Crime Prediction Using Administrative Big Data and Machine Learning Algorithms

Gyeongseok Oh, M.A., M.S.[1]; Yan Zhang, Ph.D.[1]
[1]College of Criminal Justice, Sam Houston State University, USA

Sam Houston State University

College of SHSU Criminal Justice

Abstract

This study attempts to apply machine learning techniques to predict disorder crime at the neighborhood level using administrative big data. The records of 311 service request in 2012 and official crime reports in 2012, along with 2010 US. Census data were utilized in the analysis. According to the results of random forest analysis, numbers of 311 calls regarding uncivil use of property and trash removal, rates of minority population, levels of median household income were found to be the most important factors predicting disorderly behaviors. The findings of this study revealed that the use of administrative big data may be useful to predict neighborhoods' disorder crime incidents.

Introduction

Researches in social science depend largely on conventional methods of data collection, such as survey methods, official data, and field observations. However, as data storage technologies have been remarkably developed, researchers are able to access and utilize the huge amount of data scattered online or stored in the administrative database. Further, the data mining techniques and machine learning algorithms now can efficiently process big data into meaningful data.

The purpose of this study is to explore the possibility of utilizing big data and machine learning techniques to examine social phenomena.

First, the current study uses supervised random forest analysis to produce the process of classification of the physical disorder measures based on the level of criminal disorderly conduct.

Second, the most influential factors will be selected based on the results of random forest to predict the levels of disorderly conduct at the neighborhood level.

Third, neighborhoods' level of disorderly conduct will be predicted using geographically weighted regression and compared to the level of neighborhoods' actually observed disorderly conduct.

The results of this study may corroborate the Broken Windows Theory and suggest new methods to examine spatial distribution of crime using big data.

Data and Methods

Data: Reported physical disorders via 311 service request (n=594,809), official reports of disorder crime (n=7,211) collected by Houston Police Department from 2010 to 2012 will be utilized in this study. The number of call and incidents will be counted by each census blocks in the City of Houston.

Units of analysis: 1,301 census blocks were selected as units of analysis.

Measurements:

- Disorderly conduct: Fighting in a public place, Abusive language, Noise, Cruelty to animal, Fireworks ordinance, etc
- Physical disorders: Trash removal, Uncivil use of property, Graffiti removal, Physical decay of environment, Pollution
- Social disorganization: % of Black, Hispanic, and Asian, % of vacant house unit, Median household income

Analytic Methods:

- Machine learning – Random Forest
- Geographically Weighted Regression

Random Forest

Figure 1. Random Forest Importance Plot.

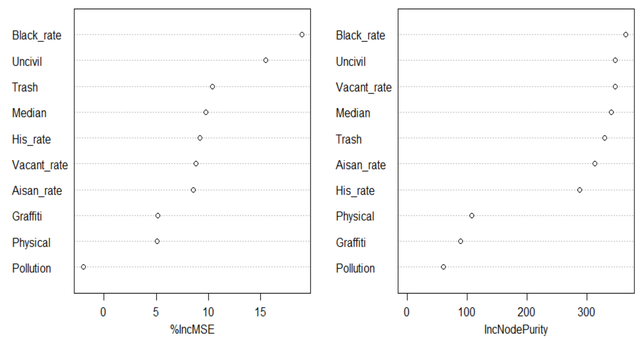
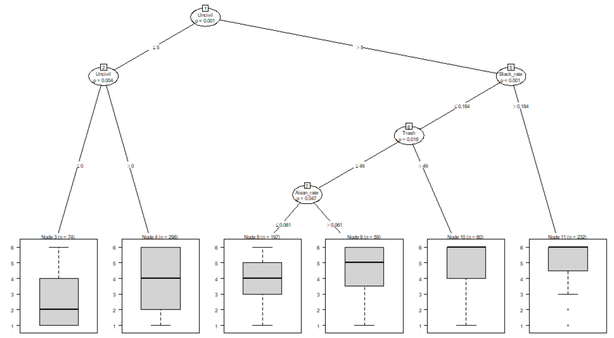


Figure 2. Decision Trees using Selected Variables.



Geographically Weighted Regression

Figure 3. Observed Disorderly Conduct

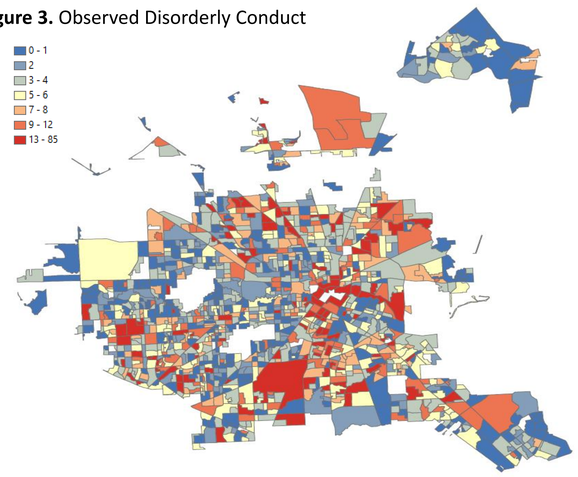
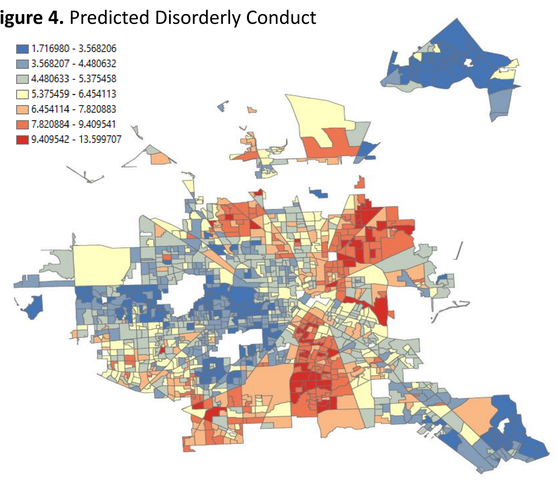


Figure 4. Predicted Disorderly Conduct



Results

The results of random forest model revealed that Uncivil use of property (cutoff 5), percentage of Blacks (cutoff 18.4%), Trash (cutoff 49), and percentage of Asians (cutoff 6.1%) are most influential covariates to classify the neighborhoods according to the level of disorderly conduct.

In terms of explained variance of outcome, 13.4% of the variance were explained in the random forest model while the GWR explained 8.4% of the variance of disorderly conduct.

Conclusions

The goal of this study was to investigate the effectiveness of administrative big data in explaining disorderly conduct in Houston, USA. Characteristics of each neighborhood and the number of calls for service regarding physical disorders of neighborhoods were employed to predict the level of disorderly crime incidents.

The findings of results revealed that Uncivil use of property, percentage of Blacks, Trash, and percentage of Asians/Hispanics successfully classify the neighborhoods according to the level of disorderly conducts. In specific, the use of administrative big data may be utilized to predict the needs of police services. Furthermore, the conventional regression analysis(GWR) also confirms that the significant predictors derived from random forest and decision tree model are also significantly associated with neighborhood-level disorder crime.

In conclusion, physical disorder measurements derived from 311 calls for service are valid to explain and predict the spatial distribution of disorderly conduct. We suggest government bodies and police administrators to utilize the real-time administrative data when making policies and decisions about allocating public resources across neighborhoods.

Future Directions

Future study should investigate the sources of difference between the results of OLS and machine learning methods. Additionally, policy implications and practical advices targeting government bodies should also be developed.

Contact Information

Gyeongseok Oh
College of Criminal Justice, Sam Houston State Univ
Email:gxo014@shsu.edu
Website:https://www.shsu.edu/~icc_www/_inc/dir.php?mode=view&item=264
Phone:936-294-4384

References

1. Gau, J. M., & Pratt, T. C. (2008). Broken windows or window dressing? Citizens'(in) ability to tell the difference between disorder and crime. *Criminology & Public Policy*, 7(2), 163-194.
2. O'Brien, D. T., & Sampson, R. J. (2015). Public and private spheres of neighborhood disorder: Assessing pathways to violence using large-scale digital records. *Journal of research in crime and delinquency*, 52(4), 486-510.
3. O'Brien, D. T., Sampson, R. J., & Winship, C. (2015). Ecometrics in the age of big data: Measuring and assessing "Broken windows" using large-scale administrative records. *Sociological Methodology*, 45(1), 101-147.
4. Sampson, R. J., & Raudenbush, S. W. (1999). Systematic social observation of public spaces: A new look at disorder in urban neighborhoods. *American journal of sociology*, 105(3), 603-651.
5. Skogan, W. G. (1986). Fear of crime and neighborhood change. *Crime and justice*, 8, 203-229.
6. Skogan, W. G. (1990). *Disorder and decline: Crime and the spiral of decay in American neighborhoods*. Univ of California Press.
7. Wen, M., Hawkley, L. C., & Cacioppo, J. T. (2006). *Objective and perceived neighborhood environment, individual SES and psychosocial factors, and self-rated health: An analysis of older adults in Cook County*, Illinois. Social science & medicine, 63(10), 2575-2590.
8. Wilson, J. Q., & Kelling, G. L. (1982). Broken windows. *Atlantic monthly*, 249(3), 29-38.

Acknowledgements

© Poster Template by Genigraphics® 1.800.790.4001 www.genigraphics.com