Predicting Violent and Nonviolent Crime Based on Location

Andrew Lee and Martin Yim

Abstract— **Using open Boston government data sets, we attempted to make a model that could accurately predict at a given location and should there have been a crime there, whether it would be likely to have been violent or not. We also hoped to generate useful visualizations over the time period of the data available (about three years). After using relative distance from given location to nearest schools, hospitals, police stations, and establishments with liquor licenses, we hoped to find a relationship between these predictors and the classification of a crime (violent or non-violent).**

I. INTRODUCTION

Our classifier was meant to be helpful in aiding discussions on urban policy and the nature of more serious crimes, such as violent crime, which we were examining in particular. We hoped that by examining proximity of violent crimes to our predictors (locations of places such as schools or police stations), we could better understand or narrow down the causes or other factors involved in the proliferation of violent crimes in a urban environment.

The data we used to build our classifier and make our visualizations, was pulled from the City of Bostons Data Portal. We used multiple data sets in order to get information on crime, as well as location of other establishments used as predictors in our model.

Our methodology was based around the assumption that violent crime has some correlation to proximity of certain establishments, and is perhaps also more likely to occur where violent crime is already an issue. As a result of this assumption, we chose to fit our model using predictors to important establishments for a urban system (where data was available), such as police stations and public schools.

We created a classifier using logistic regression because we believed it could have potential applications for predicting violent crime. This prediction capability, could then be applied to do better policing or allocate other resources to areas which may be particularly susceptible to violent crime. Violent crime is a much larger public safely issue, as compared to petty theft, larceny, or towed vehicles.

II. TECHNIQUE

The two techniques we applied to these data sets were Logistic Regression, and some simple time series visualizations.

We chose to make some visualizations through time using Heat Maps and location because we believed that it could reveal some interesting trends in location that may have been harder to spot through hard math. Also, we believed that it would be easy and useful for a layman to be able to look at and interpret quickly and efficiently. This would be paramount if this analysis was actually used in applications such as policing or government policy.

We chose to use Logistic Regression to build a binary classifier in order to make our violent-or-not predictions. Logistic Regression was our method of choice because it allowed for us to interpret the effects of each predictor on the outcome in isolation which was important given our particular predictors (different establishments and year). The ability to look at each predictor individually allowed us not only to look at its coefficient, but also to assess its statistical significance, and compare the predictors to one another. In addition to not black-boxing our model, Logistic Regression is simple, and well understood

III. DATASETS AND EXPERIMENTS

From the City of Boston Data Portal, we compiled five data sets to fit our model. Our primary data set was the Crimes Incidences data set, which included information on all types of criminal incidences from July 2012, to August 2015. We restricted all of our data to this time frame. Some of the important features for each criminal incident in the data set were: Incident Type, Date, WeaponType, and Location (given in Longitude and Latitude).

Our definition of Violent crime used the WeaponType variable in the Criminal Incidences data set. If an incidence had a value of Unarmed for WeaponType, we assumed the crime was non-violent. If otherwise, where WeaponType value was Knife, Firearm, Fist/Body Part (indicating assault) or not Unarmed, we would classify the incident as Violent.

Some observations were omitted due to missing values in the data set. For some observations, the Longitude and Latitude were missing; however the street address was given. We used a package named geopy that gave us the Longitude and Latitude when given a street address. This allowed us to avoid dropping as many data points as possible.

The other data sets had information on liquor permit, hospital, school, and police station locations. From these data sets, we primarily used the location in Longitude and Latitude to be able to calculate proximity to given locations for prediction. Some of these data sets were quite small because urban areas typically only contain a small collection of Hospitals or police stations.

Before starting to build the model, we used Excel to do basic preliminary exploratory analysis on the data sets. After examining the data sets, we began to build out our visualizations and our regression analysis.

We split our Criminal Incidences data set into training and test sets by randomizing our data set and then splitting the data in half. In the training data set, we used the location (Longitude, Latitude) and calculated it proximity to

the nearest hospital, school, police station, or establishment with a liquor permit, using the Haversine formula. Because the number of establishments in each of the data sets was relatively small, this computation was not exceedingly computationally expensive.

A. *Haversine Function*

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

In the testing data, we simply used the testing data to perform the same analysis, but withholding the WeaponType feature which was how we classified a crime as violent or not.

The visualizations were made using Year and Month attributes from the Criminal Incidents data set. Each heat map generated showed three months worth of data on violent crimes. Higher density areas of violent crime correspond to more red intensity on the heat map superimposed on the Boston map.

IV. RESULTS AND DISCUSSION

Our visual analysis of the heat maps over time (July 2012 - Aug 2015) showed variations, however, we are unsure of whether they are significantly different, or are within normal expectations. Nonetheless, they are helpful for visualizing the spread of violent crimes across the city and provide good intuition for investigating whether or not there are in fact trends over time (Fig.1 and Fig.2).

The Logistic Regression had mixed results. Our output showed that all of our predictors were statistically significant with p-values below 0.05 along with reasonable 95% Confidence Intervals that did not include 0. However, the coefficients were quite small, and the Psudeo-R-squared was also extremely low (Table 1). This suggests that although our predictors were in fact real factors in predicting violent crime, they are not the most important. Although our predictors may not have been the most power predictive variables, the directions of the coefficients for specifically proximity to Liquor Permit Establishments ($.2759$) and Schools ($-.4003$) were in line with heuristics. There was a inverse relationship between proximity to a school and a violent crime, and positive relationship between proximity to a Liquor Permit Establishment and a violent crime.

Our model would have benefitted from including other variables which may have produced a model with a higher Psuedo-R-Squared value which would increase the predictive power of our model.

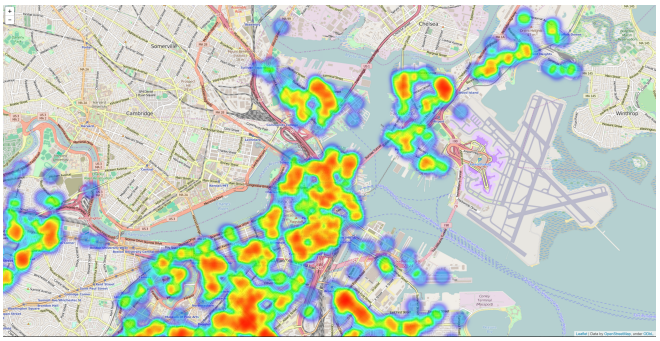

Fig. 1. Violent Crimes July 2012

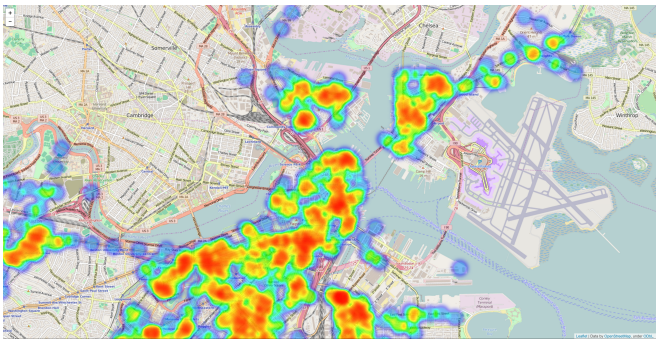

Fig. 2. Violent Crimes Oct 2012

TABLE I
LOGISTIC REGRESSION RESULTS

Dependant Variable			Violent Crimes		
Model			Logit		
No. Observations			126539		
Pseudo R-Squared			0.002584		
	coef	std err	z	$P > \|z\|$	95% Conf. Interval
pPol	0.0779	0.017	4.457	0.000	0.044-0.112
pLiq	0.2759	0.034	8.152	0.000	0.210-.0342
pHos	0.0698	0.007	10.480	0.000	0.057-.083
pSch	$-.4003$	.038	-10.648	0.000	$-.474$--$.327$
yr	$-.0006$	$8.17e-06$	-67.944	0.000	$-.001$--$.001$

V. CONCLUSION

The model we created ended up lacking as much predictive power as we anticipated finding, however, it did generate some interesting information and insights. Our results show that there is a relationship between certain types of establishments and the proximity to violent crime, but more importantly, it showed that other factors are likely to have much more impact on predictive power of classifying crime (violent or not) based on location.

The heat map visualizations we generated were not as useful as we would have hoped. Perhaps different methods for analyzing the data over time would have yielded better and more concrete insights into the nature of violent crime and its location. However, heat maps do show data in a easily digestible format.

In hindsight, we also found that perhaps doing a clustering analysis based on location of violent crimes would have been beneficial to gathering insight into understanding the spread

of violent crime in a city. We could have added the proximity to centers of violent crime to our model as a predictor, or conducted other analysis using data such as local economic data.

Overall, we discovered the proximity to locations of the establishments we used in our model were not very good at predicting whether a crime will be violent based on a given location, however our work may have interesting implications for trying to find other relevant factors for predicting violent crime in cities.

REFERENCES

[1] CriminalIncidents:https://data.cityofboston.gov/Public-Safety/Crime-Incident-Reports/7cdf-6fgx

[2] HospitalLocations:https://data.cityofboston.gov/Public-Health/Hospital-Locations/46f7-2snz

[3] LiquorPermits:https://data.cityofboston.gov/Permitting/Issued-Permits-ALL-TYPES-/2hre-tvqe

[4] PoliceStations:https://data.cityofboston.gov/Public-Safety/Boston-Police-District-Stations/23yb-cufe