Team: Opiate Risk Prediction

*Second Project Status Report*

Introduction

Progress overall continues for our group. We are currently wrapping our data wrangling and creating our authoritative data. We have cut back on some features, decided on the appropriate size value for the hexagons, and focused on grouping some data appropriately to make the project overall more manageable. The entire group has been focusing on Machine Learning via work with the UCI repository to build our skill base to work with the final data, as well as connecting to outside data sources such as weather via API. Finally, we have been focused on ensuring that we have thought through the appropriate statistical methodology to make the results compelling and informative. The entire group is focused on the next few weeks in bringing the project together.

Details are provided below.

Data Wrangling and Exploratory Analysis

* Dave- data gathering, preparation and analysis using ArcGIS/other tools
* Wranga, Caroline and Sarah- data manipulation and exploratory analysis using Jupyter Notebook, PostGreSQL and other tools
* All focused on methodology and clarification of problem

A majority of our time, since our last report, has been spent on data wrangling. The two core datasets we have received are: (1) calls for service (2015 to June 2017), and (2) Naloxone/Naloxone Administrations from 2014 to 2016. Due to potential complexity of data sources, we have chosen to limit our data to 2016. We are still determining whether we will build our models and analysis over time by month, or by week. As we complete the data wrangling in the next week, we will decide whether 52 weeks of analysis is too granular, and we will limit the data to month.

Our project focus is primarily geospatial, and is focused on predicting activity/behavior for a general, ad-hoc location within a time range (as yet to be specified). We are also looking at data across things like shifts, or other time ranges, correlations related to days of the week, and potential locations where one might want to position treatment facilities. Data wrangling currently has involved four steps. Spatial accuracy is paramount, and was a key focus. A majority of the data was correctly geo-coded by lat/long coordinate, but there were approximately 100 rows for the Naloxone data, where the data was outside the city for 2016, and had to be corrected. There were approximately 300-400 CFS where the data was located outside the city, and was corrected where possible to be brought into the city. This involved using the City’s MAR geocoding service to identify potential addresses, and re-capture those potentially missing points.

After the datasets were brought into ArcGIS, the next step was determining the appropriate cell size for the hexagons. We tried three different forms, trying to find the level of hexagon that was not too granular, nor too broad. We settled on 500M squared hexagons. This size provides roughly 789 cells that span the length and width of DC. We do have hexagons on the city’s edge that extend past the borders to insure potential overflow, or capturing on-border issues with CFS or Naloxone administration.

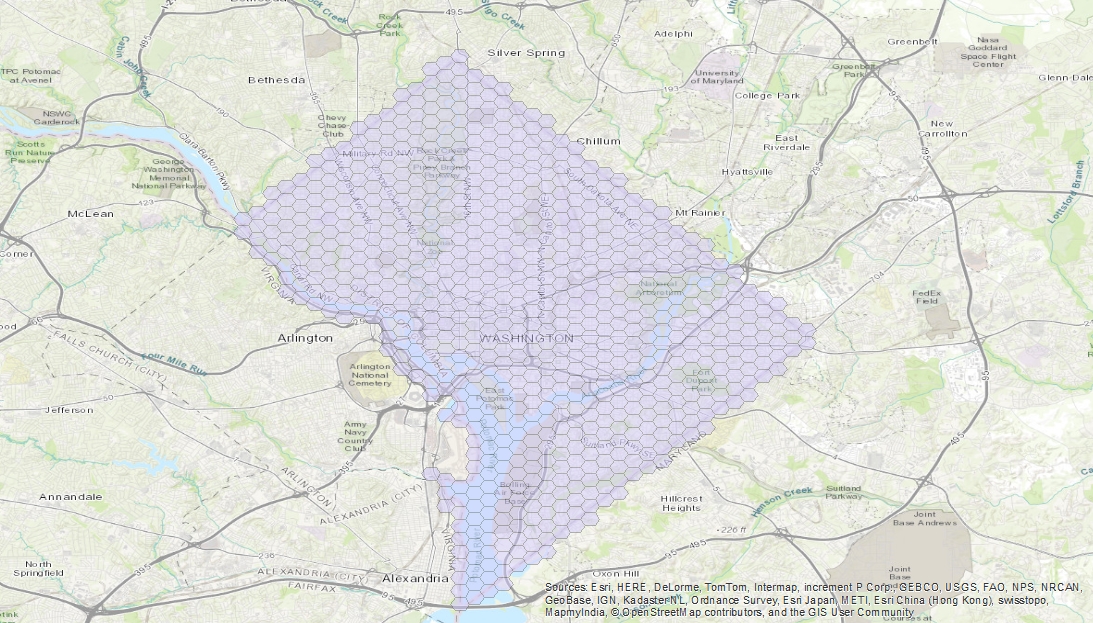
The datasets were geo-referenced and then joined to the 500M hexagon layer. Each CFS and Naloxone administration row is tied to one of the 789 hexagons. The data was joined initially, as polygon-to-point. Every point is tied to a hexagon value, which was used in our first view of summary statistics. CFS has 796,595 observations assigned to a hexagon. Administration of Naloxone has 3,030 observations tied to a hexagon.

The final stage has involved our group making some choices about how to condense the data into groups or categories that are meaningful. We have, for example, roughly 200 call types. It is not valuable to use every call types as a feature. Therefore, we are condensing the call types into groups. For example, the variety of accident types will be grouped into a single type called “accidents.” Disorderly conduct, which is one of the most common call types around a Naloxone administration, will likely be left as its own group. This simplification of the call types will help in limiting the features and making our approach overall more valuable.

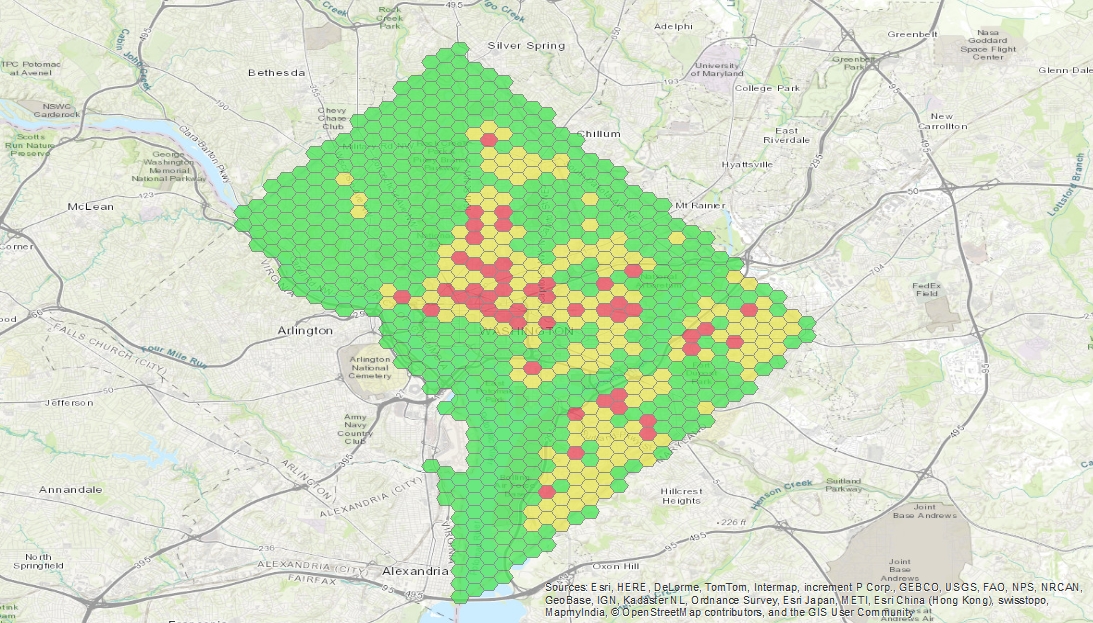
Following this grouping of the data, we will be creating a single authoritative file, by month, with as series of counts/scores per hexagon. We will be doing this for each hexagon, using the “switch selection option” in ArcGIS. In essence, we will build counts for each of our features (Gender, Incident Type, Naloxone, MPD/FEMS), separate them out by month for 2016, and then have those values for each 500M hexagon. That will be the authoritative data set, which we will then use for machine learning. Some hexagons will have minimal values, while others will have several values associated.

Some examples from the quick geospatial analysis:

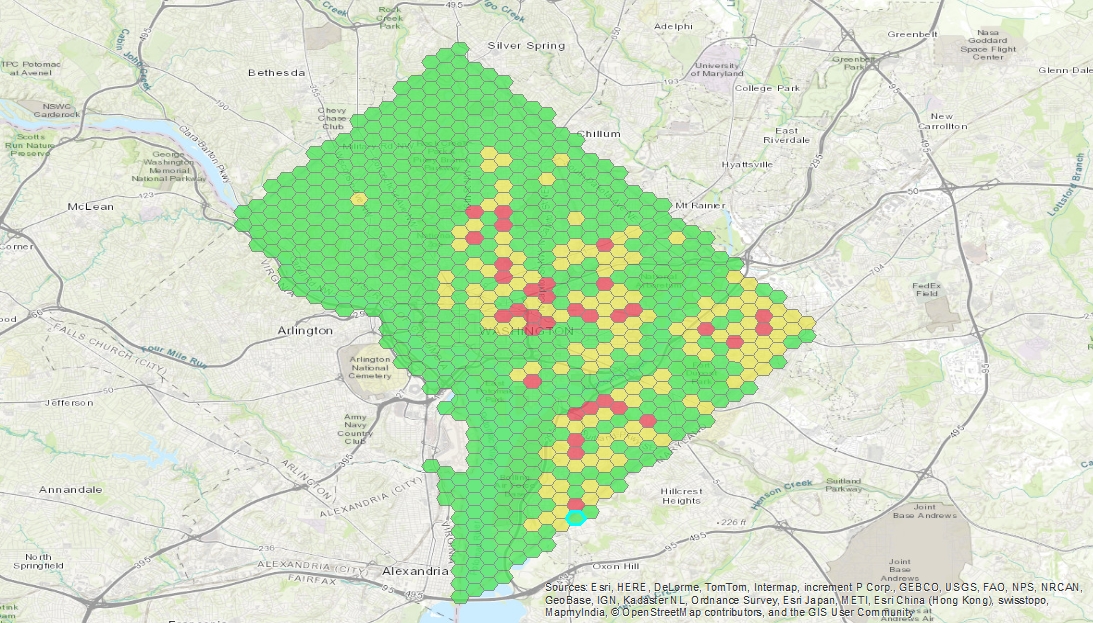
500M Grid Cells



CFS by Grid Cells



Naloxone Administration by Grid Cell



Machine Learning

* Caroline working on statistical methodology (regression types, research and writeup
* Wranga and Sarah- PostGreSQL and SciKit Learn
* All working with UCI repository and SciKit Learn

Individually everyone is working on machine learning and visualization from the UCI repository. Our plan is to apply the UCI test cases to our Capstone problem. This will allow us to be prepared for when our data is a clean, single authoritative dataframe. This dataframe will have values, by month, for each of the features we will use overall. Through this we are also considering the best approaches will be for missing values. We already know we will have missing values for age and gender. For age, using the mean value will probably work but for gender we will see how many are missing and possibly drop the rows depending on how many or look to see if we can infer the gender based on any of the other features.

Based on knowledge to specific software it has been selected to do spatial joins using that software but the team did want to make sure we understood how to use Postgres and do joins with pandas. The team has been able to read CSV to a dataframe and then save the dataframe to Postgres and also re-establish the connection to extract the Postgres table into a dataframe. We also have been able to join/merge multiple dataframes to create single dataframes.

APIs and External Data Sources- *Weather Data*

-Wranga has been working with the API

In order to gain a better understanding of outside features that may have an effect on opioid use, we have decided to add a weather element to our project. We are collecting weather data from the Dark Sky API. Our team wrote the wrapper that uses the Time Machine request to collect historic data for 2016. The data is read into a csv, which is then put into a dataframe for further analysis. At this time we are using temperature, humidity and precipitation, however we may add other elements as we continue to develop our project.