of land stretching from the mid-west to the boarder of California lacks any hue-bound indication corresponding to a notable artist density.

V. DISCUSSION

The objective of this study was to find out what spatial and demographic characteristics describe cities that have been homes to many popular music artists. In doing so, the hope was to gain further insight into the features that go into forming a strong musical community within an urban environment. In many respects, this study failed to reach this goal.

The analysis of the 2010 census data with respect to the per capita artist data from the

Echo Nest API lead to no conclusive findings, with none of the population attributes showing any significant correlation with the artist count variable. This suggests that none of these qualities have a influence on the musicality of a city. The only tested feature that did show a slight correlation, was the Occupation Percentage for Business, Management, Science and Art, a category that includes music artists. This subtle trend may be due to some underling variable, or collection of variables that influence a greater focus on these types of occupations within a city, but from the current data, no conclusions can be drawn.

While the census data failed to show any meaningful results, the spatial data analysis seems to have found a bit more success. Looking at the plots across the U.S., it appears as though the density layer roughly corresponds to population density. As discussed in the methods section, the plotted data is adjusted to normalize for the population in each city. This suggests that the trend may, in part, be indicating that population density is correlated with in the number of popular artists a city gives rise to. Even with the internet, it's plausible that it would be easier for an artists to gain popularity if they are surrounded by a larger number of people. While this postulation warrants further investigation, it would be a stretch to conclude that this is, in fact, what these maps are showing. Rather, the high density levels may be, instead, related to a high density of cities in these areas. While the density plot does account for population density, it does not account for city density. Therefore, the fluctuations in the plot are, in part, corresponding to the number of cities per unit area. Since, nearly all of the areas with a high density value, contain a large cluster of city points, any attempts to extrapolate information from the maps relevant to the artist count are merely for the sake of discussion and anticipation for what future analysis could find.

Although the results collected in this study contained a few feeble indications that there might be some trends in the musicality of American cities, the only conclusion that can be

reached, at this point, is that based on the collected data the distribution of popular music artists in U.S. cities exhibits a normal distribution pattern with respect to any concretely defined, measurable characteristic of a population.

VI. LIMITATIONS

There are several limitations within this study. First, the artist data was collected from an unofficial source and therefore its accuracy cannot be assured. Further, the methods used to collect and calculate the artist variables such as the locations and the popularity are unknown, since they are not disclosed to the public. While the 6 census characteristics chosen to compare against the artist count were carefully selected, a more in-depth investigation might look into a larger range of variables, from a greater variety of sources. Finally, as touched upon in the previous section, the method used to create the density plots were not normalized for the city density, and therefore may be misleading.

REFERENCES

- [1] <http://developer.echonest.com/docs/v4>
- [2] <http://www.nielsen.com/us/en/solutions/measurement/music-sales-measurement.html>
- [3] <http://musicmachinery.com/2012/05/20/what-is-the-most-musical-city-in-the-united-states/>
- [4] <http://www.unitedstateszipcodes.org/>
- [5] <http://musicmachinery.com/2014/02/13/age-specific-listening/>
- [6] @Article, author = David Kahle and Hadley Wickham, title = ggmap: Spatial Visualization with ggplot2, journal = The R Journal, year = 2013, volume = 5, number = 1, pages = 144–161, url = <http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf>,

Google's Expansion and The Potential Impact on Boulder's Housing Market

TOM ERICKSON

University of Colorado at Boulder
erickstm@colorado.edu

Abstract

Google has officially announced plans to expand their workforce in the city of Boulder. By the first quarter of 2017 Google wants to increase their employee count in Boulder from the current 300, to potentially 1500.¹ Boulder is an interesting city due to the fact that it is a college town that has forbidden growth of the city upwards (strict limitations on building height) and outwards. Given Boulder's limitations for space, the influx of high income employees could potentially create a strain on the city's housing market. The goal of this study was to answer the question of what could possibly happen if all 1200 of the incoming employees chose to either buy or rent property in the city of Boulder.

I. INTRODUCTION

To understand Google's origin and growth within Boulder it is important to take a quick look into their history. They originally set up their 300 person operation in 2006 once they acquired the 3-Dimensional rendering software company Sketchup headquartered in Boulder. Despite selling Sketchup in 2012, Google's office remained. They conducted a small expansion in the city when they built across the street from their current office at 26th and Pearl. This brings us to the current day in which they are planning on building a 4.29 acre lot at 30th and Pearl². When discussing their decision to expand and build a campus on this specific location, they claimed to have explored other areas of Colorado however they decided that Boulder fit their needs best because of its "business climate and infrastructure to support the needs of Google's Operations"³. Despite the company's large campus size, none of the buildings will exceed 55 feet. The campus itself could be fairly unobtrusive due to the fact that the lot they bought hosts mostly empty

buildings, however they will still be forcing out some local businesses. The most important aspect of this expansion though is the increase in population in a city that has a 3.6 percent vacancy rate among rental units⁴ and an average of 3.11 percent vacant houses⁵. Given these low figures for vacancy, my study took me in the direction of examining specifically the potential displacement that the influx of employees could cause.

II. DATASETS

For my datasets I pulled a large amount of Geographic Information Systems (GIS) data from the Boulder County Assessor's Website. This data came in the form of very large comma separated value sheets (usually half a gigabyte). The goal behind pulling these figures was to get all the property values in the City of Boulder. I found the property values, however there were quite a few adjustments that needed to be made, these changes are largely discussed in the Methods section. Overall I pulled the material which gave me all the property values

¹This according to the Denver Post

²Denver Post

³Source BizWest quote from Scott Green Google's Boulder site director

⁴page 21 *Boulder Market Housing Analysis*

⁵<http://zipatlas.com/us/co/boulder/zip-code-comparison/percentage-vacant-housing-units.htm>

in Boulder County, and once I had this data I parsed it down to cut out all the surrounding towns (Superior, Longmont, Lafayette, etc.). In addition to Boulder County's information I pulled GIS information from Zillow's API to get another idea of potential housing values in the City of Boulder. Once I had all the information on housing prices, I needed to find the range of job positions and salaries for the incoming Google employees.⁶ The property values and potential salaries of the incoming employees being settled, I needed to dictate the split of potential renters versus buyers, as well as general disposable income towards housing. BBC Research and Consulting published their analysis in their 2013 report *Boulder Market Housing Analysis* in which they explained that Boulder has a half and half split between renters and buyers. This proportion is high of course because of the University of Colorado student population, so I have run my analysis with differing proportions of the incoming employees as renters or buyers. BBC's report also postulated the figure of 30 percent of a person's income should be spent on housing if their residence is still to be considered affordable.⁷

III. METHODS

Regarding methods used in this study, I mostly employed simple functions to handle the employee's incomes versus randomly selected houses in Boulder. To start with the housing prices I took all of the assessed values of the homes according to the city of Boulder. However I soon learned that in the overwhelming majority of cities within the United States, the publicly assessed value of a home is significantly lower than the market value.⁸ In the footnote I included, the website explains how the assessed and market values can differ so greatly. The main point to take away is that the

city or county assessor is determining housing values with tax ramifications in mind, as opposed to the market value of a home being a more straightforward prediction for what a home could sell for if it were to be put on the market. To reconcile the difference between these two values I pulled data from the Zillow API which gave me the median prices from all zip codes in Boulder. Zillow separated these costs into top tier and middle tier homes which is appropriate for this study because the incoming employees can be considered the upper end of the middle tier as well as the top tier. I took all the medians of all zip codes provided then took the median of these median values to get a grand total of 376,600. I subtracted the total median of all assessed properties in the city of Boulder(19,892) from the Zillow median, to get a value of 356,708. This new median was added on to all values from the Boulder Assessor's table to get the new price of a home

$$ZillowMedian - AssessorMedian = MedianAddOn$$

$$376600 - 19892 = 356708$$

Table 1: Median Values Taken from Zillow

Break Down of Medians		
Tier	Zip Code	Median
Top	80301	506100
Top	80302	603700
Top	80303	500000
Top	80304	569700
Top	80305	435200
Middle	80301	313900
Middle	80302	315200
Middle	80303	304400
Middle	80304	296300
Middle	80305	318000

Once I had the prices of the homes I had to decide on the best way to determine whether

⁶Taken from payscale.com

⁷page 24 of *Boulder Market Housing Analysis*

⁸<http://www.home-plans-advisor.com/assessed-value-vs-market-value.html>

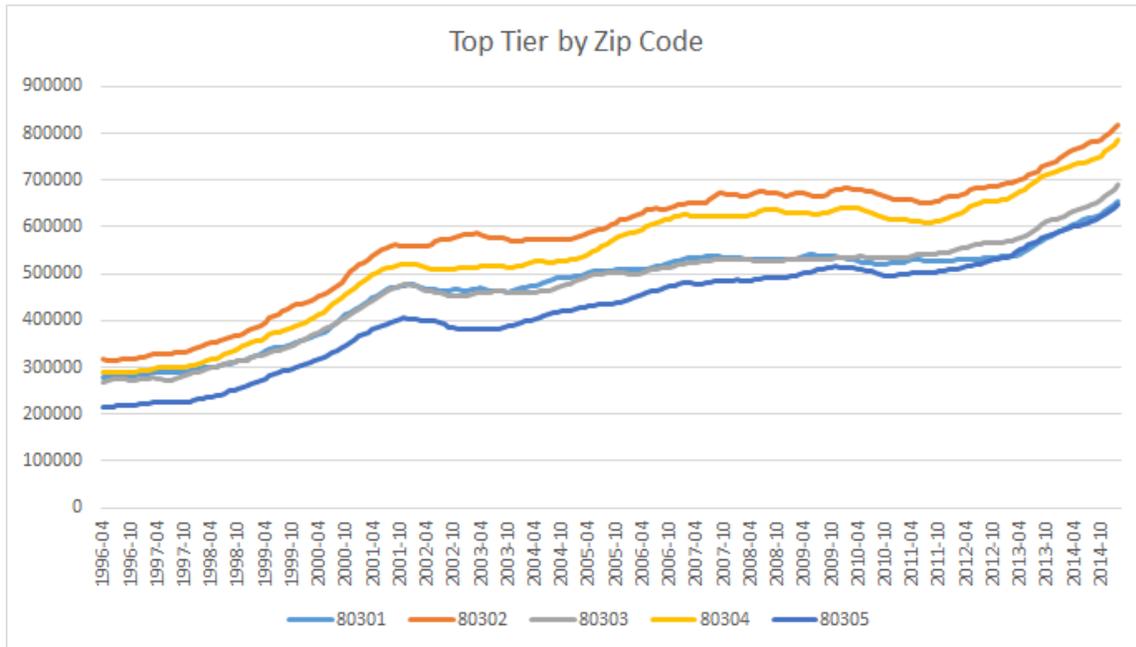


Figure 1: Top Tier Property Values by Zip Code provided by Zillow

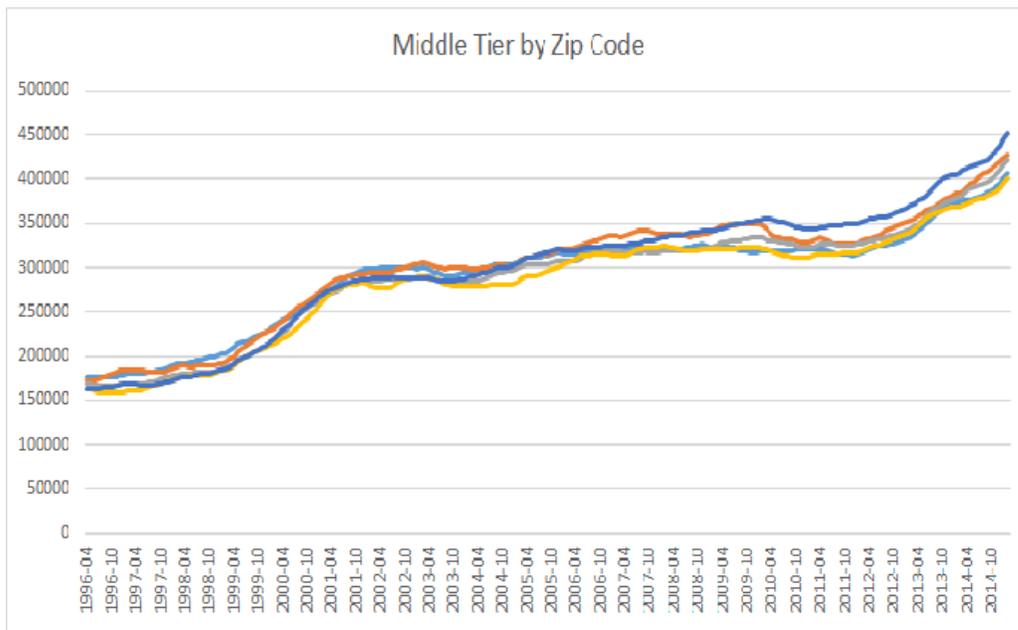


Figure 2: Middle Tier Property Values by Zip Code provided by Zillow

these incoming employees would buy or rent the homes based on its price and their income. To handle the employees I wrote a program that would generate a worker with a randomly generated job position and a salary based on that position. I created 1200 employees total, as the workforce is supposed to grow from 300 to 1500. I ran the program as a single test of 1200 employees, then I ran it ten times over generating 12000 employees to get a more solid average. Under one test analysis, I dictated half of the employees to be renters and half buyers. I also ran an analysis in which 60 percent were buyers and 40 renters. Within the differing renter buyer proportion analyses, I changed two other variables. In order for a house to be affordable for the home owner employees, they were allowed to spend 5 times their annual income in one scenario, and 4 times their annual income in another. For the rental portion of the workers I varied what is called the "price to rent ratio" as dictated by

$$\text{pricetorent} = \text{housingprice} / (\text{monthlyrent} * 12)$$

⁹. I decided to set the price to rent ratio to a high value of 10, because given BBC's figures of a 3.6 percent rental vacancy rate within the city and .2 percent vacancy rate within the university submarket, it is safe to say that Boulder is a renter's market ¹⁰. This value actually generated rental rates much lower than I would have expected but I will discuss this further in the results section. For the second run of tests I set the price-to-rent ratio to a lower 15 which would be more of an advantage to the renters and as I will argue later, more accurate to what I would expect of the results.

To revisit the proportion of employees I set to buying versus renting, I felt that a half and half split was a very high proportion of renters considering the rental market is largely students which is why I also wanted to run a 60 to 40 percent split. Even 40 percent seems to be on the higher end of renters versus buyers however I feel that both of these higher figures are justifiable due to the fact that a large

portion of employees will not have bought a house yet and will need to stay in some form of temporary housing until they are settled.

Lastly I wanted to discuss the methodology behind deciding to include all zip codes of Boulder in factoring the study. Originally I was on the fence as to whether I should exclude the 80302 zip code because it contains the predominantly student population in the area known as "The Hill". This area is not always considered the most welcoming to residents older than the traditional undergraduate age and is largely avoided by newcomers outside of that age range. While 80302 has "The Hill" it also has large portions of Folsom street and Pearl Street which are higher income, quieter, and desirable places to live. Pearl Street especially is a hub of commerce with companies like Galvanize Boulder, GNIP, SolidFire, and even Google itself. So to exclude 80302 I felt would be unjustifiable.

IV. RESULTS

The results for this study were mostly uniform with one major twist. I ran the program that generated my employees once and pulled the buying/renting rate vs the failure rate. To get an even average of the values I ran the program ten times for a larger sample size. Overall the difference in averages from running the program once to running it ten times over were minimal. Under the constraints of the housing prices being 5 times their annual income and the price to rent index being set to 15, the Google employees were overwhelmingly able to both purchase the house or rent the property they were paired up with. After multiple runs the overall average of successful purchases if the buyers were willing to spend 5 times their annual income was

$$(98.5 + 98.6 + 98.6 + 98.5) / 4 = 98.55$$

. The averages of employee's renting when the price to rent ratio was 15 was.

$$(98.5 + 98.4 + 99.3 + 98.3) / 4 = 98.6$$

⁹taken from <http://www.investopedia.com/terms/p/price-to-rent-ratio.asp>

¹⁰rates taken from pages 18-19

A 98.55 purchase rate seem suspiciously high however I felt the median method I was using was dependable so I could not find any reason to discredit the high average. Regarding the 98.6 rental rate, this result was about what I was expecting. High income employees should be able to rent in a competitive market with maybe the exception of the lowest income employee making 88,634 trying to rent out the most expensive houses. When the price to rent ratio is set at 15 the median rental rate is 1,767. Setting the price to rent ratio equal to 10 was an interesting result in that it drove down the rental pass rates more than any other metric I varied. The median rent for the price to rent ratio set at 10 was 2,650. It surprised me that this median rent could drive the average rent down to

$$(83 + 82.4 + 62.1 + 60.6) / 4 = 72.05$$

. Granted 2,650 dollars needed to be 30 percent of one's monthly income meaning

$$2650 = .3 * x$$

which leads to a monthly income of 8,833, and an annual income of 105,996. The median income of the provided employees is 135,591 meaning we should get an acceptance rate of roughly 72 percent. This is a telling result because the median annual income for the residents of Boulder was 54,539 as of 2012 ¹¹. According to BBC's analysis, the median rent as of 2012 was 1,080 ¹² which is of course much closer to the price to rent ratio of 15. So once we get an injection of 1200 employees whose median value is well over double the recent median value of Boulder's residents we could potentially see prices increase to get closer to an acceptance rate of 72 percent, hence a monthly rental rate with a median closer to 2,650.

The map on the following page is a coarse visualisation of the incoming workers with the factors of; a half and half split, housing price equal to 5 times their gross income, and 15 price to rent ratio. The yellow dots represent

purchases or rent being taken and the red is the rejection. I could not fully contain all of the points being made on the graph however While the rejected dots may seem too sparse, I believe them to be more accurate than they first appear. The dots are largely being crowded out. Clustering is not the most appropriate approach to get a cleaner view of the individually rejected properties because the purchased properties flooding the graph adds to the effect of what their potential for change could look like. Overall, this research lends itself to the simplistic conclusion that these high earning employees whose median salary is significantly higher than the rest of the currently residing population, could potentially cause massive displacement in the saturated housing market. Roughly if the 98 percent figure is taken, then we have $(1200 * .98 = 1176)$, 1176 people who would either be priced out or selling their homes come 2017. More specifically the rental market could see as much as $(600 * .98 = 588)$, 588 people being displaced. The rental market may be more of an indicator of "pricing out" or displacement in that renters are forced into competing with one another for housing as opposed to home owners who have the option to sell or not sell. This does however discount the factor of affordability of a city no longer being an option for home owners with slightly lower incomes once high income residents flood a community.

Table 2: Averages for Home Buyers

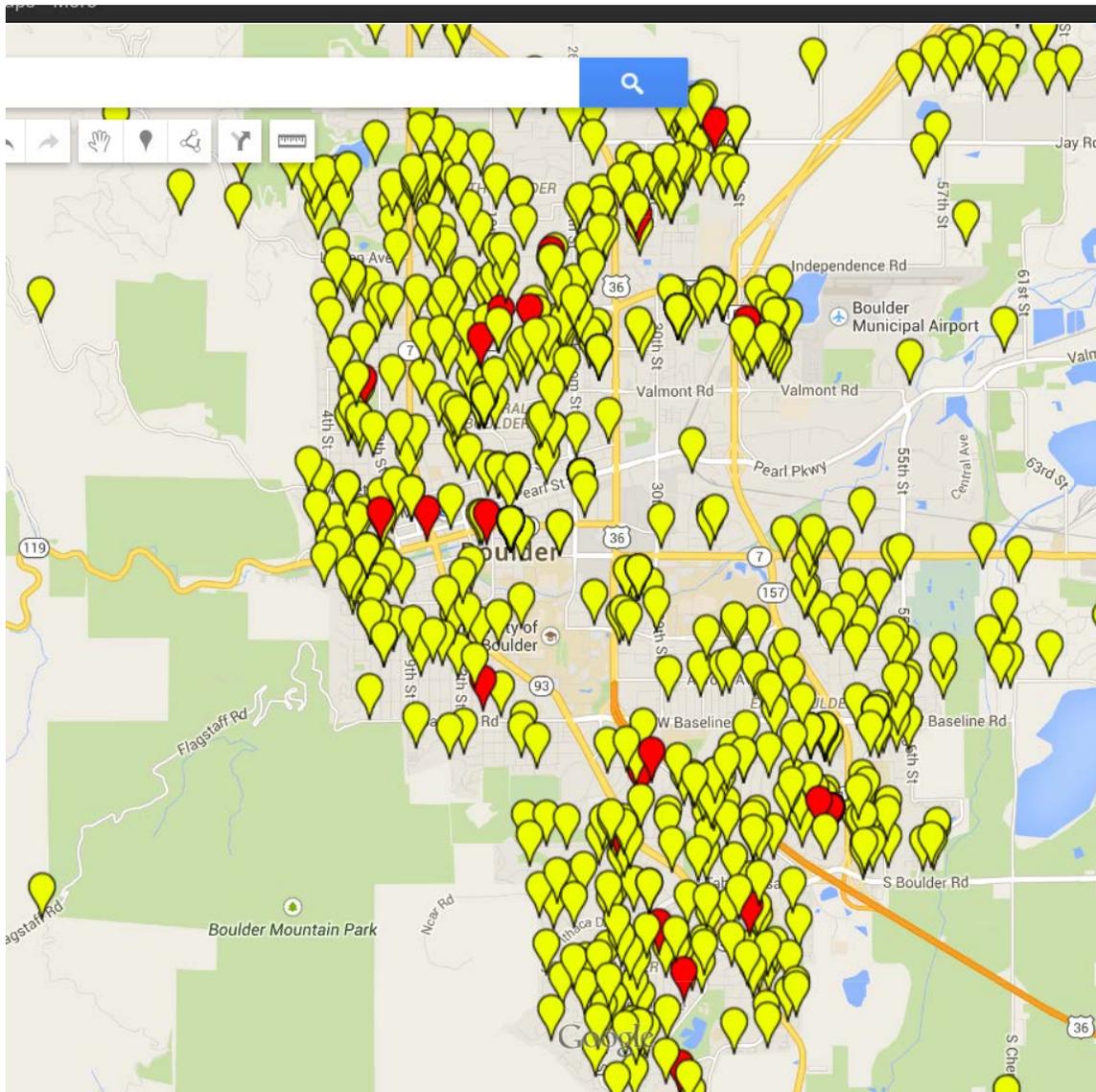
Price of House	Buyers	Houses Bought	Avg Purchase Rate
5*Gross Income	600	591	98.5
5*Gross Income	6000	5913	98.6
5*Gross Income	720	710	98.6
5*Gross Income	7200	7089	98.5
4*Gross Income	600	584	97.3
4*Gross Income	6000	5861	97.8
4*Gross Income	720	705	98
4*Gross Income	7200	7052	98

¹¹page 6 Boulder Housing Market Analysis

¹²pg 22

Table 3: Averages for Renters

Price to Rent Ratio	Renters	Houses Rented	Avg Rental Rate
10	600	498	83
10	6000	4948	82.4
10	480	298	62.1
10	4800	2908	60.6
15	600	591	98.5
15	6000	5903	98.4
15	480	477	99.3
15	4800	4718	98.3



V. DISCUSSION

I. Datasets

Reflecting on my datasets there could have been a more unified structure for the values of the homes. Pulling data from the Boulder Assessor's office distracted from the spatial component of the study because, while the data included mailing addresses it was not categorized by zip code as far as I could tell. Plus there was so much time spent on parsing out extraneous information; getting rid of surrounding areas, cutting out the mobile homes, and finally patching together massive data sets, that it was difficult to keep the progress of the study consistent. Instead of randomly selecting homes to compare against the employee, it may be better to introduce a controlled bias into the study that could see if they take to some areas more so than others. A great direction for this study to be taken in is to focus on the zip codes instead of individual properties, their market values instead of assessed values, and to use the displacement information to try to predict the future of rental and housing prices in Boulder.

II. Methods and Results

Earlier in the Dataset section I mentioned focusing on zip codes as opposed to individual properties, I would like to revisit this point as I believe it to be indicative of the main challenge regarding this study. There was so much data present in this project that the old maxim "Can't see the forest for the trees" was very appropriate. Clustering could be a great solution to this challenge because instead of focusing on individual homes scattered across the city we could examine specific areas to determine a pattern. Clustering could provide the focus needed to determine future trends regarding specific areas within the city. My method was to pull random values from the housing csv however

if future researchers could find an algorithm that in some way can determine preference we could get a much more specific mapping of displacement and how that would affect the less popular areas due to spilling over. In determining future behavior I feel the increase in median income is a good place to start but it would be interesting to see what other formulas could be more effective or which could be valuable additions. Spatially, after covering the city, moving to Boulder County as a whole to take into account the rise or decline in rent and prices of homes in the surrounding areas such as Louisville, Longmont, Superior, etcetera could be a relevant next step. Doing similar studies on these outlying areas would be interesting, because a rise in prices could be much more telling in that they have more room for expansion so arguably less of a need to raise the cost of living. Comparing and contrasting these areas with one another and seeing if the population spikes in any one of them in particular would make for an engrossing discovery into what possible areas of commerce could be up and coming. While the main focus of this particular study was to examine what effects the new Google employees could have on the housing market in the city itself, it would be even better to see how these 1200 employees could affect an entire county!

REFERENCES

- [1] Caution on Google expansion trademark of Boulder's vigilance on growth, 2015 Aguilar, John (2015)
- [2] Google made staying in Boulder priority in expansion quest, 2014 Lindenstein Joshua (2014).
- [3] Boulder Housing Market Analysis, 2013 BBC Research and Consulting (2013).
- [4] <http://zipatlas.com/us/co/boulder/zip-code-comparison/percentage-vacant-housing-units.htm>
- [5] <http://www.payscale.com/research/US/Employer=Google,inc./Salary>
<http://www.home-plans-advisor.com/assessed-value-vs-market-value.html>.
- [6] <http://www.investopedia.com/terms/p/price-to-rent-ratio.asp>
- [7] Assortative pairing and life history strategy - a cross-cultural-study.
- [8] <http://www.bouldercounty.org/dept/assessor/pages/propertydatadownload.aspx>
- [9] <http://www.zillow.com/howto/api/neighborhood-boundaries.htm>

Examining the relationship between economies of agglomeration and business performance

JOSHUA FERGE

University of Colorado, Boulder

joshua.ferge@colorado.edu

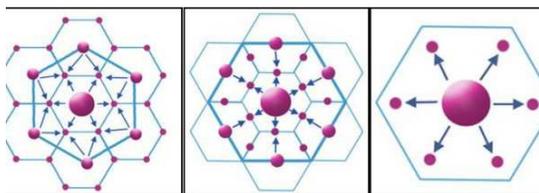
Abstract

Location, location location is a phrase that everyone knows. When starting a business, one of the most important facets to consider is the location. Currently, common metrics to consider might include property price, foot traffic, average income of people nearby, and current infrastructure to support the business. Conventionally, if there is already a similar type of business nearby, the thought is to get as far away as possible from that business, due to competition. However, some clusters, such as the diamond district in New York [7], observe that when similar businesses are clustered together, the individual businesses actually do better than if they were further away. Factors for this may include the ability to specialize – customers may go to this “hub” not exactly sure what they’re looking for, and as each diamond business can specialize due to the increased foot traffic, they can get exactly what they want. The yelp data set [1] which includes business data from ten cities in the US, will be used as the main dataset. Using number of reviews as a proxy, this paper will attempt to define a relationship between agglomeration and business success. In addition, the mathematical patterns behind the different metrics will be examined and commented on. Furthermore this paper provides as a springboard, with a section dedicated to ideas for the next steps in the process of solving this problem. This paper first contains a background section where relevant research and statistics have been summarized and put into context of this paper.

I. BACKGROUND

In relation to urban development some of the first research and thought went into Central Place Theory. Central Place Theory states that in a set of constraints, we can expect settlements to lay out in a certain way.

Figure 1: Central Place Theory $K=3,5,7$ [11]



The graphic displays a few different K principles, e.g. $K=3$ Marketing principle which states that the market area occupies a third of the area of each satellite settlement. In the

real world, many caveats arise when attempting to predict business placement using this idea, because Central Place Theory requires many unrealistic constraints, some of which are perfectly flat land, and equal purchasing power between inhabitants. However, some cities that fall in line with some of the constraints such as small cities in North Dakota exhibit features outlined by this theory, so it has at least a bit of applicability. [9]

Another economic theory which helps in understanding economies of agglomeration is Hotelling’s law [5], which states that in many markets, it benefits firms to make their products as similar as possible to other firms’ products. We can extend this idea to the location of buying a product. For example, if there were exactly two identical coffee shops in an outdoor mall that went along a street, and each were randomly placed along the street, it

would be likely that one would receive more business than the other because they are closer to the majority of people. The logical conclusion to this idea is that both coffee shops would be right next to each other in the midpoint of the mall. This idea also has some downfalls (differentiated products, street traffic, etc.) but elements of it will be shown in the analysis. [4]

Other, more practical statistics support the idea of agglomeration [2]. This paper from Cambridge compares humans to the density of matter in the universe, where the density goes from extremely high to almost 0. Businesses are the same, and within urban cities there tends to be clusters. One example of industrial clusterization is Silicon Valley where lots of massive tech companies locate themselves in order to attract more talent. Another example is "The Carpet capital of the world" in Dayton, Georgia, [3] where almost 90 percent of carpet produced internationally is in a 25 mile radius. While Silicon Valley used to be concentrated because electronics companies who used each others products could minimize shipping cost, it is now likely because of talent. However, it is likely that Dayton's carpet industry exists so close in proximity because of interrelated factories, such as yarn, machinery, and carpet tufting mills, which benefit from lower shipping costs to other factories by locating there. Around 75-80 percent of the yarn produced in the city is used by the carpet makers within Dayton. Another example of clustering that can be thought of is shopping malls, where many clothing stores lie close in proximity to each other. Ken Steif examined businesses around the New York area and deduced a few statistics that will be added upon in this paper. [4]

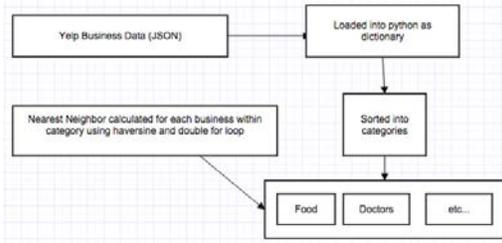
II. METHODOLOGY

To begin, the data set used, the Yelp Data Set [1], contains business data from ten cities around the US. The data for each business includes location, address, review count, reviews, average star rating, business category,

and hours open. Reviews will not be used for analysis. Another note on the data is that they include an amount of businesses that are closed. While these could be useful to some, it was not noticed at first that these businesses were being included, which skewed the results as sometime a business had replaced the closed one which resulted in the nearest neighbor of the open business to be near 0, when in actuality it was just comparing it to a business that had already been there and then closed.

The data analysis was done in Python, using Ipython as the environment, with supporting libraries `numpy`, for statistics, `Matplotlib` for plotting, and `SciPy` for more complicated statistics. The data was downloaded in a JSON format (reviews omitted), which could then be loaded into Ipython for development and analysis. Loading the data into an array did not take more than a minute. Computations also seemed to complete in a reasonable amount of time, so the use of a database was not needed. The first part of the process was to sort the businesses by categories for overview access. This was done by iterating through the businesses to create a dictionary of lists for all possible categories, then to append the business to the lists which its categories were keys. A business could have multiple categories, so a business will affect multiple categories. However, for the nearest neighbor calculations the business will use a business in the same category, so a business, depending on the category could have multiple neighbors. Ipython was extremely helpful as it runs incrementally, which means that all of the nearest neighbor calculations could be run to create the array of data used for the statistics, and they did not have to be ran more than once. Because of the decision to not use a database, this was extremely helpful. The `Matplotlib` python library was used for all of the graphs, and functioned reasonably well, while even though taking some time to generate some of the graphics.

Figure 2: Method Diagram for initial calculations



A function that calculates basic statistics were then made and each category could be analyzed by entering the corresponding dictionary key into the function. The statistics calculated include average, standard deviation, and median, for review count and nearest neighbor. Graphs generated by this function are a histogram of nearest neighbor distances, a histogram of review counts, a scatter plot of nearest neighbors, a scatter plot of review counts, and a lat-long scatter plot.

First examining the city of Phoenix, we look at the average nearest neighbor for each business type. The nearest neighbor algorithm used is the Haversine function, which calculates the distance between two points on a sphere using the latitudes and longitudes. The haversine function is given by:

Figure 3: Haversine function for distance calculation

$$2r \arcsin \left(\sqrt{\sin^2 \left(\frac{\phi_2 - \phi_1}{2} \right) + \cos(\phi_1) \cos(\phi_2) \sin^2 \left(\frac{\lambda_2 - \lambda_1}{2} \right)} \right)$$

where phi and lambda are the respective latitudes and longitudes.

This statistic gives us a good idea of what businesses are dense and what types can usually be found next to each other. This statistic works well and tails off as the distance becomes greater. Some businesses in the far left of the histogram (usually close together) include restaurants and gas stations, while ones further out include things like pest control, oncologists, and cheese shops.

The only problem with this statistic is that it does not exactly identify clusters, or if a businesses is usually clustered. For example, court

houses are in the less than .1 kilometer category, but there are only two of them. However, we can see that these two court houses are clustered together. Another problem is that it doesn't really take into account the number of businesses in the category. For instance, restaurants have a low average nearest neighbor not because they exhibit economies of agglomeration, rather that they are just more numerous and are more likely to be closer together because there are more of them. A more useful statistic may be to somehow include the number of businesses with the average nearest neighbor statistic, by dividing them. This can be seen in Table 2.

Figure 4: Haversine function for distance calculation

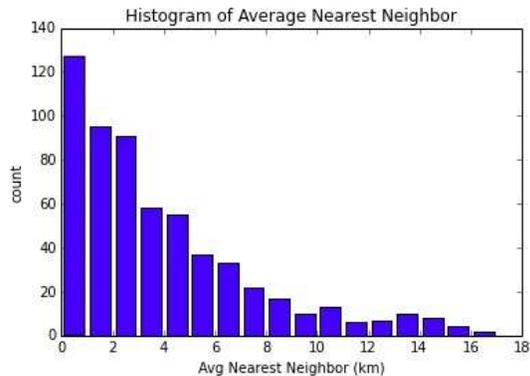


Figure 4 counts the number of business categories in each bin. The majority of categories have their average nearest neighbor less than half a kilometer away. This histogram follows a right-skewed model and as we increase average NN distance, the number of businesses falls off exponentially. It makes sense for the majority of businesses to have a nearest neighbor within a mile, as there are a lot of businesses within few categories. A bin size of one kilometer was used for NN histograms.

Table 1: Average NN Values Per Category

Food	0.323
Bars	0.523
Hotels	0.716
Doctors	0.748
Gyms	1.348
Optometrists	1.562
Computers	2.394
Hospitals	2.221
Car	2.623
Bookstores	2.831
Pediatricians	3.579
Bowling	4.651
Tax	5.352
Comfort	5.378
Radiologists	8.725
Formal	8.857
Horseback	9.611

Here we can see the average nearest neighbor for a few different categories. Food is near the bottom as there are a lot of businesses in the category, and as we zoom out into more specialized areas we can see it become greater and greater, with bowling alleys on average being 4 kilometer away from one another, and finally horseback riding being on average 9.6 kilometers apart.

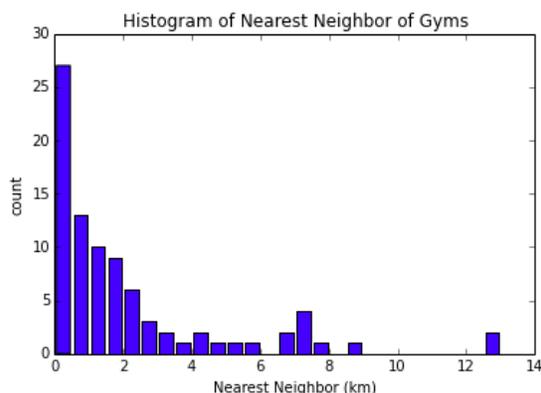
Table 2: Average NN / Business Count Values Per Category

Bars	0.019
Bookstores	0.028
Bowling	0.030
Computers	0.078
Doctors	0.130
Food	0.213
Gyms	0.316
Hospitals	0.462
Hotels	0.481
Optometrists	1.302
Pediatricians	1.725

If this average nearest neighbor is divided by the business count in a category, we get a slightly more interesting statistic. It shows that, for instance, relative to the population sizes, bars are extremely close to each other. It can be noted that in the first table gyms are 5-6 times the size of average nearest neighbor distance. However, if the size of the category is taken into account gyms are much more close to food establishments. What this means is that relative to how many gyms there are, gyms are more likely to be closer together.

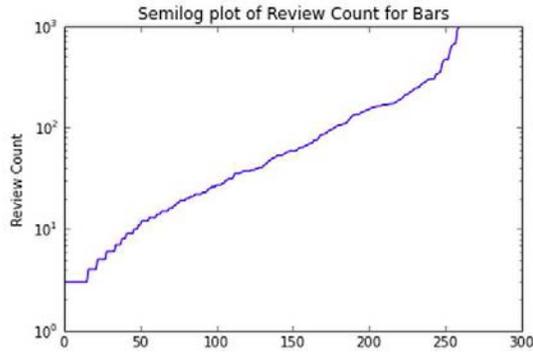
Next, we examine each category more thoroughly and test for how varied and close together the nearest neighbor is for a category.

Figure 5: Nearest Neighbor Histogram for Gyms



It can be seen for gyms that, for the most part they are always close together. This may be Hotelling’s law in effect, as gyms usually sell the same product and thus to maximize the number of customers firms choose to locate near other gyms. The histogram exhibits a right-skewed distribution that the overall NN also showed. We can further deduce that this statistic follows a power law model. This can be shown with a semilog plot.

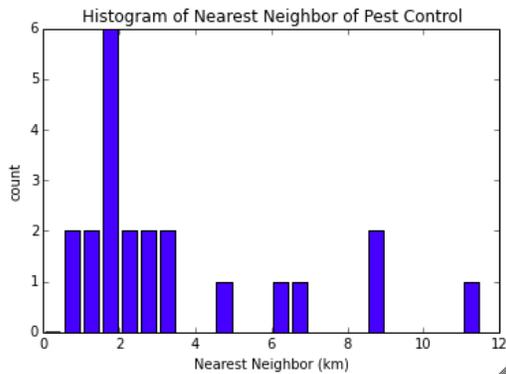
Figure 6: Bars Review Count Semilog Plot



For the most part, the semilog plot displays a straight line, which is a sign of a power law statistical model. Because of this, some statistics like mean and standard deviation lose some value as indicators of the data set because a small portion of those points impact the greatest on the entire data set. A power law function is given by

$$y = ax^k + \textit{epsilon}$$

Figure 7: Nearest Neighbor Histogram for Pest Control

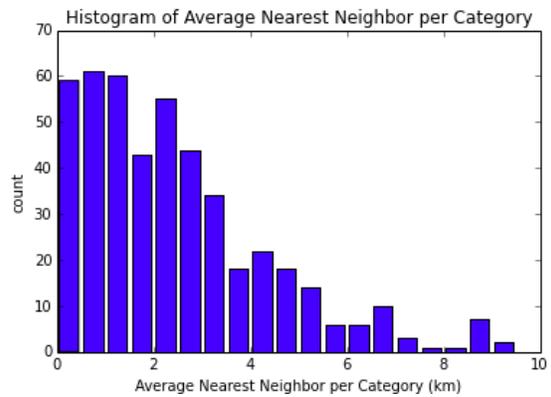


For pest control the location doesn't seem as dependent on the nearest neighbor, as it looks as though pest control businesses are more spread out in their nearest neighbor difference. One possible explanation of this is that a person never usually goes to a pest control in person, so they are never likely to pick one that is close to them (within ones that are within the same city). This means pest control operators can select properties which aren't

necessarily centrally located, and thus their average nearest neighbor is higher.

To form an even greater big picture, a histogram for the standard deviations and averages of some categories are provided. These figures and tables provide a bit more context. As well with the standard deviation it can be seen that some vary greatly.

Figure 8: Average NN Category Histogram



The histogram bars on this graph are a bit more non uniform than the individual business NN histogram. This makes sense as a lot of businesses with a particular NN distance are agglomerated into one in this histogram.

Figure 9: Standard Dev. NN Category Histogram

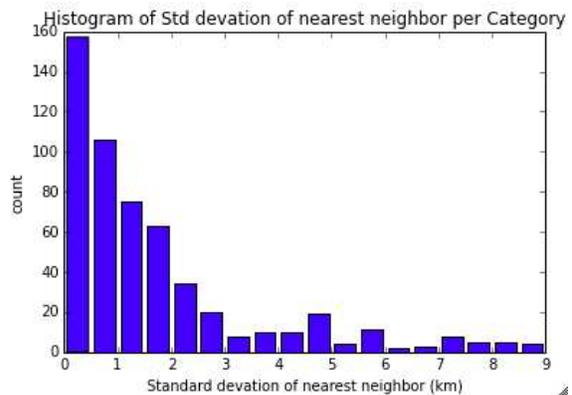


Table 3: *Standard Dev. NN Category Values*

Bookstores	0.095
Pediatricians	0.141
Optometrists	0.165
Computers	0.194
Hotels	0.243
Hospitals	0.252
Bars	0.357
Doctors	0.413
Food	0.422
Bowling	0.540
Gyms	1.112
Radiologists	7.184

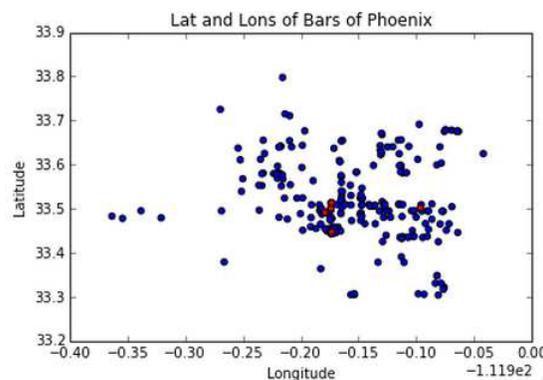
The next logical step is to attempt to correlate nearest neighbor with success. Number of reviews seems like a decent proxy, however as it will be shown there are a few problems with that as a reliable correlation member. Looking at a scatter plot of a category, it can be noticed that for the most part, there are a few outlier businesses with five to ten times the number of reviews of the average member of that category. Most businesses of a category have a minimal number of reviews while a few capture most of them. In theory this could be because yelp reviewers only review restaurants with lots of reviews because of controversy or a preference for businesses with a greater amount of yelp reviews, possibly because of the quality of the business or possibly just because those businesses come up first in a suggestion for a place to go.

Because of the non normality of the data, we cannot use simple correlation techniques like Pearson's. However, a possible alternative is to look at the top five to ten most successful businesses in a category and test whether or not their average nearest neighbor is lower than the average for the entire category. This would show that the most successful businesses tend to aggregate.

Looking at businesses with a large population, such as bars, we first extract the top five businesses ordered by reviews, and along with them their nearest neighbor. The mean is then

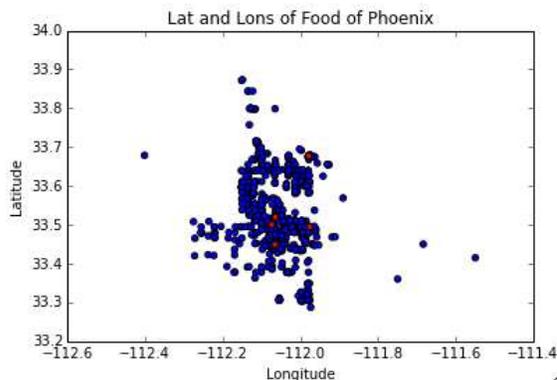
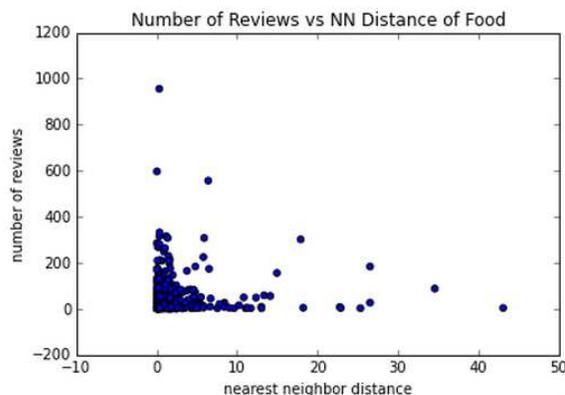
calculated. The mean and median for nearest neighbor of bars is 3 and 1.53 respectively. For the aggregate, the mean and median is .81 and .35. For the nearest neighbor data set, we have shown that there are a few outliers with very large nearest neighbor distances, so that the median may be a more accurate statistic in showing what a usual business would be. Since the median is smaller in the total population it is hard to conclude any correlations.

Plotting this data set however, shows that four of the top five bars seem to be roughly in the same area.

Figure 10: *Bar location latitudes and longitudes*

Since the points are relatively close to each other and not near the poles they have been left in latitude and longitude for simplicity.

The same map for food is also somewhat compelling, as 3 of the 5 top establishments seem to be centrally located. What could possibly be said is that successful restaurants may be more likely to locate together, rather than that they are likely to be located near other restaurants which may or not be successful (as far as number of reviews goes).

Figure 11: Food location latitudes and longitudes**Figure 12:** Scatter Plot of Nearest Neighbor Versus Distance

Looking at it visually, there doesn't seem to be any correlation between the two points. A hypothesis of negative correlation between nearest neighbor distance and number of reviews doesn't seem to be rejected by this graph. One can also note that there are some outliers with a high nearest neighbor distance and number of reviews.

A Wilcoxon Test was ran on the data set, and it returned a number extremely close to 0, which means no correlation. The data seems to be very noisy, and the use of nearest neighbor as a metric may be put into question before we completely reject our hypothesis.

III. LIMITATIONS

Due to time constraints, a few advanced analyses which could have been done were not. One of these was to compare number of reviews vs starcount, which would have validated using number of reviews as a proxy for success a bit more. In theory if a business has a lot of reviews on yelp, it means a lot of people go to it. One would think that this means the business was good, however this notion has not been validated with the data.

A small analysis in this paper showed the possibility of a business category like pest control not being clustered. A related hypothesis to explore may be that some businesses don't require a good location to be successful.

One of the most likely analyses to turn up a more successful result is to split the businesses of a category into percentage bins (possibly each 10% into their own bin) then take the nearest neighbor using only the businesses in one of the particular bins. This would refine the hypothesis to the notion that successful businesses cluster together, as it may be found that successful restaurants may be more likely to found next to each other.

There have also been many NLP (Neuro-Linguistic Processing) studies with this specific yelp data. [10] These could have been used as a supporting or main metric in showing the success of a restaurant.

Another geo-statistical tool that could have been done was Kriging, creating a semi-variogram based on non-continuous discrete values, which in this case could have been number of reviews or rating. Due to the power-law model of these statistics it would have been interesting to see if using kriging we could extract any knowledge about predictions of future business placement, or to see if reviews is correlated with location at all.

Another possible technique would be using machine learning classifications to attempt to guess where a more successful business would be. According to Kanevski et. al. in their paper [6] about geo-statistical applications of machine learning, machine learning

is well-suited for non-linear high-dimension data sets with high amounts of noise. Much more advanced variograms and correlations could be achieved using the full rich Yelp dataset, along with other supporting data sets in a large dimensional model along with machine learning.

As discussed in the background, topography of the area has a great effect on human civilization placement, and most likely business clustering. In a city where there are lots of elevation changes for example, one would hypothesize that most businesses that receive foot traffic would be located on a mostly flat area, thus resulting in a clustering based on topography. Using more in depth data of a city could have also been interesting, seeing if average nearest neighbor for a category decreased or increased for example. An original idea was to look at every major city on the Yelp data along with population statistics to analyze the relationship of clustering, population, and population density, but due to time constraints this was not done.

Central Place Theory could also be applied to these sort of business location questions. Possible questions answered using this line of reasoning might be what types of cities and businesses conform, if at all to this idea in the real world.

There might already be research on this, but examining why number of Yelp reviews conform to a power law might be interesting. It would also be interesting if a businesses' number of customers or revenue conforms to a power law. Also determining whether online websites and webpages conform to a power law might also be interesting, as this has interesting implications on the psychology and hive mind of the internet. (Someone making a new product might not want to after looking at the percentages of online traffic).

Another tool that might have been useful is a interactive web map that includes nearest neighbor info and review count, as well as other relevant statistics. Using this, a possible heuristic could have been come up with for top review businesses and business clustering.

IV. CONCLUSION

First learning about the data, it was assessed that nearest neighbor forms a power law statistical model, with a small amount of businesses that have a large nearest neighbor and a large amount with a very small nearest neighbor.

It was also shown that viewed as an aggregate from categories, the categories also show a power law model.

Number of reviews shows a power law model with most reviews coming in for a small amount of businesses.

The metric average nearest neighbor was used to analyze aggregate businesses by category. Dividing by the number of businesses in the category gave us a more accurate representation because businesses with a higher count are naturally more likely to be close together. It was shown that some nodes in the yelp data (courthouses, trampoline parks) exhibit very high clustering, relative to their population size.

Due to the power law model nature of these items, a Wilcoxon test was used, as well as a visual scatter plot to test for correlation in the data. With these tests, none were found.

There may be some truth to the idea that like-businesses tend to cluster to receive more revenue, however from this data set and analysis it is impossible to say so. Hotelling's law and Central Place Theory are very specific constructs, and they are hard to apply in this set of data for the most part.

The choice of using nearest neighbor could also be questioned, as it doesn't really tell much about how many other businesses are around a single node, just that there is at least one. Looking at maps of the data, heuristically it seems that there might be some sort of correlation, however it most likely requires more complex analysis to determine this.

While that can be deduced it is hard to say that businesses who cluster together receive more business because there are other factors outside these data that effect how businesses group and are reviewed. For example, if all of these businesses were in a downtown area,

it would be common sense that they would have more reviews than a business in a more remote area. A way to limit this a bit would be to factor business density into the calculations.

Overall, much was learned about the nearest neighbor property and review count of yelp business data.

REFERENCES

- [1] Yelp Dataset Challenge, Yelp Inc. 4/8/2015. http://www.yelp.com/dataset_challenge
- [2] Masahisa Fujita, Jacques-Francois Thisse. *Economics of Agglomeration*. 4/8/2015. <http://assets.cambridge.org/97805218/01386/sample/9780521801386ws.pdf>
- [3] Advameg, Inc. *Clusters*. 4/8/2015. <http://www.referenceforbusiness.com/small/Bo-Co/Clusters.html>
- [4] Ken Steif. Why Do Certain Retail Stores Cluster Together? 4/3/2015. <http://www.planetizen.com/node/65765>
- [5] Andrew Boyd. Hotelling's Law. 4/3/2015. <http://www.uh.edu/engines/epi2692.htm>
- [6] M. Kanevski, et al. Machine Learning Algorithms for GeoSpatial Data. 4/26/2015. http://www.iemss.org/iemss2008/uploads/Main/S04-09-Kanevski_et_al-IEMSS2008.pdf
- [7] Chana Joffe-Walt, Adam Davidson. Why Clusters Of Like Businesses Thrive. 4/1/2015. <http://www.npr.org/templates/story/story.php?storyId=121304873>
- [9] Charles University in Prague. Christaller's Central Place Theory. 4/4/2015. <http://uprav.ff.cuni.cz/?q=system/files/christaller.pdf>
- [10] Google Scholar Search, 5/3/2015. <https://goo.gl/RRAzAF>
- [11] Central Place Theory Diagram 4/4/2015. http://en.wikipedia.org/wiki/Central_place_theory

Geospatial Clustering and Classifying of Twitter Data

TAYLOR GRAHAM

University of Colorado

taylor.s.graham@colorado.edu

Spring 2015

Abstract

This paper investigates how to organize and classify a large collection of geotagged Twitter data, working with a dataset of almost 15 million tweets collected directly from Twitter. The approach described here combines spatial analysis of the GPS location of the tweet with content analysis of the text and hashtags associated with the same tweet. I use the spatial distribution of where people tweet from to define 'hot spots', which are regions across America where the largest volume of tweets come from. I then look into those specific regions to try and identify the most popular topics and hashtags of the regions, and compare differences between each area. I expect my results to vary based on where people are tweeting from. Posts on the east coast will likely be about a different subject than posts on the west coast, for example.

I. INTRODUCTION

Social media has drastically changed the way that we as individuals communicate with the world around us in the past 15 years or so. During the early years of social media, a user would be connected to their direct peers, their friends and family and colleagues (if they so wanted to be, that is). However, as the social platform grew and matured, people began interacting worldwide, and with people who they know literally nothing about. People began connecting based on similar ideologies and common discussion topics instead of simply who you knew or wished you knew.

Twitter is a social media platform loosely based on this idea. On Twitter, users are able to share their opinions and communicate with other users around the entire world, as long as you are able to keep your message short: Twitter limits users to 140 characters for any particular tweet. As a part of this message, users can include hashtags, which are distinct keywords prefixed with a # symbol. These

hashtags are used to apply some sort of theme or distinct message with each tweet. This also works well for Twitter, because they can easily track the most popular hashtags across the world, and identify the 'trending' topics, or the most used hashtags over time. I wanted to investigate the trending topics on Twitter, but with a finer granularity than what is currently reported on by Twitter.

II. DATA

My dataset was collected by downloading public tweets and their metadata using Twitter's streaming API. More information about the API can be found here: <https://dev.twitter.com/streaming/overview>. Data was collected during intervals during a two week period in late March, 2015. I had to stop collecting data for certain periods of time because I did not have a dedicated server to collect the tweets, so any time I had to move my computer, such as going to class, I had to turn off the data stream. Figure ?? is a plot of the specific times

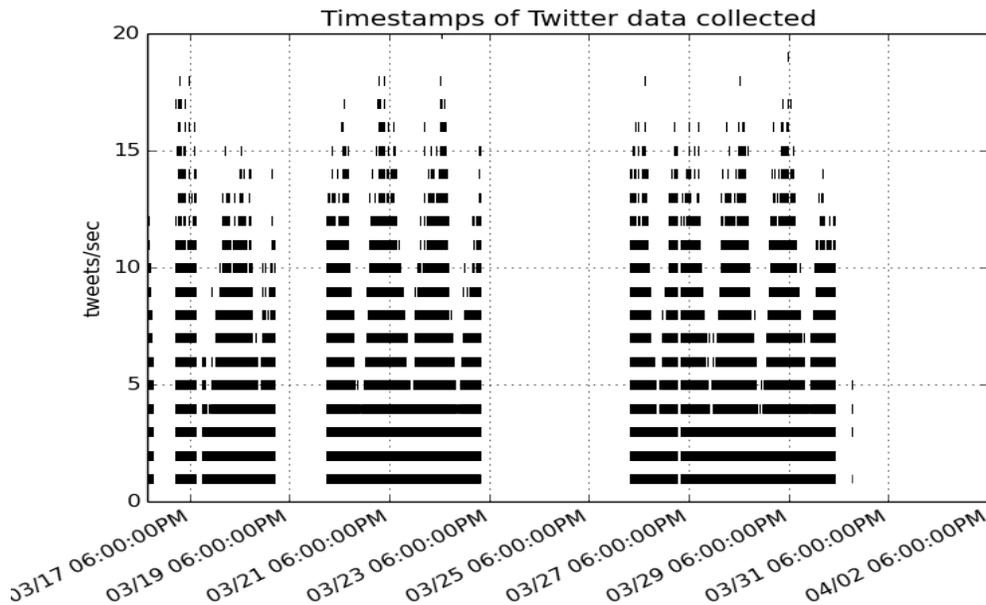


Figure 1: Data Availability

when data is available. Data that contained geolocation data and was located in the United States was kept, while the rest of the data was discarded. Luckily, Twitter's API is built so that you can query it with a bounding box of gps coordinates. The API will then only return data that is geolocated in that particular bounding box.

Data is returned from Twitter's servers formatted as JSON objects. These objects have many fields useful to this research such as the message text and hashtags as well as the gps coordinates, as mentioned earlier. However, these objects also contained many other fields that were unnecessary for the system, such as user information, retweets, and favorite counts. In order to limit the size of the database, I chose to only save the relevant fields: tweet_id, message text, timestamp, gps coordinates, and a user_id. This reduced the size of a typical tweet from around 2.5kb to 0.8kb. Once I had the final format of the data, I populated a postGIS database using python scripts. In

total, I collected 14,412,529 tweets, with 754,109 unique hashtags shared between them.

A problem I ran into using Twitter's data is that there are quite a bit of 'bots' that are set up by humans that will automatically post tweets to the system, usually spamming some advertisement or recruitment event to all of twitter. As you can imagine, these bots can greatly skew the frequency of hashtags in a particular region, or even the entire world if they send enough messages. In order to try and limit the amount of spam from a single bot, I choose to further filter the data set by ensuring that each post from a specific gps location was from a unique user_id. This takes care of the problem of a single user posting many tweets from a specific location, which is the case with these bots. By doing this filtering step, I cut the total amount of tweets I was considering from 14 million, to about 2.5 million.

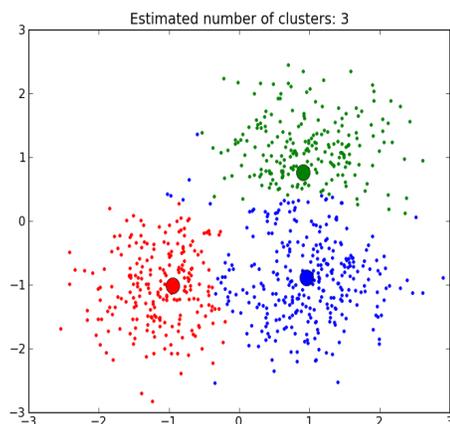
III. METHODS

Given a large collection of geotagged tweets, I wanted to automatically find popular places at which people are tweeting from. In measuring how popular a place is, I only want to consider the number of unique users who post a tweet from that particular location. This avoids things like apartment complexes and common residential areas from overwhelming the results, as a lot of users tend to tweet daily from their homes and offices. We expect this will also take into account the wide variability in tweeting rates and behaviour across different individuals. Most of the analysis methods were inspired from the paper [?], which did similar analysis on Flickr picture sets instead of Twitter tweets.

The problem of finding the 'hot spots' of the most frequently tweeted from areas in the country can be viewed as a problem of clustering points in a two-dimensional feature space. I choose to use mean-shift clustering instead of using a fixed-cluster method such as k-means, which requires the user to define the amount of clusters before hand. Mean-shift is a non-parametric method for estimating the modes of an underlying probability distribution from a set of samples, given just an estimate of the scale of the data. For my project specifically, there is an underlying unobservable probability distribution of where people are tweeting from, with modes being the most popular or frequent places to tweet from. Mean shift estimates the modes of these underlying distributions by only directly observing the locations that people are tweeting from.

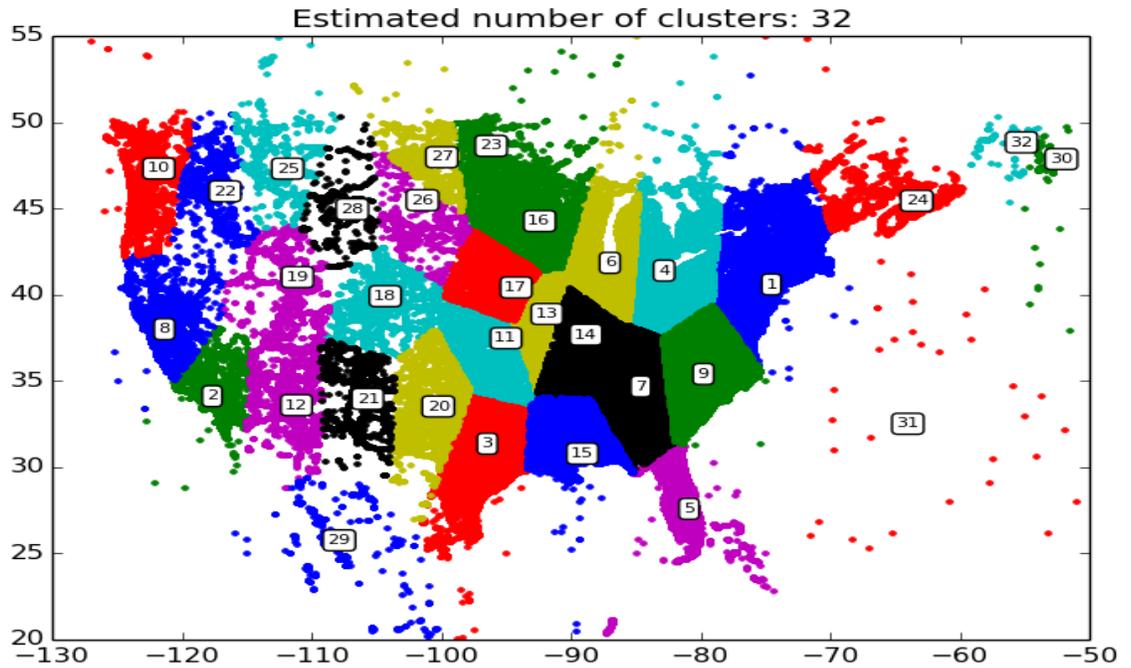
The main idea behind mean shift is to treat the gps points in the 2-d feature space as an empirical probability density function where dense regions in the feature space correspond to the local maxima or modes of the underlying distribution. For each point in the world a gradient ascent procedure is performed on the local estimated density, until that density

converges. The stationary points identified by this procedure represent the modes of the distribution. Additionally, the data points associated roughly with the same stationary point are considered members of the same cluster. Below is an example of the result of mean shift being ran on arbitrary points in space.



Once I generate clusters using mean shift, then I begin to look at the topics of each different cluster. A single query was made to my database, which returned a list of every single hashtag contained in every single tweet within that cluster. I then parsed through the list for each cluster, and sorted them based on the frequency of each hashtag. I tracked the top 5 hashtags in each cluster, since I was only interested in the most common topics in that region. Ideally, the major topics between each different cluster would change, so that only analysing the top 5 hashtags will be enough to see difference. Once topics were identified, comparisons were made between neighbouring clusters as well as clusters that were in entirely different parts of America. These comparisons highlighted some conclusions about Tweets based solely on the location where they originate from.

Figure 2: Clustered Tweets



Location	Top Hashtag	2nd Hashtag	3rd Hashtag	4th Hashtag	5th Hashtag
Cluster 1	NYC	nyc	newyork	photo	Brooklyn
Cluster 2	LA	losangeles	LosAngeles	California	la
Cluster 3	SXSW	sxsw	photo	Austin	austin
Cluster 4	photo	Job	toronto	apple	applegeeks
Cluster 5	Miami	miami	photo	miamibeach	ultra2015
Cluster 6	chicago	Chicago	photo	Job	breakfast
Cluster 7	Atlanta	Job	Jobs	TweetMyJobs	Nursing
Cluster 8	photo	SanFrancisco	sf	SF	love
Cluster 9	photo	trndnl	Job	nc	clt
Cluster 10	photo	seattle	Seattle	vancouver	Vancouver
Cluster 11	earthquake	photo	Earthquake	Sismo	USGS
Cluster 12	photo	fashion	beautiful	fashionshow	springintostyle
Cluster 13	jagstate	RubberBandYogis	FLUFFNFOLD	Missouri	PetsLoveUsDoingYoga
Cluster 14	photo	trndnl	Memphis	HAA	TheHauntedGhostTown
Cluster 15	PCB	pcb2k15	PCB2K15	NOLA	SpringBreak
Cluster 16	photo	trndnl	Job	EauClaire	minneapolis
Cluster 17	photo	Job	kc	kcfw	DesMoines
Cluster 18	photo	drums	BeatADay	denver	Denver
Cluster 19	utah	nature	photo	mountains	travel
Cluster 20	trndnl	photo	BoomerSooner	buyacar	kbb

Table 1: Most popular hashtags in each cluster

IV. RESULTS AND DISCUSSION

It turns out that the hashtags that people use when they are tweeting varies greatly depending on where those people are tweeting from. Figure ?? is a plot of one solution to the clustering algorithm ran on the entire data set, 2,469,476 tweets in total. The color of the cluster is arbitrary, and is only there so that it is easier to tell two clusters apart. The clusters are ordered by the amount of tweets in each cluster. Cluster 1 has the largest volume of tweets, Cluster 2 is the second most popular, and so on. Table ?? is a list of the top five hashtags found in the top 20 most popular clusters being analysed. I choose to only display the top 20 clusters, as they are the most frequently tweeted from areas in America. Additionally, the clusters after 20 have less variation in users, so they become dominated by the hashtags #Job, #photo, #trndl, and #TweetMyJobs.

There are a couple of things you can immediately conclude from the Figure. First of all, you can clearly see the United States, just by plotting the Tweets collected. Some interesting things I noticed were the volume of tweets in the Bahamas, and the fact that you can actually make out the great lakes. Also, you can tell by the density of the data points that there are much more Tweets being posted from the east coast and west coast compared to the midwest. This is likely due to the difference in population of those areas rather than the rate at which individuals are tweeting. I noticed that the tweets begin to disperse rapidly around Cluster 29, at the Texas - Mexico border. I speculate that this is due to the language barrier, and that people begin posting in Spanish more frequently somewhere around there.

Looking at the table you can clearly see differences in the topics between each cluster. The most posted tags were typically the closest large city or state in that cluster. In fact, in many cases, you can begin to guess what part of the United States the cluster is located in, without even looking at the map.

V. LIMITATIONS AND FUTURE WORK

A large limitation of this project was actually finding the perfect values to seed my clustering algorithm with. These values greatly impact the clusters that mean shift identifies, and I suspect that further work on the algorithm could have found some better means. There was also some variability on the size of each cluster. I sort of played around with the algorithm until the resulting clusters seemed reasonably small on a national level, but to where you can still make out the clusters. Further work could have been done to find the perfect number for this value as well. I think additional studies at a city level instead of the country wide level would be worth looking into. The clustering algorithm would have to be further tweaked to fit to the data, but I still think that popular regions within a city would be identified, and that those regions would potentially have differing tweet content.

Another huge limitation I ran into specifically with the data was the amount of posts that were clearly from a Twitter posting bot. These bots will post thousands of messages a day, each one with the same or similar hashtags. Despite my efforts to remove duplicate posts and posts from a single user in the same location, some of these hashtags still made it through. Many of the clusters were dominated by the hashtags #Job and #Jobs. This is more evident in clusters 25+, as those are the clusters with the least amount of tweets coming from them. The tag #photo was by far the most common hashtag, coming up as first in most of the clusters.

One thing I didn't take into account when doing tweet analysis was the lowercase and uppercase hashtags. According to Twitter, #NYC is different than #nyc, but for my analysis, they could easily be considered the same tag entirely. Doing so would allow for a larger variation in the results, as the 'almost duplicates' would be combined.

VI. CONCLUSION

An obvious extension of this study is looking not only at the contents of each cluster using static data collected previously, but rather looking at the data as it is consumed by the system in real time. This would allow us to not only see the differences between clusters, but also how those clusters change over time. This would be particularly interesting to watch as some new topic emerges on social media. An immediate example I thought of is the current protests happening in Baltimore. In that area of the country, I would expect the frequency of tweets to increase, making some larger clusters in the area than what was there previously. Additionally, I would expect the most frequent hashtags to not only be different than the west coast and down south, but there would also be a point where the most popular hashtags in that region would change rapidly, as the protests first began.

In this paper, I introduced several techniques for analyzing and classifying a large collection of geotagged Twitter data. The approach described combines spatial analysis of the GPS latitude and longitude of a tweet, with content analysis of the text and hashtags associated with the same tweet. I present a technique to automatically identify the places in the United States which have the most tweets being posted using the mean-shift clustering algorithm. Then, once regions are defined, I investigate the theme of each region by identifying the top 5 most used hashtags within each region. Comparisons were made between the most popular hashtags in each region. Preliminary investigation shows that the process worked, and that there are clear differences between the hashtags used in different locations across the United States. However, further analysis and refinement of the algorithms used may be necessary to accurately draw meaningful conclusions beyond that.

REFERENCES

- [Mapping the World's Photos, 2009] David Crandall, Lars Backstrom, Daniel Huttenlocher, Jon Kleinberg, (2009) Mapping The World's Photos
- [Mean Shift and other clustering algorithms] <http://scikit-learn.org/stable/modules/generated/sklearn.cluster.MeanShift.html>

Creating Distributed Temperature, Precipitation, Solar Radiation, and Humidity Maps for Boulder, CO

AMIR KASHIPAZHA

University of Colorado
amka6009@colorado.edu

Abstract

Weather Underground (WU) provides considerable climatological data. These data are obtained from meteorological and private weather stations. These stations record weather conditions of a location or a point at a specific time. To estimate climatological properties, at adjacent areas or places between stations, distributed maps of climate properties need to be created. These maps can be generated using interpolation techniques. The interpolated maps show estimations of specific climate property at a certain time.

WU provides multiple categories of climate data. For this project we used temperature, precipitation, solar radiation, and percent of humidity. The main goals of this project were to develop python script to access WU data, interpolate the data using IDW and Kriging, and visualize the interpolated maps.

IDW and Kriging generated distributed maps for the desired climate data. Kriging produced interpolated maps which were smoother. In the other words, variation of a specific climate data on distributed map was not large in a small distance like 150m. These variations were found specifically for temperature and precipitation. On the other hand, IDW generated distributed maps with rapid changes in small distances. These changes were due to large differences in observed values of adjacent stations. IDW represented potential sources of errors needed to be considered. This helps to distinguish stations that over or under measuring a specific climate property.

I. INTRODUCTION

Interpolation is a numerical analysis that calculates new data points within the range of a discrete set of known data points [Wikipedia]. In general, data is collected from specific locations and then we find ways to estimate range of that data in between measured locations. These data could be climatological, mining, or depth of water in well. Interpolation techniques have been used for this estimation. Several interpolation methods are available. Among them Inverse Distance Weighting (IDW) and Kriging are the most popular. These methods use measured value of each point and distances between points for interpolation.

IDW is a deterministic spatial method which is based on measured values or in specific mathematical formulas. This method uses values of any given pair

of points but their similarity is inversely related to the distance between them [George Y. Lu, David W. Wong 2008]. Kriging is geostatistical method based on a statistical relationship among the measured points. Kriging was initiated at the gold mines in South Africa where mine companies collected data and were interested to estimate gold ore values from several measured points before they started mining [Krige Danie and Kleingeld Wynand 2005]. In general, interpolation methods use measured values of scattered points and then create distributed maps of them over space. They produce the distributed map as a grid. Climate data are recorded in a climatological station. These stations could be considered as points. We can interpolate recorded data from these stations in order to estimate range of climatological properties between them.

In this project, IDW and kriging were used to interpolate climate data of 35 stations in Boulder, CO (Figure 1). Interpolation was performed for temperature, precipitation, solar radiation, and percent of humidity. The desired climate data were not available for all stations on specific day. Some stations do not record solar radiation or precipitation, moreover; precipitation was 0mm for some stations. Thus, interpolation performed only on stations that have the data. For example, on 4/17/2015, solar radiation distributed map was created using 5 stations instead of 35 stations.

II. MATERIALS AND METHODS

Data for this project were obtained from the Weather Underground. One of goals of this project was to generate real time interpolated maps for current day. Due to the limited API calls, which constrained the free users to 10 calls per minute and 500 calls per day, this project was divided into two separated sections. In section one, python codes were developed, using request library, to get climate data from WU. Python time library was added to this code in order to avoid over calling API numbers. This code finds climate stations near Boulder. Then, it obtains real time data from each station and save it in a text file in JSON format.

In the second section, the saved text file was used to create distributed maps for climate data. The file with JSON format is not readable for GIS software. Python shape-file (pyshp 1.2.1) library was incorporated to create point shape-file for each climate station. This library produce one shape-file for each desired climate data like precipitation. Before, creating shape-files, no data values were removed. Since the data belong to a small region, variation range of the them were not large. Due to this reason, the normalized values of the data were calculated. Then one shape-file was created for each desired climate data which contains that observed and normalized values. The pyshp inserts numbers as strings in the attribute table of shape-files and does not define

projection of the shape-files. In this regards, python Arcpy library was used to define projection and convert numerical strings to numbers. Arcpy is python library that provides access to all ArcGIS tools. We can use it to develop programs that operate on geographical data. IDW and Kriging tools from this library was incorporated to interpolate climate data for Boulder, CO.

Finally, the distributed maps were visualized by ArcMap. One of the goals of this project was to automate map visualization. Python Basemap library was used to accomplish this goal. But this library does not add legends for shape-files automatically. So we need to do it manually which is time consuming. Visualizing maps with ArcMap is manual and also time consuming, however; it has much more tools for visualization rather than the python Basemap. So ArcMap was used for visualization since it produces maps with the highest quality.

III. RESULTS

Results of interpolation methods in 4/17/2015 are shown in figures 2 to 17. As these figures show, the extent of interpolated climate properties are not the same. This is due to the availability of data from climate station.

Maximum and minimum recorded temperatures for Boulder climate stations were 6.3 and 3.4 Degrees Celsius. Figure 2 shows IDW interpolation map. This figure shows rapid temperature difference for some pretty closed stations. A difference was found was between Old North Boulder and Melody-Catalpa. These stations are located in west Boulder. Recorded temperatures for these two stations were 3.4 and 5.2 respectively. Recorded temperatures for Nobel Park and Internet station were 5.7 and 3.9. This is interesting, since the distance between these stations is 300m. There are two Paolo Park climate stations in Boulder temperature differences between them is 2.3 Degrees Celsius while they are 150m apart. Although temperature differences are not considerable for these stations, these differences affected in-

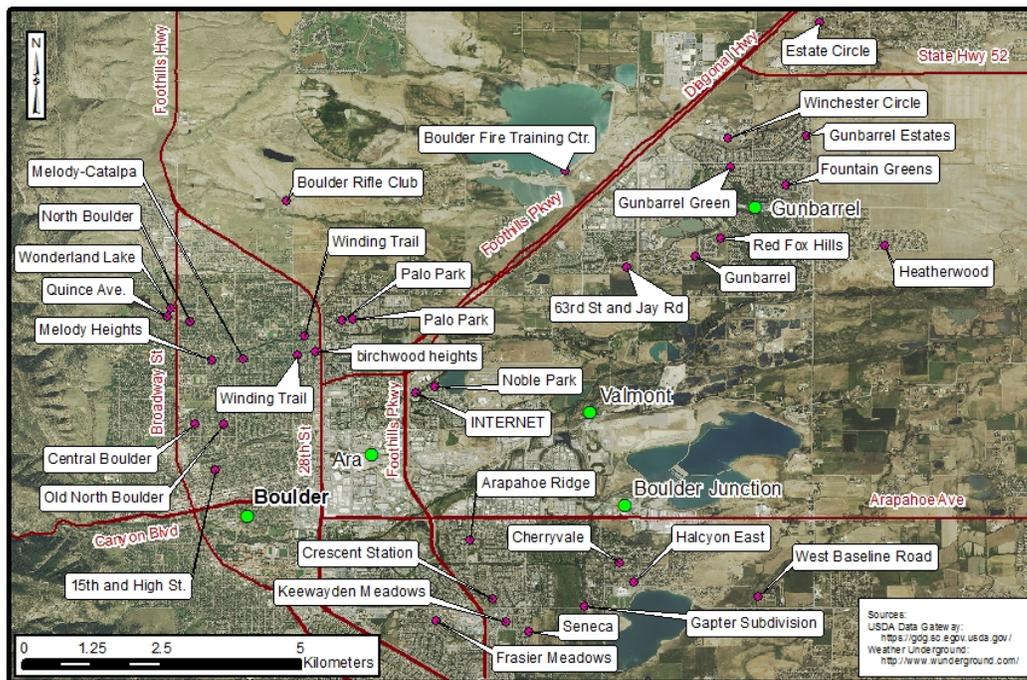


Figure 1: Location of Climatological Stations in Boulder, CO

terpolation specifically for IDW. Unlike IDW, Kriging created a smoother temperature map. It is important to find the reason for these temperature differences in short distances since they could be a source of error. Another interesting result was for Kriging. This method produced a temperature map which its minimum and maximum were 4 and 4.7 respectively (Figure 4). This is not the same as the measured temperature.

IDW interpolation generated smoother maps for precipitation compared to temperature. However there are areas that need to be considered. Maximum recorded precipitation on 4/17/2015 was 69mm for Halcyon East station. This station is located in south Boulder. Recorded precipitation of this station is considerably higher than other stations. The second largest recorded precipitation was 40mm. It is possible that the Halcyon East over measured the precipitation. Precipitation range for adjacent station of the Halcyon East varies between 30 to 40mm. One interesting result of

the precipitation was seen in the West Boulder areas. Recorded precipitation for Quince Ave was 23mm, while it was more than 28mm for its adjacent stations. As interpolations maps shows lower precipitation of the Quince Ave did not considerably affect interpolation. The reason could be the location of this station, which is in the very west side. The same result was found for Gunbarrel Estates station. Recorded precipitation for this station is 10mm higher than its adjacent station but the station does not have considerable affect on the interpolated map. In the other hand, Red Fox Hills considerably affected the interpolation map since it is not in the boarder of maps.

Figures 10 and 12 are interpolated maps of solar radiation. These maps show the solar radiation decrease from southwest to north east direction. Maximum and minimum solar radiations were 185 and 33 watt per square meter for Wonderland Lake and Boulder Fire Training center respectively. These stations are 6km apart. It is important to find the reason

for these considerable differences in a short distances. One reason could be cloud cover or cloud density since 4/17/2015 was a rainy day.

Minimum and maximum recorded humidities were 80 and 100 percent in the 4/17/2015. The humidity was relatively high since it was a rainy day. Humidity difference was 17 percent difference in Paolo Park stations even they are very close. This could be seen in the interpolation map. Kriging generated a smoother map compared to IDW for humidity.

IV. CONCLUSION

This project provides tools that could be used for real time interpolation temperature, precipitation, solar radiation, and percent of humidity for Boulder, CO. The code also could be used for a larger areas like states or a counties given that there are no API call numbers limitations. Results imply that using several interpolation methods for one specific data set will facilitate understanding the interpolated data. IDW shows rapid temperature changes in a short distance which helps to find potential sources of error. On the other hand, Kriging created smoother maps for temperature which is closer to the real world. Therefore using several interpolation methods will helps to make more accurate maps. For example, observed precipitation value of Halcyon East station was considerably higher which could be seen in Figure 6. Although it is important to find reasons of this large difference, we can remove this type of measurement from the historical data.

This project used downloaded data without controlling the quality of them. The next step, would be to replicate the interpolation after removing faulty data points. The main goal of

this project was to create real time interpolation map. As shown by the results, creating real time maps necessarily will not lead to an accurate map. In this regard, it is important to add a section to the code to make quality controlling automatically and remove climate stations which over or under measuring data.

REFERENCES

- [Krige Danie and Kleingeld Wynand 2005]
Krige Danie and Kleingeld Wynand (2005)
The genesis of geostatistics in gold and diamond industries Space, Structure and Randomness Lecture Notes in Statistics Volume 183, 2005, pp 5-16.
- [George Y. Lu, David W. Wong 2008] George Y. Lu, David W. Wong 2008 An adaptive inverse-distance weighting spatial interpolation technique Computers and Geosciences archive. Volume 34 Issue 9, September, 2008. Pages 1044-1055
- [Wikipedia] <http://en.wikipedia.org/wiki/Interpolation>
- [USDA Data Gateway]
<https://gdg.sc.egov.usda.gov/GDGOrder.aspx>
- [ArcGIS Interpolation]
<http://help.arcgis.com/en/arcgisdesktop/>
- [Python Shape-file Library]
<https://pypi.python.org/pypi/pyshp>
- [Weather Underground]
<http://www.wunderground.com/>

Figures

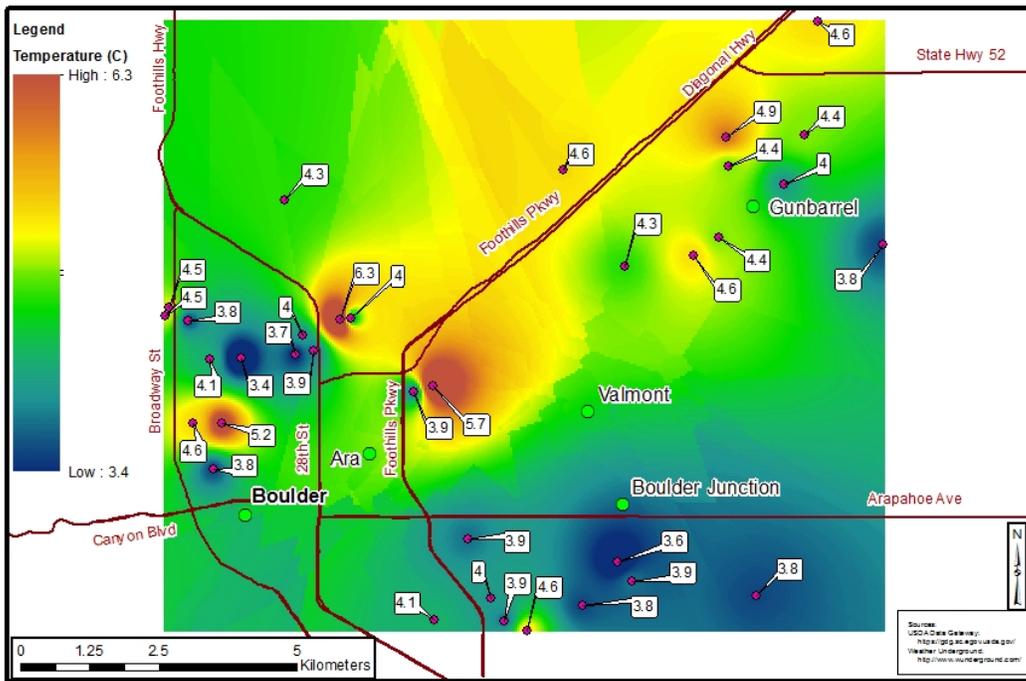


Figure 2: IDW interpolation of temperature

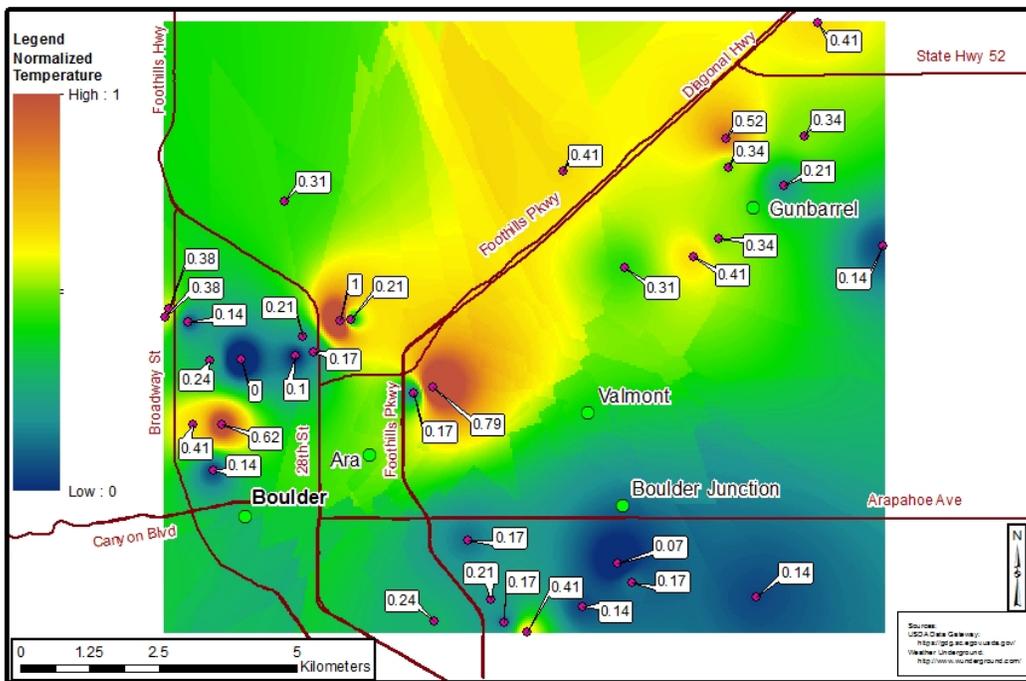


Figure 3: IDW interpolation of normalized temperature

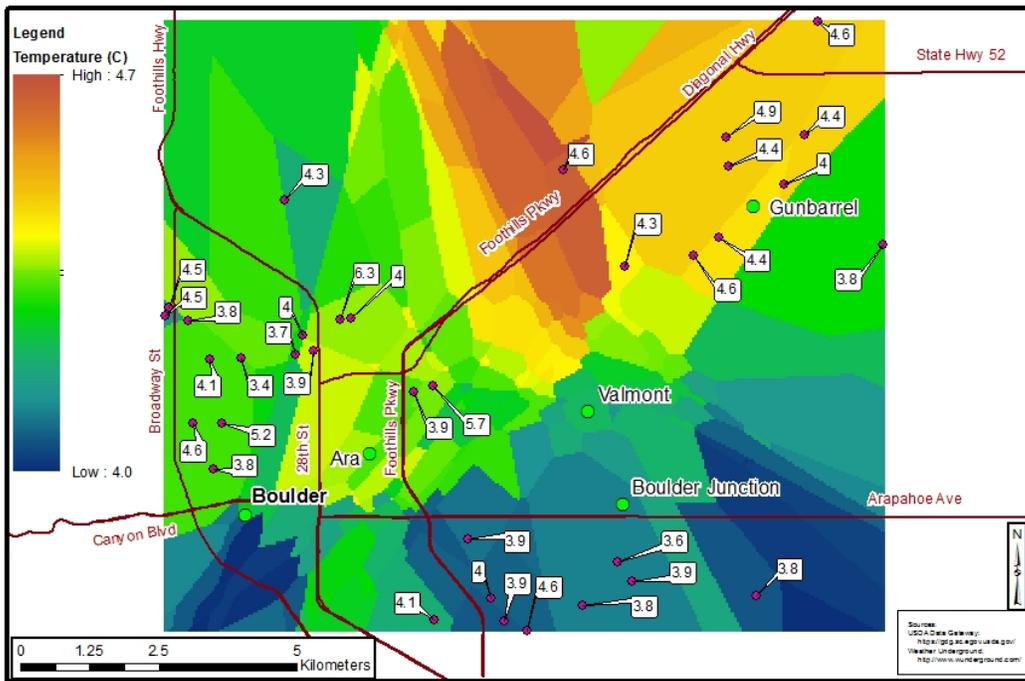


Figure 4: Kriging interpolation of temperature

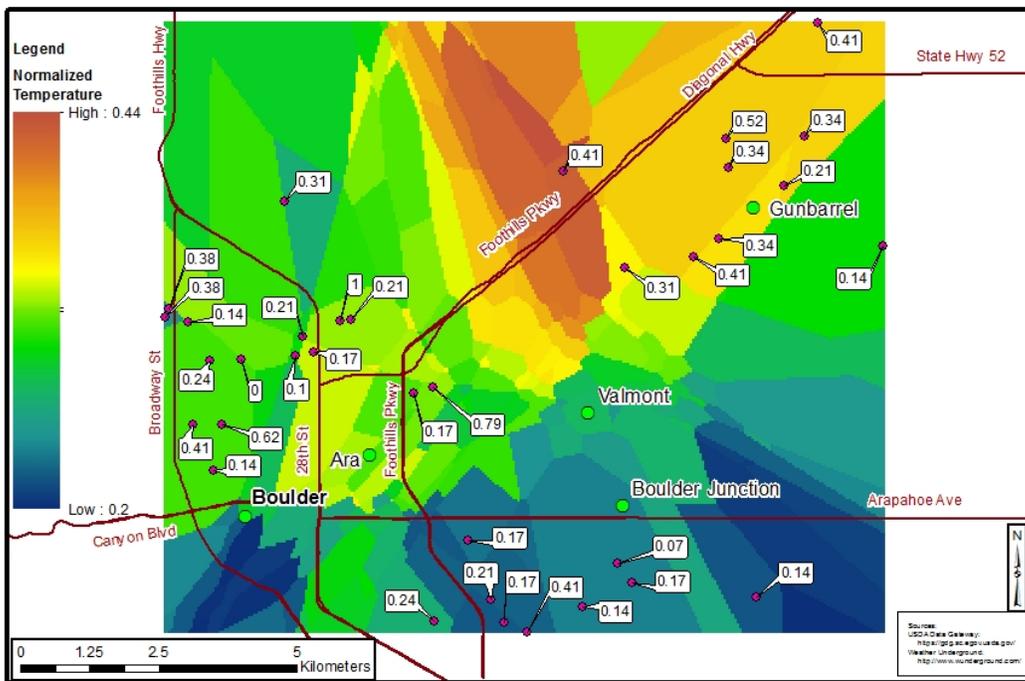


Figure 5: Kriging interpolation of normalized temperature

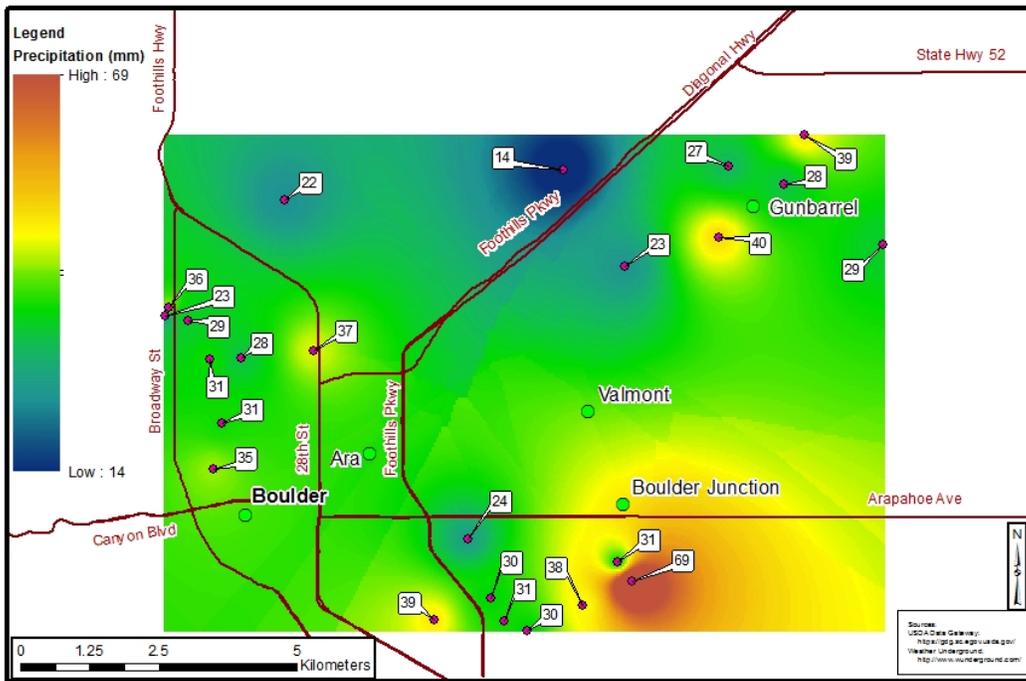


Figure 6: IDW interpolation of precipitation

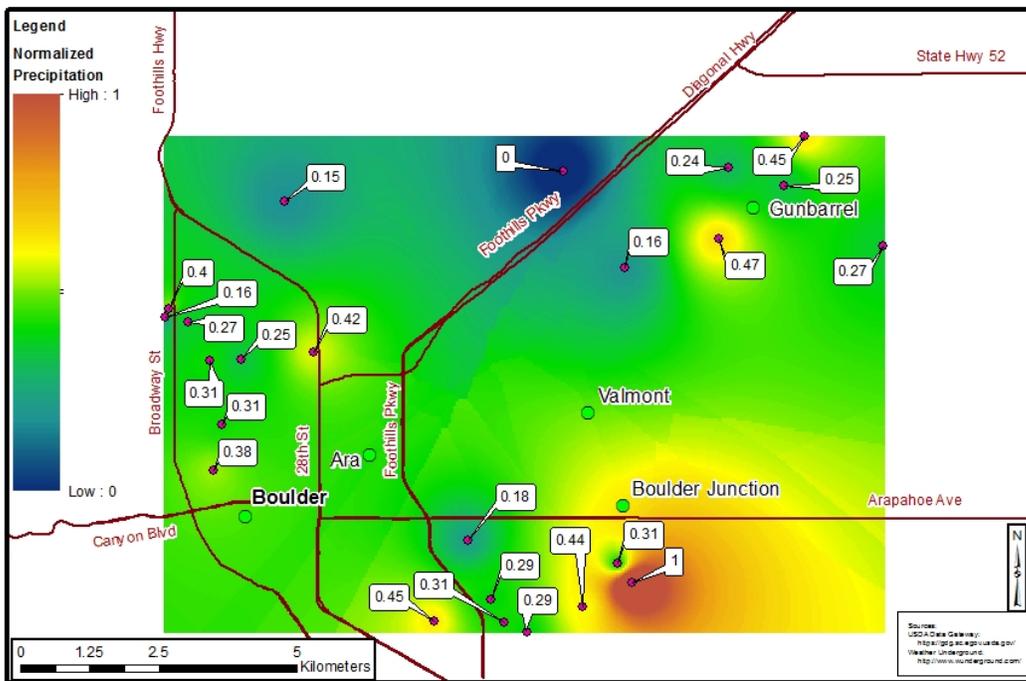


Figure 7: IDW interpolation of normalized precipitatin

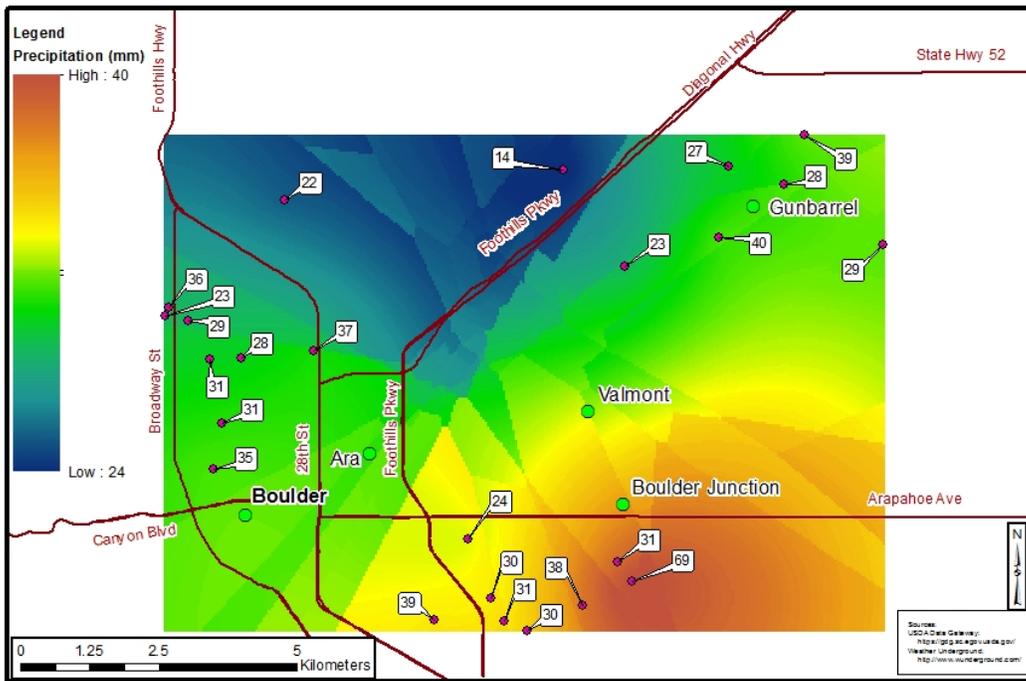


Figure 8: Kriging interpolation of precipitation

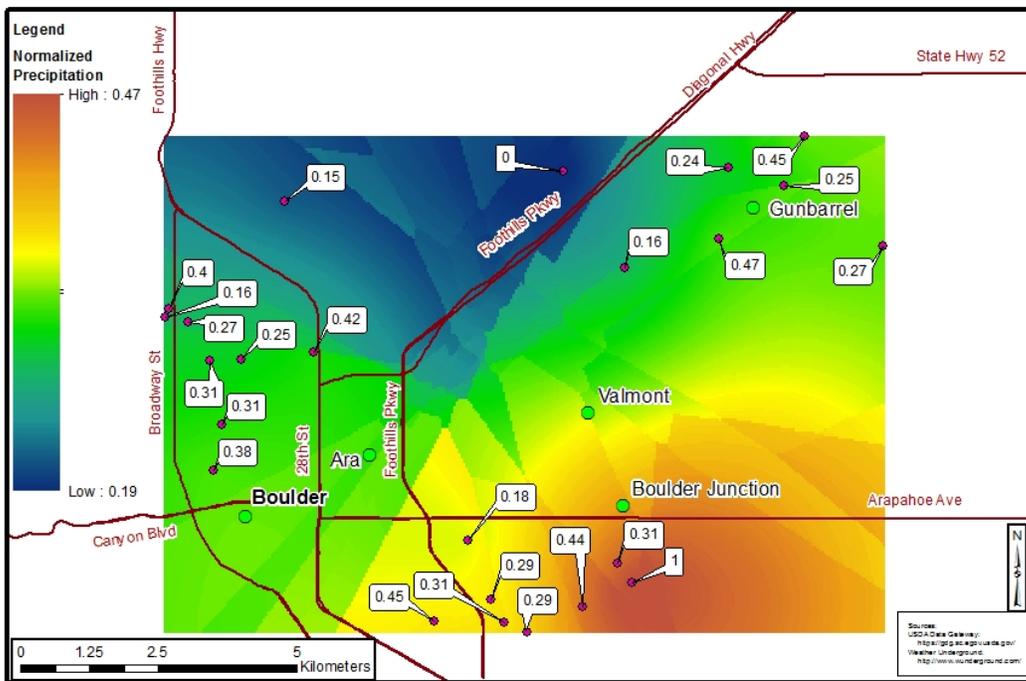


Figure 9: Kriging interpolation of normalized precipitation

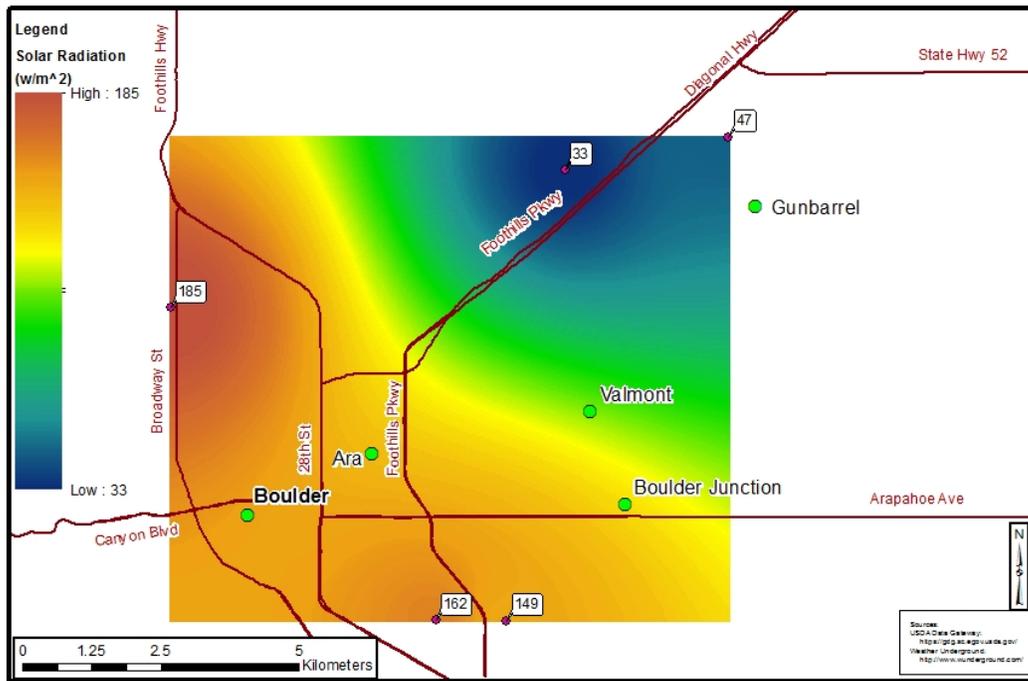


Figure 10: IDW interpolation of solar radiation

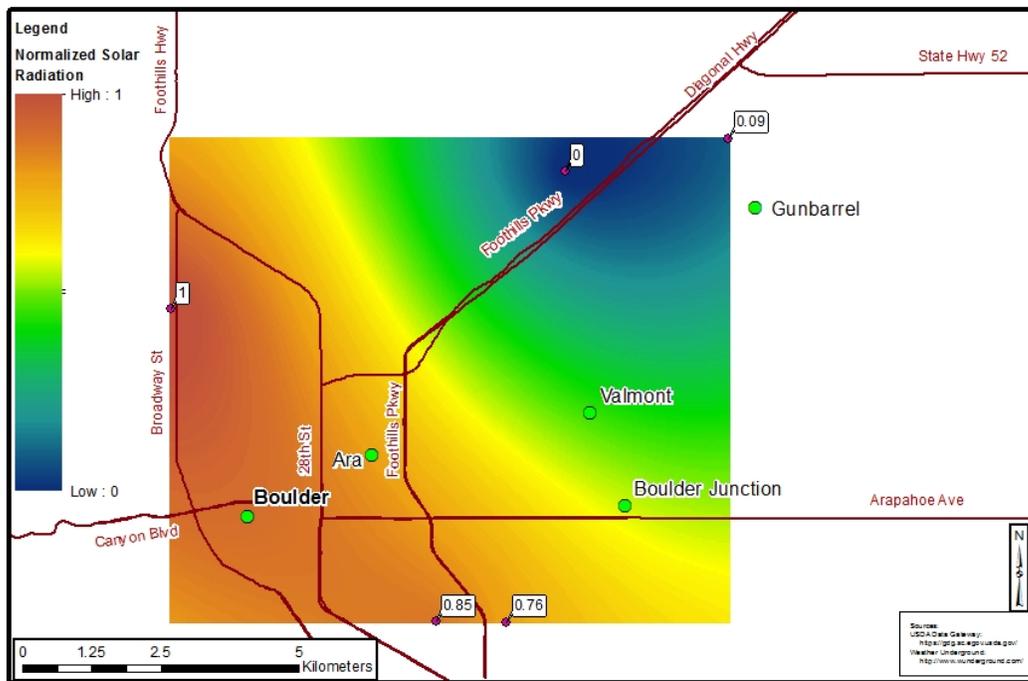


Figure 11: IDW interpolation of normalized solar radiation

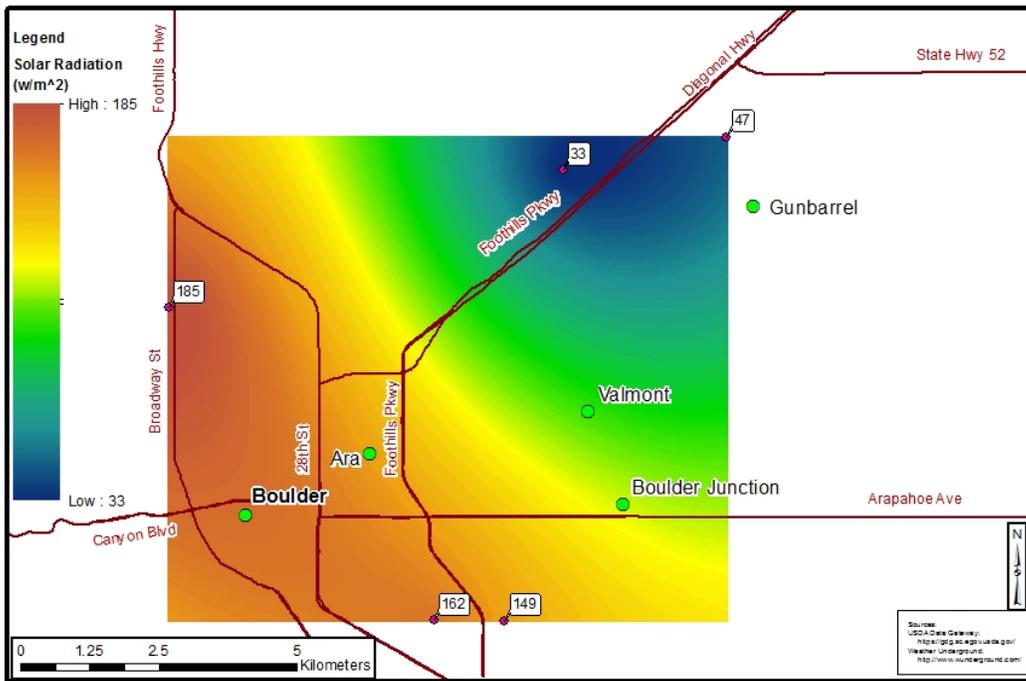


Figure 12: Kriging interpolation of solar radiation

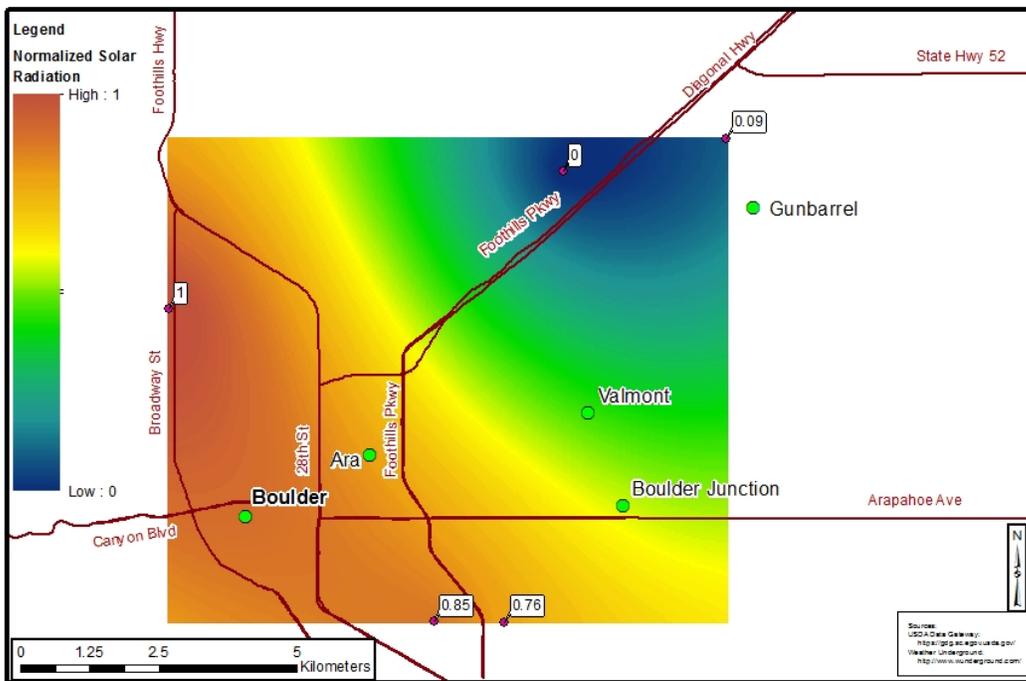


Figure 13: Kriging interpolation of normalized solar radiation

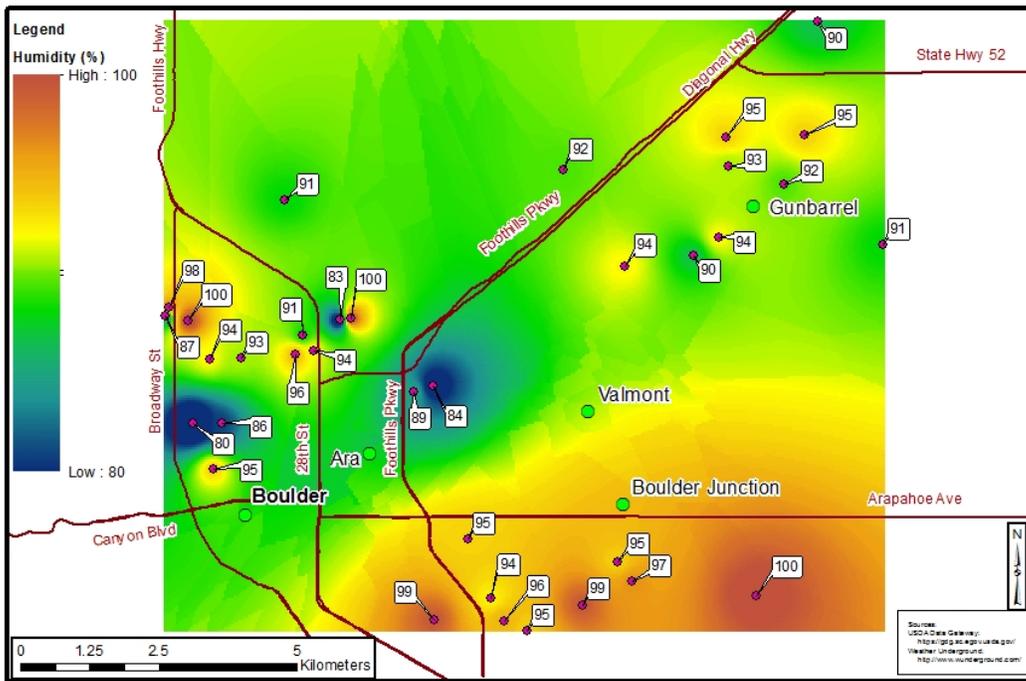


Figure 14: IDW interpolation of humidity

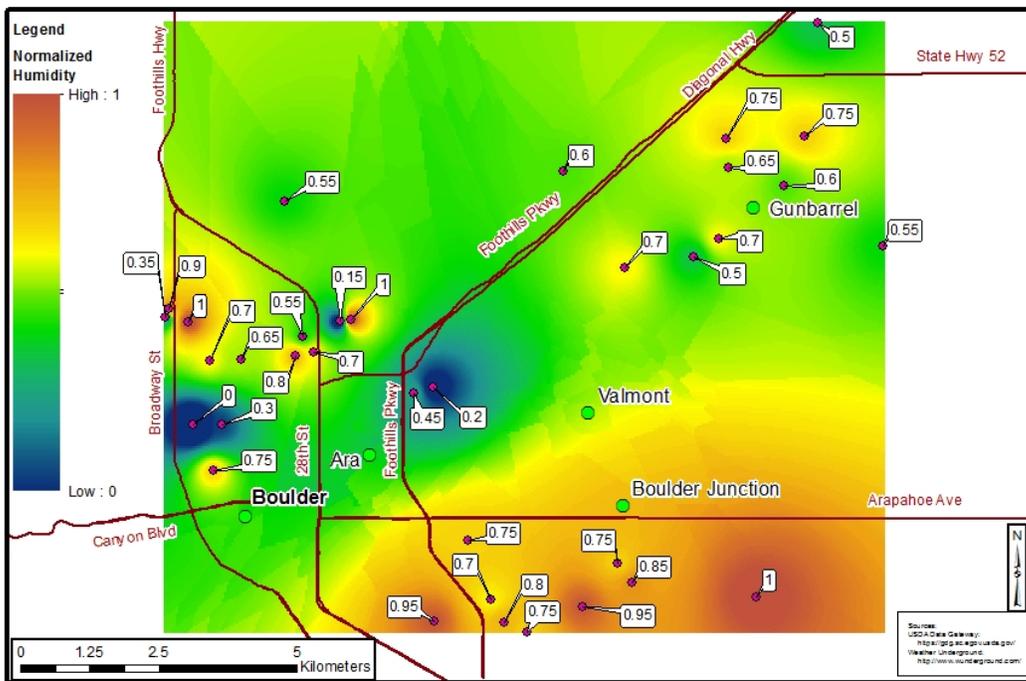


Figure 15: IDW interpolation of normalized humidity

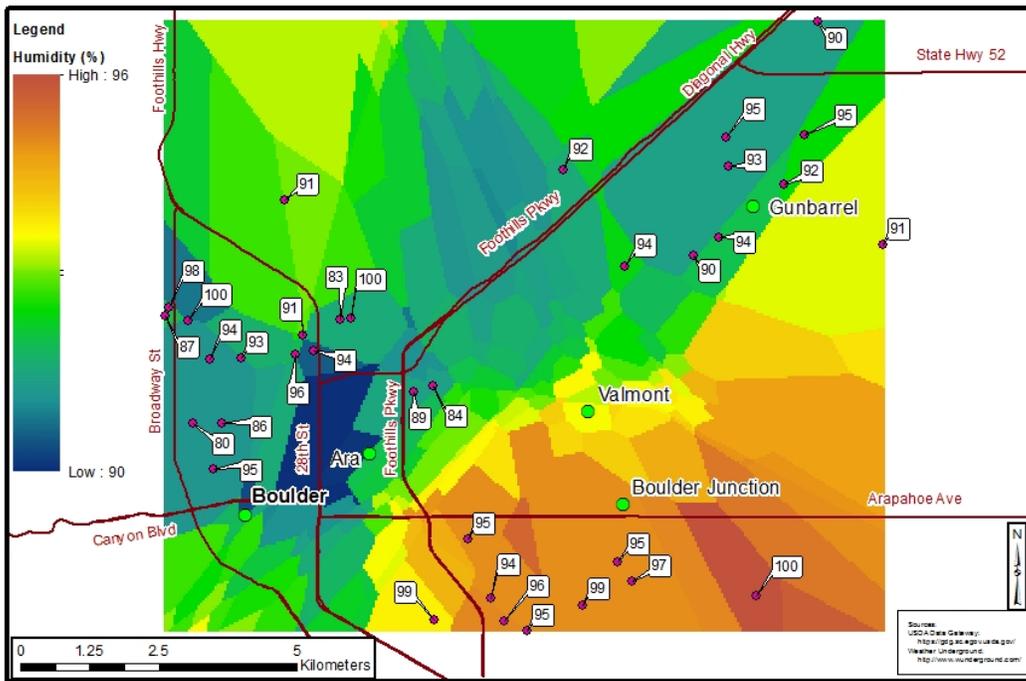


Figure 16: Kriging interpolation of humidity

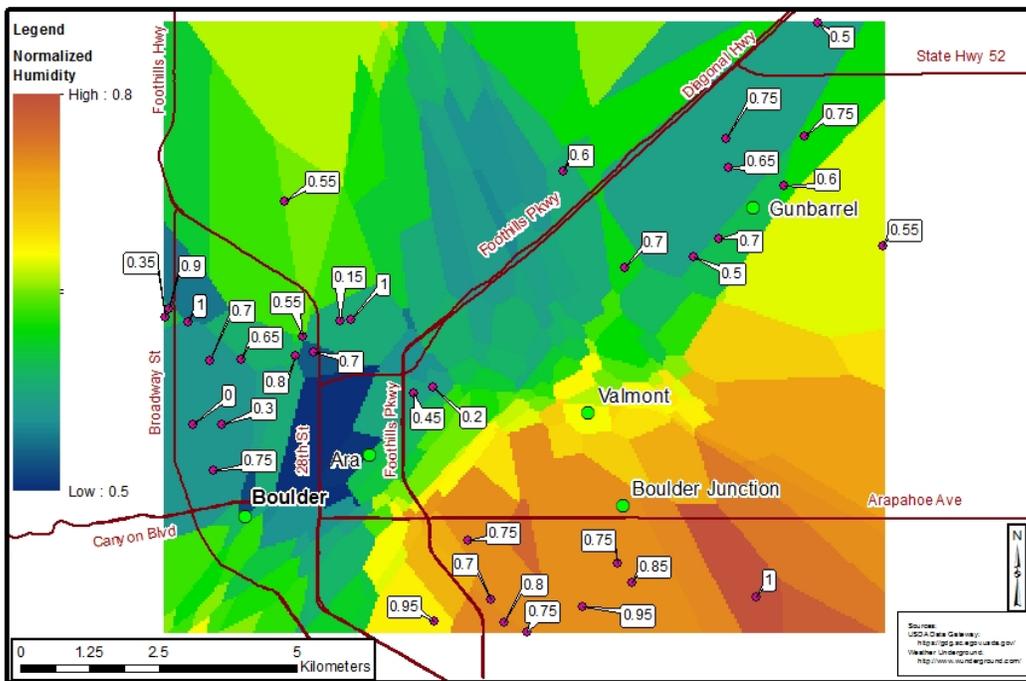


Figure 17: Kriging interpolation of normalized humidity

A geospatial analysis of climate change and its effect on the American Pika's habitat in the Great Basin

Hannah Keller

University of Colorado

hannah.keller@colorado.edu

Abstract

American Pikas are small mammals that live in high mountain ecosystems and due to global warming, pikas are rapidly losing their habitat. Pikas generally live above tree line, where they thrive in the colder climate, but with global temperatures rising, pikas are beginning to die out due to their sensitivity to heat. In Oregon and Nevada, pikas have already disappeared in over one-third of their previously known habitat¹. Using climate data collected from the North American Regional Climate Change Assessment Program and the USDA as well as conservation data from NatureServe, this paper examines the effects of global warming on the range of habitat available to pikas. Pikas are a particularly interesting species to examine in terms of the effects of climate change as there is little else endangering their habitat other than rising temperatures. This fact has caused controversy around the U.S. Fish and Wildlife Service's ruling that pikas are not threatened with extinction, because of the direct correlation to climate change². This project aimed to analyze climate data focused on known pika habitats within the Great Basin watershed and examine possible thresholds between known extinction areas and current habitats.

I. INTRODUCTION

American Pikas are currently running out of options. Pikas, small mammals and cousins to the rabbit, live in cold alpine climates. Over the past 12,000 years or so, pikas have been retreating upslope due to rising global temperatures¹. A characteristic that sets them apart is unlike many mammals that live in cold climates, pikas do not hibernate. As a result they have very thick fur and high metabolism rate in order to live during the cold season, therefore Pikas are particularly sensitive to milder temperatures. What may not seem a large global temperature increase to some, can cause huge habitat changes for pikas. For instance, a mere two degrees Celsius can force pikas up an extra 1300 feet in order to find cooler temperatures to live². With global temperature continuing to rise at increasing speed, pikas are quite literally running out of places to go.

Just in the past decade extinction rates have increased nearly five-fold and with projections showing potential warming of 4.5 to 14.4 degrees Fahrenheit during the next 100 years, Pikas are in the midst of a crisis¹. In order to examine the relationship between the increasing extinction of pikas and climate change, the Great Basin, an arid area that includes parts of Nevada, Utah, Idaho, Wyoming, Oregon, and California, will be the focus of this project. The Great Basin is home to the majority of the pika population and has been the focal point for many previous studies of pikas. Due to the high altitudes pikas inhabit, they are a good proxy for studying the impact global warming as it is the most pervasive threat affecting the American Pika⁴. Using current predictions of pika extinction of known habitats within the Great Basin and climate model data, this study aimed to find a habitat temperature threshold for pikas.



Figure 1: *American Pika*

II. DATA

I. Pika Habitats in the Great Basin

To begin we used a habitat distribution shapefile of the American Pika obtained from NatureServe, a non-profit organization that provides high-quality scientific expertise for conservation³. In order to narrow down the scope of the study, we decided to focus on habitats within the Great Basin, as it holds the majority of pika habitat. From USGS, United States Geological Survey we obtained a shapefile with the outline of the Great Basin which we planned to use to clip the NatureServe shapefile. Within QGIS, the shapefiles were converted to a common projection of WGS84, the World Geodetic System reference system, as it is a fairly standard projection. Unfortunately the scales of the two shapefiles did not match so using the Affine Transformation plugin for QGIS, we were able to overlap the two vectors to create an intersection.

Unfortunately due to the climate data being

used for the project, the vector shape was not suitable for subsetting the data, as was the original plan. What we ended up doing was using an overlaid image of predicted pika extinctions from Dr. Scott Loarie, a modeler at Carnegie Institution for Science, and the NatureServe shapefile, as shown in figure 2, as a classifier to create two sub areas, one for predicted extinction and one for predicted survival in the next century². Again, because of the temperamental formats and projections of the climate data we decided to use bounding polygons of the two areas as our final outlines for habitat study.

II. Subset NARCCAP Data

Although the climate data obtained and used for this project is a bit unpredictable in its formatting, making it hard to subset via a shapefile, we went with using NARCCAP data, North American Regional Climate Change Assessment Program⁵, because of the availability of future modeled climate data. As NARC-

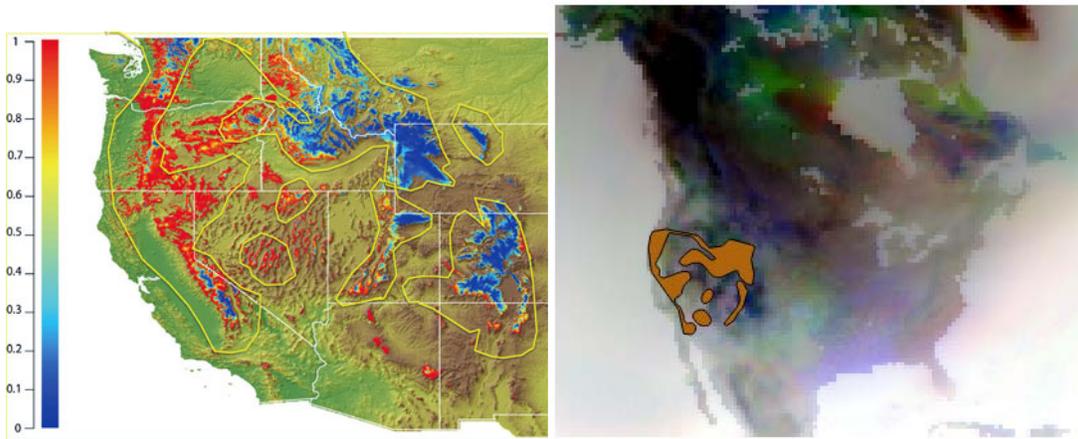


Figure 2: *Pika Habitat Predictions²*

Figure 3: *Ideal Subset Method of NARCCAP data*

CAP is a program, which incorporates different global and regional climate models from around the world, we decided to focus on using data with the same map projections. Therefore, our final selection of model data was using CRCM, Canadian Regional Climate Model, as the regional model, using either CCSM, Community Climate System Model, or CGCM3, Third Generation Coupled Global Climate Model, as the model driver. The CRCM model uses a polar stereographic map projection, as defined by the National Snow and Ice Data Center. Both model combinations of regional model and driver had data associated with model runs for the past (1968 to 2001), using observed data, and model run for the future (2038 to 2070).

Given the vast number of variables that incorporate climate data, in order to narrow the scope for this project, only temperature data was analyzed. Specifically maximum temperature data was collected from NARCCAP and used for analyzing a threshold as, although snow and snowpack can affect pikas, they are most sensitive to warm temperatures⁴. NARCCAP data is stored as NetCDF files, Network Common Data Form, which is a very common data format for climate data, but are often very large in size. The best method for querying and downloading NetCDF data is using

OPeNDAP, a web protocol specifically design to simplify downloading NetCDF data. Using NCL, NCAR Command Language, the bounding shapefiles of the predicted surviving and exterminated pika habitats were used to subset, via an OPeNDAP call, maximum daily temperatures from the past and future models as laid out previously. NCL is a very powerful language for working with NetCDF files and such we were able to use a method that found the nearest coordinates in the data's native polar stereographic projection from the WGS84 coordinates of the shapefiles. Lastly, we converted the subset data from degrees Kelvin to degrees Celsius as we found it an easier unit to work with.

III. METHOD

With the initial subset of the data done, we used two methods of aggregation to aggregate the daily data into average monthly data over the entire model simulations and average yearly data. Both were performed using NCL. Both aggregations started by reading the time variable from the NetCDF file and translating it to a more human readable format via the NCL function `utccalendar`. NARCCAP data is stored with the time format of number of 'days since (1968/2038)-01-01,' so in order to

logically be able to perform calculations, it is best to transform the time into a more logical format. `Utccalendar` returns the native NARCCAP format (ex: 6039) to an array of years, months and days⁵. Using this new array and the interval of the aggregation, a new array is created to hold the aggregate data.

When aggregating the data by year, which was the easier of the calculations, after the data format had been updated, a new array was created to the size of the interval as returned from the `utccalendar` function. Variables are created to capture missing data and create new cell bound data for the output NetCDF file. We then loop through the data, only reading the data from the input file as needed, until the output data's dimensions match those expected. Each loop calls the NCL function called `dimavgn`, which averages a variable's given dimension at all other dimensions, in our case time, latitude, and longitude. This function is what returns the actual aggregated data over one time interval, in this case one year.

In order to average cyclically over years of a given subset file, the process was a bit more involved. After the updated date format is returned, a new array is made the size of number of years within the file. The aggregated data is stored in a new variable and similar to the yearly aggregation we loop over, using the NCL function `dimavgn`, but wait to add the data to the output array until an extra set of iterations have been performed. For both aggregations, the output time metadata is updated and rest of the metadata is copied from the original files to be used in further calculations, such as for plotting.

With the data aggregated and converted to degrees Celsius, it was decided to output the data in csv format in order to take advantage of other tools for further processing and visualizing. We did create plotting scripts in NCL, but as it only plots a given time interval we decided not to integrate that within the scope of this project. To output the data to csv, an NCL script was written and each variant of our data, extinction, survival, current, future, yearly aggregate, and monthly aggregate

were output to csv. Using Excel, it was simple to add further monthly averages, as the NCL script averaged over every data point within the boundary shapefile. Those averages were then used to create plot graphs to visualize the differences between the predicted extinction areas and survival areas, but also the predicted future climates.

IV. RESULTS

Overall, both the future and current simulation data that was compiled showed differences between the areas Dr. Loarie's model predicted for extinction in the next century and survival in terms of maximum temperatures. What is interesting about our findings is that it seems the areas predicted for survival tend to have higher temperatures in the summer months than the areas of predicted extinction, but lower temperatures in the winter. This does not quite follow what we had expected, which was lower temperatures all around, but both the current and future models show higher summer values for the expected survival areas. It could be speculated that because of the colder temperatures during the winter, the ground does not heat as fast in the summer, as pikas are known to occupy rock crevices and areas shaded from the sun. It also could be that the temperatures at night in the summer in the areas predicted to survive are cooler as they have been known to exhibit nocturnal behavior to avoid particularly warm temperatures⁴. What is seen from the results is that the areas predicted for extinction have, although slightly less extreme peak, a longer period of hotter temperatures, whereas the areas predicted for survival have a shorter period.

As for a threshold limit, it was hard to determine from the results found as it was expected higher temperatures and larger bell curves would be factors to the predicted extinction areas. Pikas have been known to die in temperatures from as low as 25.5°C and from the current labeled model data, temperatures have consistently reached well over that in all areas pikas are known to live in the Great

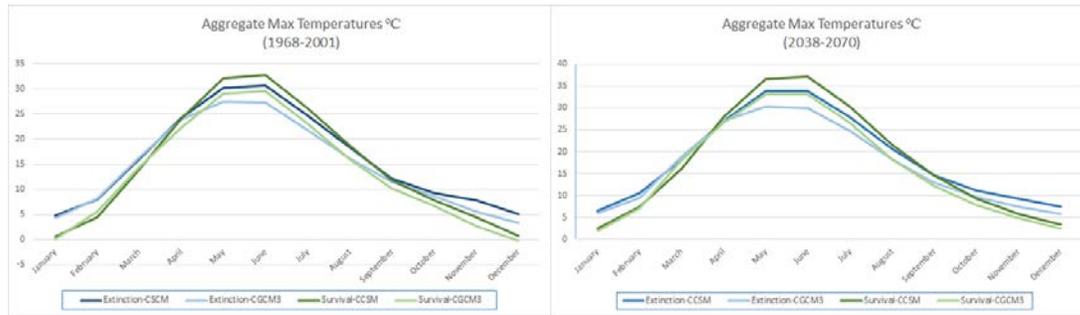


Figure 4: Current Temperature Monthly Averages

Figure 5: Future Temperature Monthly Averages

Basin. Therefore, a threshold limit from our study was inconclusive, more data and variables should probably be added to the study in the future in order to better determine a possible threshold limit as it is a fact that the American Pika is a species that is dying off and extinction has already been seen in multiple areas. What could be said from the results of this project is that given known areas of extinction and modeled predictions with far more variables than this study, that pikas thrive better in areas that have a less prolonged period of heat in the summer, even if the peak temperatures may be higher.

V. LIMITATIONS

A large limitation to this project was the subsetting of the NARCCAP data. We ran into a lot of roadblocks and spent more time than necessary trying to convert the shapefile projections to match that of the climate data. There is also a script for subsetting NetCDF files via shapefiles in NCL, which we ended up using the basic functionality of, but given a non-rectangular vector, we could not get the NARCCAP data subsetting properly. Therefore our results do span more area than laid out for pika habitat by the NatureServer vector and are much more stereotyped by Dr. Loarie's predictions than originally hoped. It also would have been great to span the area being studied beyond that of the Great Basin, since pikas reside as far north as into British Columbia, Canada and NARC-

CAP spans all of North America. To further the analysis, adding more variables than simply temperature data would have definitely added more insight to the prediction of future pika habitat. From all that was researched about the American Pika, sensitivity to heat is definitely the species' largest vulnerability, as they do not hibernate in the winter so have a high metabolism, making it very difficult to keep themselves cool during the summer, which is why temperature was the focus of the study, but considering they do burrow themselves within the snow during the winter adding precipitation and snow pack as variables could have added some very interesting insight.

At the beginning of the study, we obtained some crowd sourced data from the Front Range Pika Project, a citizen science effort to gather baseline data on the current distribution of pikas and pika habitat. It was our hope to incorporate recent citizen observation data with our findings, unfortunately the crowd sourced data was very limited in scope and not part of the Great Basin area.

VI. CONCLUSION

The American Pika is a species, whose options for habitat is fairly narrow and is getting very smaller due to rising global temperatures. Considering pikas have a very high metabolism to keep them going during winter as they do not hibernate, such as many mammals that live in alpine terrain, they are very sensitive to

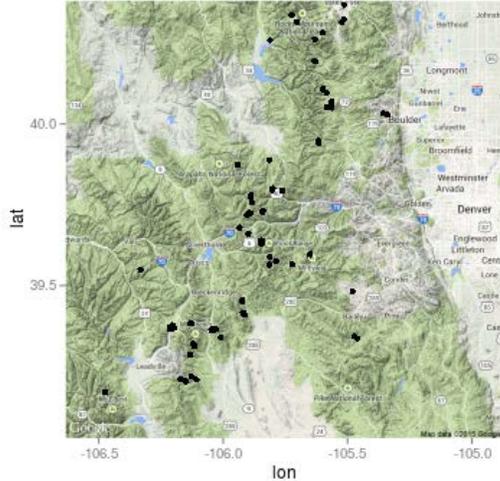


Figure 6: Front Range Pika Project Observations

higher temperatures. With the growing global temperature, the areas of extinction of pikas is also growing. Pikas are an interesting species to focus on climate change as it is seemingly the species' greatest threat⁴ and their natural habitats are fairly contained to one area in the world, the Great Basin. Using climate model data from simulations using observed data and simulations of future climate, we were

able to examine areas of predicted extinction and survival for pikas to try and determine temperature thresholds for pika habitat. This study tapped into the potential of integrating climate model data along with conservation data to help predict and understand how climate change is affecting different species, such as the American Pika.

REFERENCES

- [1] "American Pika". *National Wildlife Federation*. National Wildlife Federation is a voice for wildlife, dedicated to protecting wildlife and habitat and inspiring the future generation of conservationists. www.nwf.org.
- [2] Pimm, Stuart L. "High-living Pika Can Help Us Understand Our Climate Fate." *Blog National Geographic*. NatGeo News Watch, 05 Feb. 2010. Web. 19 Mar. 2015.
- [3] Patterson, B. D., G. Ceballos, W. Sechrest, M. F. Tognelli, T. Brooks, L. Luna, P. Ortega, I. Salazar, and B. E. Young. 2007. Digital Distribution Maps of the Mammals of the Western Hemisphere, version 3.0. NatureServe, Arlington, Virginia, USA.
- [4] Beever, E. and Smith, A.T. 2011. *Ochotona princeps*. The IUCN Red List of Threatened Species. Version 2014.3. www.iucnredlist.org. Downloaded on 05 May 2015
- [5] Mearns, L.O., et al., 2007, updated 2012. *The North American Regional Climate Change Assessment Program dataset*, National Center for Atmospheric Research Earth System Grid data portal, Boulder, CO. Data downloaded 2015-05-05. www.narccap.ucar.edu/index.html.

Identifying landslide prone areas of Colorado

HUI SOON KIM*

University of Colorado Boulder

huki2996@coloraod.edu

Abstract

Landslides are devastating phenomena that causes huge damage around the world. It results in estimated tens of deaths and \$1 - 2 billion in economic losses per year in the US alone [0]. Landslides can be triggered by a number of factors including rainfall, earthquakes, volcanic activity, slope disturbance, human land use, geology or change of slope [1]. Landslide susceptibility can be mapped using a number of different methods depending on the data available. The focus of this project is to analyze the landslide prone areas of Colorado by combining some factors from DEM (digital elevation model), rainfall, geology data. Landslide prone areas can be identified using prior knowledge from literature survey by assigning the proper weight for each factors. For each of the factors that influence landslide susceptibility, several classes will be defined to describe the variance of that factor. Each class will be given a rank with a highest value given to the class with the highest susceptibility to landslides. The factors are then each assigned a weight in order to account for the varying influence. Finally, two methods are applied to those data to estimate landslide risk: Heuristic analysis and Classification. The combination of these factors are calculated as a landslide risk . Moreover, these preprocessed factors are fitted into maximum likelihood classifier to do partially supervised classification. Each result is verified partially by comparing them with historical landslide data. The results of this study would suggest that Qualitative susceptibility zoning method(Heuristic Analysis) is an effective tool in examining landslide susceptibility problems when we do not have prior knowledge about the landslide history for some regions. The rough assessments for landslide risk can be made without landslide history data. Furthermore, Classification method may be used to estimate the landslide risk under the conditions that the accurate and detailed training data (landslide history data) can be acquired.

I. INTRODUCTION

This paper analyzes the landslide prone area of Colorado. Some may think that Colorado is not susceptible to landslides because most of the area consists of massive rocks. However, landslides can be caused by various factors other than geology so Colorado may be not safe area for landslide. Landslide-related casualty reports for Colorado indicates that nine people were killed and three were injured in 2010 and Total direct costs of landslides in Colorado for the year 2010 were approximately \$9,149,335 [2].

Besides the effects landslides can have on the life, landslides can have great effect on natural environments. Identifying areas prone to landslides has great potential value and could allow better management of its territory.

Even though landslides are local and complex phenomena that are controlled by large variety of internal and external factors, this paper assumes that the attributes contributing to the landslide activity in Colorado can be steep slopes, geology and rainfall. In that sense, the west part of Colorado can be prone to landslide compared to the east part because of high slope.

*MS student in Computer Science

In this study, two kinds of experiments were carried out in order to identify locations that are at high risk for a landslide. The first one is qualitative susceptibility zoning (Heuristic analysis) which combines each values of factors that affects the landslides and gives some qualitative outcome of landslide risk for area. The second one is classification method to predict the landslide area. A maximum likelihood classifier were partially trained by area data with landslide history and without landslide history, then the learned classifier predicts the landslide risk for new area. In order to do carry out these experiments, some GIS related software such as QGIS, GRASS and PYTHON are used.

II. METHODS

i. Study Area Location and Description

Colorado is notable for its diverse geography, ranging from alpine mountains, arid plains and deserts with huge sand dunes, deep canyons, sandstone and granite rock formations, rivers, lakes, and lush forests. The borders of Colorado were originally defined to be lines of latitude and longitude, making its shape a latitude-longitude quadrangle which stretches from 37N to 41N latitude and from 102:03W to 109:03W longitude.

East, northeast and southeast Colorado are mostly the high plains, while Northern Colorado is a mix of high plains, foothills, and mountains. Southwest and southern Colorado are a complex mixture of desert and mountain areas. Northwest and west Colorado are predominantly mountainous, with some desert lands mixed in. Landslides in Colorado are concentrated along the Front Range, central mountains, and western part of the State and are typically associated with areas of significant slope[3].

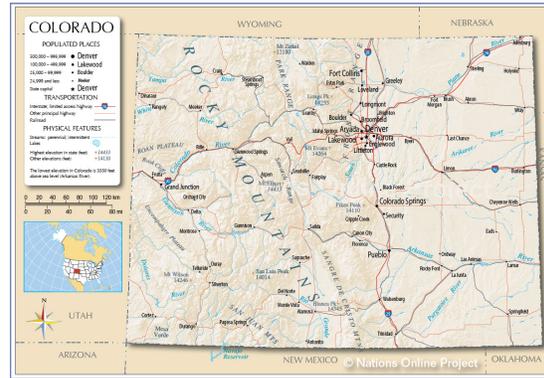


Figure 1: Study of Area Colorado State

ii. Collecting DataSets

The dataset needed for this project was collected by downloading imagery data and vector shape files from www.coloradoview.org and United States Geological Survey database. The entire 10-meter Colorado State Continuous DEM files(10.6GB) which are derived by aerial remote sensing imagery and Geology shape files are retrieved from coloradoview.org.

The average annual precipitation vector data and landslide incidence-susceptibility in the conterminous United State vector shape files were obtained from USGS website. Both websites provide these data directly without any restriction.

Each of the downloaded data used different spatial reference system. So NAD27, DATUM["North American Datum 1927"] spatial reference system were used for consistent analysis.

iii. Data Preprocessing

1) Slope:

The slope data for Colorado was generated from a 10-meter DEM raster data which were acquired from aerial remote sensing imagery. The DEM raster file was projected into NAD27 spatial reference system and loaded into QGIS. The raster data was then clipped to the outline of Colorado using the Colorado boundary

(polygon) shape file and the slope was generated using raster analysis tool of QGIS.

The data was then reclassified on a risk scale or values from 0 - 4, this was done in using the raster calculation tool of QGIS. Mass-movement processes such as landslides are more likely to occur on steeper slopes than on shallower gradients therefore highest slope values were assigned a risk of 4 and the lowest 0 [4] . As we can see in the below figure 2 and 3, the slope of west Colorado is much higher compared to the east side.

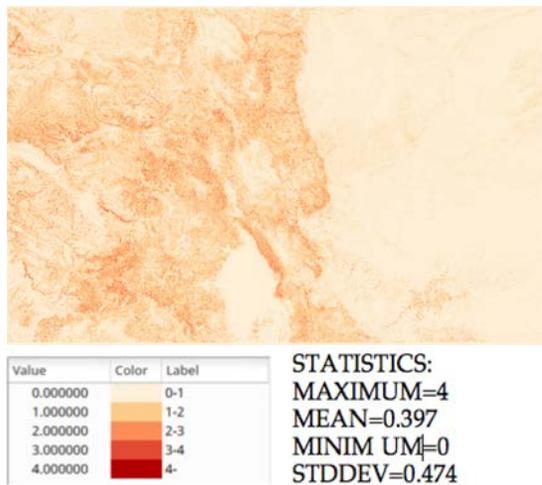


Figure 2: Rescaled 2D slope Raster data

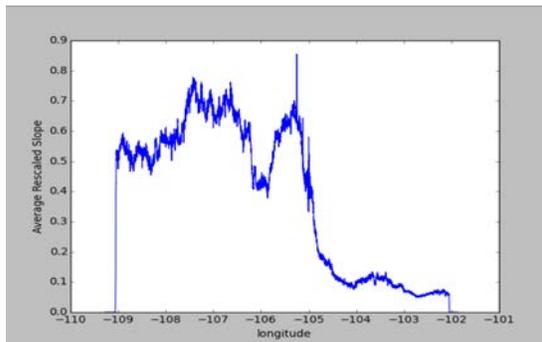


Figure 3: Rescaled average slope along longitude

2) RainFall:

The precipitation vector shape file was projected into EPSG:4267 spatial reference system

and uploaded into QGIS. The vector data was then clipped to the outline of Colorado using the Colorado boundary (polygon) shape file. The average annual precipitation of Colorado ranges from 10 to 70 inches. Since landslides are more likely to occur on high rainfall area than on low rainfall area, highest rainfall area were assigned a risk of 4 and the lowest 0 [5].

This rescaling work was done by adding new dimension in .dbf file of the vector data using Python. And then, the vector data was converted into raster using raster operation of QGIS. As we can see in the below Figure 4, the rainfall of west Colorado is greater than that of east side.



Figure 4: Average annual Rainfall Vector

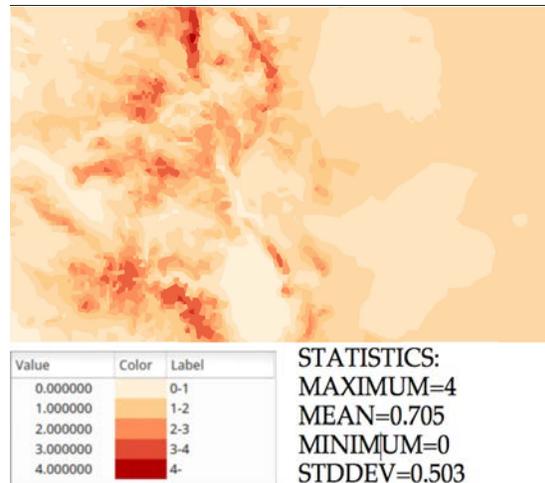


Figure 5: Average annual Rainfall Raster

3) Geology:

The same process which has been done for the rainfall data was conducted for Colorado geology vector shape file. This data has the attribute of rock type with 29 classes. According to the variability to affect the landslide, the new values between 0 - 4 was assigned in the .dbf file like below table [6] . And then the vector data converted into raster. Unlike the above two cases, the geology characteristics to influence landslide is not clearly distinguished between west and east Colorado.

Assigned value	Rock type
0	No data/ water
1	Basement
2	Sandstones
3	Mudstones
4	Landslide

Figure 6: 2D slope Raster data

In this table each categories includes following rock types. (0)Water: no data and water (1) Basalt: felsic gneiss, biotite gneiss, granite intermediate volcanic rock, mixed-clastic/volcanic, plutonic-rock andesite, ash-flow tuff, granitoid and rhyolite (2) Sandstones: alluvium, dune sand, glacial drift, gravel, quartz latite and quartzite (3) Mudstones:claystone, shale, siltstone, limestone and clastic (4) Landslide: landslide



Figure 7: Geology Vector

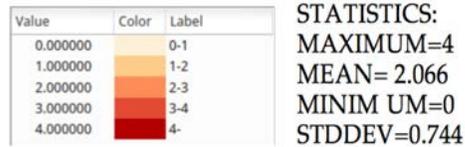
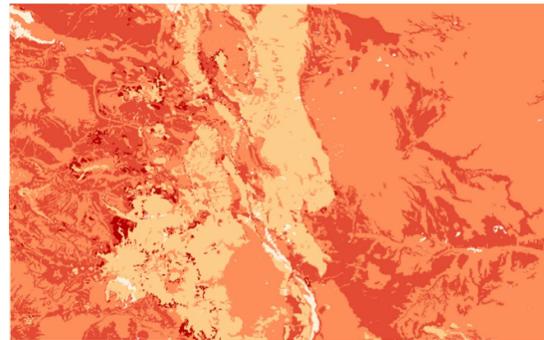


Figure 8: Geology Raster

iv. Experiment -1: Heuristic analysis

- Qualitative susceptibility zoning without landslide data

Qualitative susceptibility zoning method without landslide data is one of the heuristic analysis which uses expert opinion when carrying out the zoning to assess the hazard. These methods combine the mapping of the landslides and their geomorphologic setting as the main input factors for assessing the hazard. This zoning method uses prior knowledge to assign weighting values to a series of input parameters. These are summed according to these weights, leading to landslide susceptibility and hazard classes. These methods are common, but it is difficult to determine the weighting of the input parameters [7].

In this experiment three factors were applied in order to assign the risk of landslide events in various parts of the Colorado: slope, rainfall and geology. It is obvious that the factors to have great effect on landslide is different according to the specific geographic characteristics for some area so different weight to specific factor has to be applied. ie, It is important to note different areas will have different predominant factors.

But this experiment deals with the entire

Colorado state whose area is enormously large. Hence, inducing a proper weight to specific factor for the entire area may be impossible. Therefore, the calculation of the landslide risk was done using simple equal-weighted sum for the factors like following equation.

$$\text{Landslide Risk} = 1 \times \text{Slope value} + 1 \times \text{Geology value} + 1 \times \text{Rainfall value}$$

Each parameters converted into raster files which are all scaled between 0 and 4 according to varying influence to the landslide are summed with equal weight using QGIS raster calculation tool. The final result was generated as a form of raster imagery file. The result is shown in Figure 9 and 10.

v. Experiment -2: Classification

- Maximum Likelihood Classifier

Partially supervised Maximum Likelihood (ML) classification algorithm is basically a multi-class classifier which uses the Bayes Theorem [8]. Given the classes W_i where $i = 1 \dots n$, we define the probability that a pixel belongs to each class W_i as $P(W_i | X)$ where X is the vector of the each factor (band, attribute) values.

We assign the pixel to the class W_i if $P(W_i | X) > P(W_j | X)$ for all $j \neq i$. So we have to estimate $P(W_i | X)$ for all $i = 1 \dots n$. From the training areas, we can estimate the probability that a pixel of a class W_i has values X , .ie., $P(X | W_i)$. Bayes Theorem shows that

$$P(W_i | X) = P(X | W_i)P(W_i) / (P(X))$$

where $P(W_i)$ is the a-priori probability of the class (% pixel in the class). Since $P(X)$ is same for all classes, a pixel belongs to a class W_i when $P(X | W_i)P(W_i) > P(X | W_j)P(W_j)$ for all $i \neq j$.

We assume that measured samples, such as vector X , are Gaussian distributed as started by the central limit theorem[7](Marc Bartels and Hong Wei). So the the probability distribution of the given class W_i ,

$$P(X | W_i) = (2\pi | C_i|)^{-n/2} \times \exp([-1/2(X - m_i)' C_i^{-1}(X - m_i)])$$

where C_i is the covariance matrix and m_i is the mean of the class W_i . Taking log for the equation $P(X | W_i)P(W_i)$, then

$$\begin{aligned} G_i(X) &= \ln(P(X | W_i)P(W_i)) \\ &= \ln(P(W_i)) + \ln((2\pi | C_i|)^{-n/2} \\ &\times \exp([-1/2(X - m_i)' C_i^{-1}(X - m_i)])) \\ &= \ln(P(W_i)) - n/2 \ln(2\pi | C_i|) \\ &\quad - 1/2(X - m_i)' C_i^{-1}(X - m_i). \end{aligned}$$

Hence, the pixel belongs to the class W_i when $G_i(X) > G_j(X)$ for all $j \neq i$. This model classifies all the pixels, even those with a probability very low.

Even though ML- classifier supports the multi-class classification, this experiment implements binary classification because the available training data (USGS landslide history shape file) only consists of binary label whether there were landslide history for specific area or not. So, Binary ML (maximum likelihood) classifier was modeled using GRASS Tool. Landslide incidence-susceptibility in the conterminous United State vector shape file were used to train the classifier. Two distinctive classes are defined: area with landslide history and area without landslide history. For each parameters (slope, rainfall, geology), two partial area with a history of landslide and with no history of landslide were fed into ML classifier in order to learn the optimal parameters and then the landslide risk of entire area of Colorado were classified. The result is shown in Figure 14.

III. RESULTS

i. Result of Experiment -1: Heuristic analysis

- Qualitative susceptibility zoning without landslide data

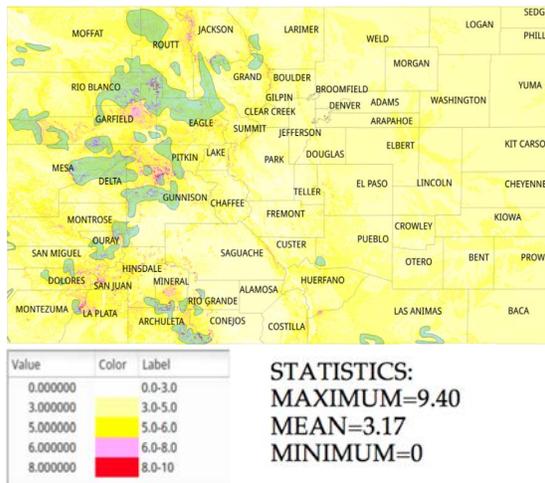


Figure 9: Combined Imagery : Landslide Risk of entire Colorado

involved in landslides (high landslide incidence area). Comparing them to the highest risked area (red area), it seems that the calculated high risked areas includes most of the polygon and neighborhood area.

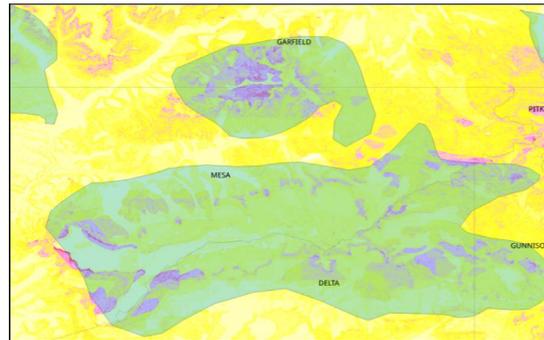


Figure 11: Calculated Landslide Risk in Grand Mesa

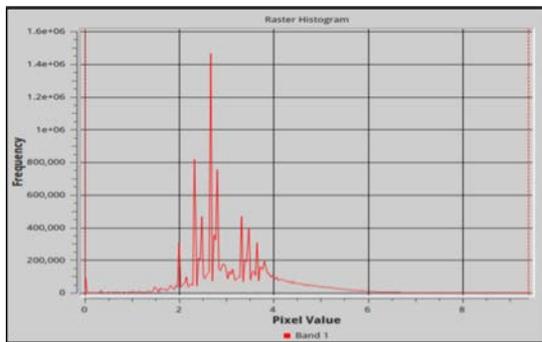


Figure 10: Histogram of Combined Imagery



Figure 12: Aerial Landslide Imagery in Grand Mesa

The calculated Landslide Risk map (Figure 9) shows a distribution of landslide risk across Colorado but there appears to be a relative concentration of highest risked areas where calculated values are greater than 6 (red region) in west Colorado. The highest risked areas appear to be located in areas of relatively high rainfall and much steeper slopes. Based on the visual comparison of the all above figures, the geology appears to have less of an influence than the other factors.

Locations of historical landslide events in US have been gathered from USGS web site are shown in Figure 9. The BLUE Polygon area means that more than 15% of the area is in-

Recently, There have been landslide events in the calculated high-risk zone like Figure 11 and 12, on the north slope of Grand Mesa. On Sunday, May 25, 2014, a large landslide rushed down a Colorado mountain near the town of Collbran covering an area three miles long and one-half to three quarters of a mile wide. It claimed the lives of three ranchers and triggered a small earthquake[9]. As we can see, this study would suggest that the employed method and calculation did capture the areas that have a high potential for land-sliding.

There are also some area where the calculated landslide risk is low but the more than 15% of area inside the polygon has landslide

history like Figure 13. This may suggest that the natural environment is a bit more dynamic than this static model or that there may be risks below the resolution (spatially) of this calculation which have not been captured and may have been averaged in this calculation.

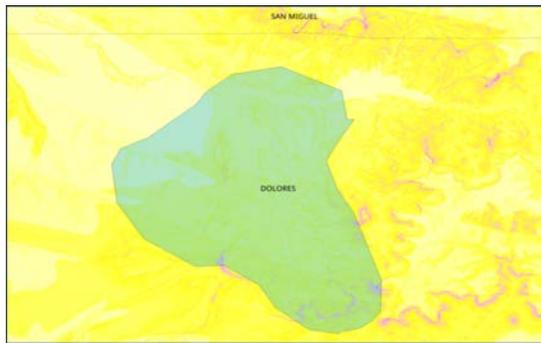


Figure 13: *Calculated Landslide Risk in Dolores*

ii. Result of Experiment -2: Classification

- Maximum Likelihood Classification

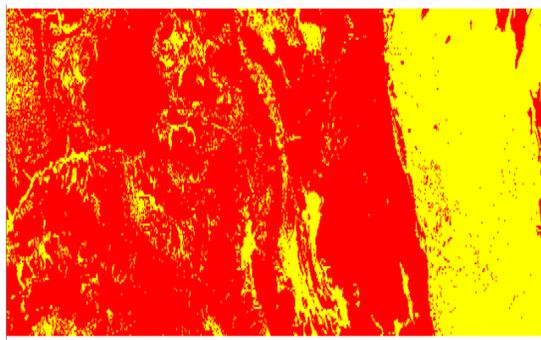


Figure 14: *Result of ML Classification*

According to the result, most of area (red region, 65.22%) of Colorado is classified as a high risk area for landslide. It seems that the result of classification is greatly affected by the slope data comparing the result with the slope figure. Also, comparing the this result with experiment-1(Heuristic Analysis), the method of classification seems to be very rough way to estimate the landslide risk.

When implementing classification method, there are some limitations about training data from USGS. There are two reasons for this rough result. The first reason is to use binary classifier. In this experiment, the ML classifier was trained with only two classes which are area with landslide history and without landslide history from USGS database because of lack of detailed landslide data. Therefore, it only predicts two classes and it is impossible to see the detailed landslide risk values like experiment1.

The second reason is that the training data is not accurate. The training data used for each landslide high and low risk region is Landslide incidence-susceptibility in the continuous United State shape file from USGS database. From this file, some each partial area are clipped and these area were fed into classifier to train the ML classifier. However, the landslide high risk area of the shape file means that more than 15% has landslide history. Therefore, it may also include a lot of area without no landslide history. This may result in inaccurate prediction.

Nevertheless, It might be possible to get a better result using classification method, if more accurate and detailed(multi-leveled) training data were used.

IV. CONCLUSION

The results of this study would suggest that Qualitative susceptibility zoning method(Heuristic Analysis) is an effective tool in examining landslide susceptibility problems when we do not have prior knowledge about the landslide history of some region. The rough assessments for landslide risk can be made without landslide history data. The results of this method will however be improved by consideration of actual empirical data which may guide the definitions of classes more robustly as well as the consideration of the weighting of the respective factors.

Moreover, this project shows that classifica-

tion method may be another option to estimate the landslide risk under the conditions that the accurate and detailed training data (landslide history data) can be acquired.

The project was successful in delineating high risk landslide areas of Colorado, especially using Heuristic Analysis, which is mainly western part of the State and are typically associated with areas of significant slope. The final result of this project relied on various forms of data retrieved from multiple sources. And the result can be improved using more relevant input data with landslide.

REFERENCES

- [0] Socioeconomic and Environmental Impacts of Landslides in the western hemisphere (Robert L. Schuster and Lynn M. Highland, U.S. Geological Survey, U.S.A)
- [1] A Simple Definition of a Landslide: Bulletin of the International Association of Engineering Geology, v. 43, p. 27-29. (Cruden, D. M., 1991)
- [2] Landslides in Colorado, USA: Impacts and Loss Estimation for the Year 2010
- [3] Colorado Geography from NETSTATE (http://www.netstate.com/states/geography/co_geography.htm)
- [4] Landslide characteristics and slope instability modeling using GIS (Lantau Island, Hong Kong (F.C. Dai))
- [5] Rainfall-induced landslide susceptibility zonation of Puerto Rico (Chiara Lepore, Sameer A. Kamal, Peter Shanahan, Rafael L. Bras)
- [6] The Relationship Between Geology and Landslide Hazards of Atchison, Kansas, and Vicinity (Gregory C. Ohlmacher)
- [7] Guidelines for landslide susceptibility, hazard and risk zoning for land-use planning (Robin Fell)
- [8] Maximum Likelihood Classification of LIDAR Data incorporating multiple co-registered Bands (Marc Bartels and Hong Wei)
- [9] Earth Observatory, NASA (<http://earthobservatory.nasa.gov/IOTD/view.php?id=83883>)
- [10] Landslide susceptibility from topography in Guatemala (J.A. Coe, J.W)
- [11] QGIS Training material (<https://www.qgis.org/>)
- [12] Grass GIS (<http://grass.osgeo.org/>)
- [13] Dbfpy, Python module for reading and writing DBF files (<http://sourceforge.net/projects/dbfpy>)

A.
APPENDIX : ENLARGED FIGURES

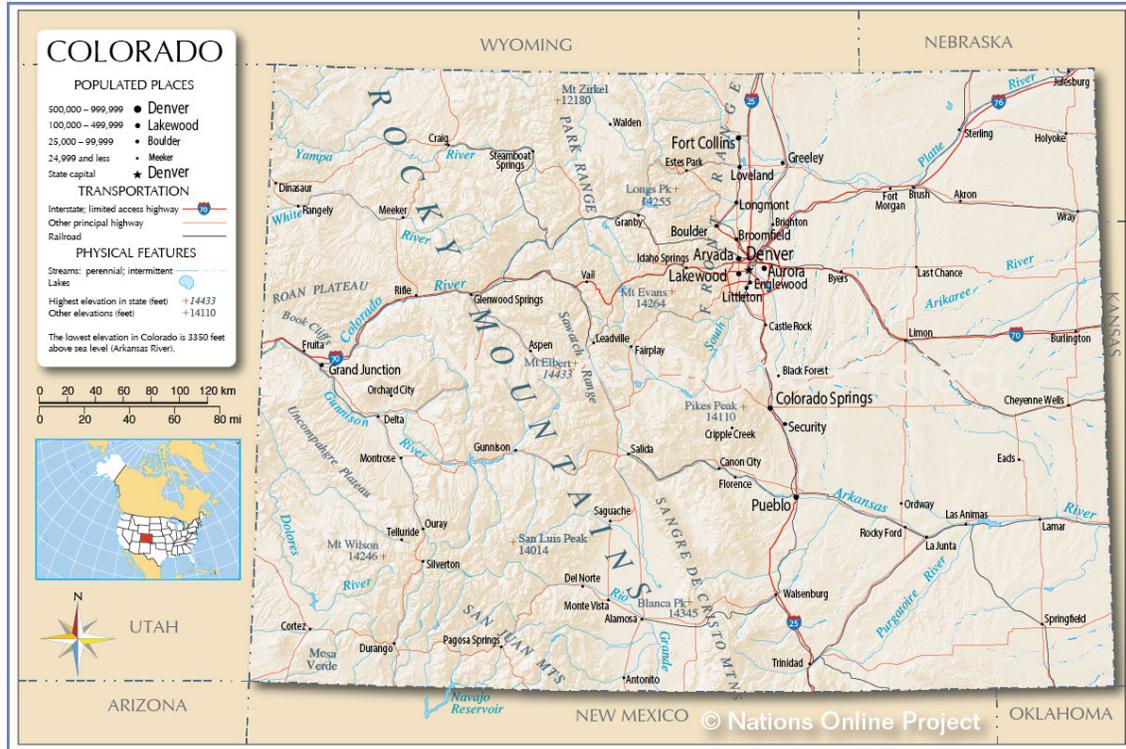


Figure 15: Study of Area Colorado State

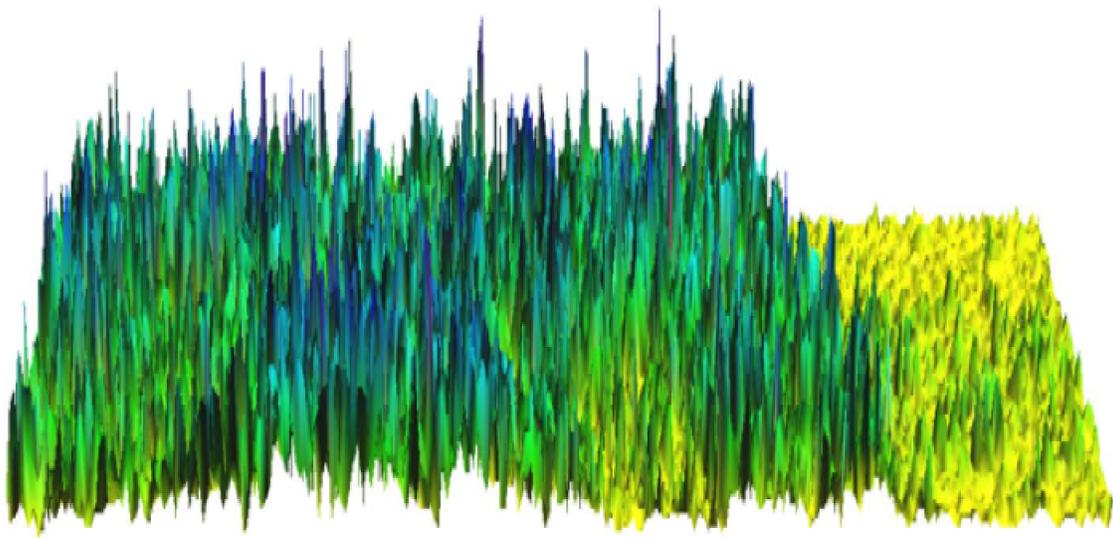


Figure 16: *3D Visualization of Slope*

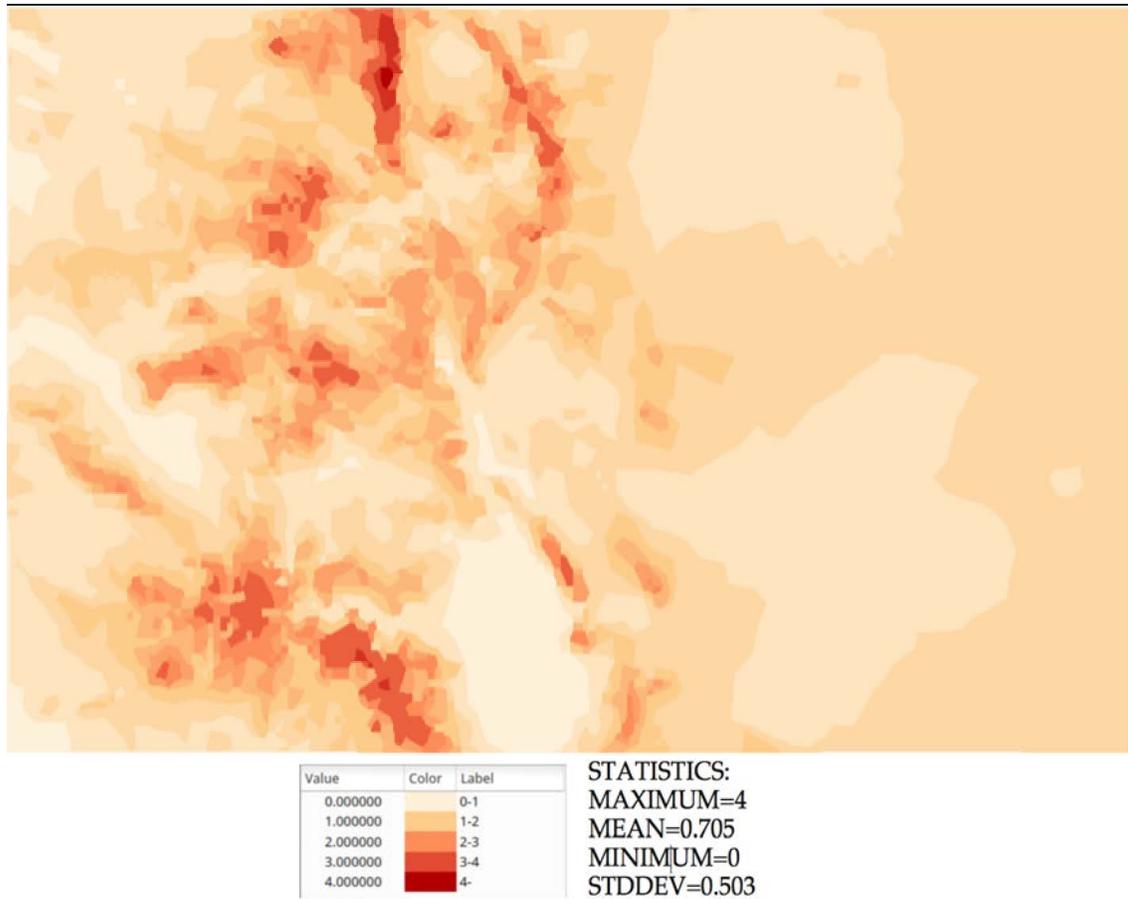


Figure 17: Average annual Rainfall Raster

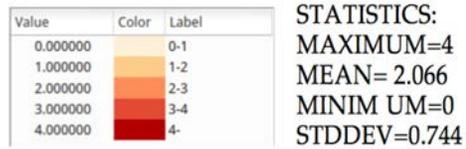
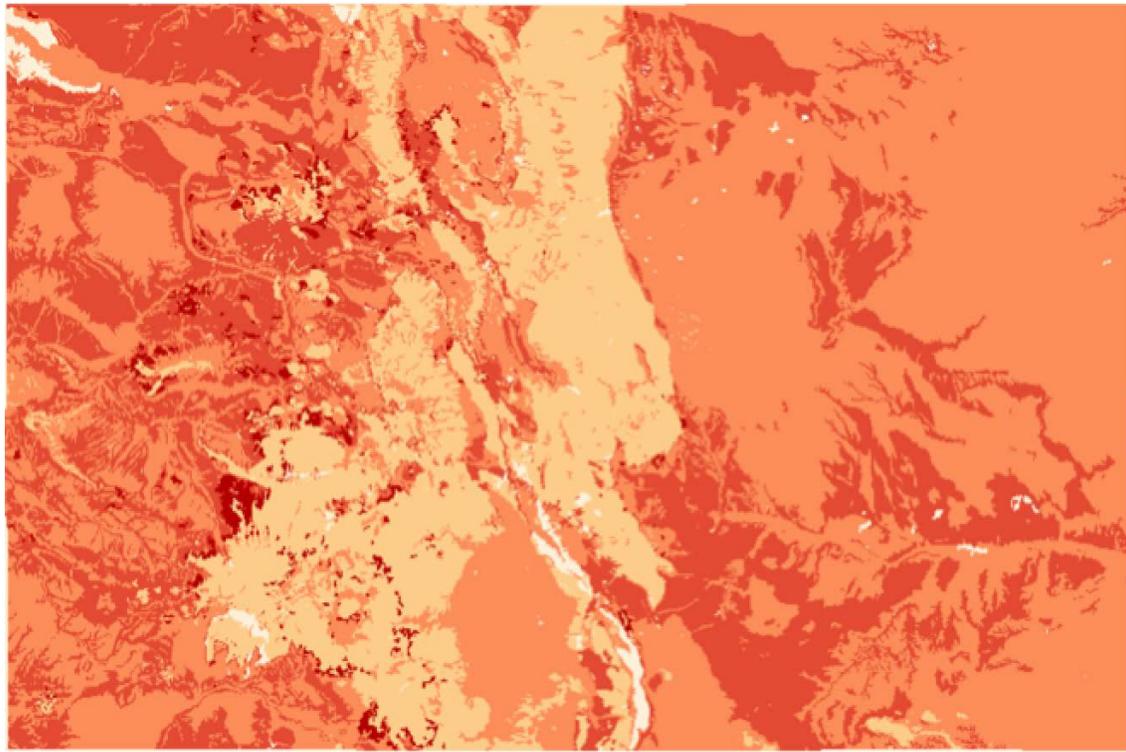


Figure 18: *Geology Raster*

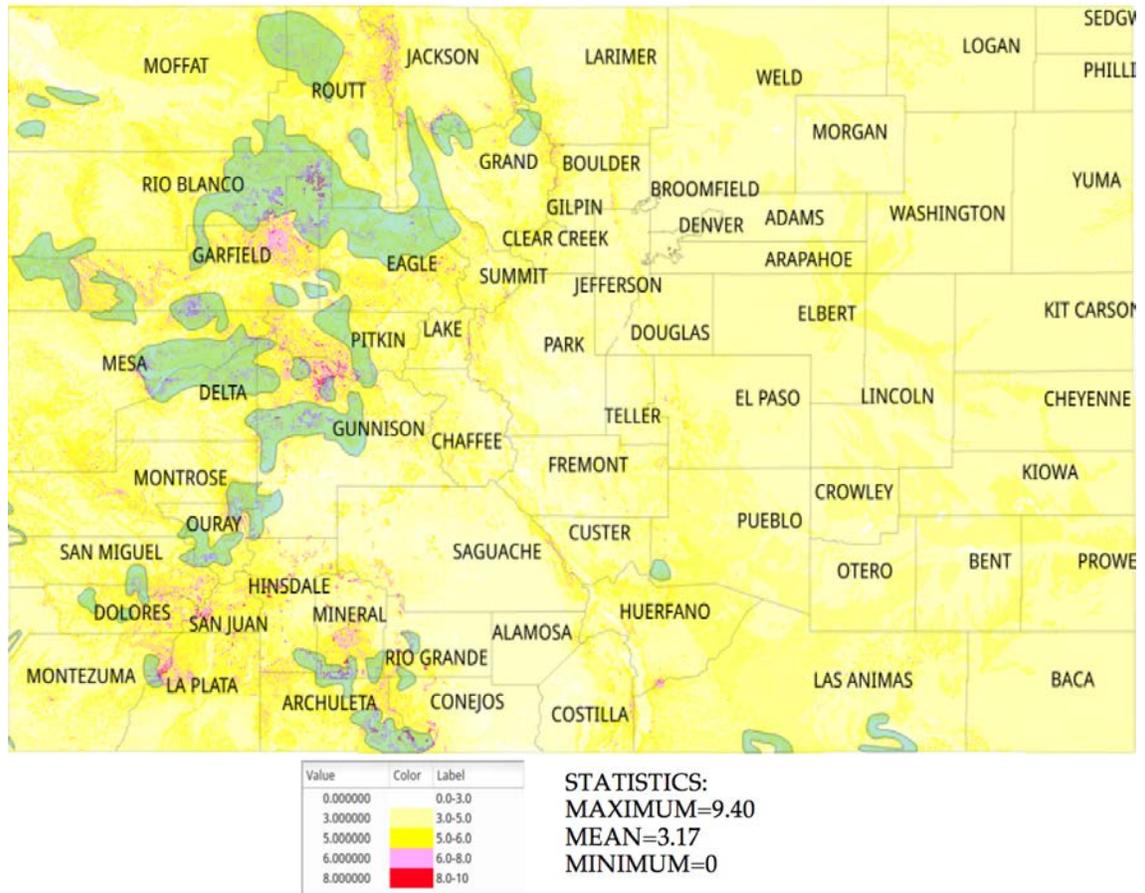


Figure 19: Combined Imagery : Landslide Risk of entire Colorado

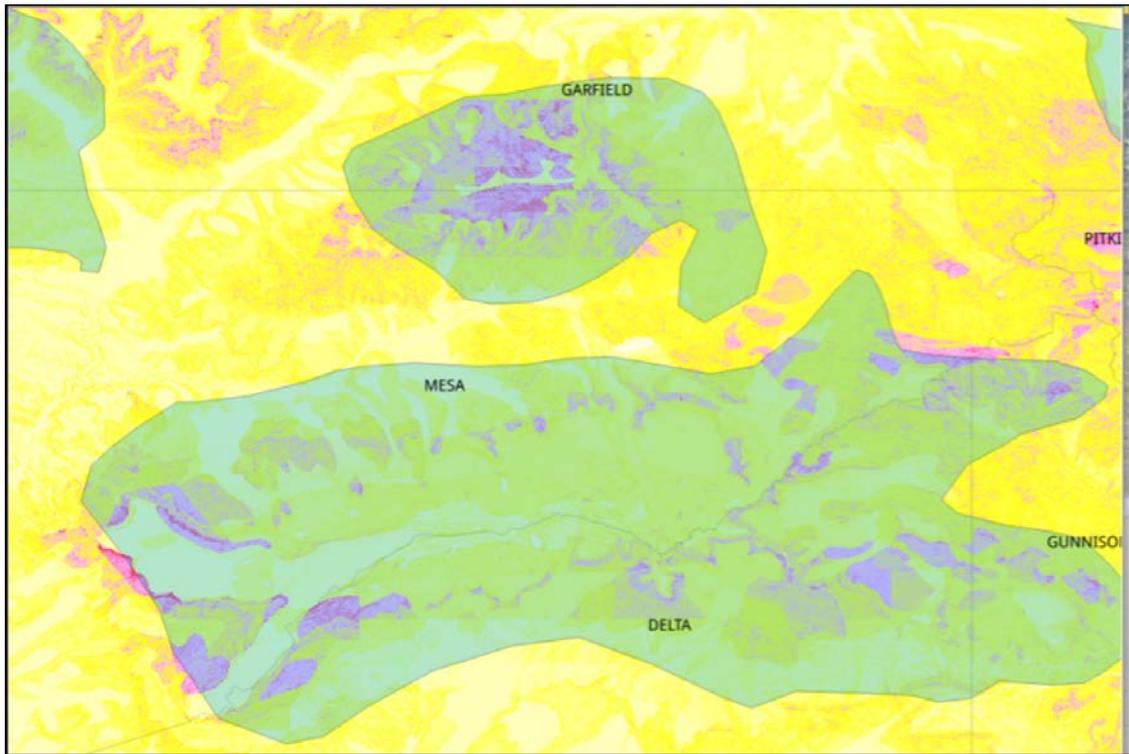


Figure 20: *Calculated Landslide Risk in Grand Mesa*

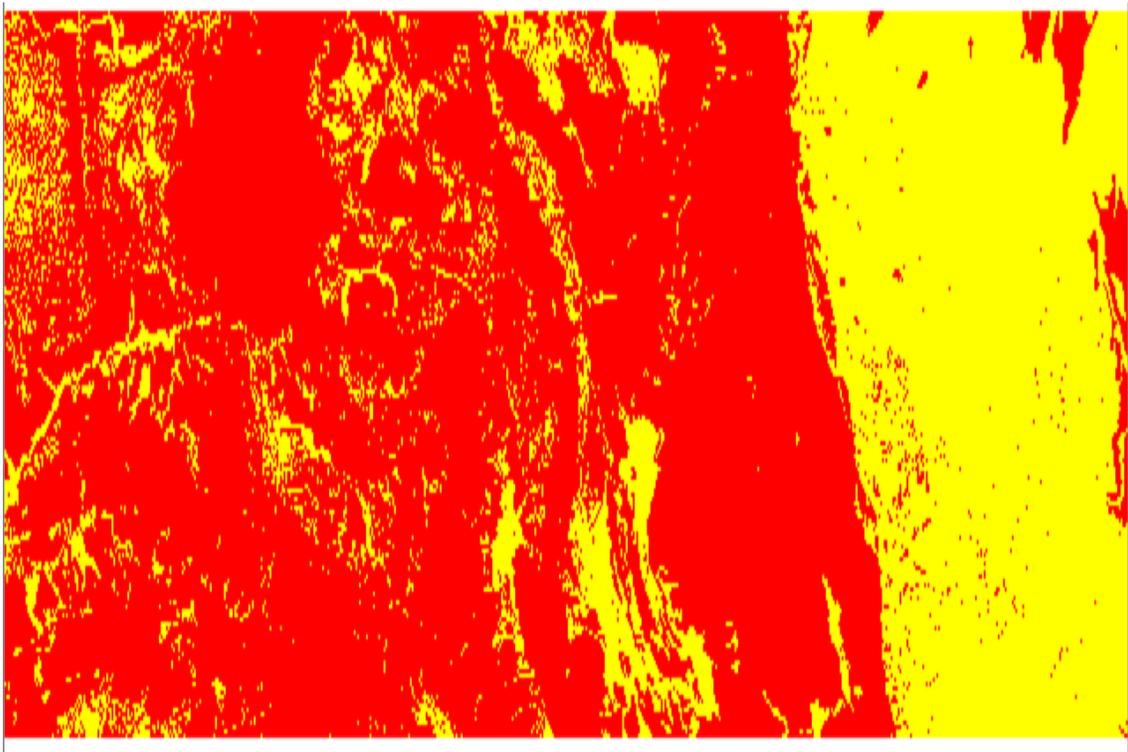


Figure 21: *Result of ML Classification*

Boulder Neighborhood Price Index

DAVID LEVIN*

University of Colorado Boulder
dale0007@colorado.edu

Abstract

The relative price of common food items and household goods is a major contributor to the cost of living in a particular area. Along with metrics such as property value and median household income, the consumer price index offers insight into just how expensive a region is. Often, this price index is measured on a broad, regional level such as state or county-wide which provides us a general understanding of the cost of living on a national scale. However, this generalized metric lacks the resolution necessary to determine relative affluence of a neighborhood within a city. In this study, we attempt to determine the relative cost of living for different neighborhoods within the city of Boulder, CO by studying the variance in the price of consumer goods within city limits. In addition, we use the data collected to draw comparisons between the different supermarkets based upon various factors.

I. INTRODUCTION

Of all the factors that influence consumer purchases of goods, cost is at the forefront of the consumer mind. However, there are a myriad of other factors that influence the consumer decision of where to purchase goods. In particular, the factors of convenience and quality are among the most influential, in addition to cost.

The tendency of consumers to align their decisions with these three important factors has given rise to many different classes of stores that cater to each of these considerations. The consumers who in the moment value ease of access to a product over cost, will choose one of the many millions of convenience stores across the country for their purchases. These businesses are aware of their role in the food chain, and will mark up their prices accordingly as they know their typical client is far less interested in cost.

Conventional grocery stores, on the other hand, are ones where low cost is at the forefront of the business model. These stores are typically large corporations that have a huge buying power, and as such can offer attractive low prices that will drive consumers to shop

with them. However these businesses are fewer and farther between than convenience stores in most cases, and lack the public perception of offering high-quality, healthy products.

In recent years as the public has become more conscious of the quality and source of their food, specialty health-food stores have cropped up around the nation which tout the products they carry as being healthier and superior in quality. However, it is commonly thought that with the supposedly higher-quality food of a health-food store, also comes higher prices.

While these three classes of retail markets exist for separate purposes and adhere to separate business models, they will most certainly have overlap in the types of products they carry. Basic items such as candy bars and toilet paper are ones to be carried in any of these stores, regardless of which type.

However, the relative cost of common household items found in these classes of stores, varies greatly. This is most certainly due to the confluence of many separate factors including size of store, corporate vs. local, location, hours open, etc. For an informed and thoughtful consumer, these are the factors that

*Big thanks to Prof. Caleb Phillips for all of his guidance throughout this project

influence their decision as to which store to go to, and for which purpose.

In our study, we examine the relationship between these factors and the relative cost of particular items, in order to find the factors which are most influential.

II. DATA

I. Determination of items

An integral part of our study was to populate a list of items that were the most ideal to show price differences between stores, in order to draw the most meaningful conclusions.

In our consideration for which items to choose, we looked to find items that are most readily available across all stores, that have most propensity for price variation, and that cover a large swath of the items typically purchased by a household.

Furthermore, we soon realized it was necessary to price some items by weight (eg. vegetables), some items by exact item (eg. B&J Pint), and some by lowest price (eg. milk), since each item was particular in how it ought to be measured. Care was taken to choose items from each class, in particular the 'exact' type which will be most useful for comparison.

As a final consideration, we looked to choose items that could be found in both grocery and convenience stores, so we may draw a comparison between the two types.

Our final shopping list, shown in (Figure 1), was a diverse list of 18 items with a variety of price classes.

Item	Price Class	Item	Price Class
Milk	Lowest	Eggs	Lowest
Peppers	Lowest/lb	Potatoes	Lowest/lb
Bananas	Lowest/lb	Oranges	Lowest/lb
Oranges	Lowest/lb	Bread	Exact
GF Bread	Exact	Clif Bar	Exact
Kind Bar	Exact	Turkey	Exact
Ice Cream	Exact	Water	Exact
Toilet Ppr	Lowest	Deodorant	Exact
Orange Jce	Lowest	Avocado	Lowest/lb

Figure 1: Final shopping list with price classification.

II. Method of measurement

Our measurement scheme was developed with the uniformity of recorded items in mind. We followed the precedent set by the United States' Consumer Price Index, which is very specific on finding the price of recorded products.

In the case of exact items the price recording was trivial, though for the other classes of items it was not so simple.

Items that were classified as 'lowest' were tied to a very specific definition of what that item was. For instance, we measured milk specifically as a conventional gallon of whole milk (non-organic). In the case that a similar but not very well-matched item was encountered, we sided in favor of not recording it.

For the produce items, classified as 'lowest/lb', we again targeted specific kinds of items as mentioned above. If these items were found as sold /ea when we were to record it as /lb, we weighed three of the items and recorded the weight, in order to later calculate a price per pound.

Another consideration we made was to include sale and "club card" prices into our data. If we found a price to be measured that also had some discount associated, we recorded both the original price and discount price.

III. Locations & Collection Method

Considering our group is based out of Boulder, Colorado, we chose our home city as the setting for our study. Our colleague Caleb Phillips, an associate of the Boulder Food Rescue, provided us a list of the grocery stores of Boulder, while the locations of convenience stores were pulled from the Factual API [Factual].

Once we received the data, it underwent another stage of refinement, coming to include 16 stores that we consider "supermarkets". The main criteria for this classification is a store that has a sizeable produce section.

The street addresses of each store were then geocoded and added to an online Google Maps map creator, which generated a map of the stores to be surveyed (Figure 2).

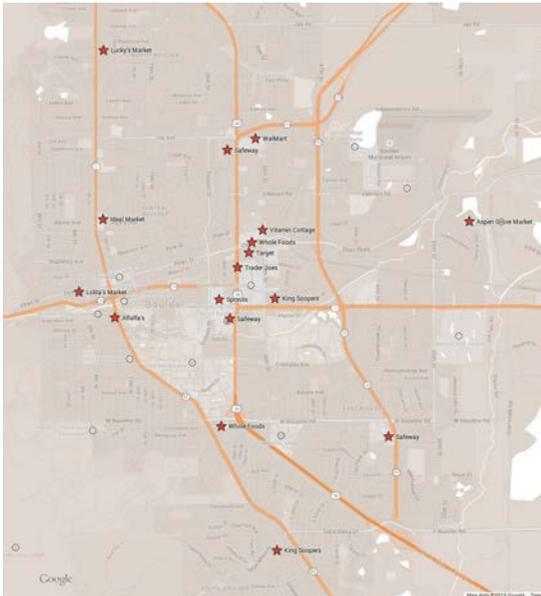


Figure 2: Tiny screenshot of supermarket map

IV. Method of Collection

The price of goods is certain to fluctuate over time. Thus, our team resolved to measure all of the prices across the various stores within the course of a few days, ensuring that ambient price fluctuations don't spoil our data.

As a gesture of good faith to our carbon cycle, our team has resolved to not directly emit any carbon emissions in the course of this data analysis. Therefore, we have chosen to bicycle around the city in order to collect our data. Our survey team will be equipped with a helmet and a clipboard with a sheet listing the items to price and which method to record each.

V. Other Sources of Data

In order to draw further comparison in between each of the stores, we drew more information about each location from alternate sources.

Using the Boulder County Property Viewer [County Assessor's Office], we drew rough bounding boxes around each store's footprint which gave us an estimate as to the square footage for each building.

The supermarket data we received had no zip code attributes, so we cross-referenced the stores in the list with the Factual API.

Again using the Factual API, and filtering the results, we obtain the amount of hours each store is open per week.

Finally, we use our personal intuition cross-referenced with the website of each store, to classify stores into local, chain, health-food and non-health food categories.

III. METHODS

Our first approach at analyzing the data is to simply look at it with the naked eye, to see if any obvious anomalies popped out immediately.

From here we digitize our data and import it into R for statistical analysis. By visualizing facets of the data with various types of graphs, we can draw some conclusions as to the relationships between the factors we are studying.

Once we have our data sorted, we move to begin the geospatial analysis using the lon/lat coordinates extracted from the Factual data. With various tests we attempt to determine if there is any correlation between the price of items and location of the marketplaces.

In addition to drawing conclusions from graphical representations and the spatial component, we use correlation tests and t-tests to determine if any non-spatial factors (Square Footage, etc.) are related to price.

IV. RESULTS

Upon initial inspection of the data, one thing was clear: the prices of items in chain stores were (almost exactly) uniform throughout their many stores in the city. The assumption which went along with our hypothesis, was that each store in the city would have unique prices for their items, which is now proven to be incorrect.

With the vast majority of stores (12/16) belonging to a chain, that causes our distribution to overall be spatially invariant; enough to ren-



Figure 3: Health food stores (green) vs. Non-health food stores (orange)



Figure 4: Chain stores (green) vs. Local stores (orange)



Figure 5: Price of items as a function of zip code (80301 in orange)

der useless the calculations we initially sought to perform.

I. Interpretation of Graphs

In light of this immediately apparent revelation, we were still able to test for correlation between price and our other factors. Using a scatter plot, we examined the relationship between whether a store is listed as a "health food" store and the prices of that stores' items (Figure 3).

Overall, there seems to be certain items where health food stores are more expensive than non-health food stores, and vice versa. Conventional milk, eggs, orange juice, and household products are consistently more expensive at health food stores, while produce is generally less expensive.

Applying this same analysis to the case of chain vs. local stores (Figure 4) does not yield nearly as fruitful results. The distinctions between which type of store is more expensive is apparent on few but not many of the items. This is aided by the fact that many locally-owned stores in our city also carry often did not carry any conventional products, so we lack the benefit of that comparison.

Looking at the same plot using the zipcode of each store as a factor (Figure 5), one can easily notice that the stores in zip code 80301 are by far the least expensive overall. This zip code encapsulates much of the dense commercial activity in our city, so this may be an indication that more commercial areas tend to be less expensive overall.

II. Numerical Analysis

To supplement the conclusions that we derived from the graphs, we also calculated the median value of the prices in each class of stores, in order find out which classes of stores were overall most expensive.

By selecting a subset of our data to include only the only 6 items that were found at each store, and then calculating the median, we obtained the following results: local stores are 30% more expensive than chain stores and

health food stores are 25% more expensive than non-health food stores.

Using a correlation test to find out if there is any direct relationship between square footage of a store and its price, or also its zip code and its price, we found a negative result which suggests that there is no direct relation.

Store Type	Median	% Difference
Local	4.59	29.6%
Chain	3.54	
Health	3.99	25.1%
Non-health	3.19	

Figure 6: Median price for each store class with price difference

V. DISCUSSION

While it is clear that the price variance across different stores in our city is not directly a function of the stores' location, our data still showed some interesting results.

In particular, our study has affirmed the general notion of consumers, that health food stores and locally-owned stores, are generally more expensive than their non-health food and chain store counterparts.

Furthermore, one implication of our results is that the most intelligent consumer will not choose which supermarket to shop at based upon geographic location, but rather based upon whether or not a store is a chain store and/or a health-food store.

However, it should be noted that in the case of our particular small city, the majority of supermarkets are part of a chain, whereas for other places in the country and the rest of the world the percentage of locally-owned stores could be higher. If this were the case, the prices of various stores would be unique in which case the price variation could very well be a function of geographic location.

Another interesting bit that came about with the data collection, was that of the price similarities in between stores of the same chain in a local region. Clearly this fact was not anticipated by us but was confirmed by one of the

managers of the Safeway stores. Furthermore, we learned from them that prices are changed once per week in their store.

VI. LIMITATIONS

By far the most limiting factor in our analysis, was the time constraint. To establish a sampling protocol and collect data, and then analyze that data within the span of finals month, was not enough time to perform the most robust analysis.

The most evident consequence of not having enough time, was the omission of convenience store prices in our data. An ideal vision of this study would have included prices from these businesses in order to draw comparisons, but the task of collecting data from supermarkets proved to be laborious so we were forced to cut out convenience store prices from our collection scheme.

Furthermore, due to the inexperience of our team in all aspects of our study including sampling design and data analysis, the results obtained took far longer than would have taken for a group of experienced data scientists.

One trouble that our team ran into was in some of the products that we initially chose for our metric. For example, our team initially thought to measure a 1lb bag of rice as one of the items, but it soon became apparent that the many different kinds of rice and bag sizes would be very difficult to compare between stores.

In addition, we initially wished to measure the lowest price for a whole wheat loaf, but the variety of different weights that bread loafs come in along with the sheer amount of different kinds of bread, rendered that metric too difficult to calculate and we instead switched to find an exact match.

On the data processing side, our study was plagued by lack of experience of its authors, in calculating additional metrics with our data in addition to the ones we performed. We would have liked to incorporate the sale vs. non-sale data and also to categorize the items (ex. produce, household goods...) in order to perform

numerical analysis on each group instead of simply deriving our data from the graphs.

Finally, we were not able to pull the data necessary for finding the hours per week that each store is open, due to a complication with the Factual API. This limited our analysis to the metrics that we were able to collect information for.

VII. CONCLUSION

In this study, we set out to complete a novel geospatial analysis on supermarket prices but instead affirmed commonly perceived notions of consumers as to which types of stores are most expensive. The disparity between what we intended to find and what we did find, was due to an error on our part by assuming supermarket prices would be unique for each store.

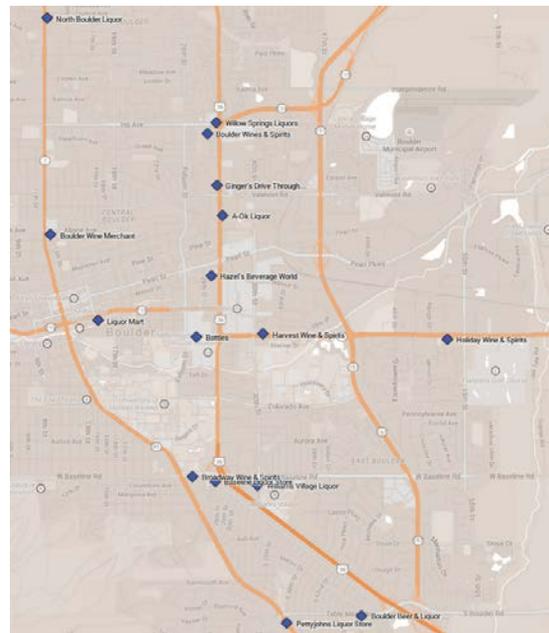


Figure 7: Tiny map of locally-owned liquor stores

As a consideration that could only have come about with a trial such as this, in the future we look to study the prices in other classes of stores that are predominantly locally-owned so that their prices will be far more unique on average, in order to perform the spatial analy-

sis we initially intended to perform.

With the Factual API data for businesses in Boulder, we found that all of the liquor stores in our city are not part of a chain, making them ideal for a similar study in the future. With the distribution of these businesses covering a large portion of our city (Figure 7), we could most readily draw some meaningful geospatial analysis with data from these locations.

Furthermore, since liquor stores are all locally owned, their hours open per week and store sizes will vary more than stores of a chain, further adding to the benefit to studying liquor

stores over supermarkets.

In addition, as a virtue of convenience for future studies, we shall write down the aisle number of each product so that if we wish to return we may locate the products more readily, cutting down on the time needed for data collection.

Overall, our study was initiated with good intentions but was hampered by a faulty assumption, though we were still able to draw some meaningful conclusions and surely shall use the lessons learned from this experience to inform future studies.

REFERENCES

[County Assessor's Office] <http://maps.bouldercounty.org/boco/propertyviewer/>

[Factual] <http://factual.com>

Boulder Flood Impacts Pre-flood and Post-flood - 2013

JOHN RAESLY*

University of Colorado - Boulder

john.raesly@gmail.com

3/29/2015

Abstract

The 2013 flood in Boulder, called the 100 year flood, was a natural disaster for the century in Boulder. The floods' power changed much of the ecosystem that surrounds Boulder, CO. The purpose of this project is to examine the 2013 flood in Boulder, CO and its impacts on the river morphology of the creeks and rivers surrounding Boulder, CO and the change in land cover that occurred. Direct correlations between the rivers and the amount of rainfall in those areas feeding into the rivers are of particular interest. This study will investigate how river morphology has changed from pre to post flood? Do areas designated as floodplains affect the river change? What is the correlation between rainfall amounts and river density change? What is the land cover compared to water cover for pre versus post flood? QGIS will be utilized to perform file conversions and other usable formats for Grass GIS. Then Grass GIS will be used to integrate an analysis method for the raster data and dissolve it as necessary. Previous techniques from other flood studies that have shown to be effective will be analyzed. Sources will include the Boulder Open GIS, which has some from the NOAA Disaster Data to analyze similar conditions (rainfall amount). This study will focus on Boulder County in Colorado.

I. INTRODUCTION

Understanding the river morphology for the Boulder 2013 Flood is important for assessing the floodplains that are at the top concern if such an event were to happen again. The 2013 flood in Boulder was a natural disaster that could be cause for concern. The flood changed the stream channel within Boulder County, CO. The flood caused a total of 3 fatalities during its 8 day rain period. 306 people are still reported as missing (as of September 19th 2013). The flood had affected an area estimated to be 4,500 square miles — roughly the size of Jamaica. 12,118 people were under mandatory evacuation orders and 1,000 people had to be airlifted to safety from remote locations, the largest airlift rescue operation in USA since Katrina. 1,502 homes destroyed, 17,494 homes damaged (Colorado

Office of Emergency Management estimates) and 30 (state-maintained) bridges destroyed, 20 (state-maintained) bridges damaged (according to Colorado Department of Transportation). [1] This study focuses specifically on Boulder County. In this study, the 2013 flood in Boulder, CO is examined and its impacts on the river morphology of the creeks and rivers surrounding Boulder, CO and the change in land cover that occurred. Direct correlations between the rivers the amount of rainfall in those areas feeding into the rivers are of interest. In particular, this study will investigate: How river morphology has changed from pre to post flood?; Do areas designated as floodplains affect the river change?; What is the correlation between rainfall amounts and river density change?; What is the land cover compared to water cover for pre versus post flood?

*A thank you or further information

II. DATA

The data for this study was gathered from the Boulder Open GIS (Geographic Information Systems) on the Boulder County government website. This data is free to use and be distributed. Some of this data is sourced from a make up of NOAA's Disaster Data. NOAA (National Oceanic and Atmospheric Administration) is an American scientific agency within the United States Department of Commerce focused on the conditions of the oceans and the atmosphere. [2] The rainfall amounts on the Boulder Open GIS website were a compilation of all the days added into one dataset.

III. METHODS

Several measures were developed to assess the stream channel - pre flood against the stream channel - post flood shapefiles.

- Clip the shapefiles
- Find the difference in the shapefiles
- Branch shapefiles and find flow patterns
- Assign rainfall minimum and maximum attribute for Boulder County
- Assign general rainfall minimum and maximum to separated areas
- Correlate the rainfall amounts against the post-flood stream channel

I. Clip the Shapefiles

QGIS was used for most operations during this research. Using the built-in function, Clip, was able to show the similarities in both of the shapefiles. It created a new shape based on the area of the input layer that is overlapped by the clipping layer. The attributes of the chosen layer only were copied to the new feature. With the new information of which parts stayed the same, it was then useful to examine the portion that did not change. It was then discovered that only a small portion of the stream channel did not change. In figure 1 (image file not created yet), you can view the similar stream channel across Boulder County.

II. Find the Difference in Shapefiles

The difference function in QGIS was then used to determine the difference in the two shapefiles being examined. Both of the shapefiles had the same attributes which made it simpler to compare and contrast. The difference function created a new feature based on the area of the input layer that was not overlapped by the clipping layer.

III. Assign Rainfall Attribute for Boulder County

The purpose of this part was to assign a general legend for each area that fell into similar rainfall minimum and maximum ranges. The numbers can be viewed in figure 1. It was based upon finding a general rule for each set of values for the designated areas.

IV. Correlate Rainfall and Post-flood River Morphology

The correlation was created by implementing Spatial Autocorrelation within QGIS. It was also easily viewed by overlapping the rainfall data with the post-flood stream channel shapefile. Which can be viewed at figure 4.

IV. RESULTS

Only 48.4% of the stayed consistent with the pre-flood stream channel during this disaster. The mass of the findings concluded that most of the now stream channel was newly created. The post-flood stream channel is mostly represented by the rainfall that happened during the flood. Where there was less rain, most of those channels have disappeared and dispersed into the areas where there was the most rainfall. There was very little stream channel in Boulder, CO pre-flood, but now there are many new streams that have been created because of this flood and it mostly amounts to the mass rainfall that was recorded in the Southwestern part of the county.

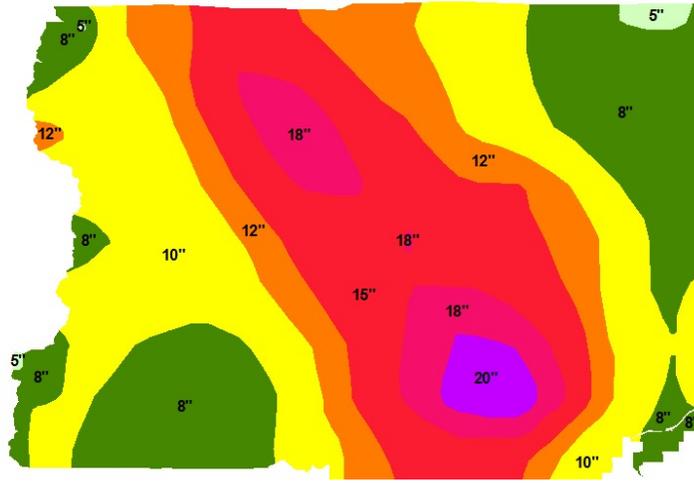


Figure 1: Rainfall Distribution

V. DISCUSSION

The entirety of this disaster created many new stream channels throughout Boulder County. Whereas most of the stream channel occurred in the Northwestern part of the county it is now distributed evenly throughout the county with more of it leaning towards the eastern border. Large amounts of rainfall can be extremely effective in rerouting a whole river system. This causes many unknown side effects. There could be a loss of habitat in the area that for so long had relied on the river system, but the newly created streams could thrive and welcome all new sorts of habitat. This does mess up the ecological balance that nature depends on. It is like taking the wolves out of Yellowstone.

VI. LIMITATIONS

Since there was not a lot of time allotted for this study, some limitations have come about. The nature of the vectors provided by Boulder Open GIS did not include any information about the river density which was a subject of interest. It would have been beneficial to look at multiple other sources for data to see

if there were any conflicting results. As most weather companies have their own rain catchers', some of the resulting rainfall data could have changed from a day to day basis. As there are now many reasons to attribute to habitat loss, ie global warming, the main focus should be upon the species that whole heartedly relied upon the previous river system. Another focus of study should be if there are new disasters that were previously averted because of the river morphology. One of the disasters could be the snow melt and the floodplains. Does the new stream channel cause concern for residents situated around these floodplains that were previously unaffected?

REFERENCES

- [1] "Colorado Flood Facts and Figures — FloodList." FloodList. September 19, 2013. Accessed May 3, 2015. <http://floodlist.com/america/usa/colorado-flood-facts-figures>.
- [2] "National Oceanic and Atmospheric Administration." Wikipedia. Accessed May 3, 2015. http://en.wikipedia.org/wiki/National_Oceanic_and_Atmospheric_Administration.

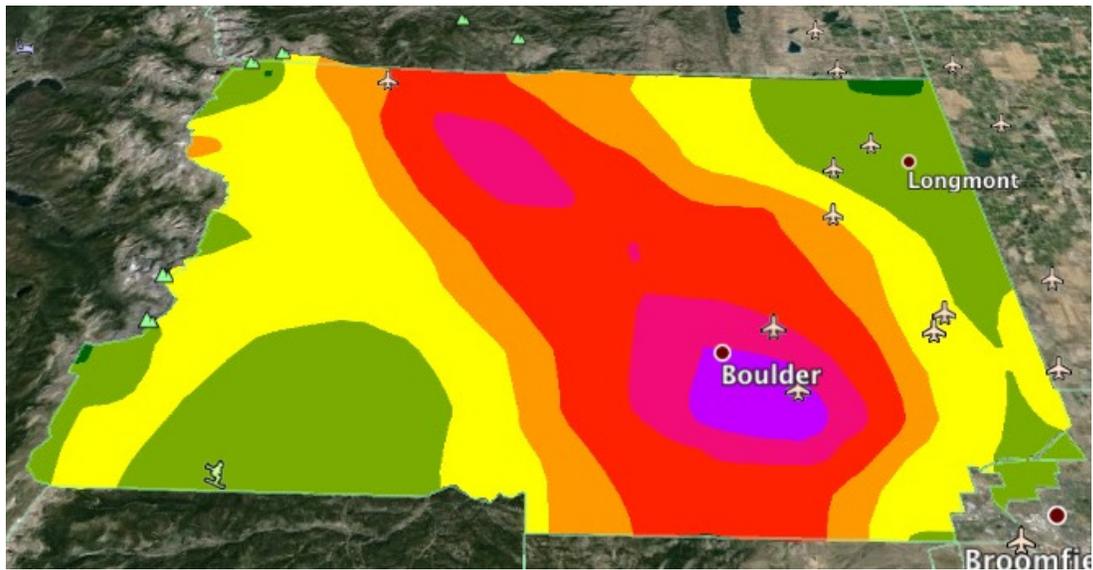


Figure 2: Rainfall Distribution over Boulder County

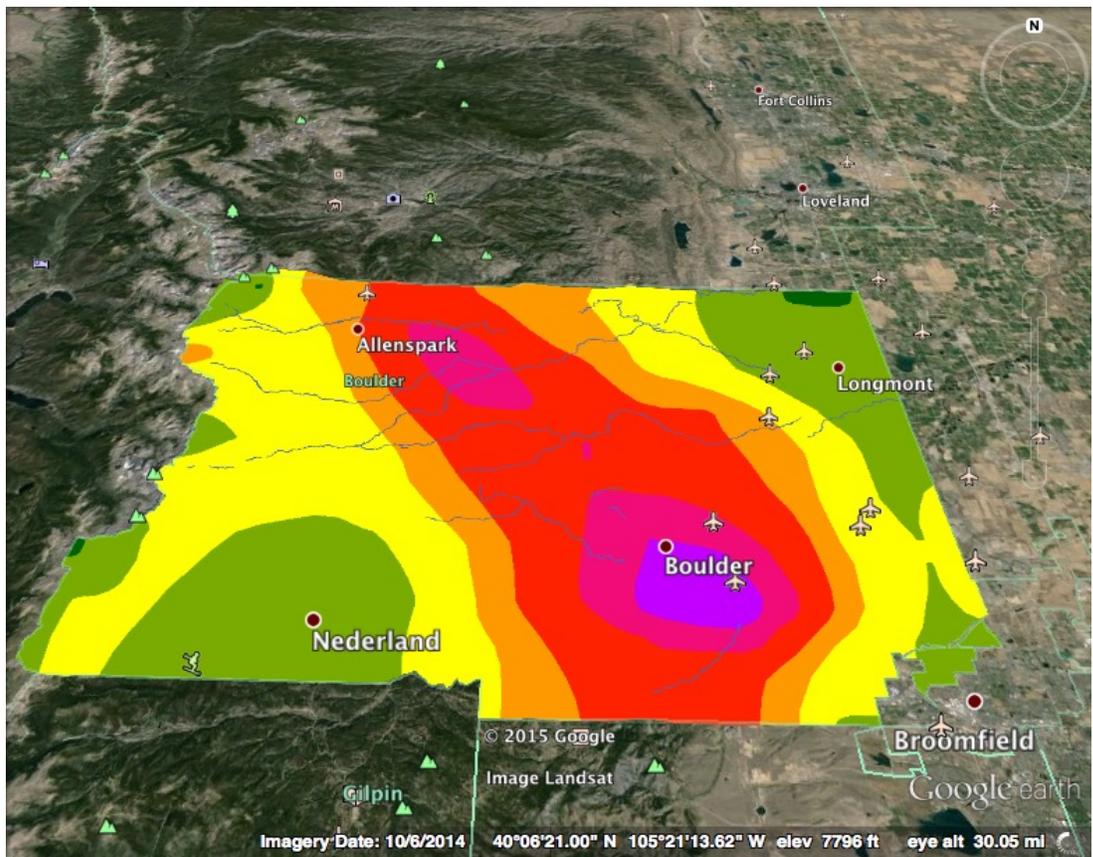


Figure 3: Pre-Flood Stream Channel and Rainfall Distribution

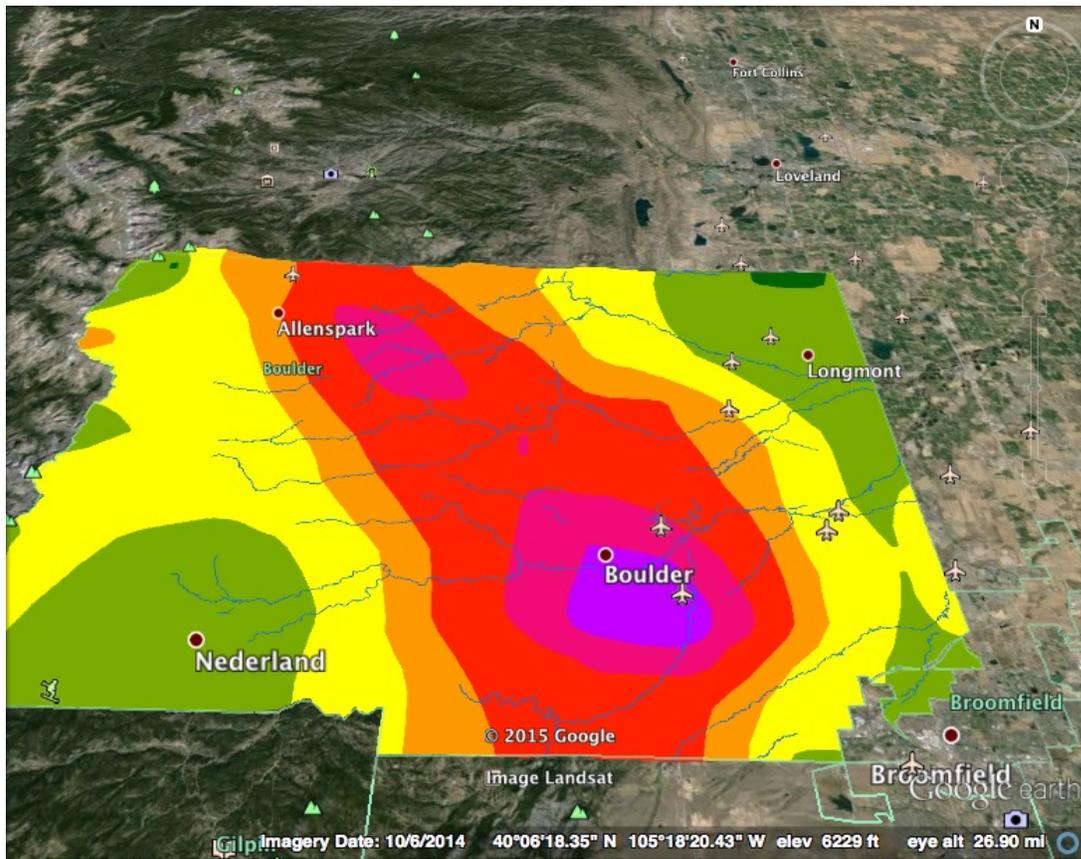


Figure 4: Post-Flood Stream Channel and Rainfall Distribution

A Geospatial Analysis of Oppressive Language On Twitter

FORREST TAGG RIDLER

University of Colorado Boulder
forrest.ridler@colorado.edu

Project STYX hosted at:

<http://ec2-52-10-55-168.us-west-2.compute.amazonaws.com:8080/>

September 10, 2015

Abstract

The use of slurs in any society is a contentious subject. The meanings and connotations of words that are or have been used to express hate toward a specific group of people constantly offend, change, and are reinterpreted. In this project, a system called STYX was developed to track the use of a subset of such words with derogatory connotations in the United States. It tracked words related to racism, homophobia, and misogyny which were returned by the TRACK endpoint of the Twitter streaming API. Totals were calculated from all Tweets returned, regardless of whether or not the tweet was geo-tagged. Map data was generated using all geo-tagged tweets within the USA. The spacial results of STYX seem to indicate that the use of derogatory slurs is directly correlated with population density, and that use is widespread without any clear correlation in space. The summation statistics indicate that misogynistic words are the most prevalent, homophobic words are moderately used, and that racist words are the least used. It is important to realize that STYX is only relevant for analysis of Twitter users with geo-tagging enabled, not society as a whole. The developed framework could be used to track the spacial dynamics of any keyword-category topic in the USA, not just derogatory slurs.

I. MOTIVATION

Words can have divisive and confusing affects on a society, thus it is important to understand the types of such words which are used and what contexts they're used in. Project STYX attempts to find some subset of answers to the question of where derogatory slurs are used, and how prevalent certain types of slurs are.

II. METHOD

STYX was implemented in a full-stack JavaScript environment involving Node.js, MongoDB, Express.js, Socket.io, Leaflet.js, and Angular.js. It was built in a way so that it is real-time, asynchronous, and always up-to-date with the data it generates.

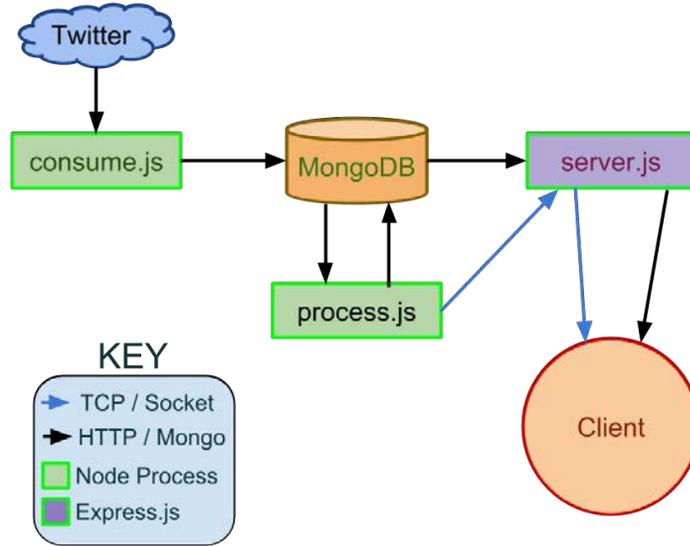


Figure 1: STYX System Architecture

STYX is built with Node.js. One Node process, called "consume," connects to the Twitter Streaming API and waits for data to be returned. Once data is returned, it immediately writes the data to a MongoDB database hosted by Mongolab. No intermediate processing is done on the tweets in order to comply with Twitter's requirement that users keep up with the streaming API.[1]

A second Node process, called "process" connects to the database and reads all Tweets saved in the database every five seconds. For each Tweet read, it extracts coordinates (if there are any), and generates counts based on the category that the Tweet's text falls into. It saves a statistics object, coordinates objects, and Tweet objects back to different collections in the database. It simultaneously writes any new data to a socket controlled by Express. Finally, "process" deletes all the Tweets saved by "consume" from the last five seconds. For further processing, the Tweets could be saved indefinitely, however STYX relies on free-tier cloud technologies, so persisted data is kept to a minimum.

To enable visualization and interaction with the data, STYX makes use of Express.js,

Socket.io, Angular.js, and Leaflet.js. Express.js acts as a web server for Node. It controls two main components: a REST API to load data onto a map and a series of sockets to push data as it is into the web page. Once a user loads the page, angular makes a REST call to Express, and populates the main map showing the 5000 most recent geo-tagged Tweets color-coded by category. It makes another REST call to get the 500 most recent geo-tagged tweets to display on an "Interactive Map" where the user can actually view the Tweets' text and find the user who Tweeted such a Tweet on Twitter. Socket.io simultaneously connects to the TCP socket controlled in Express to update the statistics shown on the web page every five seconds, as well as display the most recent geo-tagged tweet on the "Interactive Map" seconds after it is posted. Leaflet.js is what does all of the mapping, and it uses Open Street Map as its "Tile Layer."

A list of the (offensive) words being tracked can be viewed at the Project STYX URL.

III. RESULTS

III.1 Spatial Analysis

When considering only the USA, it appears that derogatory slurs are highly correlated with

population density. Following are samples of maps displaying occurrences of different categories of Tweets. Racist Tweets are red dots, homophobic Tweets are orange, and misogynistic Tweets are yellow.

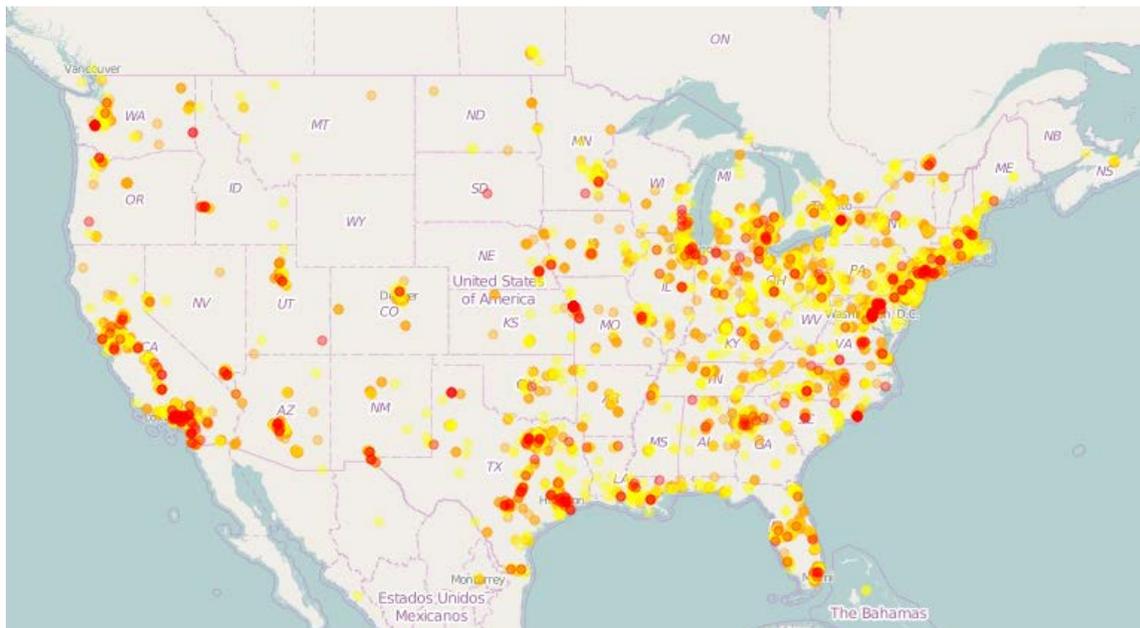


Figure 2: Combined Tweets

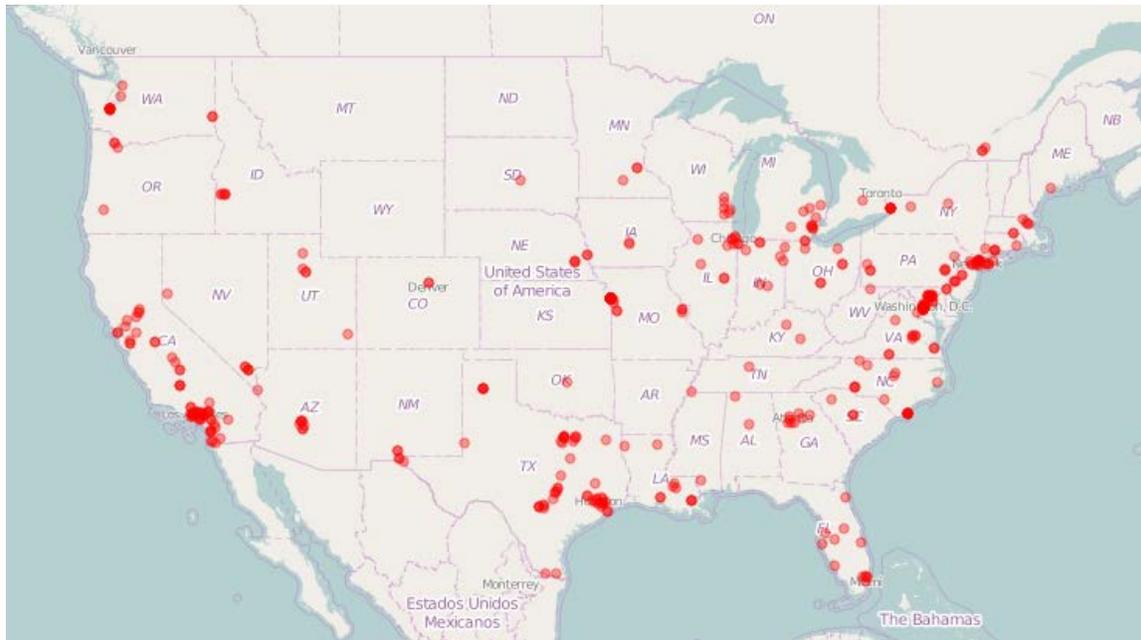


Figure 3: Racist Tweets

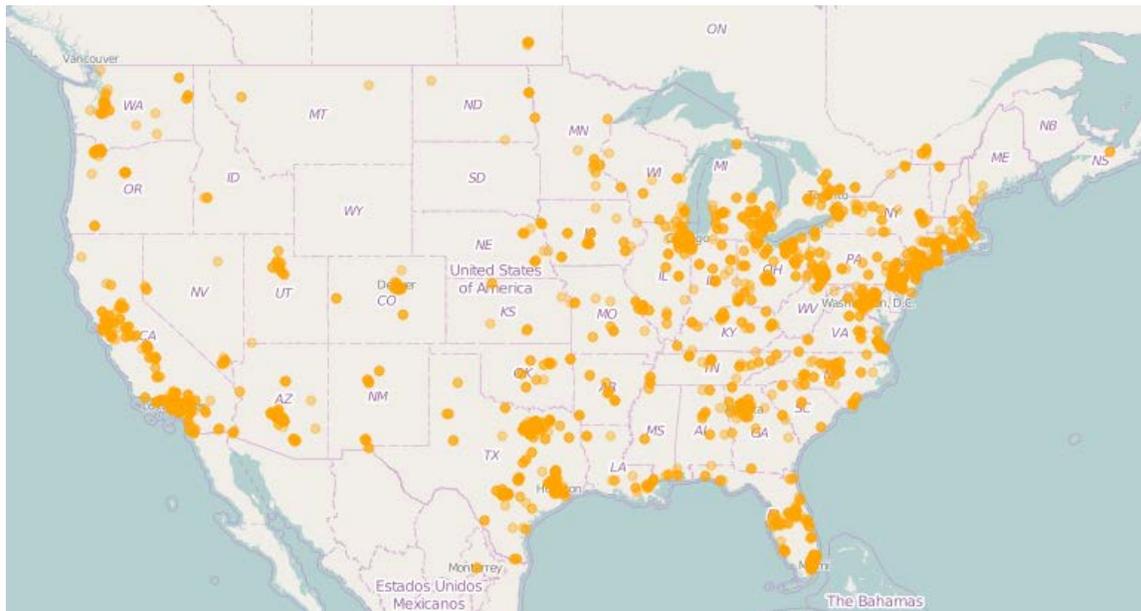


Figure 4: Homophobic Tweets

Without further analysis, maps like these do not reveal much other than that people who use derogatory slurs are well-distributed in the U.S. population. Further, more useful analysis

is discussed later in this paper.

At the start of the project, STYX was not focused on the US. Analyzing the whole world is out of scope for STYX, but the following

map is interesting. While the United States appears to be well-mixed in its use of derogatory slurs, the UK has a very high density of misogynistic occurrences and South Africa has

a quite high number of racist occurrences. This map definitely shows a cultural factor related to the popularity of derogatory slurs in casual language.

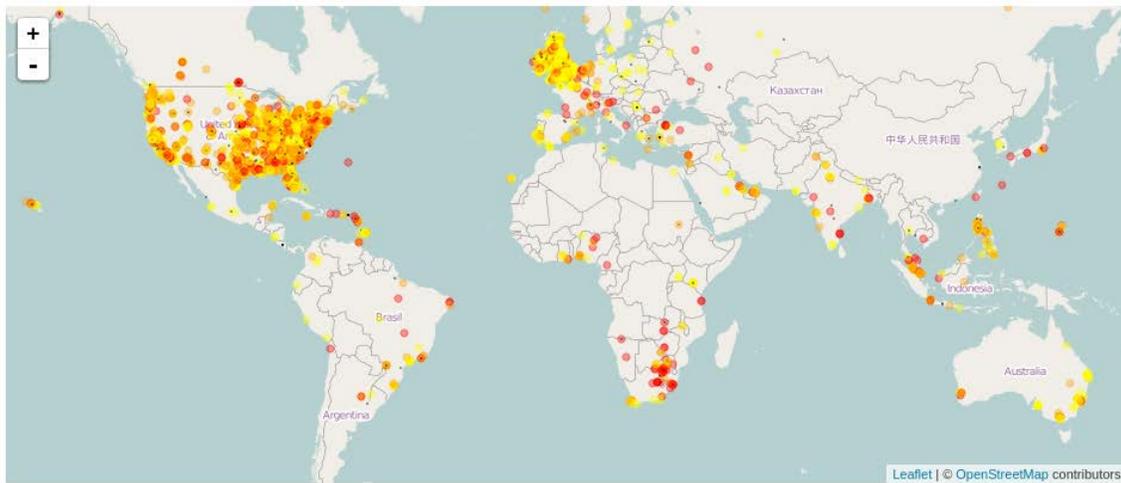


Figure 5: Tweets Around the World

In order to make more use of the generated data, another round of batch processing was done, separate from the streaming analysis in STYX. GeoJSON files containing shapes for all U.S. counties and population estimates for each U.S. county were obtained from the Census Bureau and fused together to create a new GeoJSON file where the population estimates are a property of each county. Using Turf.js this file was then matched against each geo-tagged tweet in the STYX database, and a "Tweet Count" was added as another property to the GeoJSON file. Finally, a rating metric was calculated and added as a property to the

file using the following formula:

$$\frac{100,000 \times t}{p}$$

where t is the number of tweets and p is the estimated population for each county. The metric is scaled by 100,000 in order to maintain resolution across all visualization layers, since there are very few tweets compared to population.

With this final GeoJSON file containing all of the fused data complete, a heat map can be generated with Leaflet.js. It shows counties that have a relatively large number of derogatory slurs for its population.

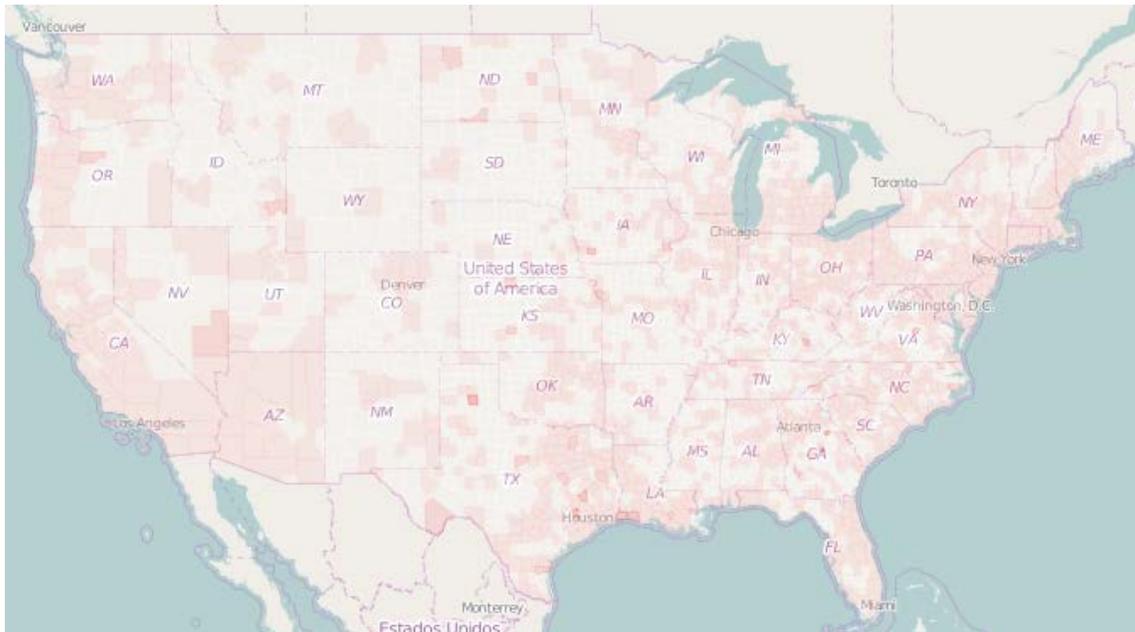


Figure 6: U.S. Counties

The above map shows each U.S. county colored red in proportion to its derogatory slur rank described above. The map can be viewed in higher resolution with details about each individual county at the project STYX URL.

The map is quite interesting, but an overall frequency distribution of the rankings was calculated to provide a more complete picture of the data.

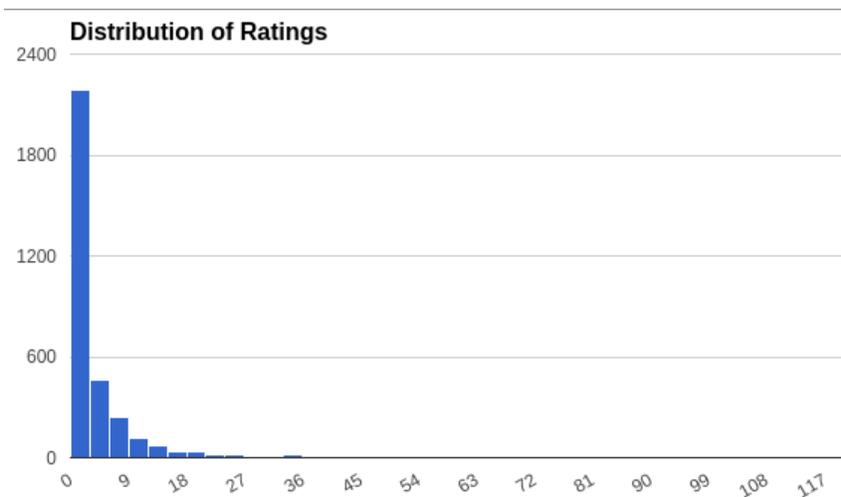


Figure 7: Complete Histogram

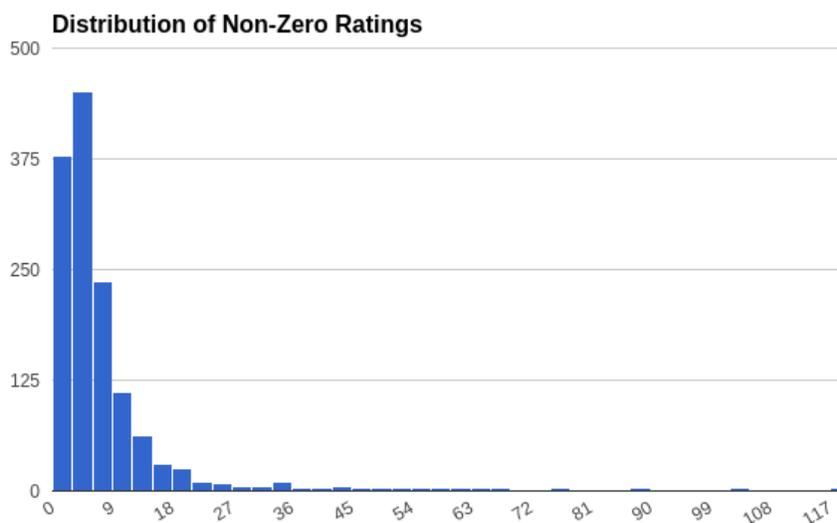


Figure 8: Adjusted Histogram

An overwhelming majority of counties have no tweets with derogatory slurs in them, and many counties have very few Tweets relative to the population there. There are a few counties, however with very high proportions of derogatory tweets compared to population.

III.2 Aggregation

A running total was calculated every five seconds reflecting every tweet returned by the streaming API. After running for about three days, it returned the following numbers:

Category	Total	Ratio
homophobic	136,185	0.31
misogynistic	287,220	0.65
racist	15,193	0.03

The ratio column is the ratio of the given category versus the total number of derogatory Tweets received, not the total number of Tweets posted on all of Twitter. We did not have access to determine the exact number of all Tweets posted, so an estimate was generated: The percentage of Tweets that contain the words STYX was tracking was estimated under the assumption that Twitter receives about 6000 tweets per second, or 30000 tweets per five seconds. STYX received between 3 and 90 tweets per five seconds, which indicates that

between 0.01 and 0.3 percent of Tweets contain derogatory terms. On average, 24 tweets were received every 5 seconds, which indicates an that average of 0.08 percent of the total Tweet population contain derogatory slurs.

IV. CONCLUSIONS AND FUTURE WORK

Currently, STYX is only capable of finding keywords and categorizing them. A possible extension to the project would be an implementation of a form of NLP to determine the severity of the context in which a derogatory term is used. Facetious uses of these words are far different from someone making physical threats to someone who they directly call one of these words. Both of these types of messages were displayed by STYX.

A similar analysis was done at Humboldt state university. Their findings are not very detailed, but they generated a map that sheds more light on hate speech on twitter.[2]

Another possible use for STYX would be to analyze any categorizable topic. For example, if someone wanted to study the spacial proliferation of Colorado’s top microbreweries, all they would need to do is provide new configuration files, and after a few days, STYX would show the results. On April 27th,

2015, we ran STYX with keywords related to civil unrest such as "BlackLivesMatter," "riot," "mob," "ISIS," "Lynch" etc. categorized as "social" and "terrorist." There was a relatively violent protest occurring in Baltimore that day, and there was a clearly high density of "so-

cial" points on the map near Baltimore.[3] This demonstration of STYX's other abilities was brief, but it showed how STYX could be used as an application to a much different categorizable topic than derogatory slurs. Tweets related to the Baltimore riots are mapped below.

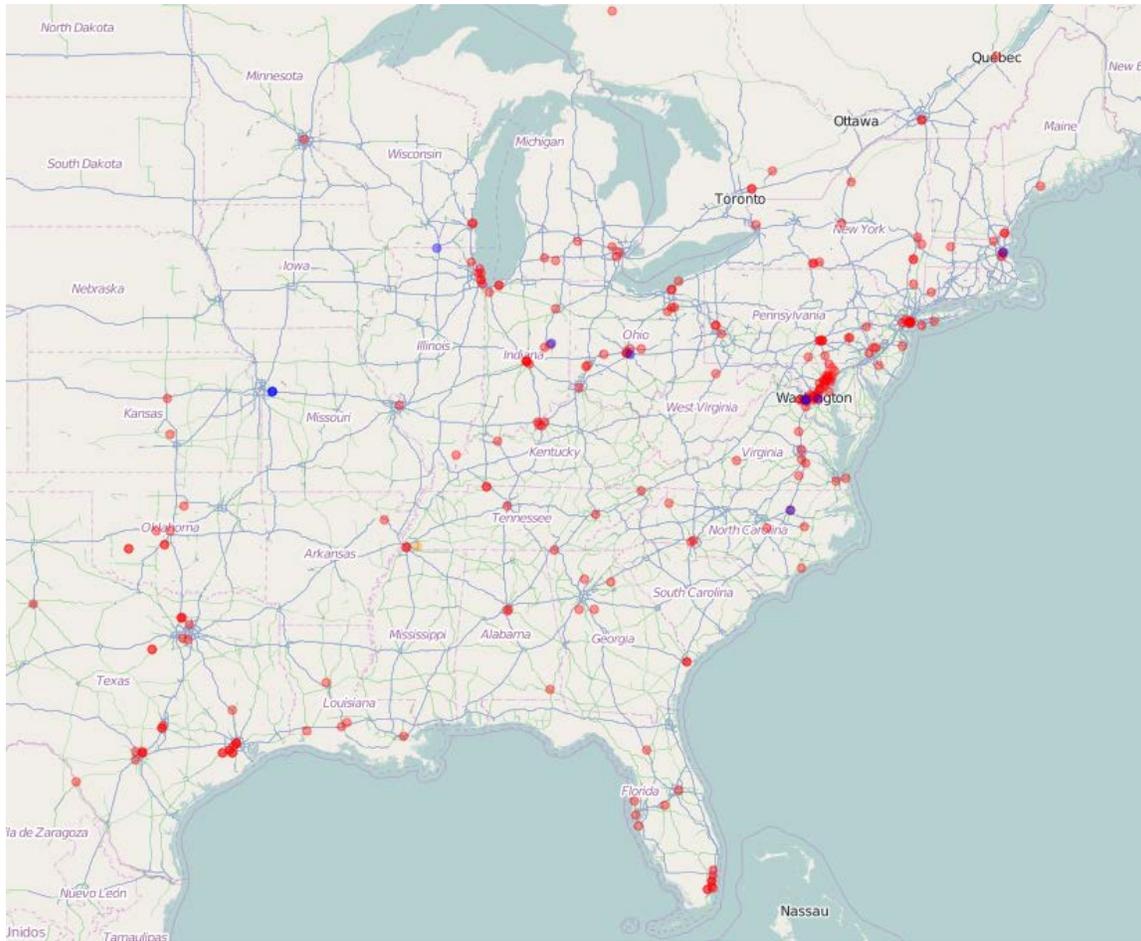


Figure 9: Tweets Related to Civil Unrest on 4/27/2015

REFERENCES

- [1] <https://dev.twitter.com/streaming/overview/processing>
- [2] http://users.humboldt.edu/mstephens/hate/hate_map.html
- [3] http://en.wikipedia.org/wiki/2015_Baltimore_riots

Evaluating Potential Correlation Between Water Deficit and Mountain Pine Beetle Infestation in Whitebark Pine

STEPHEN THOMA

University of Colorado, Boulder
stephen.thoma@colorado.edu

Abstract

The lands of the greater Yellowstone area contain ever growing swatches of brown. Many of these dead zones are the result of the current outbreak of mountain pine beetles. One particularly hard hit species of pine is the White Bark Pine: 90% of its population in the greater Yellowstone area has been killed in the last 20 years. Preserving whitebark pine is vital as they are an ecologically important keystone species that prevents erosion by providing structure for snowpack and soil at treeline. Pines that are facing a lower relative water deficit have a higher chance of surviving mountain pine beetle attacks. This project investigated tools that land managers could use to identify areas to plant whitebark pine that offer the highest chances of survival. These areas were identified by considering changes in relative deficit, and other microclimatic conditions including precipitation and solar gain as influenced by local elevation, aspect and day length. I found that relative water deficit is likely not a factor influencing the mortality of white bark pine in Glacier National Park as very few areas within the trees' habitat experience significant amounts of water deficit.

I. INTRODUCTION

White bark pine is an ecologically important keystone species: the seeds of their cones provide vital nutrients to many animals living at high altitudes including Clark's nutcrackers – a small bird that plays a key role in spreading the seeds of pine trees. Additionally, these trees create a stable environment for other trees and plants to grow by consolidating soil and preventing erosion. However, changes in climate are causing these trees to be killed in startling quantities (Keane & Arno, 1993). Currently, whitebark pine are a candidate species for federal listing under the Endangered Species Act. Their habitat is rapidly shrinking, and they're being faced with an onslaught of mountain pine beetles that conservation managers have limited resources to combat.

Because of this, it is important to identify

locations where replanting has a very high chance of long term survivability for the trees. Recent research of the factors that influence a tree's ability to survive beetle attacks has found that trees that face less drought stress are more capable of fending off beetle attacks. This may be because trees facing a shortage of water are less capable of producing protective chemicals for resin production which are critical to mounting an effective biological defense (Arango-Velez et al., 2011). To aid land managers in their efforts to help this species persist, I have calculated the point by point relative water deficit of a section of Glacier National Park in order to locate these areas.

II. METHODS

I. Area of Study

I chose to perform my analysis on a section of Glacier National Park. Glacier National Park is a large area in North Western Montana. This park has a large amount of alpine terrain with its highest peaks reaching over 3000 meters in elevation. These high altitude areas are conducive to growth of whitebark pine – and groves of the trees can be found throughout the park. The park’s climate is highly variable, but USGS researchers have noted a warming trend in recent years. Average annual rainfall ranges from 23 inches in the driest areas of the park, to 30 inches in the most wet. Until recently, forest fires within the park were suppressed, which has contributed to the trend of whitebark pine being replaced by subalpine pines.

My research focused on a roughly 30 by 50 kilometer rectangle contained in the park’s boundaries containing an acceptable representation of the park’s varied climates. A subsection of the park was used instead of the park in its entirety in order to decrease the processing requirements of running the model (discussed in greater depth in the limitations section).



Figure 1: *The DEM of Glacier National Park with the area of study overlaid in green.*

II. Data

A digital elevation model (DEM) was created from mosaicked USGS 10-meter DEM data, clipped to the boundaries of the national park. This model was then upscaled to a 1 kilometer resolution. The DEM was necessary for the calculation of slope and aspect values. Elevations from the model were also input into Cli-

mateWNA (Wang, Hamann, & Murdock, 2012) for the calculation of small scale temperature and precipitation estimates.

The water model is capable of incorporating soil data: specifically the water holding capacity of the top 150 cm of soil. But, in order to further reduce the complexity of the modeling task, water holding capacity was assumed to be a constant 100 mm throughout the area of study.

Due to the hardware constraints involved in collecting statistics for localized weather, climate data is not available at high resolutions. PRISM monthly climate data is available in 2.5 by 2.5 arc minute sections. To increase the resolution of this, data the program ClimateWNA was used to extract and downscale the PRISM climate data. ClimateWNA was then used to calculate monthly climate variables for specific locations in the area of study at centroids of a 1 kilometer grid.

Monthly solar radiation values were calculated for points on the same 1 kilometer grid using ArcMap’s solar gain tool with its default values.

III. Analysis

I calculated the water balance of my area of study with a modified implementation of Thornthwaite’s water balance model (Thornthwaite, 1948) (Lutz, van Wagendonk, & Franklin, 2010), which analyzes the allocation of water among the components of a system. The model begins by calculating the melt factor F_m

$$\begin{aligned} T_a \leq 0^\circ\text{C} : F_m &= 0 \\ 0^\circ\text{C} < T_a < 6^\circ\text{C} : F_m &= 0.167 * T_a \\ T_a \geq 6^{\text{circ}}\text{C} : F_m &= 1 \end{aligned}$$

with T_a being defined as the mean monthly temperature in degrees celsius. The melt factor represents the rate at which a given cell will lose snow, and is used to determine the fraction of snow storage that will melt in a month. From here I calculated estimates for rain and snowfall amounts using the known monthly precipitation values.

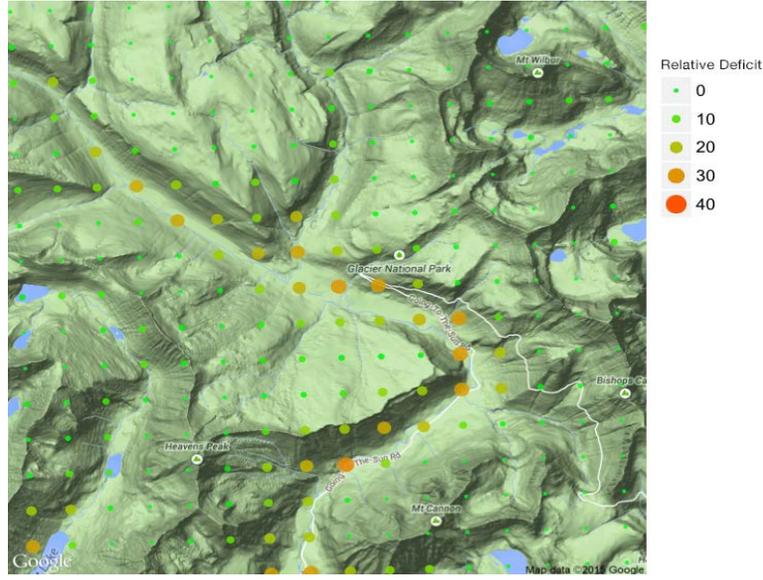


Figure 2: Mapping of the computed grid's relative deficit values.

$$\begin{aligned} RAIN_m &= F_m * P_m \\ SNOW_m &= (1 - F_m) * P_m \end{aligned}$$

Snowpack $PACK_m$, and melt water $MELT_m$ input was calculated beginning in August with a snowpack of 0, then calculated for the following 29 months in order to generate a base snowpack to work from. The last 12 months calculated were used for the final result.

$$\begin{aligned} PACK_m &= (1 - F_m)^2 * P_m + (1 - F_m) * PACK_{m-1} \\ MELT_m &= F_m(SNOW_m + PACK_{m-1}) \end{aligned}$$

Snowpack and melt water values are used to determine the amount of water available to a centroid, as well as how much water it is expected to lose for any given month. Next, the calculated solar gain data was used to generate potential evapotranspiration (PET).

$$PET_m = 29.8 * days * dayLength * \frac{VSP(T_a)}{T_a + 273.2}$$

In this water model, monthly PET is calculated using the Hammon equation (Hamon, 1963). The $dayLength$ value used in this calculation is determined using the Gavin method (Forsythe, Rykiel, Stahl, Wu, & Schoolfield, 1995), which

uses a latitude and day of the year to calculate a day length accurate to 1 minute. VSP is the vapor saturation pressure at a given T_a . Finally, with the monthly PET values, the soil water balance can be calculated. Similar to the snowpack calculation, the soil water balance is run for 30 months, starting with fully saturated soil in May, and only the last 12 months of the simulation are used. The soil water balance is necessary to determine the amount of actual evapotranspiration (AET) occurring.

$$\begin{aligned} WAT_m \geq PET_m : SOIL_m &= WAT_m - PET_m + SOIL_{m-1} \\ WAT_m < PET_m : SOIL_m &= SOIL_{m-1} - FRAC_m \end{aligned}$$

$$SOIL_m > SOIL_{max} : LOSS_m = SOIL_{m-1} * (1 - \exp(-\frac{PET_m - WAT_m}{SOIL_{max}}))$$

$SOIL_{max}$ was set to a constant 100 mm. Monthly change in soil moisture is a given month minus the following month.

$$SOILCH_m = SOIL_m - SOIL_{m+1}$$

A revised PET_m can be calculated such that:

$$PET_{mod_m} > WAT_m \wedge T_a > 5 : PET_m = WAT_m + SOILCH_m$$

The AET is chosen $AET = \text{minimum}(PET_{mod}, PET)$. From here, the monthly water deficit can be easily calculated as

$$DEFICIT_m = PET_m - AET_m$$

AET and Deficit values were written to a CSV file containing monthly values for each, as well as the UTM coordinates associated with the values. Using R's ggplot2 and ggmap (Kahle & Wickham, 2013) libraries, I was able to further analyze the water balance model's results.

III. RESULTS

I found a strong inverse correlation between elevation and deficit (see III). In the area I examined, my analysis found no points with a deficit greater than 50 mm. Further, there were no points with a deficit greater than 10 mm above an elevation of 2250 m. Therefore, based on this analysis it seems likely that the loss of whitebark pine in Glacier National Park is not strongly driven by mountain pine beetles, and may instead be the result of other factors such as white pine blister rust (WPBR).

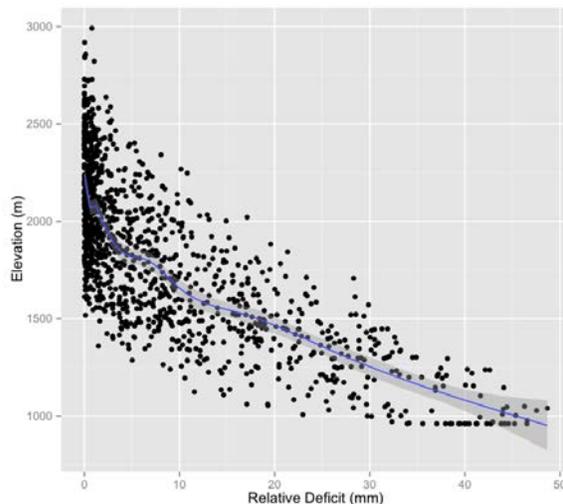


Figure 3: Comparison of elevation to relative water deficit.

IV. DISCUSSION

Although the results of this analysis do not support the hypothesis that water deficit is a driving factor in the decline of Glacier National Park's whitebark pine, the map of water deficit may still be a useful tool for managers,

for example, in conjunction with fire management. Actual evapotranspiration and relative deficit data is not generally available at this resolution, and can provide additional information to park managers.

Typically dry weather is unfavorable for the spread of WPBR: it can spread most effectively in cool, humid air (Maloy, 1997). For the most part, Glacier National Park's climate fits this description. The park receives, on average, 30 in of precipitation, and is cold for a large amount of the year. These factors combined with my calculated deficit values strongly support the conclusion that a majority of the dying white bark pine in Glacier National Park are being affected by WPBR, not mountain pine beetles. If correct, this information could inform management efforts by selecting for trees that are more resistant to WPBR.

The deficit values were calculated using 30 year norm climate data which shows a clear long term warming trend. However, because of this long time span, it is not possible to account for important annual variations in climate. These shorter term fluctuations can have a large impact on vegetation: plants adapted to cool and wet conditions will experience stress on warmer years. The calculated deficit values show a significant increase during the warmer summer months, signifying that a warmer year would have a noticeable impact on deficit values.

V. LIMITATIONS

This research was conducted over the span of a month and a half. This time constraint made it necessary to limit the research's scope. The water holding capacity of the soil a tree is growing in is a significant factor in how long the tree can sustain itself without water input. My calculations assumed the water holding capacity of the soils in the area of study were constant, which would have resulted in soils with a below ideal water holding capacity to be inaccurately favored. The resolution of my calculations was also not ideal. Similar research was performed using 90 m centroids – a reso-

lution more than 10 times more accurate than mine. But, running the model with millions more points would have taken too much time per run to be feasible for me to use. I would have also liked to calculate the heat load in R based off of the slope and aspect derived from the DEM, instead of relying on ArcMap's solar gain tool.

VI. CONCLUSION

The climate is noticeably warming: higher elevations are experiencing increasingly warmer temperatures. These changes are exposing whitebark pine to increasing numbers of threats. It is vital that these threats be iden-

tified and addressed early on, or these trees will likely become extirpated. As it stands, the species is in danger of extinction in two to three generations (Fish & Service, 2011). While I was unable to confirm my hypothesis, the deficit information resulting from my calculations provides interesting insights into Glacier National Park's ecosystems that deserve further exploration. A higher resolution analysis that incorporates soil water holding information would be a necessary next step towards gathering further insights. Water balance is also a powerful tool for examining vegetation distributions. Future research could incorporate existing vegetation mappings in order to provide a baseline for drawing conclusions.

REFERENCES

- Arango-Velez, A., Meents, M., Linsky, J., El Kayal, W., Adams, E., Galindo, L., & Cooke, J. (2011). Influence of water deficit on the induced and constitutive responses of pines to infection by mountain pine beetle fungal associates. *BMC Proceedings*, 5(7), 29.
- Fish, U., & Service, W. (2011). *Whitebark pine to be designated a candidate for endangered species protection*. Retrieved from <http://www.fws.gov/mountain-prairie/species/plants/whitebarkpine/PressRelease07182011.pdf>
- Forsythe, W. C., Rykiel, E. J., Stahl, R. S., Wu, H.-i., & Schoolfield, R. M. (1995). A model comparison for daylength as a function of latitude and day of year. *Ecological Modelling*, 80(1), 87–95.
- Hamon, W. R. (1963). *Computation of direct runoff amounts from storm rainfall*. publisher not identified.
- Kahle, D., & Wickham, H. (2013). ggmap: Spatial visualization with ggplot2. *The R Journal*, 5(1), 144–161. Retrieved from <http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf>
- Keane, R. E., & Arno, S. F. (1993). Rapid decline of whitebark pine in western montana: evidence from 20-year remeasurements. *Western Journal of Applied Forestry*, 8(2), 44–47.
- Lutz, J. A., van Wagendonk, J. W., & Franklin, J. F. (2010). Climatic water deficit, tree species ranges, and climate change in yosemite national park. *Journal of Biogeography*, 37(5), 936–950.
- Maloy, O. C. (1997). White pine blister rust control in north america: a case history. *Annual Review of Phytopathology*, 35(1), 87–109.
- Thornthwaite, C. W. (1948). An approach toward a rational classification of climate. *Geographical Review*, 38(1), pp. 55-94. Retrieved from <http://www.jstor.org/stable/210739>
- Wang, T., Hamann, D., A. Spittlehouse, & Murdock, T. (2012). Climatewna - high-resolution spatial climate data for western north america. *Journal of Applied Meteorology and Climatology*, 51, 16–29.