Case Study #1

Matthew DaVolio, Kerry Jones, Gabriel Rushin, Jingyue Xiong

Identify the Problem

Descriptive Analysis:

Crime has a profound impact on people's well-being, both socio-psychologically and financially. While studies have shown crime rates have actually been decreasing over the past several decades, improvements must continue to be made to ensure these rates do not stagnate. With recent technological advances, spatial and temporal data mining techniques were made readily available to law enforcement officials and more engineers skilled in the specific area have emerged to further improve these techniques. Along with these technological advances, policy makers and social scientists have continued to receive public pressure to ensure crime rates continue to decrease, leading to further improvements in the infrastructure of cities as well as new regulations within government controlled agencies.

These data mining techniques take into account data related to both the nature of the crime, as well as events and conditions that occur outside the crime itself. Conditions which are observed that may affect the probability of a crime occurring or the severity of the crime include the weather and lighting conditions in the nearby location, along with the geospatial and physical aspects of this location such as the demographics, the type of neighborhood surrounding the location, and its relative location to police stations and other related structures. Time related data is also taken into account such as the time of day, whether the crime occurs on a weekend or weekday, and any relations to holidays.

Normative Analysis:

In order to decrease crime rates in major cities, results from data analysis allow policy makers and law enforcement agencies to understand patterns in criminal behavior to both predict and anticipate crimes as well as create conditions in which criminals are less likely to attempt to carry out a crime. By studying these results, policy makers can know whether infrastructure improvements such as street lights or <> will help deter criminals. Law enforcement agencies can use results from this data analysis to know where the should have their patrols stationed to attempt to prevent criminals from taking action as well as know when they need to increase patrols if they are expecting a higher level of crime due to weather or time.

However, since it is realistically impossible to completely eliminate crime, a secondary goal is to decrease response times of police in order to increase the

probability of arriving while the crime is still taking place, or at a minimum increasing the odds of finding the perpetrator after the fact and allowing for prosecution.

Stakeholders:

As previously discussed, crime rate information is vital to both law enforcement agencies and political leaders in order to allow them to make an effort at decreasing crime rates in their cities. However, many other groups can find this information useful and informative to allow them to make decisions. Those who commute to work, especially by public transportation or by walking will be able to use this analysis to know the safest routes to travel to and from work. Similarly tourists can be knowledgeable about unsafe areas or routes. Another group of interest is potential home buyers or people moving to a new city and looking to rent as they can use this information to decide if an area is the right choice for them and know if they may need to purchase additional home security systems or protection. This can help them make the right choice not only for their safety, but can take into account any additional financial costs that may arise from crime related to their new area of residence.

Financial costs related to crime affect many other groups as well. Insurance companies can reference to the crime information when making household insurance policies and decide whether or not to charge higher premiums for specific geographical locations. Crime-related expenditures also causes a strain on state and federal budgets. The loosening of the crime situation could alleviate the budget constraint and lead to more spending on education and policing instead of incarcerating the criminal offenders. In addition, police departments can make more informed public safety decision regarding allocating resources to hot-spot areas and arranging efficient police patrolling time.

Impact:

A systematic crime prediction learning model could provide information on explanatory variables which can be used to identify underlying causes of the type of the crime in order for the police and policy-makers to make more informed decisions on regulations and allocation of police forces. Crime also comes with financial cost. Crime rate reduction also positively correlates with the improvement of social welfare. Expenses on incarcerating criminal offenders could be better spent on education and other areas of interest.

Objectives and Metrics

The objective of this paper is to understand the factors that decrease or increase crime and to optimize these factors in order to decrease crime rates. An attempt to

optimize police presence, and decrease response times can assist in achieving lower crime rates. Optimizing police presents includes placing police officers in locations in hotspots, or locations where crime occurs the most often. It is supported by numerous research, that placing police officers in these hotspots decrease crime rates in these locations. For example, "*An Ex Post Facto Evaluation of Tactical Police Response in Residential Theft from Vehicle Micro-time Hot Spots*" explains that micro-time-hot spots policing had led to a nearly 20% reduction in residential theft from vehicle crime. However, "*When police patrols matter. The effect of police proximity on citizens' crime risk perception* " states that an increase in police interaction in the community also decreases crime rates. Additionally, the presence of police officers can leave a lingering effect (even when they are not present) to deter crime. Also putting police officers in better location could decrease response times. Decreasing response times lead into decrease crime rates due to the higher possibility of criminal being caught.

In order to measure these a decrease in crime rates, crime occurrence should be measured and categorized as violent and nonviolent. Average response times for crimes occurring in hotspots should also be measured. The amount of taxpayer dollars going towards defense lawyers and criminal court cases should also be measured. A decrease in crime rates should decrease the amount of court cases and defense lawyers. The number of 911 calls should be categorize between crimes and general emergencies. Also the number of reports to police should also be measured. These metrics will tell whether crime is decreasing in areas.

State of the Art

Clustering:

DBSCAN, Density Based Spatial Clustering Application with Noise, is a clustering technique that groups together points closely related. Using a collection of historical crime data, administrative boundary, and police data, DBSCAN is used to cluster states with similar crime trends. Additionally, using "next year's" cluster data in combination with state poverty data, the author uses a C4.5 decision tree, to predict future patterns in upcoming years. In this research, the author used crime data between 1954-2006 collected by the Integrated Network for Social Conflict Research. A major problem associated with this solution is processing large amounts of data(zota bytes)(Malathi 2011).

Similar to the increased likelihood of aftershocks after a major earthquake took place, specific forms of criminal behavior could spread through neighboring areas via a contagion-like process(Johnson 2008). In light of this discovery, *G.O.Mohler et al.* proposed the self-exciting point processes, originally defined in seismology, to model

the spatial-temporal clustering patterns of criminal behaviors. Their approach combined the stochastic declustering with Kernel Density Estimation in a novel way. By comparing the predictive accuracy of their point process model with the traditional crime hotspot visualization, they showed the merit of incorporating the point process methods to improve crime hotspot maps by taking into account the background rate of crime that subsequently triggered future events so as to increase the predictive power.

Pattern Discovery:

Cascading Spatial Temporal Pattern Mining- Pradeep et al., developed an algorithm to discover ordered subsets of event types that are located near each other. For instance, consider an event such as a bar closing. This event is a considered a crime generator, because a bar closing leads to cascading criminal activity (drunk driving or assault). This approach was tested on 2007 crime data from of events that took place in Lincoln, Nebraska. Using the Kolmogorov-Smirnov test to tests the equality of statistical distributions, was used to compare the candidate distributions of criminal activity on all nights vs. on Saturday nights, and after game nights. They observed that bar closings on Saturday nights and bar closings after football games are crime generators. Specifically, looking at crimes such as larceny, vandalism, and assaults. One issue that remains unsolved is accounting for spatial(local vs. global) and temporal aspects of patterns(time semantics) (Pradeep 2012).

Spatial Modeling:

Hotspot mapping is another analytical framework used to target high-crime activity. A majority of this research focuses on methods used to identify hotspots not necessarily how accurate they are in predicting where future crime is likely to occur. Chaney et al. explore investigate the accuracy of the following hotspot techniques: Spatial ellipses, Thematic boundary mapping, Grid thematic mapping, and Kernel Density Estimation(KDE). Using a Prediction Accuracy Index, defined by number of crimes accurately predicted with respect of total number of crime events within the size of the study area. This study was done utilizing 2002-2003 GIS crime data(point data) from Central/North London. The results suggests that KDE is the outperforms other techniques used for predicting crime patterns and predicting specific types of crimes. Investigating how models handle different volumes of input data and selecting ideal user thresholds(parameters) are two unsolved areas of research(Pradeep 2012).

Another variation of spatial modeling was performed by Xiaofeng Wang and Donald E. Brown by looking at breaking and entering incidents in the Charlottesville, VA area in their paper *The Spatio-Temporal Modeling for Criminal Incidents*. Wang and Brown improve on the basic hotspot mapping techniques by creating generalized additive models, based on spatial modeling, and furthermore a localized version that allows for specific regional factors to influence the predictive crime model. Using these additive models rather than basic hotspot modeling allows for a more diverse set of data

to be considered for predicting future crimes. These models also greatly improved previous models that were used as predictors (Wang 2012).

Exploring Social Media:

One way to understand crime risk is understanding the environmental conditions and behavior of the community members surrounding areas where crime is most likely to occur. Using a combination of points-of-interest (banks, churches, community, parks,etc.) as representations of spatial conditions, social media data (Twitter), and crime event data, Bendler aims to explain how social conditions can contribute to crime. Initially, Bendler uses a zero-inflated poisson regression to take into account the data cannot have non-negative values and areas in the grid where crime doesn't exist. Evaluated using Vuang Closeness Test, AIC and BIC. The preliminary results suggests twitter data is highly correlated with Motor Vehicle Theft and Larceny. More specifically that Motor Vehicle Thefts occur when there aren't many twitter messages (indicating no public presence or activity near crime occurrence). Oppositely, larceny crimes are more likely to occur when people are present. (Bendler 2014).

Since regression doesn't take into account spatial dependency, it's unclear, whether this relationship between twitter and crime incidents is a global or local effect, thus Bendler employs Geographically Weighted Regression. Using this algorithm, Bendler applies regression to cells independently, to assess how different or similar, the estimated values in each cell are to one another. Thus providing an understanding how specific crime event types are dependent on factors defined by space, also known as spatial dependency. Finding varying coefficient estimates across the region, the author was able to conclude "that places with a high Social Media activity yield considerably negative Twitter coefficient estimates for Motor Vehicle Theft and positive estimates for Theft/Larceny"(Bendler 2014)..

One extension of the research should adjust for spatial boundaries,instead of a grid-based approach would provide a more accurate representation of the spatial conditions of a given area (Bendler 2014).

In related work, Gerber investigated the additional value of using twitter data in a combination approach using Latent Dirchlet Allocation (LDA) with Kernel Density Estimation (KDE). Gerber collected over 60,000 records of crimes observed between January 1, 2013 and March 31, 2013. Each observation consisted of: a timestamp, the latitude and longitude of crime at the city block level, and crime type. Additionally,he collected geolocated twitter data from a bounding box using Twitter Streaming API (Gerber). Using LDA, a probabilistic model used to generate topics from a collection of documents (tweets), and was able to identify topics occurring in Chicago and incorporate that information into KDE. KDE provides an assessment of how frequent an event, in this case, crime event, occurs within a given region. This assessment was evaluated using the Area Under the Curve (AUC); a way to evaluate true positive vs false positive rates of classifying datasets. Further investigation into the semantics of tweets, social networks, and modeling for temporal aspects could give further improve these models performance(Gerber 2014).

Statistical Models:

Micro-time hot spots of theft from vehicle crime in Port St. Lucie, Fl were researched in this article. Micro-time hot spots are defined to be the appearance of multiple and closely related crimes within a close proximity. The research utilized logistic regression and propensity score matching (PSM). The Logistic regression was used in conjunction with a 1-1 matching without replacements and a .10 caliper width of the logit standard deviations. This had eliminated observations outside the region of interest. The covariates utilized in the logistic model were year, season, district, radius, targets, known offenders, crime, and time span. As a consequence, 172 micro-time hot spots with 86 matching pairs were discovered and researched over the five years. The researched concluded that tactical response to these micro-time hot spots lead to an almost 20% reduction in theft from vehicle crime. This conclusion had excluded spatial displacement of the crime.

However, the article had failed to address macro-time hot spots. In addition, this paper states that this method is best for suburban cities or small towns. Santos explains these areas have a more manageable amount of crime. Thus, further research would have to be performed to see if micro-time hot-spots are viable for larger areas. Also research was done on only one crime type and one police department. Further research would need to be conducted with multiple police departments and crime types.

Citations

Bendler J. Brandt T. Wagner S. and Neumann D . (2014), 'Investigating Crime-to-Twitter Relationships in Urban Environments—Facilitating a Virtual Neighbourhood Watch', Proceedings of the European Conference on Information Systems (ECIS), Tel Aviv, June 9–11.

Gerber, M. S. (2014). Predicting crime using Twitter and kernel density estimation. *Decision Support Systems, 61*, 115-125.

Malathi., A., & Baboo, S. S. (2011). An Enhanced Algorithm to Predict a Future Crime using Data Mining. International Journal of Computer Applications IJCA,21(1), 1-6.

Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., & Tita, G. E. (2011). Self-Exciting Point Process Modeling of Crime. Journal of the American Statistical Association, 106(493), 100-108.

Pradeep, M., Shekhar, S., Shine, J. A., & Rogers, J. P. (2010). Cascading spatio-temporal pattern discovery: A summary of results. Proceedings of the 2010 SIAM International Conference on Data Mining, 327-338.

Santos, R. G., & Santos, R. B. (2015). An Ex Post Facto Evaluation of Tactical Police Response in Residential Theft from Vehicle Micro-time Hot Spots. Journal of Quantitative Criminology J Quant Criminol, 31(4), 679-698.

Wang, X., & Brown, D. E. (2012). The spatio-temporal modeling for criminal incidents. Security Informatics, 1(1), 2.

Unfamiliar Items:

- **Vuong's Closeness test**
- **Kernel Density Estimation**
- **Bayesian Information Criterion(BIC)**
- **Clearer Definition on how LDA creates topics**
- **Other types of Decision Trees(C4.5?)**
- **Point Process**
- **1-1 matching without replacement**
- **Calipher width**