Decision Support System for Effective Police Patrolling

Final Capstone Report

DATS 6501

August 16, 2019

By:

Varsha Waingankar

Contents **Page**

1. Introduction

1.1 Definition Problem

Although the numbers of sworn officers increased by 52,000 (up by 8 percent), the report found that from 1997 to 2016, the rate of full-time sworn officers per 1,000 decreased by 11 percent. During the same period, the general population in the U.S. increased by 21 percent. The severity of the officer shortage varies from department to department, but the national trend is clear: Since 2013, the total number of working sworn officers has fallen by about 23,000. The number of officers per capita is down even more sharply, from 2.42 per 1,000 residents in 1997 to 2.17 officers per 1,000 in 2016.

1.2 Research Question

How to help decision makers to more efficiently allocate police patrols and officer time, which are very expensive and scarce for many jurisdictions?

1.3 Practical Objective

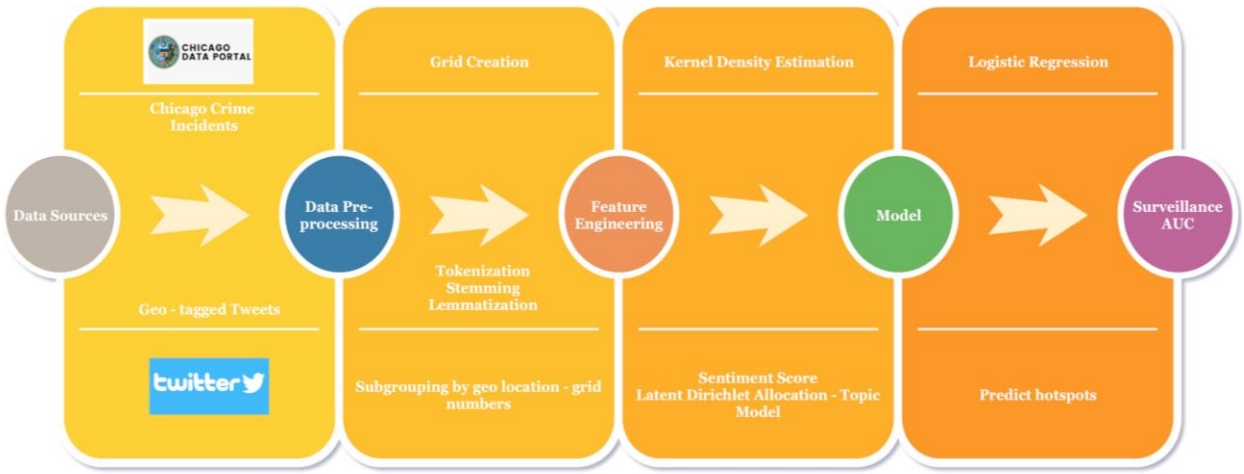
Predictive policing, in essence, is taking data from disparate sources, analyzing them and then using the results to anticipate, prevent and respond more effectively to future crime.

Predictive policing entails becoming less reactive. The predictive vision moves law enforcement from focusing on what happened to focusing on what will happen and how to effectively deploy resources in front of crime, thereby changing outcomes. It borrows from the principles of problem-oriented policing, community policing, evidence-based policing, intelligence-led policing and other proven policing models. Current analytic tools and techniques like hot spots, data mining, crime mapping, geospatial prediction and social network analysis can be applied to a broad range of criminal justice problems. For instance, they can be used to anticipate localized crime spikes, inform city and neighborhood planning and aid in police management decisions.

For example, **Reducing Random Gunfire in Richmond.** Every New Year's Eve, Richmond, Va., would experience an increase in random gunfire. Police began looking at data gathered over the years, and based on that information, they were able to anticipate the time, location and nature of future incidents. On New Year's Eve 2003, Richmond police placed officers at those locations to prevent crime and respond more rapidly. The result was a 47 percent decrease in random gunfire and a 246 percent increase in weapons seized. The department saved $15,000 in personnel costs.

2. Methodology

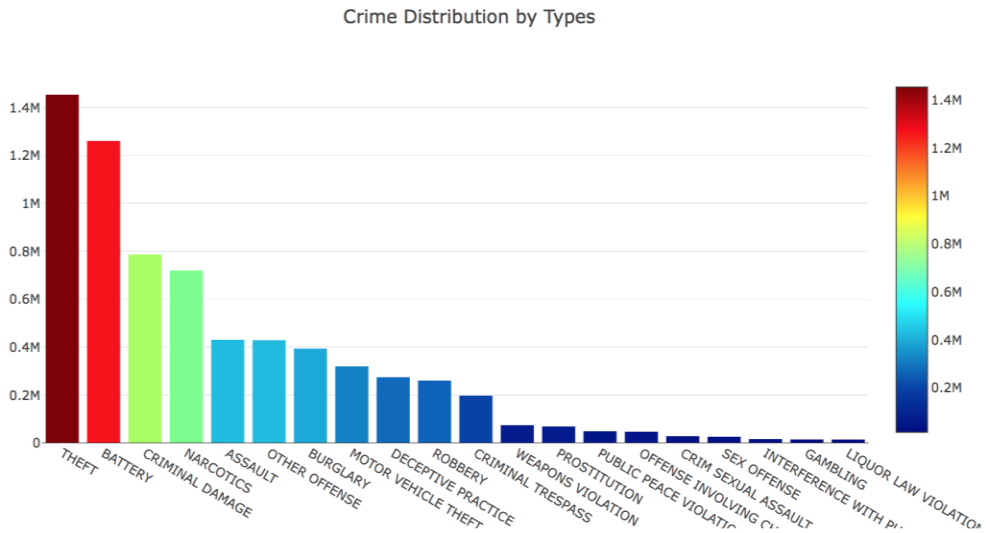
In the following section we will discuss the methodology applied to accomplish our research goal. First, we will describe the data collection process of obtaining the tweets and crime information we needed, utilizing tweepy API and Chicago crime data. Next, we will explain some data pre-processing steps we applied, tokenization, stemming, lemmatization, grouping data based on location. Additionally, we will discuss our data modeling process, sentiment analysis, topic modeling. Lastly, we will analyze our results through surveillance plots.



2.1 Data Collection

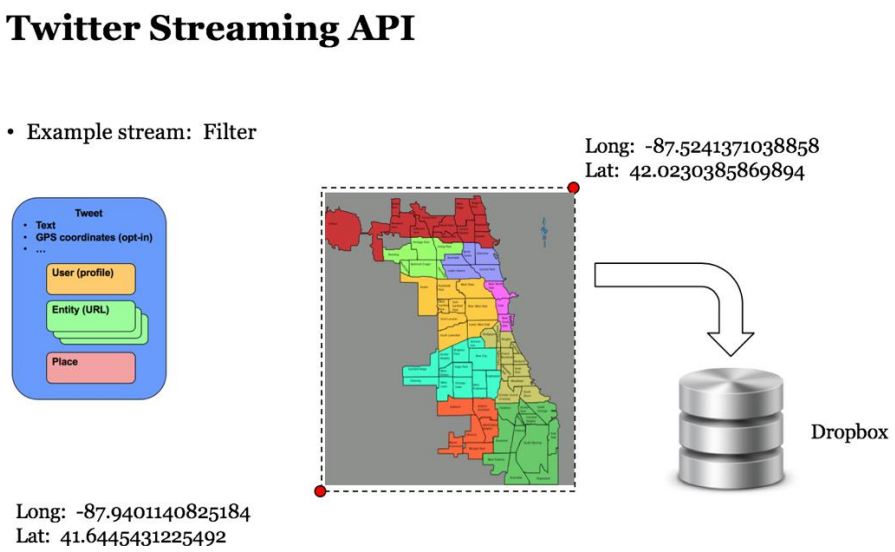
Crime Data

Chicago, Illinois ranks third in the United States in population (2.7 million), second in the categories of total murders, robberies, aggravated assaults, property crimes, and burglaries, and first in total motor vehicle thefts (June–August 2019). In addition to its large population and high crime rates, Chicago maintains a rich data portal containing, among other things, a complete listing of crimes documented by the Chicago Police Department. Using the Data Portal, we collected information on all crimes documented between June 2019 and August 1, 2019 (n = 60,876). Each crime record in our subset contained a timestamp of occurrence, latitude/longitude coordinates of the crime at the city block level, and one of 27 types (e.g., ASSAULT and THEFT). Figure shows the frequency of each crime type in our subset.



Twitter data

During the same time period, we also collected tweets tagged with GPS coordinates falling within the city limits of Chicago, Illinois (n = 1,528,184). We did this using the official Twitter Streaming API, defining a collection bounding box with coordinates [− 87.94011,41.64454] (lower-left corner) and [− 87.52413,42.02303] (upper-right corner). Most GPS-tagged tweets are posted in the down- town area of Chicago.



2.2 Data Pre-processing

Grid Creation

From the locations of known crimes of type T within the training window (these points received a label T). From a grid of evenly spaced points at 200- meter intervals, not coinciding with points from the first set (these points received a label NONE). It will be converted into UTM (Universal Transverse Mercator) co-ordinates for obtaining grid of a locality.

2.3 Data Modeling

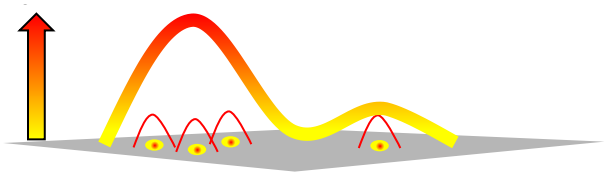
Kernel Density Estimation

Kernel Density Estimation (KDE) is a non-parametric technique for density estimation. A known density function (the kernel) is averaged across the observed data points to create a smooth approximation. Estimation data is historical crime record.

$$f_1(p) = k(p,h) = \frac{1}{Ph}\sum_{j=1}^{P} K\left(\frac{||p-p_j||}{h}\right)$$

$K(x) = \phi(x)$ (standard normal density)

$h =$ bandwidth (smoothing parameter)



Topic Analysis

Given the neighborhoods defined above, the problem reduces to estimating the n-1 topic importance values for each neighborhood given, the tweets posted by users in each neighborhood. We used latent Dirichlet allocation (LDA) for this purpose. LDA is a generative

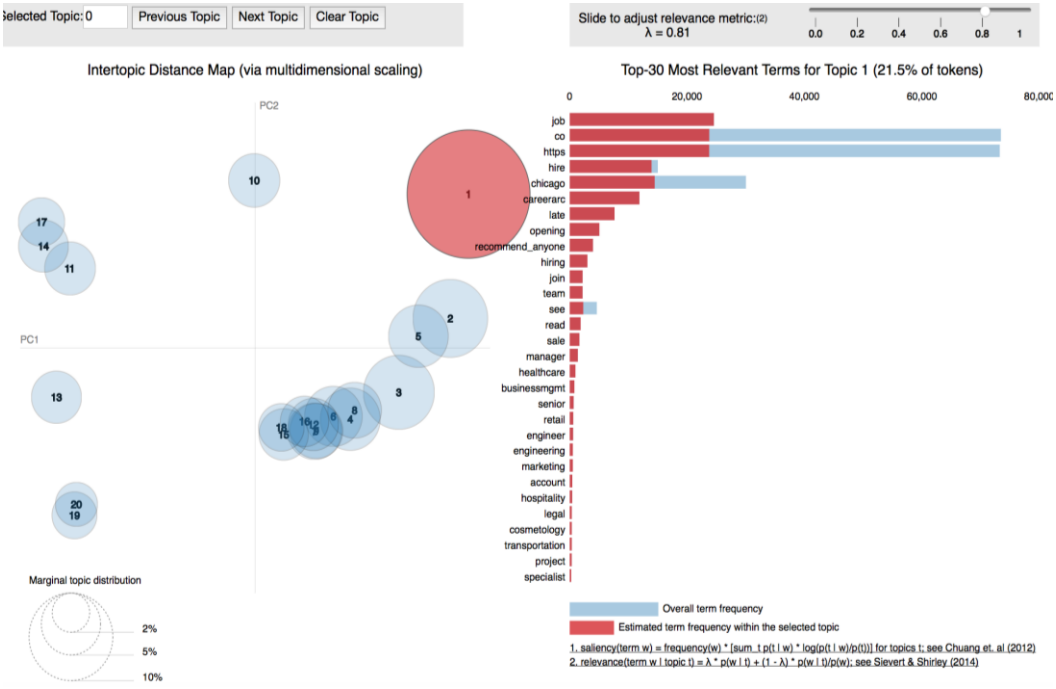
probabilistic model of textual content that identifies coherent topics of discussion within document collections. It is described as follows:

Word distribution of topics: the probability that each word belongs to (or defines) a topic. Topic distribution of documents: the probability that each topic belongs to (or defines) a document d. In order to see how geo tagged tweets actually contribute to crime prediction, we first needed to categorize our tweets into main topic areas for a particular location. This is how the matrix representation looks like:

	Topic 1	Topic 2	Topic 3
Document 1	1	0	1
Document 2	0	0	0
Document 3	0	0	1
Document 3	0	1	0

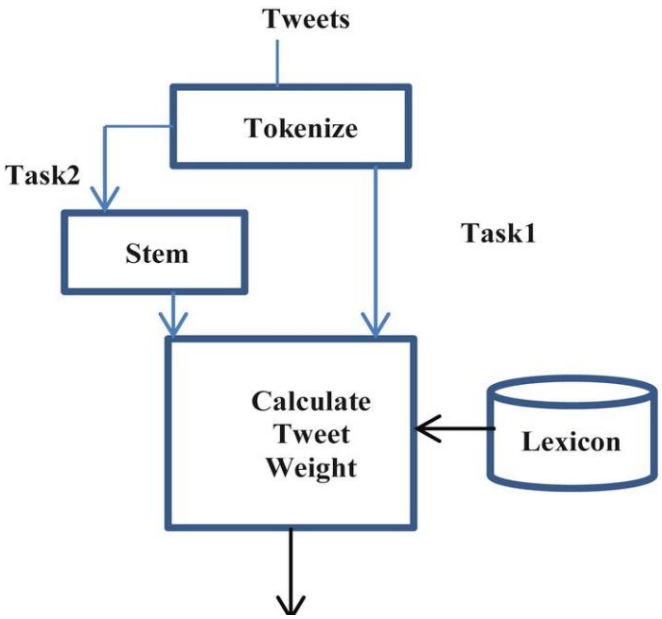
	Word 1	Word 2	Word 3
Topic 1	1	1	0
Topic 2	0	1	1
Topic 3	1	0	1

Visualization of topic Modeling



Sentiment Analysis on Tweets:

There are two major approaches: Supervised machine learning approaches/ Unsupervised lexicon-based approaches. Unsupervised method for predicting the sentiment by using lexicons that have detailed information, specially curated and prepared just for sentiment analysis was used. AFINN lexicon for sentiment extraction and analysis was implemented



Gives the final sentiment score.

Model Execution and Performance Evaluation

Model

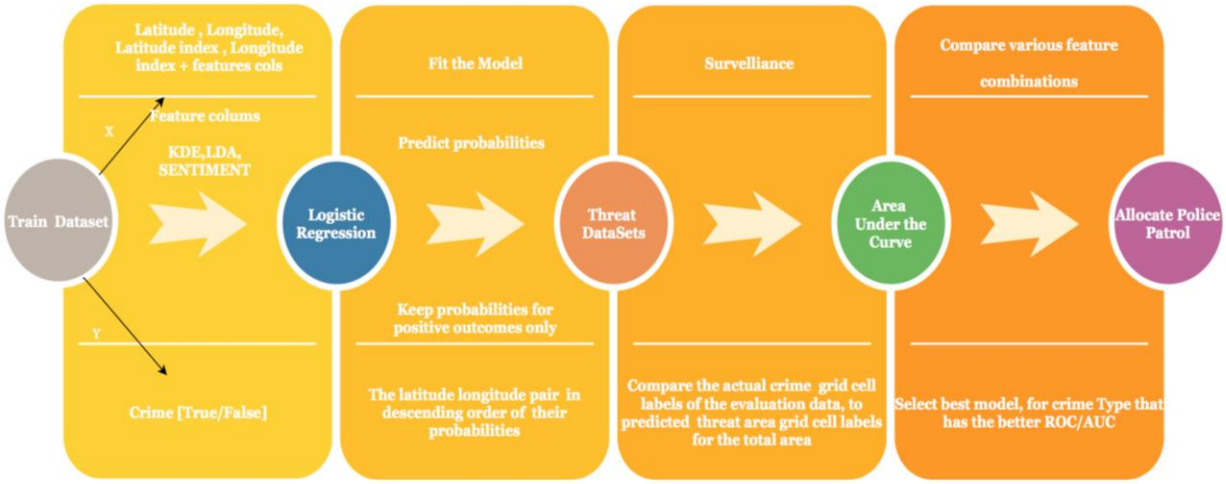
Training the model on 31-day window for crime type T. Making T predictions for the first day following the training window, sliding one day into the future and repeating.

Evaluation

Project the surveillance plot that measures the percentage of true T crimes during prediction window, that occur within the most threatened area, according to model's prediction for T.

Lastly, because each model execution produced a series of surveillance plots for crime type T, one for each prediction day, aggregate the plots to measure overall performance

Execution Architecture

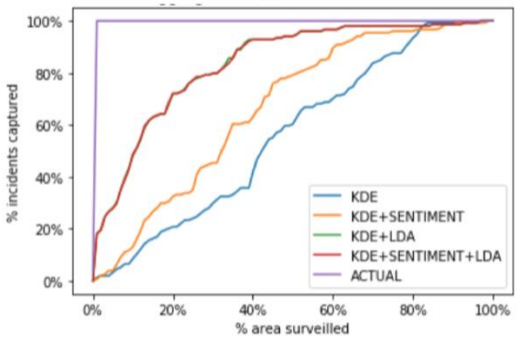


The area under the curve for different categories of crime, on daily basis vs aggregated performance of 10 days.

2.4 Visual Analysis

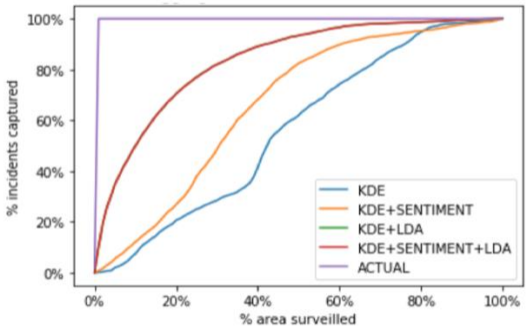
Crime Type - Theft

AUC – One day Prediction




```
56.538961038961034 KDE
67.18831168831167 KDE+SENTIMENT
84.07142857142856 KDE+LDA
84.04545454545455 KDE+SENTIMENT+LDA
100.03246753246754 ACTUAL
```

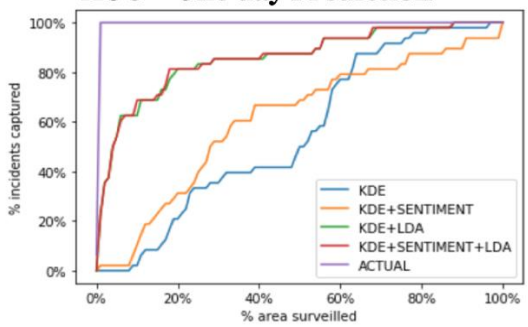
AUC – Aggregated Plot for 10 days prediction




```
56.8164610444482 KDE
66.96569207368599 KDE+SENTIMENT
84.40949805644753 KDE+LDA
84.38837248605712 KDE+SENTIMENT+LDA
```

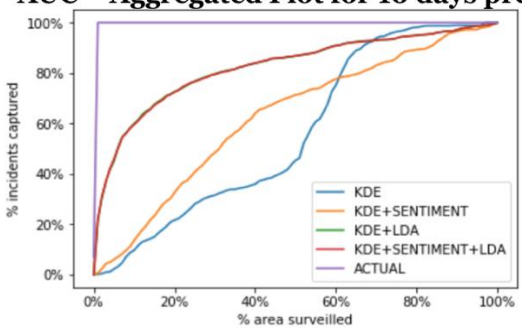
Crime Type - Narcotics

AUC – One day Prediction




```
56.395833333333336 KDE
61.541666666666664 KDE+SENTIMENT
86.1875 KDE+LDA
86.33333333333331 KDE+SENTIMENT+LDA
100.0625 ACTUAL
```

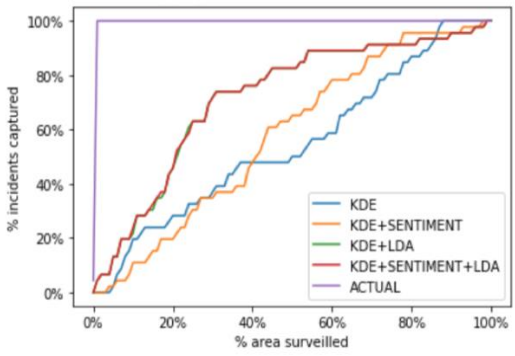
AUC – Aggregated Plot for 10 days prediction




```
56.63587684069611 KDE
62.99732262382865 KDE+SENTIMENT
82.75033467202141 KDE+LDA
82.71820615796518 KDE+SENTIMENT+LDA
100.06827309236948 ACTUAL
```

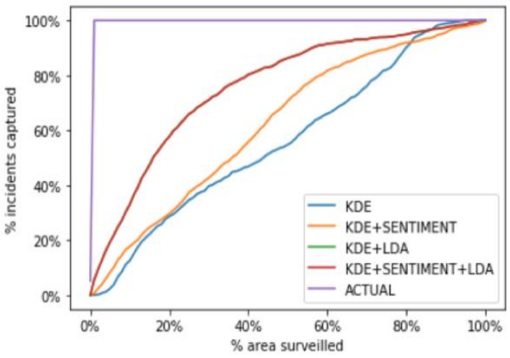
Crime Type - Burglary

AUC – One day Prediction




```
55.21739130434783 KDE
58.45652173913044 KDE+SENTIMENT
73.95652173913044 KDE+LDA
73.8913043478261 KDE+SENTIMENT+LDA
100.04347826086956 ACTUAL
```

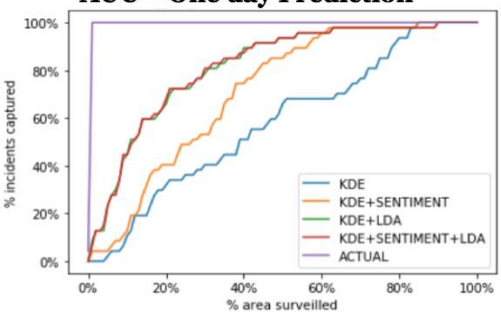
AUC – Aggregated Plot for 10 days prediction




```
56.898360655737704 KDE
62.90409836065574 KDE+SENTIMENT
76.51803278688526 KDE+LDA
76.50819672131149 KDE+SENTIMENT+LDA
100.05245901639344 ACTUAL
```

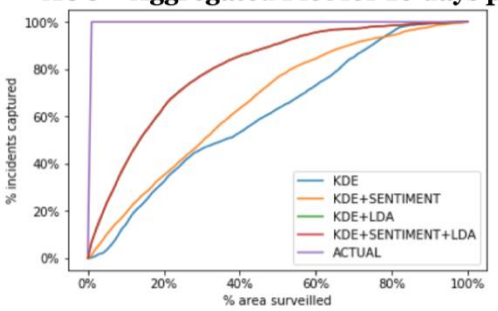
Crime Type - Battery

AUC – One day Prediction




```
58.72340425531915 KDE
71.8936170212766 KDE+SENTIMENT
82.1063829787234 KDE+LDA
82.25531914893617 KDE+SENTIMENT+LDA
100.04255319148936 ACTUAL
```

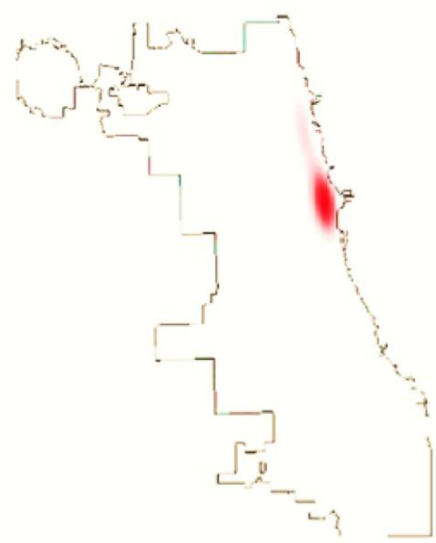
AUC – Aggregated Plot for 10 days prediction



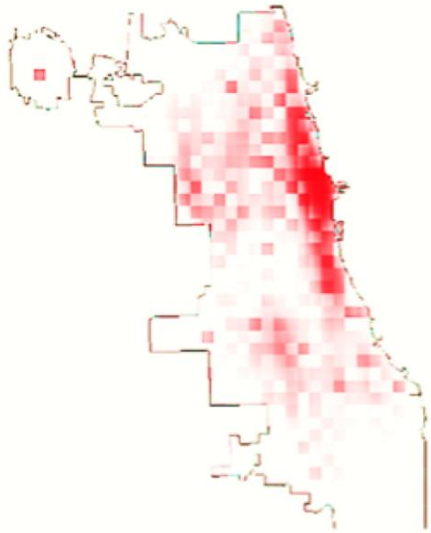

```
61.76088314934388 KDE
66.78816913143095 KDE+SENTIMENT
80.8117058946053 KDE+LDA
80.836075817538 KDE+SENTIMENT+LDA
100.0202041241408 ACTUAL
```

Crime prediction (Theft)

Predicted hotspots using KDE

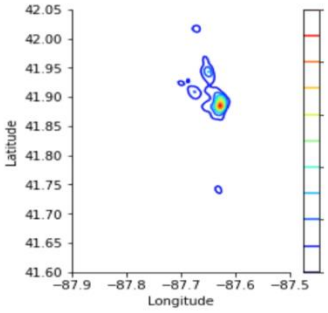


Predicted hotspots using KDE+LDA

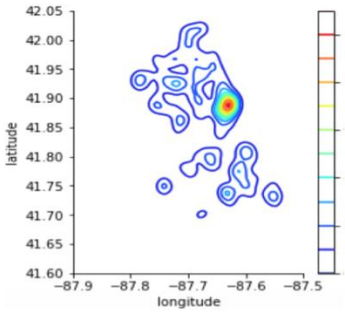


Visualization of actual hotspots / model generated

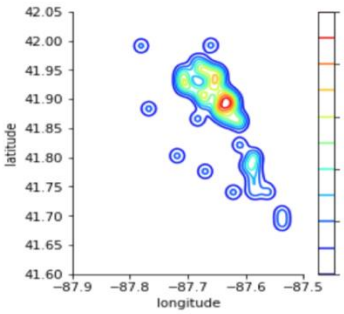
KDE generated Hotspots



Actual Hotspots



KDE+LDA generated Hotspots



3. Conclusion

3.1Results

The best model for crime hotspot prediction is using Kernel density estimation and LDA topic models, as it yields the best AUC (Area Under the Curve). Using the best model, decision makers can allocate scarce resources (e.g., police patrols) across the geographic space, more efficiently, leading to reduction in wasted effort and decrease in crime response time. Only for 19 of the 25 crime types studied, the addition of Twitter data improves crime prediction performance versus a standard approach based on kernel density estimation.

3.2 Future Research and Applications

Digging deeper into the semantics of tweets could provide performance improvements compared to the models that were implemented. For example, it would be interesting to analyze the predicate, argument structure of tweets in order to extract the events they describe and the actors in those events, using network analysis. Analyzing the networks between follower and followee might facilitate the anticipation of events (e.g., parties) that are known to correlate with criminal activity.

Using temporal modeling, it makes sense that crime patterns could exhibit these behaviors and that Twitter content might be more predictive when message timestamps are taken into account. For example, one could identify trends within the topic proportions for a neighborhood and incorporate a trend variable (e.g., magnitude of increase or decrease) into the model. One could also allow for delayed effects of Twitter topics, the intuition being that Twitter users often anticipate crime-correlated events.

4. References

This project has been implemented based on this paper.

https://www.sciencedirect.com/science/article/pii/S0167923614000268