International Journal of Data Modelling and Knowledge Management
Vol. 3 No. 1 (January-June, 2018)

Crime Prediction Using Public Transportation Data and the Random Forest Algorithm

Peng Zhao*, Natasha Kapoor*, Manaswini Thakur*, Tyron Ty*, Edward Moskal*, Gregory Michaelson and Sylvain Jaume*****

**Saint Peter's University, Jersey City, NJ 07306*
***DataRobot Inc., Boston, MA 02110*

Authors emails
sjaume@saintpeters.edu

Abstract

In this paper, we propose a new methodology to predict burglary rates in a particular neighborhood using the Random Forest algorithm. Data sets of criminal records, transportation, and neighborhoods were employed and appropriate features were extracted and used in the classification problem. The criminal data set included geo-spatial locations (expressed in terms of longitude and latitude) of all the reported crimes in the city of Bogota along with the date of crime. The transportation data set included geo-location of each subway station and temporal measures of the number of passengers entering and exiting from the stations, and the neighborhood data included geographic division of different neighborhoods, number of houses, households, population, and house values. OnStreetMap data, consisting of the geo-spatial locations of different kinds of amenities (such as schools, police stations, and banks/ATMs) were also utilized along with the three aforementioned data sets. The Random Forest algorithm was tested with different models/feature sets to predict burglary rates. Our best performing proposed model obtained an accuracy up to 73.15% when predicting whether a specific area in the city will be prone to burglary or not.

Keywords: *Predictive analytics, Machine Learning, Data mining, Crime prediction, Public transportation, Big data, Big Data analytics, Smart cities, Latin America*

INTRODUCTION

Burglary is a social problem affecting the quality of life and the economic development of a society. Recent research has shown that burglary tends to be correlated with slower economic growth at both the national level and the local level; for example in cities (Cullen & Levitt, 1999) and metropolitan areas (Mehlum, Moene, & Torvik, 2005). Burglary related information has always attracted the attention of criminal law and sociology researchers. These scholars have focused on studying the behavioral evolution of burglaries and its relations with specific characteristics of the neighborhoods. Factors such as neighborhood features, social networks, and poverty issues have been related to burglary. Existing studies have also explored the relationships between burglary and socio-economic variables such as education (Ehrlich, 1975), ethnicity (Braithwaite, 1989), income level (Patterson, 1991), and unemployment (Freeman, 1999). Other scholars have focused on providing the evidence of significant concentrations of burglary at micro levels of geography with the geographical features (P. L. Brantingham & Brantingham, 1999; Weisburd & Green, 1994). It has been noted that crime events are correlated with the public transit system. Various studies have profiled transportation crime, occurring mostly in large cities such as New York (Clarke, Belanger, & Eastman, 1996), Chicago (Block & Davis, 1996), and Washington D.C.

(La Vigne, 1996). These cities with high population densities are more susceptible to crime. The majority of these crime events consist of less serious incivilities such as burglary and theft. Brantingham *et al.* (P. J. Brantingham, Brantingham, & Wong, 1991) made the hypothesis that public transit tends to provide opportunities for crime, since it potentially moves high-risk people around the city. One research group (Phillips & Sandler, 2015) has tested the hypothesis of whether public transit can spread crime in cities by using a strategy based on temporary closures of rail stations for maintenance in the Washington D.C. rail transit system. When closing one rail station, crime reduces by 5% at other stations on the same train line. This study demonstrates that crime is not random and it is worthwhile to study the relationships between public transit and crime. Another study has been conducted to determine the impact that rail transit has on crime, utilizing neighborhood crime data in Atlanta, Georgia (Millman, 2015). Using a random-effects estimator, the results indicated that neighborhoods in close proximity to rail stations experienced higher crime rates than neighborhoods that were further away from rail stations. Also, neighborhoods that are closer to rail stations with higher housing values experience higher crime rates. Recently, a number of authors (Letouze, Meier, & Vinck, 2013; Morozov, 2014; Toole, Eagle, & Plotkin, 2011; Wang, Rudin, Wagner, & Sevieri, 2013; Ferrara, De Meo, Catanese, & Fiumara, 2014; Traunmueller, Quattrone, & Capra, 2014; Song, Qu, Blumm, & Barabasi, 2010; Krumme, Llorente, Cebrian, Moro, et al., 2013; Singh, Murthy, & Gonsalves, 2010; Biau, 2012; Powers, 2011; Provost & Fawcett, 2013; Mohler, Short, Brantingham, Schoenberg, & Tita, 2012; Tarling & Morris, 2010; Caplan, & Kennedy, 2010; Bock, 2012) have taken advantage of the deluge of data collected in cities to verify these hypothesis based upon actual data (Jayaweera, Sajeewa, Liyanage, Wijewardane, Perera, & Wijayasiri, 2015). Bus and subway stations equipped with cameras, X-ray scanners, and automated gates allow for recording the flow of passengers going to and leaving from different areas of the city. Mining these very detailed geographic and temporal data points about the population of a city could help understand the relation between the flow of passengers and the occurrence of burglaries, and predict where burglaries are most likely to happen. These data-driven predictive models could then enable police departments to deploy preventative forces in the newly identified high crime intensity spots in the city. However, transportation burglary is still an understudied research topic in urban areas. In this paper, we use the Random Forest algorithm (Ho, 1995) and select a number of features, such as passenger count, location of subway stations and amenities, burglary events, real-estate values, and demographic features derived from available data sets to predict burglary rates in a particular neighborhood of Bogota. The main contributions of this paper are:

1. Testing the Random Forest algorithm with different models/feature sets to predict burglary rates.
2. Utilizing various data sets to extract features and build the models.
3. Discussions of the theoretical and practical applications of our approach.

The paper is structured as follows: In the Data section, we delineate the data sets utilized for our experiments. The Methods section describes the methodology employed for our experiments and in particular, the development of a Random Forest algorithm. The Results section presents how well the algorithm predicts burglaries in Bogota districts. This is followed by a discussion of our analysis and a conclusion.

DATA

Data Pop provided us with a number of data sets with information about crime records and transportation activity in the City of Bogota for a period of time ranging from 2007 to 2012.

crime records	Daily
trans portation	every 15 sec
neighborhood	Static

Figure 1: The figure shows the data sets that we used and their time granularity. Since the records in each data set have different time stamps, we created a mapping across the different data sets.

We describe in detail the data sets (from different sources) utilized in our study. The first data set contains crime records taken by the police of Bogota. Most crime records have a time stamp and a geo-spatial location. The crime records include crime cases from January 2007 to March 2013. The specific data attributes include crime ID, date of crime, geography (longitude, latitude, and address where the crime took place), and the crime type (homicide, burglary, vehicle lifting, shoplifting, theft of trade, personal injuries, etc.). We assessed the distribution of crime types and observed that burglary had the highest number of crime events. We also aggregated the total number of burglaries by each neighborhood.

An example of a criminal case record with the corresponding data attributes are summarized in Table

Attribute	Data
Crime ID	0000001
Date	06/04/11
Crime Type	Burglary
Latitude	-74.118
Longitude	4.615
City	Bogota
Address	KR 41C CL 3B 37

Figure 2: An example of a criminal case record based on the crime record data set.

2. The second data set contains transportation information, and records the passenger counts in different subway stations in the city every 15 seconds. The data set also consists of longitude and latitude coordinates of each subway station. (Note that the temporal granularity of each data set is different and a method to map one data set onto the other is necessary). The third data set utilized was the neighborhood data, which consisted of geographic information for the city of Bogota. The actual shape of the neighborhood was provided along with the surface areas of spatial structure. The shape of the neighborhood can be presented on the map and can be referenced by a unique ID. For each of the neighborhoods in Bogota, information about the number of houses, households, and population were provided.

Figure 4 represents the geo-spatial location of each subway station across the Bogota neighborhoods and the monthly passenger flow from 2007 through 2013. There are a total of 46,898 neighborhoods.

Information Type	Data
Bogota Neighborhood	Total number of houses
	Total number of households
	Total Population
Real Estate Value	House values

Figure 3: Basic statistics of the different neighborhoods in Bogota were used in our crime prediction study.

To complement the information contained in the data sets provided, we created a data set about the real estate value of different neighborhoods in the city. The data was extracted from fincaraiz.com.co. The real estate data set was created at the time of this study, i.e. fall of 2015, and was merged with the Bogota neighborhood data set. The neighborhood data is summarized in Table 3.

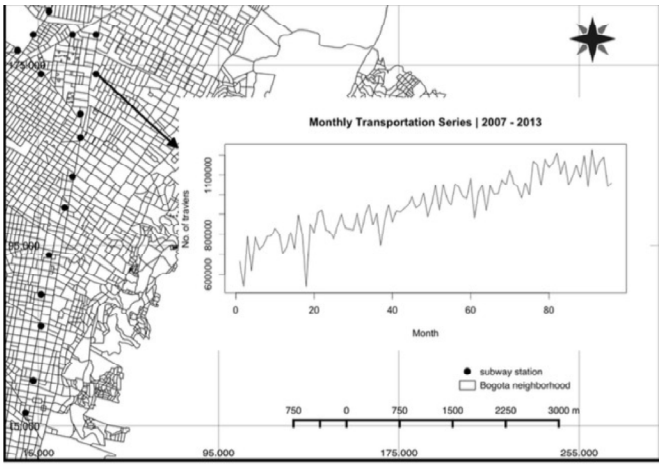

Figure 4: A map of the spatial and temporal structure of the transportation data. The time-series plot shows an increase of passengers entering and exiting the subway station over time from 2007 to 2013. This can indicate an increase in the likelihood of burglary events.

The OnStreetMap dataset consisted of geo-locations of different kinds of amenities. For each amenity, the data provided the amenity name, type, and geographic location of the amenity. Figure 5 represents the geo-spatial location of each amenity (such as schools, police stations, and banks/ATMs). The amenities were extracted from the OnStreetMap data and allocated to each neighborhood centroid with the distances calculated between them (see below).

In order to link all data sets provided from different sources, we mapped all spatial data sets to a Bogota neighborhood centroid. The centroid was calculated using the Bogota neighborhood data set. For example, we mapped each burglary event location (longitude and latitude coordinates) to the nearest neighborhood

Figure 5: A map of the geo-spatial location of different kinds of amenities, such as schools, police stations, banks/ATMs, in specific areas of the city. The geo-spatial relationships between each amenity and neighborhood centroid assisted with the prediction of the burglary rate in a specific area of the city.

centroid it belonged to by a measure of distance. For subways stations, we mapped each subway station location to the nearest neighborhood centroid it belonged to by a measure of distance. Similarly, we repeated the process for the amenity locations.

We further detected relationships between each amenity and neighborhood centroid in terms of crime prediction. A neighborhood that was highly commercialized experienced a higher level of burglary rates. The most challenging part of our study was to clean and map the data from different sources in order to build the models for predicting crime. To yield the best prediction accuracy, we introduce a modeling strategy using the Random Forest algorithm in the next section.

METHOD

We considered the problem of burglary classification as a binary problem. For each neighborhood of the city, we classified whether it will show a high or low level of burglary by performing a feature transformation. This feature transformation was achieved by splitting the burglary rate into two classes: a low burglary rate (class 0) and a high burglary rate (class 1). The threshold was set to the median burglary rate. The number of burglaries less than or equal to the median is assigned a class 0 and the number of burglaries greater than the median is assigned a class 1. Figure 6 is a heat map that visualizes the number of burglaries in different areas of the neighborhoods in Bogota. The highest number of burglaries is in Center Chapinero and the Northeast area respectively. Center Chapinero, located in downtown Bogota, has the highest real estate value in the city followed by the Northeast area. The West and South regions of the city have a low concentration of burglaries and are constant over time. Crime increased over time in the Northwest region from 2007 to 2012. Burglaries are more likely to happen in high population density areas like Center Chapinero, while residential areas with low population density experience less crime. We randomly split all data into training (80% of data) and testing (20% of data) sets. The models were built on 80% of the training data set and then used for predicting the dependent variable on the 20% testing data set. We trained the Random Forest algorithm with different models created on the training data set following a 5-fold cross validation strategy. We tested the different models created using the allocated testing data set for prediction of burglary rates. The purpose of this practice is to evaluate how the model will perform with new data.

As shown by Bogomolov *et al.* (Bogomolov *et al.*, 2014), the Random Forest algorithm constructs many

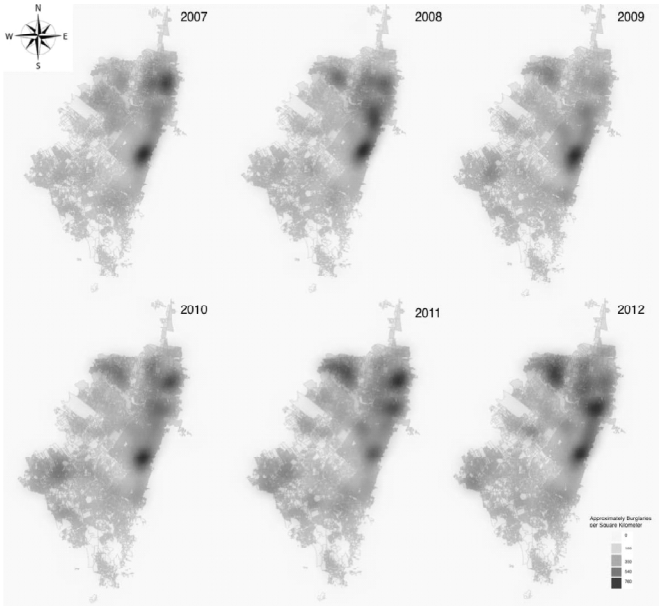

Figure 6: The heat map represents the number of burglaries in the city of Bogota from 2007 to 2012. The darker the area, the higher the number of burglaries during the year. The area that had the highest concentration of burglaries is Center Chapinero (central east part of the map), which is located in downtown Bogota.

classification decision trees and has several advantages over other classification algorithms. The Random Forest algorithm can produce a highly accurate classifier with a high speed learning process, and can handle many input variables without deleting any of them. It develops many decision trees based on random selection of data and random selection of variables. It provides the class of the dependent variable based on many trees. We also considered using a Support Vector Machine (SVM) algorithm (Meyer & Wien, 2015), which is often used for binary classification and developed by Cortes and Vapnik. The algorithm can be used for both linear and nonlinear classification and regression. For linear classification, a SVM constructs an optimal separating hyperplane between two different classes by maximizing the margin between the classes' closest points. The points lying on the boundaries are called support vectors. For non-linear classification, a kernel technique is performed when the data points are not linearly separable. We chose the Random Forest algorithm to construct our models for predicting crime, since it has several advantages compared to the Support Vector Machine algorithm.

We trained the Random Forest algorithm based on different models, which included a selection of features from the criminal case records, neighborhood, transportation, and OnStreetMap data sets. The calculated features included demographic, temporal measures, and geographical relationships. The output of the different models would be the binary

classification of the dependent variable, indicating whether a specific area in the city will be prone to burglary or not.

We defined the model based on the neighborhood data set as the baseline classifier. The neighborhood-based model included basic features, such as number of houses, number of households, population, and the price per square meter for a house. The neighborhood-based model is commonly used for prediction and is the most basic predictive model used in our method.

We recorded spatial relationships for burglary classification by using geographical features including nearest distance between each neighborhood centroid to an amenity type (such as subway station, school, police station, bank, or ATM). We defined this approach to be the geographic-based model.

We used all features from the transportation data set, which included spatial and calculated temporal measures. We created a model trained with the subset of these temporal measures that included the mean, median, and standard deviation of yearly, quarterly, monthly, and weekly passenger flow of the subway stations. We defined this approach and called it the standard deviation-based model.

Each tree in the Random Forest classifier is constructed by using the following algorithm:

1. Let the number of training cases be N, and the number of variables be M.
2. Let m be the number of input variables to be used to determine the decision at a node of the tree, where m must be less than M.
3. Select a training set for this tree by choosing N times with replacement from all N available training cases by taking a bootstrap sample. Use the rest of the cases to estimate the error of the tree by predicting their classes.
4. For each node in the tree, randomly select m variables on which to base the decision at that node. Calculate the best split based on these m variables in the training set.
5. Each tree is fully grown and not pruned.

Equation 1 and 2 represent the classification decision tree impurity measures in the Random Forest algorithm. Equation 1 defines impurity, which is measured by entropy. Entropy is a probability based measure used to calculate the amount of uncertainty.

$$i(A) = -\sum_{j} P(c_j)\log_2 P(c_j) \tag{1}$$

where P(.) is the fraction of samples at node A that are in class via j = 1,2,...,N. Gini index is measured by

$$i(A) = -\sum_{i \neq j} P(c_i)\,P(c_j) = \frac{1}{2}\left[1 - \sum_{j} P^2(c_j)\right] \tag{2}$$

Equation 2 is another way to define impurity and is a more general measure used in classification decision trees. The performance metrics that were used to evaluate the different models were accuracy, F1, and the area under the ROC curve (AUC). The metrics are used to measure the model's test prediction accuracy level. The performance metrics will be discussed in the results section of this paper.

RESULTS

In this section we report the experimental results obtained by the Random Forest algorithm. The performance metrics that were used to evaluate the different models' crime prediction performance were area under the ROC curve (AUC), accuracy, and F1. Area under the receiver operating characteristic (ROC) curve provides an insight to the predictive ability of the model. The receiver operating characteristic curve is a plot of 1-specificity against sensitivity. The false positive rate represents the x-axis and the true positive rate represents the y-axis. The steeper the slope of the curve, the higher the true positive rate. This will yield a higher AUC measure closer to one. An AUC measure of 0.5 indicates the model predicts at random.

We first used all features based on the neighborhood data set. These features consisted of number of houses, households, geographic division, and home price per squared meter. We defined this model to be the neighborhood-based approach, determined as the baseline classifier. To create the spatial relationships for burglary classification, we used geographical features such as nearest distance between each neighborhood centroid to a subway station, school, police station, and bank/ATM. We determined this model as a geographic based approach since the features are generated all from spatial relationships.

In order to understand the value added by the transportation data set, we compared the prediction performance of the Random Forest algorithm using all features with three different models trained with

1. Only the subset of features derived from the temporal relationships using the mean output of passenger flow (i.e. mean of yearly, quarterly, monthly, and weekly based output).
2. Only the subset of features derived from the temporal relationships using the median output of passenger flow (i.e. median of yearly, quarterly, monthly and weekly-based output).

3. Only the subset of features derived from the temporal relationships using the standard deviation output of passenger flow (i.e. standard deviation of yearly, quarterly, monthly, and weekly based output).

We define the three models as mean, median, and standard deviation based approaches, respectively. However, we only reported the standard deviation-based model in the performance results table, since it yielded the highest AUC performance measure.

The performance results on the crime prediction accuracy of the different models created with the training data set are reported in Table 7. The standard deviation-based model yielded the highest accuracy of 73.15%, when predicting burglary rates.

Predicti ve Model	Accuracy %	CI	F1 %	AUC
neighbor	67.94	(0.6563, 0.7019)	70.20	0.7273
geographic	72.91	(0.7069, 0.7504)	74.56	0.7291
std dev-based	73.15	(0.7094, 0.7528)	75.46	0.7315

Figure 7: A summary of the performance results for each of the different predictive models. The performance metrics that were utilized to evaluate the models were accuracy, the Confidence Interval (CI) at 95%, F1, and the AUC score (i.e. Area Under ROC Curve).

Comparing the different models that were built, the neighborhood-based model yielded the lowest accuracy of 67.94%. The model was determined as the baseline classifier. The features that were used in this model included basic statistics like number of houses, number of households, population, and the price per square meter for a house. The geographic-based model yielded an accuracy of 72.91%, which is a better performing model than the baseline classifier. The standard deviation-based model yielded an increase in accuracy of 5.21% when compared with the baseline classifier (73.15% vs. 67.94% accuracy). Figure 8 is a graphical representation of the area under the receiver operating characteristic curve. It visualizes the crime prediction accuracy of the standard deviation-based model compared to the neighborhood-based model, which was determined as the baseline classifier. The standard deviation-based model represents the curve line and the baseline classifier represents the diagonal straight line. The standard deviation-based model yielded an area under the ROC curve of 0.7315, which is better at predicting crime in certain areas of the city compared to the baseline classifier. The greater the area under the curve (closer to 1.0), the better the predictive ability of the model.

DISCUSSION

The results reported in the previous section indicate that the utilization of transportation data improved the level of accuracy when it came to predicting bur- glary rates in a particular neighborhood. The transportation data contains the geo-location (expressed in terms of longitude and latitude) of each subway station and the temporal measures of number of passengers entering and exiting from the subway stations.

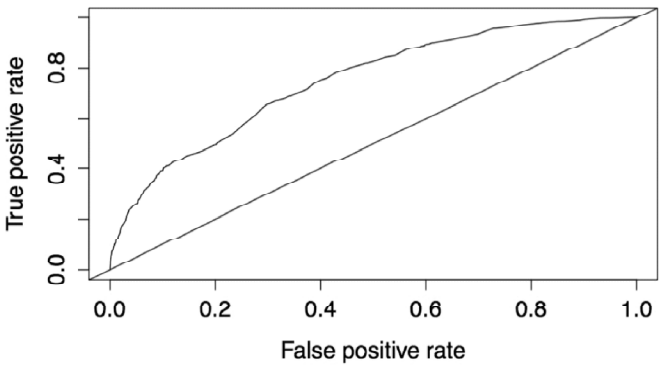

Figure 8: The graphical representation is a ROC curve plot, which illustrates the area under the receiver operating characteristic curve. The area under the curve for the best model is 0.7315, which is represented by a curve line. As shown in the figure, the steeper the slope, the higher the true positive rate.

By extracting human behavioral features from the transportation data set, the prediction accuracy of the models did improve when compared to the baseline classifier. The Bogota neighborhood data set contains information only about the number of houses and households and the population of the neighborhood for a given year.

The features extracted from geographical relationships between each amenity (such as schools, police stations, and banks/ATMs) and the neighborhood surrounding improved the model prediction accuracy significantly as well. This could provide an insight when for allocating police officers to high burglary geographic spaces. Geographical features in combination with temporal features are the best features to have as input in a model in order to predict burglary rates in a particular neighborhood, since a place with high population of passengers entering and exiting subway stations would have a variety of travelers visiting the place either on a yearly, quarterly, monthly, or weekly basis.

We have outlined and tested different predictive models to automatically classify at a 73.15% level of accuracy, whether a specific area in the city will be prone to burglary or not. The proposed approach could

have clear practical implications by informing police departments and city governments on how and where to invest their efforts and on how to react to burglary events with quicker response times. From a proactive perspective, the ability to predict the safety of a geographical area may provide information on explanatory variables that can be used to identify underlying causes of these burglary occurrence areas and therefore enable police officers to intervene in specific neighborhoods. The distinctive characteristic of our approach relies on the usage of features computed from transportation activity data combined with geographic and demographic information. Previous research efforts in criminology have tackled similar problems using background historical knowledge about crime events in specific areas, criminal profiling, and wide description of areas using socioeconomic and demo- graphic indicators. Our study provides evidence that aggregated data extracted from the transportation data set consists of relevant information to describe a geographical area in order to predict its burglary level. In particular, the features extracted from the transportation data set are dynamic and related to human activities.

Our study had several limitations due to the constraints of the data sets utilized. We only had access to the Bogota neighborhood data set for a given year, thus we cannot capture the demographic features over time. In addition, the transportation data set provides very limited information about the age and gender of travelers, which we could not utilize in our study.

CONCLUSION

We have proposed a new methodology to predict burglary rates in a particular neighborhood using the Random Forest algorithm. We have proposed a model that predicts the burglary rate at a 73.15% level of accuracy based on human behavioral features derived from transportation data in combination with geo- graphic information. We have shown that our approach significantly improved prediction accuracy when compared with using traditional statistical data about a neighborhood. Moreover, we have provided insights about the most predictive features, which are extracted from spatial and temporal data. We believe that our findings open the door to exciting avenues of research in computational approaches to deal with a well-known social problem such as burglary.

References

[1] Biau, G. (2012). Analysis of a random forests model. The Journal of Machine Learning Research , 13 (1), 1063–1095.

[2] Block, R., & Davis, S. (1996). The environs of rapid transit stations: A focus for street crime or just another risky place. Crime prevention studies, 6, 237–57.

[3] Bock, J. G. (2012). The technology of nonviolence: Social media and violence prevention. MIT Press.

[4] Bogomolov, A., Lepri, B., Staiano, J., Oliver, N., Pianesi, F., & Pentland, A. (2014). Once upon a crime: towards crime prediction from demographics and mobile data. In Proceedings of the 16th international conference on multimodal interaction (pp. 427–434).

[5] Braithwaite, J. (1989). Crime, shame and reintegration. Cambridge University Press.

[6] Brantingham, P. J., Brantingham, P. J., & Wong, P. S. (1991). How public transit feeds private crime: notes on the Vancouver skytrain experience. Security Journal, 2 (2), 91–95.

[7] Brantingham, P. L., & Brantingham, P. J. (1999). A theoretical model of crime hot spot generation. Studies on Crime & Crime Prevention.

[8] Caplan, J. M., & Kennedy, L. W. (2010). Risk terrain modeling manual: Theoretical framework and technical steps of spatial risk assessment for crime analysis. Rutgers Center on Public Security.

[9] Clarke, R. V., Belanger, M., & Eastman, J. (1996). Where angels fear to tread a test in the new york city subway of the robbery/density hypothesis. Preventing mass transit crime, 6, 217–36.

[10] Cullen, J. B., & Levitt, S. D. (1999). Crime, urban flight, and the consequences for cities. Review of economics and statistics, 81 (2), 159–169.

[11] Ehrlich, I. (1975). On the relation between education and crime. In Education, income, and human behavior (pp. 313–338). NBER.

[12] Ferrara, E., De Meo, P., Catanese, S., & Fiumara, G. (2014). Detecting criminal organizations in mobile phone networks. Expert Systems with Applications, 41 (13), 5733–5750.

[13] Freeman, R. B. (1999). The economics of crime. Handbook of labor economics, 3, 3529–3571.

[14] Ho, T. K. (1995). Random decision forests. In Document analysis and recognition, 1995. proceedings of the third international conference on (Vol. 1, pp. 278–282).

[15] Jayaweera, I., Sajeewa, C., Liyanage, S., Wijewardane, T., Perera, I., & Wijayasiri, A. (2015, April). Crime analytics: Analysis of crimes through newspaper articles. In Moratuwa Engineering Research Conference (MERCon), pp. 277-282. IEEE.

[16] Krumme, C., Llorente, A., Cebrian, M., Moro, E., et al. (2013). The predictability of consumer visitation patterns. Scientific reports, 3.

[17] La Vigne, N. G. (1996). Safe transport: Security by design on the washington metro. Preventing mass transit crime, 6, 163–197.

[18] Letouzé, E., Meier, P., & Vinck, P. (2013). Big data for conflict prevention: New oil and old fires. New

Technology and the Prevention of Violence and Conflict, New York: International Peace Institute.

[19] Mehlum, H., Moene, K., & Torvik, R. (2005). Crime induced poverty traps. Journal of Development Economics , 77 (2), 325–340.

[20] Meyer, D., & Wien, F. T. (2015). Support vector machines. The Interface to libsvm in package e1071.

[21] Millman, S. (2015). Determining the effects of rail transit on crime.

[22] Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., & Tita, G. E. (2012). Self-exciting point process modeling of crime. Journal of the American Statistical Association.

[23] Morozov, E. (2014). To save everything, click here: The folly of technological solutionism. PublicAffairs.

[24] Patterson, E. B. (1991). Poverty, income inequality, and community crime rates. Criminology , 29 (4), 755–776.

[25] Phillips, D. C., & Sandler, D. (2015). Does public transit spread crime? evi- dence from temporary rail station closures. Regional Science and Urban Economics , 52 , 13–26.

[26] Powers, D. M. (2011). Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation.

[27] Provost, F., & Fawcett, T. (2013). Data science for business: What you need to know about data mining and data-analytic thinking. O'Reilly Media, Inc..

[28] Singh, S. R., Murthy, H. A., & Gonsalves, T. A. (2010). Feature selection for text classification based on gini coefficient of inequality. FSDM, 10, 76–85.

[29] Song, C., Qu, Z., Blumm, N., & Barab´asi, A.-L. (2010). Limits of predictability in human mobility. Science , 327 (5968), 1018–1021.

[30] Tarling, R., & Morris, K. (2010). Reporting crime to the police. British Journal of Criminology, 50 (3), 474–490.

[31] Toole, J. L., Eagle, N., & Plotkin, J. B. (2011). Spatiotemporal correlations in criminal offense records. ACM Transactions on Intelligent Systems and Technology (TIST), 2 (4), 38.

[32] Traunmueller, M., Quattrone, G., & Capra, L. (2014). Mining mobile phone data to investigate urban crime theories at scale. In Social informatics (pp. 396–411). Springer.

[33] Wang, T., Rudin, C., Wagner, D., & Sevieri, R. (2013). Learning to detect pat- terns of crime. In Machine learning and knowledge discovery in databases (pp. 515–530). Springer.

[34] Weisburd, D., & Green, L. (1994). Defining the street-level drug market.