Survey of Crime Analysis and Prediction

Lenin Mookiah, William Eberle and **Ambareen Siraj**

lmookiah42@students.tntech.edu and {weberle,asiraj}@tntech.edu

Tennessee Technological University

Cookeville, TN 38505.

Abstract

Crime analytics and prediction have long been studied among research communities. In recent years, crime data from different heterogeneous sources have given immense opportunities to the research community to effectively study crime pattern and prediction tasks in actual real data. In this survey paper, we will discuss research that takes into account a variety of crime related variables, and shows where in some cases, information that has been widely accepted as influencing the crime rate, actually does not have an effect.

Introduction

For many years, research in the area of crime analysis has been used towards the mitigation of crime and public safety. In the past decade, with the advent of Big Data, public data availability, and e-governance there has been an almost exponential increase in the creation of data analytics and visualization tools for the policing and handling of public-safety systems.

In this survey, we see various aspects of past and recent research directions in crime analytics. Few researchers have studied the causation of one or more variables in terms of the crime rate. Researchers have studied the effects of variables such as age, alcohol, etc. to crime rate, while some research has examined the underlying factors or co-relation associated with the crime rate, such as the abortion rate and increasing incarceration rate. Also, researchers have studied the problem of finding hot spots for crime and using that for prediction. Prediction of future crime events is a non-trivial task. Some researchers are trying to predict crimes from news feeds, census data and online resources. In addition, some research work has focused on the bias of media regarding crime reporting towards a particular group involving categories such as race and gender. Table 1 shows list of important work discussed in this paper.

Effects

In this section, we will discuss the effect that television and news papers have on the fear of crime, as well as its effect on crime and the incarceration rate. Cultivation theory is a social theory which examines the long term effects of television. Early researchers studied study crime cultivation through television viewing and the fear of victimization (Doob and Macdonald 1979). The authors conclude that the amount of television viewing has a significant impact on one's fear of victimization.

Other research suggests that the crime risk may be smaller than one would think (Kemshall 1997), while others did quantitative studies on threatening trends in the news and concludes news makes the situation worse (Berger 2001). There have also been studies between the correlation of crime rate and prison populations with regression (Listokin 2003). The author concludes that once the endogeneity is addressed, incarceration rates could be predicted. Others study the correlation between age and crime (Hirschi and Gottfredson 1983). The authors conclude that though correlated, age is not useful for predicting involvement over the life time of offenders.

In summary, media might just exaggerate the perception of actual crime events. Actual crime predicted will be more accurate by using statistics of incarceration rate than age.

Factors and Policies

Researchers have also studied the various influencing factors that impact the dynamics of the crime rate. In one study, a researcher examined the phenomenon of the sharp crime fall in the 1990s (Levitt 2004). The author conclude that four factors - increased incarceration, more police, the decline of crack and legalized abortion - played significant roles in crime reduction. Other factors, such as a strong economy, which are generally viewed as a major factor, didn't appear to have an impact.

One researcher studied booms and busts of crimes using wavelength and shape studies to identify factors of influences (LaFree 1999). The author concludes that the 1960s to mid-1970s was a crime-boom period, while the 1990s was a crime-bust period. Other research has also studied the effect of rising middle-eastern crimes in Australia (Priest 2004). Another researcher studies the correlation of street light with crime (Pease 1999). The author suggests that lighting places could help bring a potentially criminal situation under control, especially for the arms race problem in the streets. Another proposes effective reward and penalty methods to enhance overall policies in the criminal justice system (Kahan 1998). While other researchers study *Caucasian* attitudes to-

Article	Approach	Data Source
(Doob and Macdonald 1979)	Regression	Survey
(Listokin 2003)	Least Square estimates and Regression	DoJ[2] and UCR[1]
(Hirschi and Gottfredson 1983)	Statistical	DoJ[2]
(LaFree 1999)	Graph plot study	UCR[1]
(Green, Staerklé, and Sears 2006)	Statistics and Principle Component Analysis	Survey (849 respondents)
(Gruenewald et al. 2006)	Statistical	Heterogeneous Data[6]
(Gerber 2014)	Kernel Density Estimation	TwD[5] (3 months)
(Shingleton 2012)	Regression	Salinas PDD[4]
(Featherstone 2013)	Statistical	South African PDD[4] and TwD[5]
(Liao et al. 2010)	Bayesian model	PDD[4]
(Radinsky and Horvitz 2013)	Clustering and Entropy	NA (23 years)
(Ali et al. 2011)	Visualization	Malaysian NA[3]
(Mohd 2010)	Clustering (k-means)	Malaysian NA[3]
(Alruily 2012)	Self-Organizing Map	NA[3]
(Pritchard and Hughes 1997)	Statistical	560 NA[3]
(Weiss and Chermak 1998)	Statistical	227 NA[3]

Table 1: Table of Survey

[1]Uniform Crime Reports
[2]U.S. Department of Justice
[3]News Articles
[4]Police Department Data
[5]Twitter Data
[6]Alcohol Beverage Control and Hospital Discharge Data

wards crime policies (Green, Staerklé, and Sears 2006). The authors conclude that there is a positive effect of symbolic racism on both preventive and punitive penalties, especially punitive ones.

HotSpots and Prediction

In this section, we will discuss the research of marking hot spot crime zones and the subsequent task of predicting crime. Studies show that hot spots are one of the main factors in reducing crime (Levitt 2004). *Hot spots* are defined as the areas that have a higher concentration of crime events. Researchers show that there is correlation between the characteristics of a population and the rate of violent crimes (Gruenewald et al. 2006). The authors uses data of hospital discharge records, availability of alcohol outlets, and census information. The author employs a statistical approach and concludes that the socio-economic factor of location such as unstable poor, economically stable, immigrant Hispanic and rural majority population are the strongest correlations to violent crimes.

One researcher has examined the social network Twitter to predict crime (Gerber 2014). The author collected GPS-tagged tweets (n = 60, 876) mentioning crime events between January 1, 2013 and March 31, 2013. The author uses Latent Dirichlet Allocation for learning topics and related terms from tweets and eschews deep semantic analysis in favor of shallower analysis via topic modeling. Out of 25 crime types, 19 show improvement in Area Under Curve (AUC) method after incorporating topics from tweets along with density-based estimation. Some researchers also have studied the trend of using web-based crime mapping from the 100 highest GDP cities of the world (Leong and Chan 2013). The authors conclude that a web-based mapping showing crime visualization is the most common tool. Also, the main factors that drives numerous crime mapping is e-governance and community policing.

Another researcher employs regression models to predict crime trends in Salinas, California (Shingleton 2012). The author uses three methods - Ordinary Least Squares, Poisson Regression, and Violence Prediction using Negative Binomial Regression. He concludes that the ordinary least squares approach is adequate enough to predict three crime types since predicted crime values from both Ordinary Least Square and Poisson regression are comparable.

Another study uses vehicular description from microblogs to predict crime (Featherstone 2013). The author employs topic modeling (Latent Dirichlet) on common useful terms to find patterns. In addition, the author is able to discover a correlation between reported crime census statistics from the South African Police Service and crime events discussed in tweets. Others propose a novel Bayesian based prediction model (Liao et al. 2010). The authors use the model to predict the accurate location of the next crime scene in a serial crime. Results are effective with predicting three out of four crime locations (i.e., accuracy of 75%).

News Feed

In this section, we will discuss the analysis of crime patterns and predicting crimes from news feeds. In one instance, researchers use 20 years (1986-2008) of New York Times news corpus to predict future events such as disease outbreak and floods (Radinsky and Horvitz 2013). The authors build a prediction model named *full-model with likelihood* from history data to predict possible occurrence of future events. For example, the authors are able to predict a cholera outbreak in Bangladesh from past data. The authors compares the results with two other baselines models named frequency based and co-occurrence based. The proposed full-model outperforms with 24% precision compared to frequency based with lesser

than 1% and co-occurrence with 7%.

Researchers employ k-means and affinity propagation for clustering crime news from Malaysian newspapers (Bsoul, Salim, and Zakaria 2013). In addition, they provide an interesting comparison of past research work. Others propose a news retrieval system with visualization techniques for a Malaysian newspaper (Ali et al. 2011). The primary goal is to improve the user interface for better user interaction and information flow. The authors try to implement front-end system visualization components by studying the needs of targeted users. Another researcher has studied various clustering algorithms on crime topics (Mohd 2010). In this case, the author proposes an enhanced k-means for finding the optimal initial centroid. The proposed algorithm proves to work much better than single-pass clustering on the Malaysian newspaper corpus. The enhanced k-means achieved a better F-measure of 0.915 compared to a single pass approach with 0.712.

Another researcher built a crime policing system to extract information such as crime type and location from reports to provide for a more effective crime analysis (Alruily 2012). The author employs an unsupervised Sequential Minimization Optimization for clustering. Additionally, the author uses visualization techniques for analysis. Other researchers studied 100 homicide committed over 560 articles from Milwaukee's newspapers to analyze the deviation involved in news reporting (Pritchard and Hughes 1997). The authors use different deviation factors based on statistics (e.g., white suspect/victim, child/senior-citizen victims), status (e.g., wealthy, white suspect/victim), and culture (e.g., female suspect). The authors conclude status and cultural deviance are much stronger predictors than statistical deviation to study the effect of a suspect's gender. Some researchers study crime modeling and mapping in Easter Nairobi, Kenya (Mburu 2014). The author uses 1,422 solved crimes constituting 346 series of 10 different crime types. The author uses regression to study the relationship between the socioeconomic factors and crime events. The author also uses visualization techniques as a supplementary approach.

Another researcher has studied the value of associated terms in news entities to reveal that the underlying pattern is equivalent to that of citation networks from research paper publication networks (Tseng et al. 2012). The author employs associate weighted term analysis using modified *dice coefficients*. 85000 article are used in the analysis. Other researchers employ incident threading from news articles (Feng and Allan 2007). The authors propose two approaches. First, clustering articles and links using predefined rules that use tf-idf and similarity metrics. Second, proposing an optimization problem with simulated annealing for defining a global function. The latter performs better in link generation than the former.

Other researchers study and build a global crisis monitoring system for real-time news events named Europe Media Monitor (EMM) (Tanev, Piskorski, and Atkinson 2008). EMM collects 40,000 news events from 1400 news sources. The authors implement event extraction and clustering for crisis and violent news from different news sources. From each cluster EMM tries to extract only the main event information and aggregate them. For 31 clusters of violent and disastrous events, 23 were found (74 % precision). Others have also employed a news agent that learns and explains personalized news (Billsus and Pazzani 1999). The author implements a model to learn users' short-term and long-term interests from the feedback.

Bias

In this section, we will discuss the bias of media towards victims and perpetrators in news reporting. One researcher studies hundreds of news stories from 260 international and national newspapers involving 100 countries (Van Dijk 2013). The author attempts to infer if newspapers from different countries and political ideology would provide equally variable types of descriptions. The author studies bias in media such as newspapers, televisions and radios. For example, in one of the violent events that occurred in Amsterdam between police and squatters, the author concludes that the press forced the journalists to focus on negative violent consequences, that attributes only to the squatters.

Other researchers study the relationship between depiction of race such as tone and style in sports news (Mastro, Blecha, and Atwell Seate 2011). The authors conclude that black athletes are overrepresented as criminals compared to their white counterparts. Also, crimes committed by Black athletes (vs. Whites) are explained in more detail. Others also study the news value of African-American victims in media using metrics such as number of articles and words (Weiss and Chermak 1998). The author concludes news organizations provide less importance to these crimes and perpetuate the general public devaluation of African-American victims.

White Papers

In this section, we cite a few of the white papers that have been published. Some researchers discuss current basic and sophisticated methods deployed and developed by researchers (Groff and La Vigne 2002). The authors provide a clear picture of different methods with its advantages, disadvantages, and skills required for each method from Table 1 in this work. Another researcher also studies Smarter Public Safety approaches that apply optimization techniques, intelligence processing, and integrating/collaborating methods for fighting crime (Cotton 2012). The author discusses deployment and crime analytics software from IBM for a few police departments including Memphis, Tennessee and Vancouver Police Departments, Canada.

Future Challenges

While this short paper provides a survey of past and current research on crime, more research is needed for investigators and policy makers. Some future work that we will be investigating is the extraction of *patterns* based upon multiple, heterogeneous data attributes, such as crime news stories, user profiles, and social media. Analyzing such a diverse set of "big data" can not only aide in the prediction of crime, but perhaps provide law enforcement individuals with tools that can secure areas that are subject to future criminal activities, like senior-living homes and community centers. In

addition, we will investigate methods for extracting the *reasoning* for events for which users are interested. For example, if a user is interesting in swimming, they might be especially interested in safety at a local beach. Finally, we plan on implementing novel *visualization* techniques that will allow users to the study the evolution of patterns.

References

Ali, N. M.; Mohd, M.; Lee, H.; Smeaton, A. F.; Crestani, F.; and Noah, S. A. M. 2011. i-jen: visual interactive malaysia crime news retrieval system. In *Visual Informatics: Sustaining Research and Innovations*. Springer. 284–294.

Alruily, M. 2012. Using text mining to identify crime patterns from arabic crime news report corpus. *De Montfort University*.

Berger, C. R. 2001. Making it worse than it is: Quantitative depictions of threatening trends in the news. *Journal of Communication* 51(4):655–677.

Billsus, D., and Pazzani, M. J. 1999. A personal news agent that talks, learns and explains. In *Proceedings of the third annual conference on Autonomous Agents*, 268–275. ACM.

Bsoul, Q.; Salim, J.; and Zakaria, L. Q. 2013. An intelligent document clustering approach to detect crime patterns. *Procedia Technology* 11:1181–1187.

Cotton, B. 2012. Smarter public safety: Enabling and enhancing crime fighting capabilities. *A Frost & Sullivan Whitepaper*.

Doob, A. N., and Macdonald, G. E. 1979. Television viewing and fear of victimization: Is the relationship causal? *Journal of Personality and Social Psychology* 37(2):170.

Featherstone, C. 2013. Identifying vehicle descriptions in microblogging text with the aim of reducing or predicting crime. In *Adaptive Science and Technology, 2013 International Conference*, 1–8. IEEE.

Feng, A., and Allan, J. 2007. Finding and linking incidents in news. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, 821–830. ACM.

Gerber, M. S. 2014. Predicting crime using twitter and kernel density estimation. *Decision Support Systems* 61:115–125.

Green, E. G.; Staerklé, C.; and Sears, D. O. 2006. Symbolic racism and whites' attitudes towards punitive and preventive crime policies. *Law and Human Behavior* 30(4):435.

Groff, E. R., and La Vigne, N. G. 2002. Forecasting the future of predictive crime mapping. *Crime Prevention Studies* 13:29–58.

Gruenewald, P. J.; Freisthler, B.; Remer, L.; LaScala, E. A.; and Treno, A. 2006. Ecological models of alcohol outlets and violent assaults: crime potentials and geospatial analysis. *Addiction* 101(5):666–677.

Hirschi, T., and Gottfredson, M. 1983. Age and the explanation of crime. *American journal of sociology* 552–584.

Kahan, D. M. 1998. Social meaning and the economic analysis of crime. *The Journal of Legal Studies* 27(S2):609–622.

Kemshall, H. 1997. Sleep safely: crime risks may be smaller than you think. *Social Policy & Administration* 31(3):247–259.

LaFree, G. 1999. Declining violent crime rates in the 1990s: predicting crime booms and busts. *Annual Review of Sociology* 145–168.

Leong, K., and Chan, S. C. 2013. A content analysis of web-based crime mapping in the world's top 100 highest gdp cities. *Crime Prevention & Community Safety* 15(1):1–22.

Levitt, S. D. 2004. Understanding why crime fell in the 1990s: Four factors that explain the decline and six that do not. *Journal of Economic perspectives* 163–190.

Liao, R.; Wang, X.; Li, L.; and Qin, Z. 2010. A novel serial crime prediction model based on bayesian learning theory. In *Machine Learning and Cybernetics , 2010 International Conference on*, volume 4, 1757–1762. IEEE.

Listokin, Y. 2003. Does more crime mean more prisoners? an instrumental variables approach. *Journal of Law and Economics* 46(1):181–206.

Mastro, D. E.; Blecha, E.; and Atwell Seate, A. 2011. Characterizations of criminal athletes: A systematic examination of sports news depictions of race and crime. *Journal of Broadcasting & Electronic Media* 55(4):526–542.

Mburu, L. W. 2014. Modeling and mapping crime in eastern nairobi, kenya. *University of Heidelberg*.

Mohd, M. 2010. Optimal initial centroid in k-means for crime topic. *Journal of Theoretical and Applied Information Technology*.

Pease, K. 1999. A review of street lighting evaluations: Crime reduction effects. *Criminal Justice Press*.

Priest, T. 2004. The rise of middle eastern crime in australia. *QUADRANT-SYDNEY THEN VICTORIA THEN BALMAIN-* 48(1/2):9–17.

Pritchard, D., and Hughes, K. D. 1997. Patterns of deviance in crime news. *Journal of Communication* 47(3):49–67.

Radinsky, K., and Horvitz, E. 2013. Mining the web to predict future events. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, 255–264. New York, NY, USA: ACM.

Shingleton, J. S. 2012. Crime trend prediction using regression models for salinas, california. *Naval Postgraduate School*.

Tanev, H.; Piskorski, J.; and Atkinson, M. 2008. Real-time news event extraction for global crisis monitoring. In *Natural Language and Information Systems*. Springer. 207–218.

Tseng, Y.-H.; Ho, Z.-P.; Yang, K.-S.; and Chen, C.-C. 2012. Mining term networks from text collections for crime investigation. *Expert Systems with Applications* 39(11):10082–10090.

Van Dijk, T. A. 2013. *News analysis: Case studies of international and national news in the press*. Routledge.

Weiss, A., and Chermak, S. M. 1998. The news value of african-american victims: An examination of the media's presentation of homicide. *Journal of Crime and Justice* 21(2):71–88.