Spatial Data Mining for Disease Mapping & Analysis: Flu Spread in University of Arizona 2014

-Amit Juneja

Contents

1) Introduction

About the Project

This paper is intended as an experiment in data mining process of the social network. The intention is to mine useful data related to spread of a particular disease for the purpose of further analysis and mapping. While the authenticity of the data cannot be 100% guaranteed, a lot of validations and cross-checking is performed to keep the data as accurate as possible. Additionally, mining data opens up a lot of opportunities to classify the data in to meaningful features. For the simplicity of this project all valid and meaningful data is assumed to be associated with the spread of flu, and each symptom is not classified into individual features. Data mining is a vast field, and almost all social networks can be mined with the help of correct resources. For the purpose of this study, the social network data mining is restricted to just twitter.

Also, the data is limited in it's capacity, hence the results are meaningful but not the best estimation of disease spread.

Objective

The objective of this study is to analyze and map the spread of flu for University of Arizona students in the last 3-4 months based on the data mined from social networking website, Twitter. The study is limited to University of Arizona, and about 2-3 miles from each direction from the center of the university to correctly account for all the students. Due to the limited time frame for mining data, the study does not include regions beyond the 2-3 mile radius, and hence any data from beyond those points is not considered while mapping flu spread.

For the purpose of analyzing data, a point pattern analysis followed by K-function test is carried out to confirm the spatial pattern of the mapped data. For risk estimation of the disease, ordinary kriging is carried out towards the end.

While the data is not classified into features, each useful data is given a weight based on many factors to estimate the intensity of disease for the individual in that particular area,.

Motivation

Motivation for this study comes from the growing popularity of social networking websites among individuals of all age. Kids as young as 12, and adults as old as 80 use popular social networking websites these days, thus contributing to the vast amount of data generated by these websites each day.

The most attractive feature about the use of social networking websites is keeping your friends or people in your network updated about the events in your life. Most people share the information about their well being on websites like Facebook, Twitter. This study takes advantage of this fact to mine data related to flu in the last 3 months. Given the fact that the target audience for this study is students, it is easier to gather data due to the popularity of social networking websites among students.

2) Methodology

Mining Data from Twitter

Twitter makes it easy for it's developers to extract data via various Application Programming Interfaces (APIs) written in many different programming languages. For this study, the *Tweepy* api written in **Python** was used. The API provides functionality to search for keywords, add geographic filters, like specific coordinates or city, and also gives the option of filtering results by date. There are many ways in which these APIs can be further used to search individual users, their followers and collect their tweets.

To gather data related to flu , it was important to look for the right keywords. There are many different ways in which an individual can share their illness on Twitter. They can use certain hash tags, phrases or can just be direct.

The following keywords were used while searching for relevant data:

1. **#fever**
2. **#sick**
3. **#flu**
4. **sick**
5. **flu**
6. **fever**
7. **cold**
8. **Nyquil**

9. **Tylenol**
10. **ill**
11. **under the weather**

After gathering results, the most challenging part is to clean the data and filter out noise. Most of the times, the above-mentioned keywords are used in completely different context, thus generating a lot of useless data. A specific branch of computer science, Natural Language Processing, deals with situations like this. For this study, the data was manually read and filtered. Since the data being analyzed was limited in its capacity, it made the manual filtering possible but definitely not easy.
Data from about **2,250** users was mined, and was limited to **200** tweets per user. This resulted in about **450,000** tweets in total.
These tweets were then filtered based upon the above given keywords. The filtering usually resulted in about **1000-2000** matches, and then these tweets were mined for useful data.

There were a lot of ways in which the useful data was sometimes discarded. For instance, if the data was older than September 2014 it was not considered.
If the data was useful but lacked coordinates, it was discarded because it becomes hard to map data without actual coordinates.
If the data was useful and had coordinates, but was not from the decided range (2-3mile radius) it was discarded to avoid any outliers and keep the data normalized for efficient mapping.

Lastly, if the data was not from a university student, it was discarded as the study specifically targets the students of University of Arizona.

As mentioned earlier, while the data is thoroughly checked for its authenticity, it cannot be 100% verified. Luckily most useful data was straightforward and not ambiguous, but when mining data from the social network, one cannot ignore the magnitude of informality in data. Also, almost all the data was user's own opinion about their health, and in some cases it may just be false alarm for a temporary infection. There is no way to account for false negatives and false positives in this study, and hence it is assumed that the data found is accurate for analysis purposes.

Challenges

While Twitter is friendly with data extraction from their database, they impose a lot of limits on the APIs to avoid abuse of their resources.

Each API has rate limits, which are different based upon what the API is used for. For searching for specific keywords and obtaining tweets for just those keywords, the rate limit is **150 requests** per **15-minute** window. The API will start throwing errors after 150 requests are exceeded or after 15minutes end from the time of its use (whichever happens first). The API will then be unavailable for close to 15 minutes. If you are searching for a specific user's tweets, the API poses a rate limit of **180 requests** per 15minute window. Approximately **3,200 tweets** can be mined per user without exceeding the rate limit.

To overcome these issues, multiple twitter accounts were created, and a python script was run as a daemon in the background to make sure a change of account happened after the rate limit exceeded. If meanwhile the program is put to sleep (using sleep functionality available in time module in Python), the searching does not restart and continues from the exact place it was left off at.

Each account is authenticated using **OAuth,** and is given a specific access token and consumer key along with access token password and consumer secret token. To access the API each account needs to create an app , after which the **OAuth** credentials are granted

Another problem with the Twitter API is that it does not return results older than 7 days. At the most, it can return old results that were popular. Since this study

requires analyzing data at least upto 3 months old, a different technique was adopted.

Twitter API allows scanning up to **3,200** tweets per user, and this was used as an advantage to scan each user who came up in the results for a specific coordinate specified in the filters, along with the radius. In this case the coordinates were of the center of university with radius kept to 3miles.

A blank search resulted in **18,246** tweets from users who had tweeted around those coordinates. These tweets were analyzed for duplicates user ids, and all the unique twitter ids were extracted. Multiple searches with random and popular were keywords were also carried out to extract as many users as possible. In the end the total count of unique users from the given bounding box was **2,652**.

Considering there are **40,000** students in the university, this sample is very less. A reason for such less number of users is

1) Not a lot of people have geotagging on for their twitter accounts, which is why they are never detected.
2) Not a lot of students are on Twitter, and prefer using other social networking websites.

Twitter APIs are infamous for not resulting all the tweets, and filtering out most tweets. Since the search was mainly focused on extracting twitter ids, only getting sampled tweets was not an issue because the APIs returned at most one tweet by each user, which is sufficient.

About **200** tweets were extracted from each user per request. This number was chosen to avoid data beyond the 3 month limit, and most users tweet about 200-300 times on an average in 2-3 months

Using search engine in a popular text-editor called **Sublime- Text**, these tweets were searched with specific keywords and data was extracted, cleaned and analyzed.

3) Analysis

Mapping mined data

All the clean data was mapped and created into shapefiles.

The maps below show data, which is up to 4 month old.

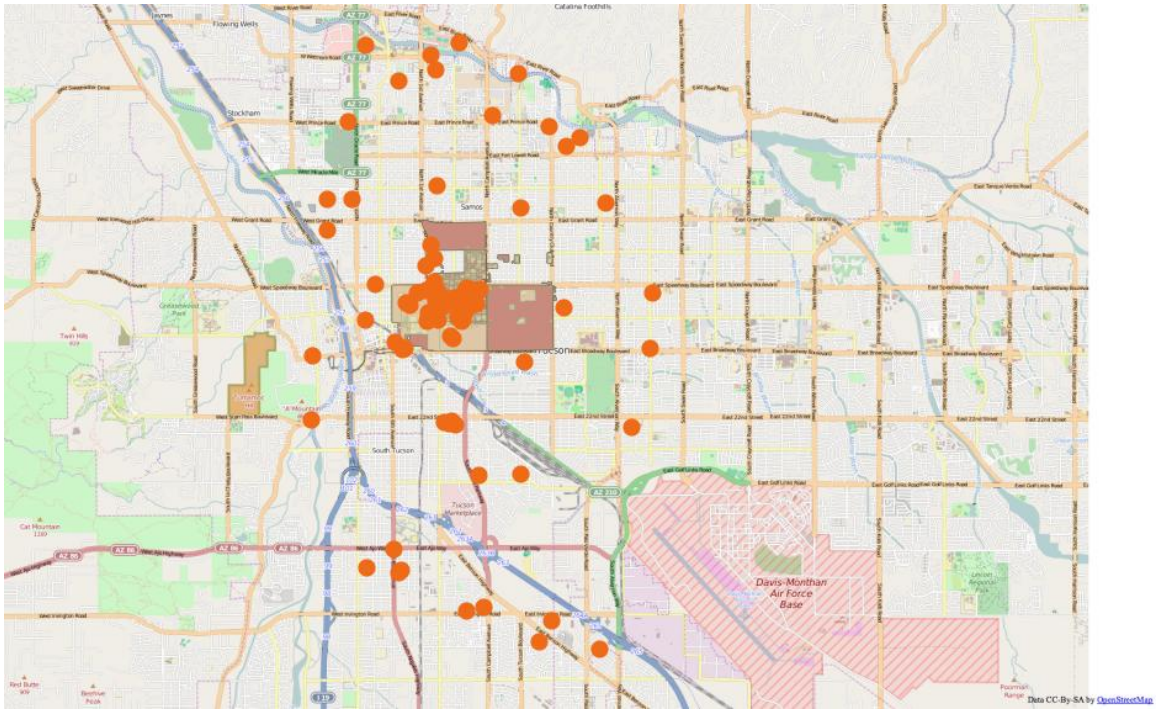

Figure 3.1: Data collected from tweets mined on map of Tucson. Brown polygon is the University area

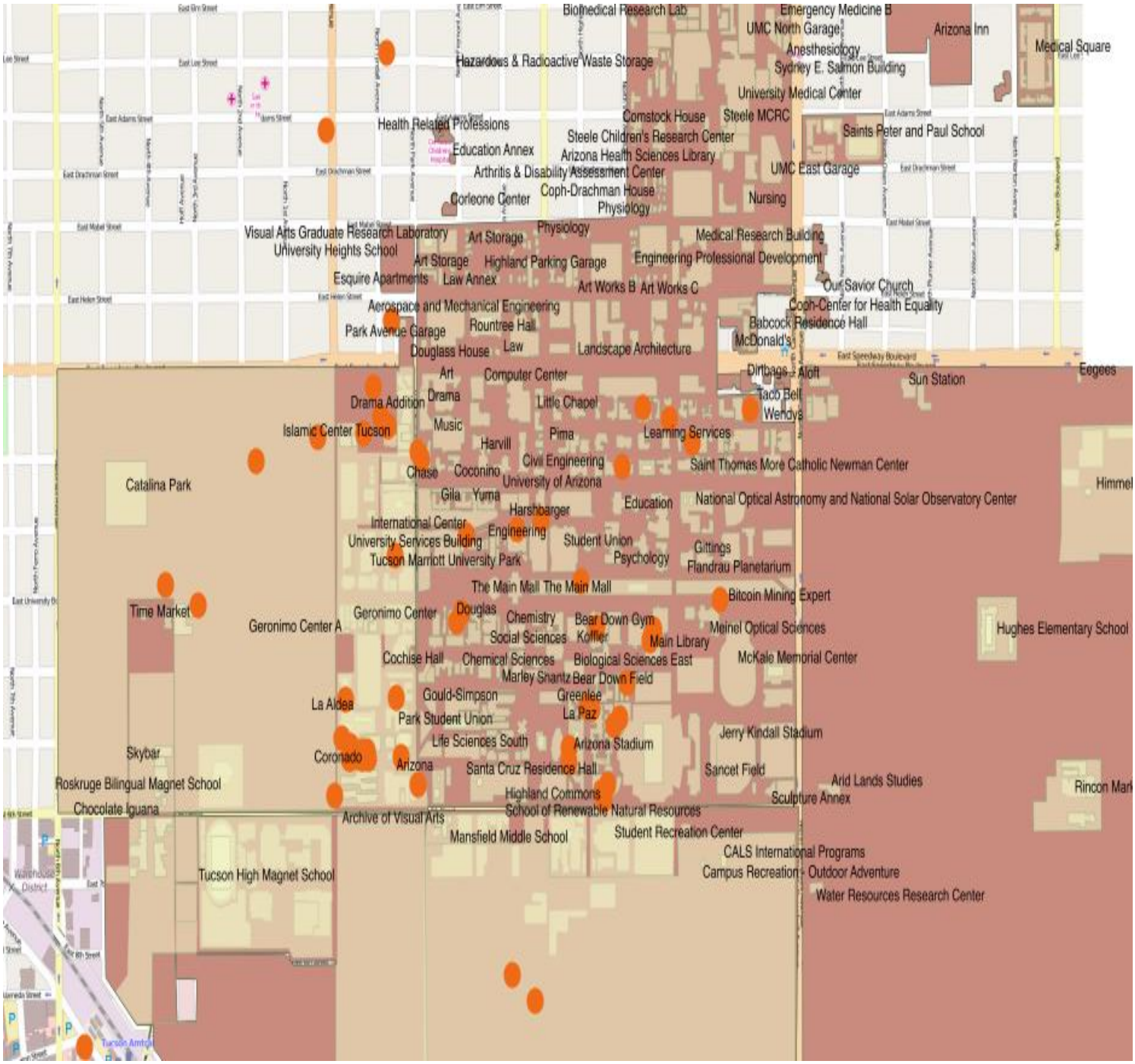

Figure 3.2: Map zoomed in closer to University area with labelled buildings

Figure 3.3: Map zoomed in on cluster newar the university area

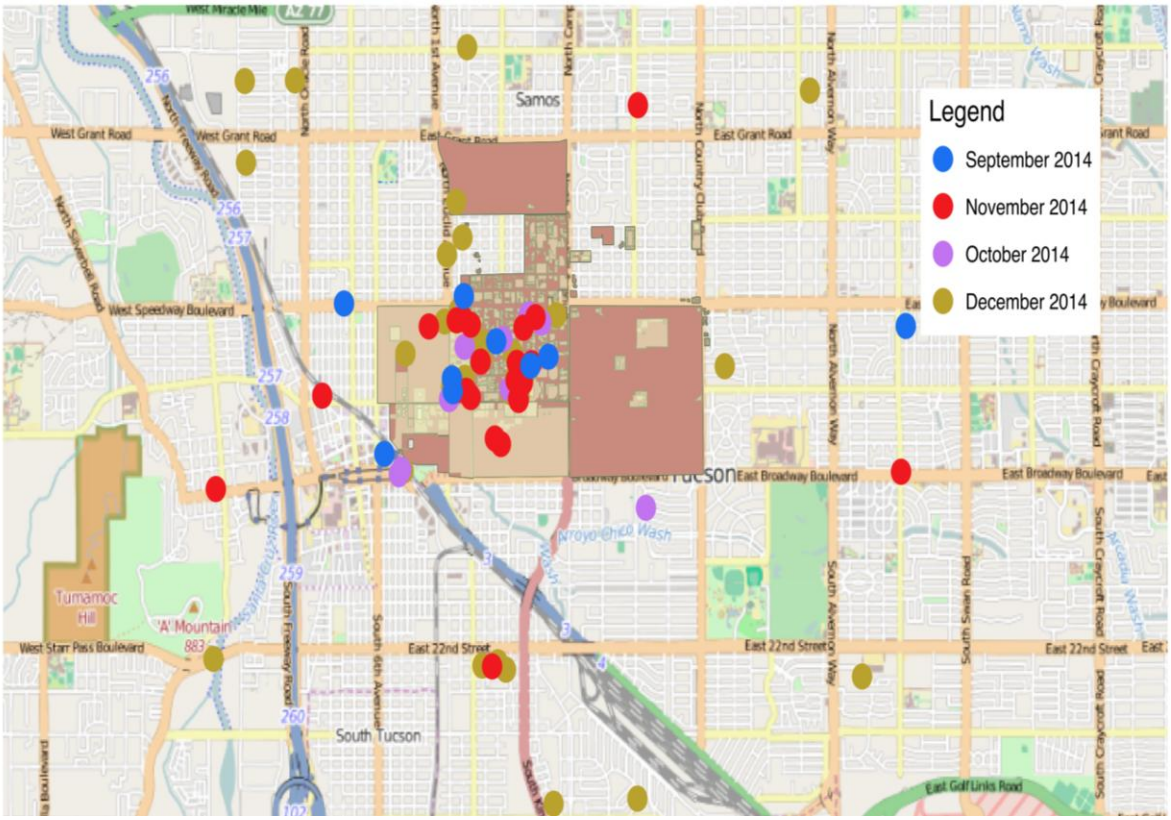

Figure 3.4: Map of data mined from tweets separated by creation date

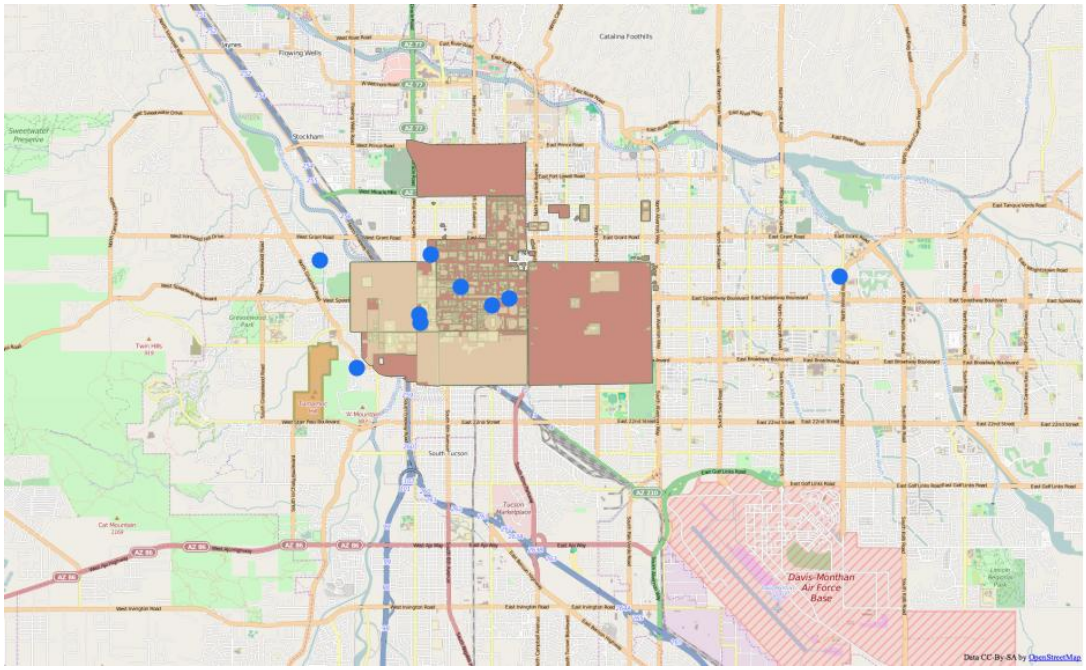

Figure 3.5: Data from September 2014

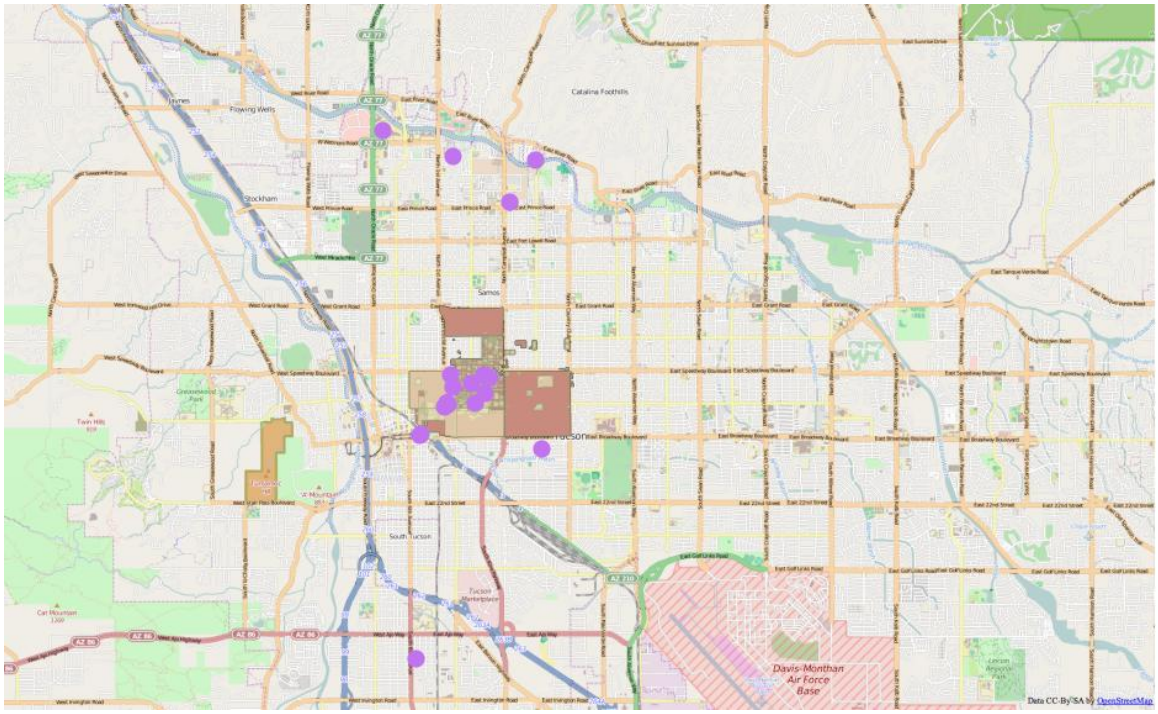

Figure 3.6: Data from October 2014

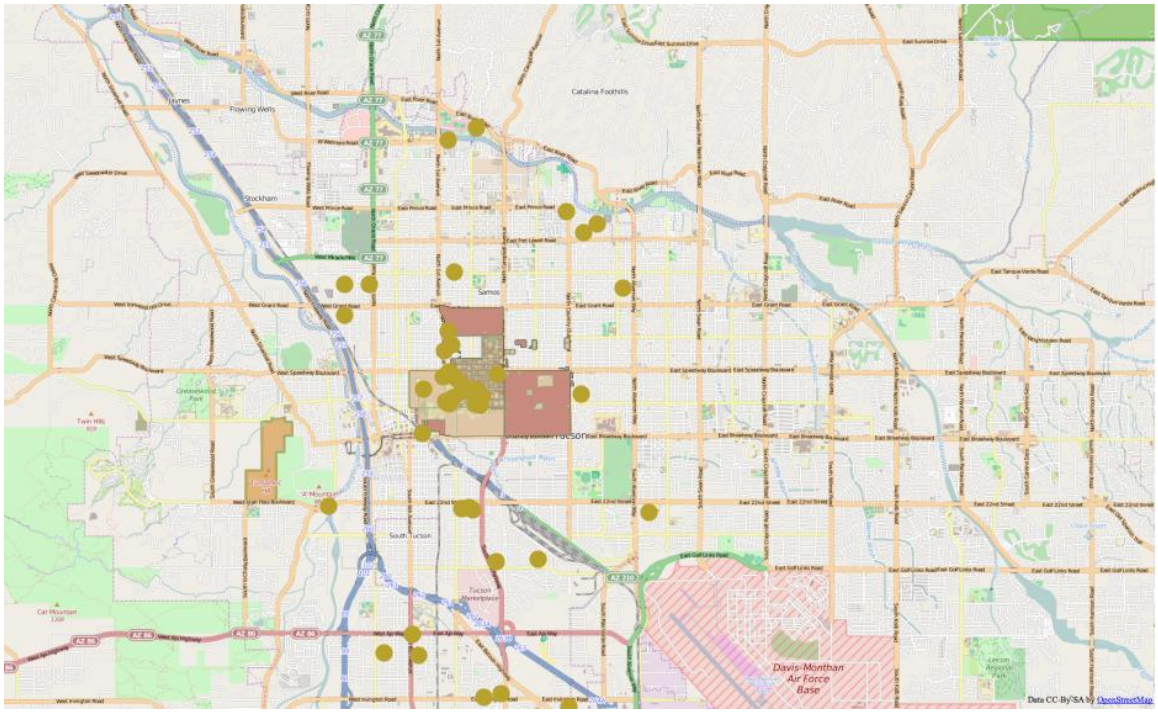

Figure 3.7: Data from November 2014

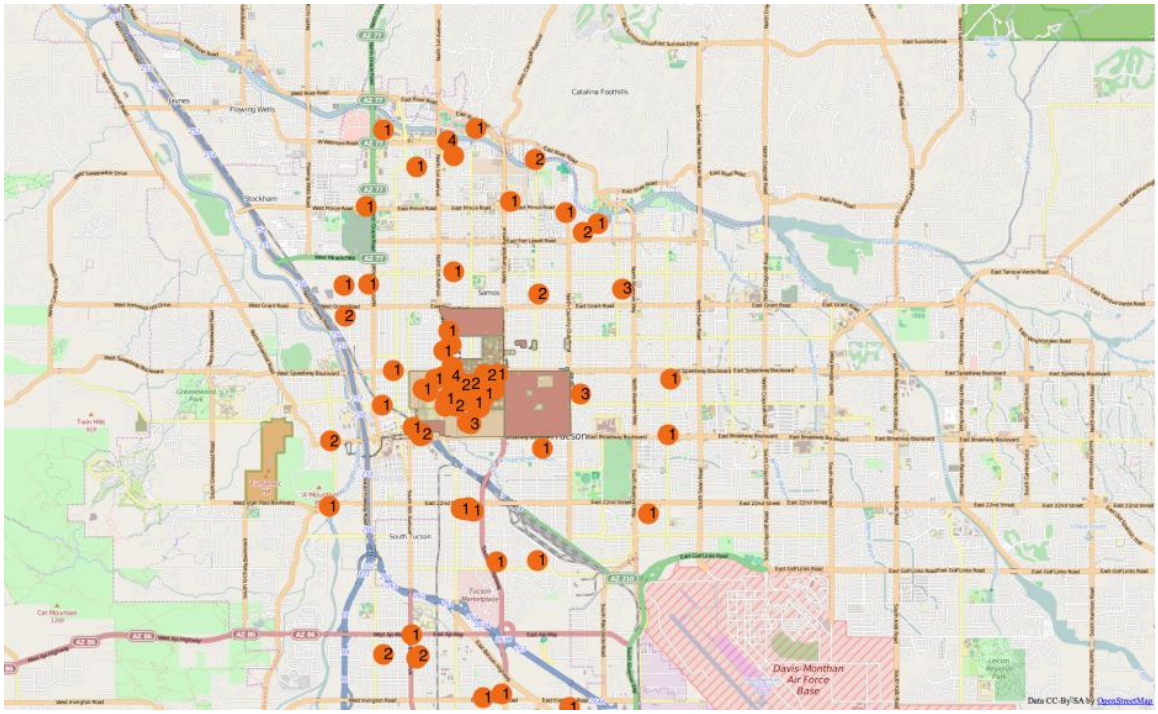

Figure 3.8: Data from December 2014 labeled with intensities

Point Pattern Analysis

The point pattern analysis of the data was carried out in R, and the results are pasted below. They will be discussed at length in the next section.

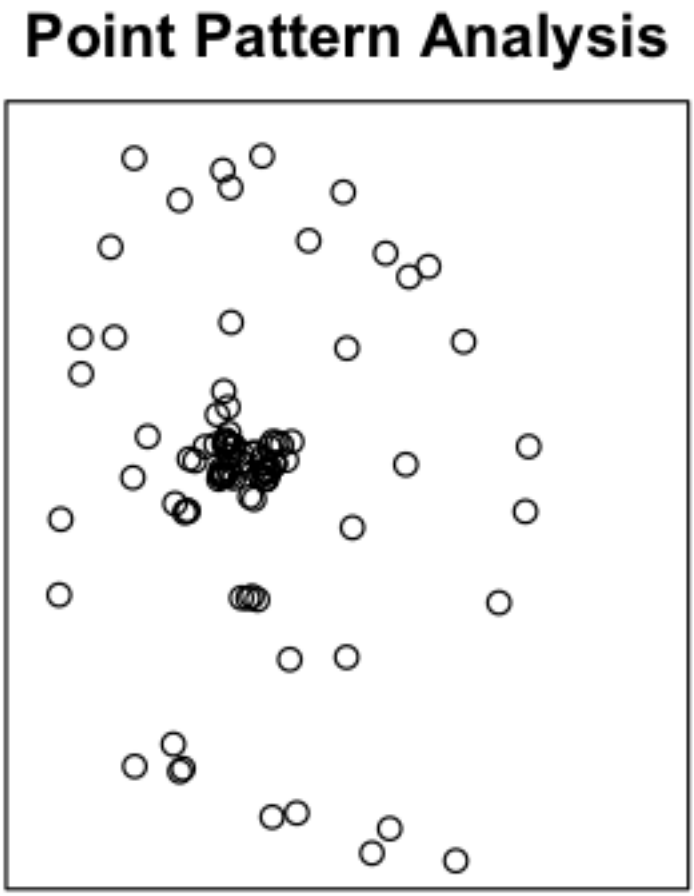

Figure 3.9: Point Pattern Analysis of the data

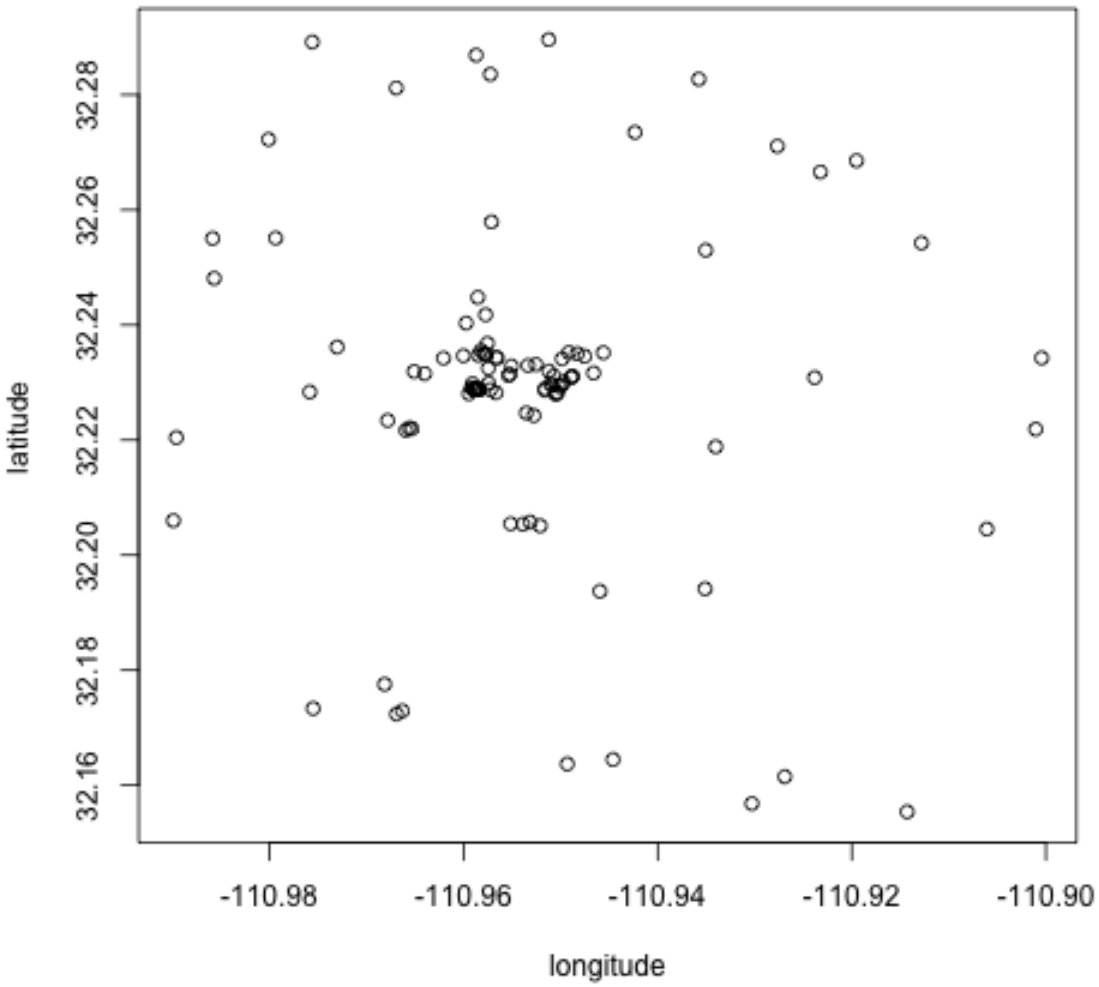

Figure 3.10: Point Pattern Analysis on a smaller range

K-Function Test

The point pattern analysis of the data was carried out in R, and the results are pasted below. They will be discussed at length in the next section.

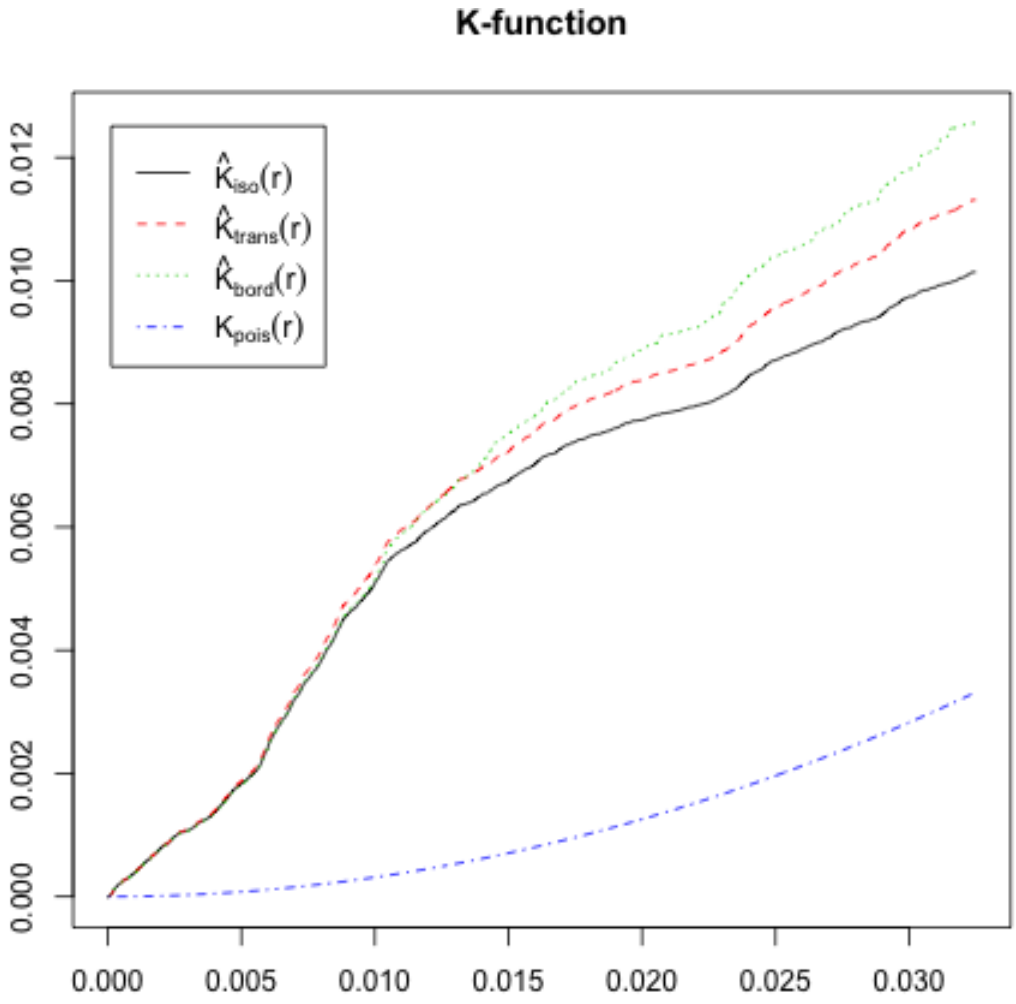

Figure 3.11: K-function test

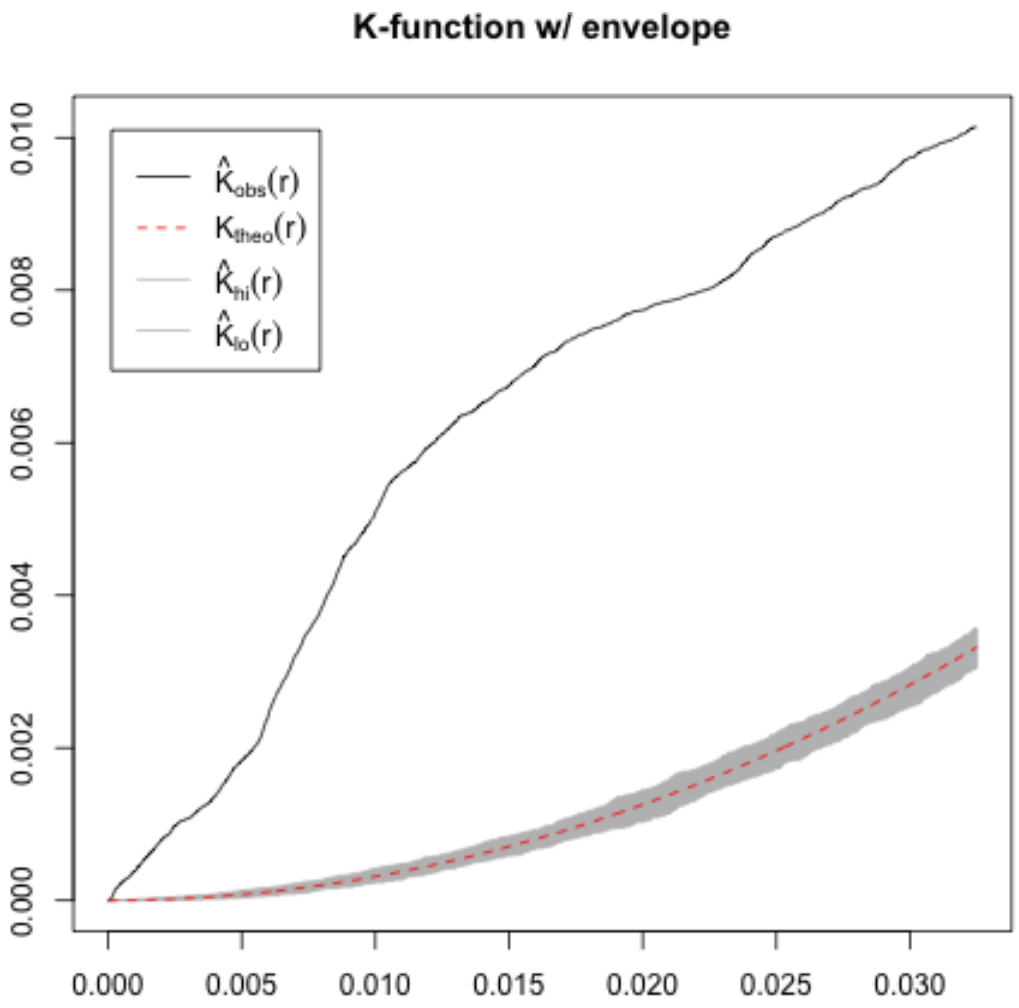

Figure 3.12: K-test with envelopes

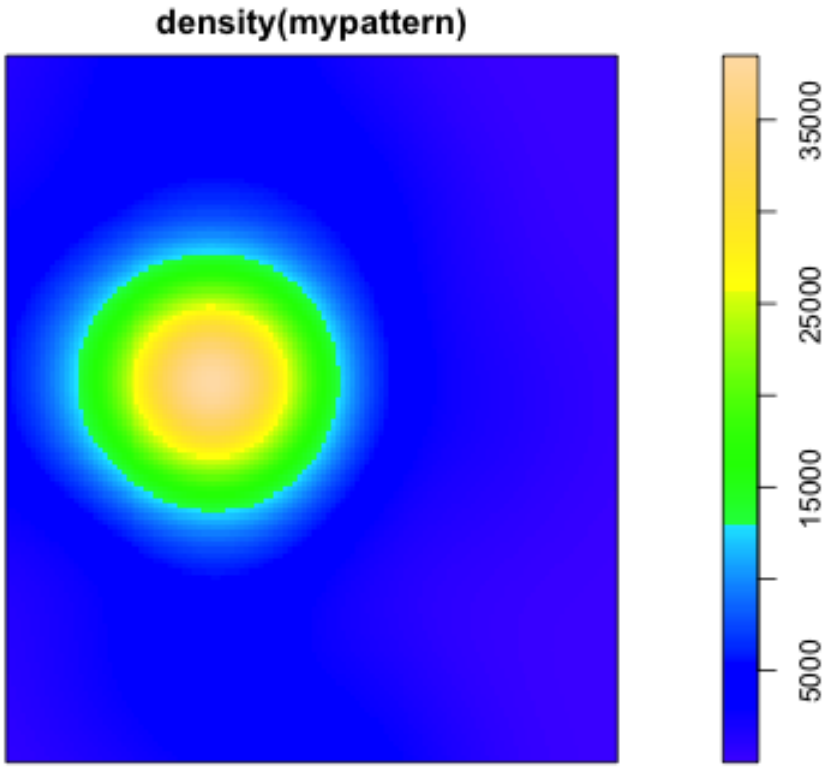

Figure 3.13: Density map of data

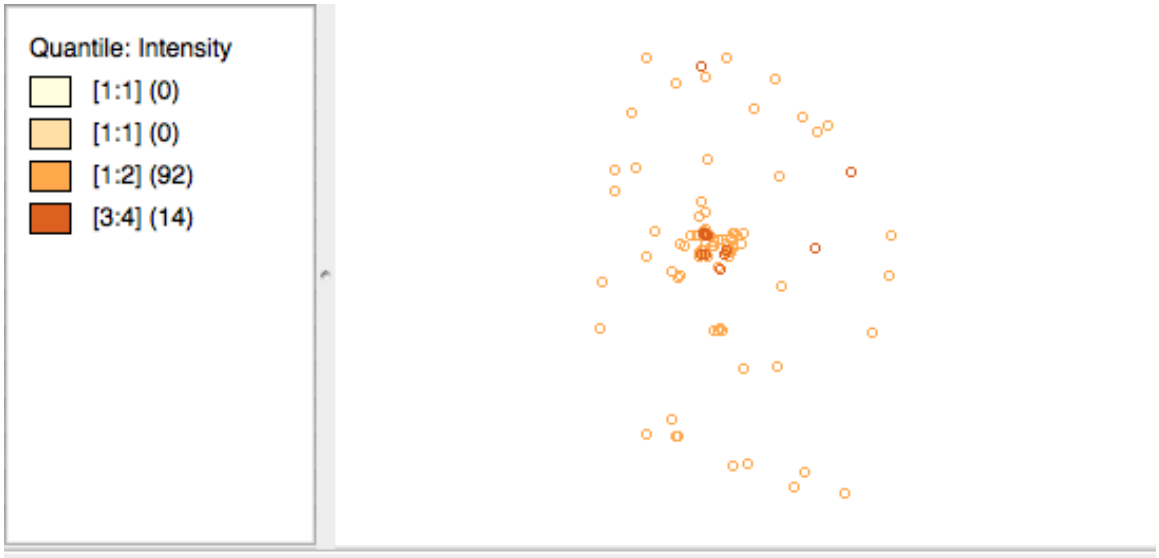

Figure 3.14: Quantile map based on intensity

Ordinary Kriging

While Baye's estimation was the first choice for risk estimation for flu spread, the acquired data lacks important features like population count per region. Hence, Ordinary Kriging is carried out for interpolation after regression.

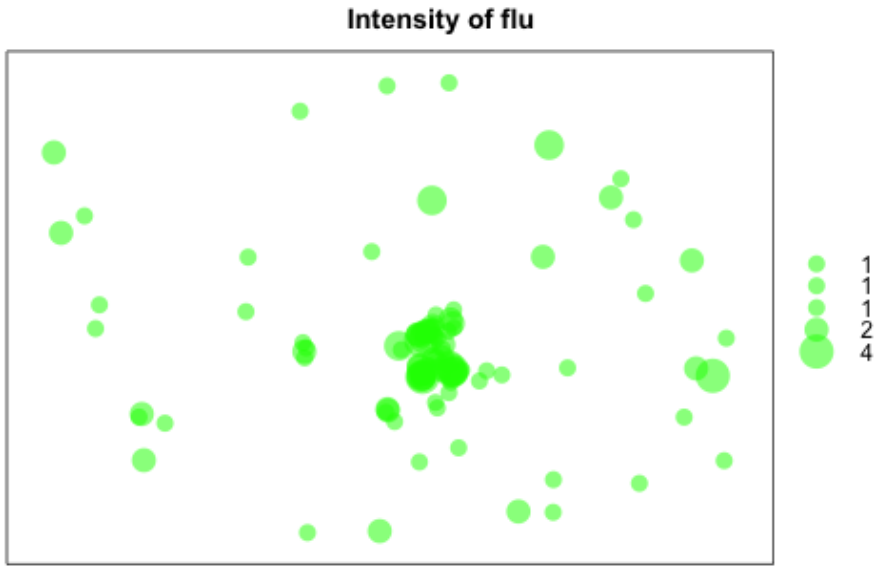

Figure 3.15: Bubble sort based on intensity

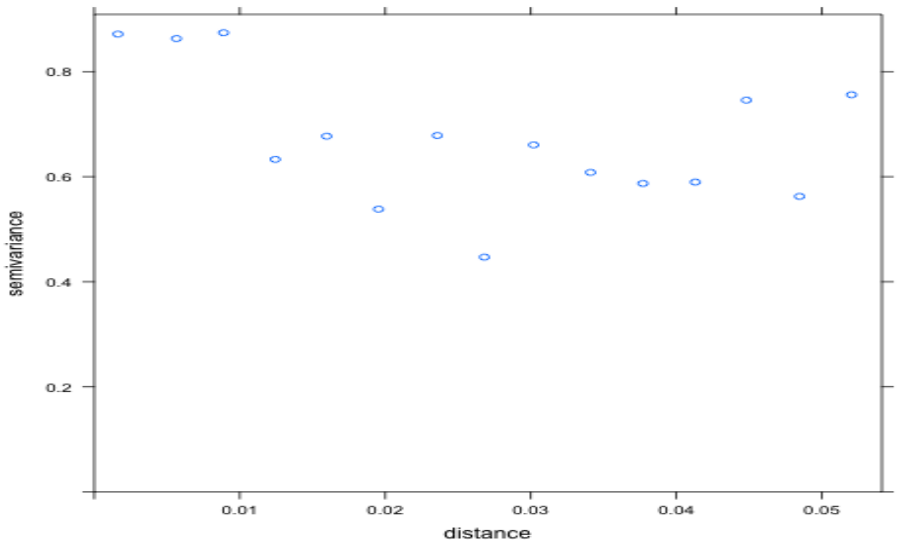

Figure 3.16: Variogram based on intensity

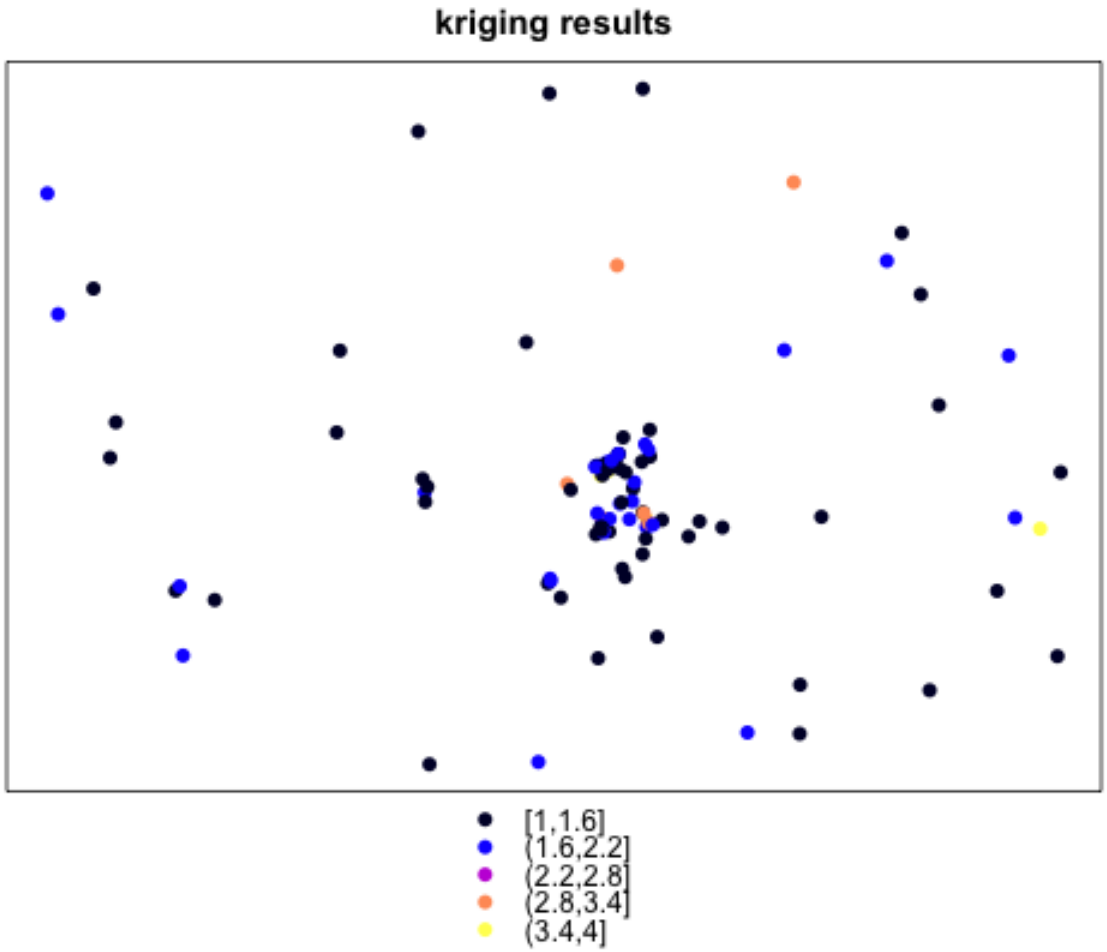

Figure 3.17: Oridinary Kriging results based on intensity

4) Conclusion

Conclusions based on Analysis and Mapping

From figure 3.1 it is clear that the points are more concentrated near the university area, and are fairly spread out as we move away from the university area. This is because most people tweeted from around the university, which also helped confirm their identity as students. Additionally most points were discarded if they were not confirmed to be students (after going through their tweet history). Zooming in on the university area shows that the points are actually much more spread out around the university.

One would expect more data to be around dorms, but most dorms did not generate much data. The major cluster near the university area (fig 3.3) comes from the area around Cornardo dorm. This could be a result of 3 dorms in that area, and the student population which frequents the area due to Park Student Union and volleyball courts.

Surprisingly not a lot of data came from highland avenue dorms, which is closer to Campus Health and has more dorms than the Cornardo area. Eitherway, just by looking at the maps it can be said that most sick students came from the Cornardo area.

Looking at the data time in Fig3.4 wise shows how the tweets increased progressively in a span of 3 months. This could be a result of weather changing from extremely hot in September to extremely cold in December. The September data is

not much reliable intensity wise as the tweets mostly came from October, November and December, which was also a result of taking only 200 tweets per user.
These results are analytically proved by the Point pattern analysis in figure 3.9 and 3.10. Fig 3.10 shows a zoomed in point pattern analyses. The overlapping points are around the cornardo dorm and neighboring areas. They look clustered due to shorter range on the plot.

K function estiamtes show perceived clustering around the university area as well. The desnity function is another proof.
The Quantile map is based on the intensity values. Intensity for a particular data is calculated based upon certain factors which involve – multiple problems like fever, cold, headache earn an intensity of 4, recurring problems earn an intensity of 3 unless they have been going on for a very long time. Problems, which cause a lot of discomfort like not being able to sleep at night because of fever, earn an intensity of 2 and mild to less troublesome problems are 1. These intensities have been carefully calculated after evaluating each and every data source from their previous tweets. The quantile maps shows larger densities around university area, and the intensities decrease as we move away from university. This can also be confirmed by the bubble sort which shows higher radius in the center (fig 3.15)

Estimation of risk from disease is analyzed by ordinary kriging, by essentially interpolating the data based on given intensities. The variogram plotted with the intensity variable is a singular matrix, which shows a very vague horizontal line

implying an almost negative correlation, which states that intensities are mostly random.

The kriging shows predicted intensities. It is easy to observe that most intensities are predicted to be between 1 and 2 , and are clustered around the university area. Rarely does the predicton go over 2, and most of it is not even near the university area. The highest intensity is very rare in the prediction map, and is away from the university.

Kriging shows that most intensities around the university area were 1 or 2, and hence there is chance of people getting mild to medium flu like symptoms around the university in coming months.

Once again, the data is credible but only to an extent, and these analysis and estimations are based on the same credibility.

Future Work

Future work for disease mapping by mining data from social network includes exploring more social networking websites like Facebook, Google+, etc.. and gathering as much data as possible for establishing credibility. Of course there are limitations on these social networking websites, so we also need to rely on advanced technology for hacks.

Disease mapping can be useful for a variety of reasons, but most importantly it is necessary for people to be aware of the disease-spread rate around them and their chances of catching the infection. Hence, it is important to enhance technology in disease mapping.