A Theory-Driven Algorithm for Real-Time Crime Hot Spot Forecasting

YongJei Lee (ylee@uccs.edu, School of Public Affairs, University of Colorado)
SooHyun O (osn@mail.uc.edu, School of Criminal Justice, University of Cincinnati)
John E. Eck (john.eck@uc.edu, School of Criminal Justice, University of Cincinnati)

With the advent of computer mapping and geographic information systems, real-time crime hot spot forecasting has become an important interest to policing (Groff and La Vigne, 2002). Nevertheless, crime hot spot forecasting presents challenges to policing. There is a high volume of crime hot spot misclassifications and a lack of theoretical support for existing forecasting algorithms. Additionally, many algorithms are complex and proprietary. Thus, they require expensive contracts with private vendors, and they do not have the transparency necessary. Transparency is particularly important as police use of stop-question-frisk tactics in hot spots has been linked to racially disparate treatment of non-whites, and excessive interference in the daily lives of people who are not involved in crime.

To fill these challenges, we created an algorithm based on two theories. We use population heterogeneity theory to find places that are consistently experiencing crimes in the forecasted month. This narrows the focus to places with consistently high levels of crime. Second, we apply a state dependence theory to address short-term elevated risk at these places. We implement this algorithm in Microsoft-Excel, making it extremely simple to apply and completely transparent. It does not need highly specialized expertise to implement, can be modified by agencies as needed, and can be examined should its application be linked to racially disparate outcomes. Experiments show high accuracy and high efficiency in hot spot forecasting. These results also demonstrate how basic theories could lead to build a sound algorithm for hot spot forecasting. We finally discuss the implications of this simple, theory-driven forecasting method for policing practices.

Population heterogeneity framework is quite common in explanations of criminality. It assumes among people there are different populations of people, based on their enduring characteristics that predict crime. The clearest example is offender's gender: boys and men are more likely to be deviant than girls and women. Variations in criminality are partially explained by this variation in the populations of people (Nagin and Paternoster, 2000).

Population heterogeneity can be applied to places. For example, among bars there are characteristics that are associated with high violence and characteristics associated with low violence (Madensen and Eck, 2008).

In our algorithm, we employed population heterogeneity framework to classify places where the hot spot forecasting is consistently successful over months. First, we calculate the probability of the occurrence of crime in the target month based on the Poisson distribution of crime in prior months. If the Poisson probability is greater than a specific threshold *and* the forecasted month experiences crime, we consider the forecast has produced a true positive case.

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

The more true positive cases over months, the more a place is consistently experiencing crime. Thus, our algorithm selects places with high true positives over the entire study period, and screens out places with no crime, random crime, and low probabilities of crime. Some of these persistent true positive places will turn into hot spots in the forecasted month.

Researchers have used state dependence to study criminal victimization and repeat victimization (Nagin and Paternoster, 2000; Nagin and Land, 1993). One explanation for repeat victimization is that a crime elevates the victim's chances of further victimization in the short term. There is considerable evidence for this (Farrell, 2005), particularly for burglary. Consider the following example. Person A has a very low chance of having their house broken into, and this is true of all other houses in the population. By some fluke (say the teenage daughter in the family chances to meet a ne'er-do-well at the high school football game and he learns that her house has valuable items and will be vacant on Friday afternoon) the house gets broken into. After this burglary, the chances of another have increased. Unlike the proverbial lightning bolt, once struck the chances of being struck again go up. For this reason, state dependence model is sometimes called a boost model (Pease, 1998).

Among the places selected by population heterogeneity portion of our algorithm, some will have exceptionally high chances of crime in the next month, because of the high crime levels in the current month. So we use the state dependence part of our algorithm to identify the most currently active members of the population of the high crime places.

Methods

Data Set

We use the calls-for-service (CFS) records provided by the Portland Police Bureau (PPB) in Oregon for the period of March 1, 2012 through February 28, 2017. CFS records include all CFS, street crime, burglary, and theft of auto cases. National Institute of Justice (NIJ) initially released these data set for the period of March 1, 2012 through July 31, 2016, and then released updated PPB's CFS data over a six-month period for the real-time crime forecasting challenge.[1]

Geo-processing

We use the X-Y coordinates to geocode all types of CFS records on the map document using ArcGIS 10.3. We created a grid system covering the entire area of Portland. Each grid cell has 500 feet on a side to contain approximately one block inside the grid cell. Some grid cells located at the edges of the city were trimmed to have their shape within the boundaries of the study area. Finally, we created 17,163 grid cells with their size varying between 7.31 ft^2 and 250,000 ft^2 spatially overlaid on the Portland.

We spatially joined all CFS records on the grid system and counted the number of CFS records per grid cell. We found that 11,548 grid cells or 67.3 percent of the entire grid cells experienced at least one crime during the study period.

[1] https://www.nij.gov/funding/Pages/fy16-crime-forecasting-challenge-document.aspx.

Models

To operationalize population heterogeneity framework, we calculate the Poisson probability of crime for every month based on the distributions of crime in the past 12-months. If the Poisson probability of crime is greater than .5 (for burglary, we use 0), and crime occurred in the forecasted month, we coded the grid cell as 1 (otherwise 0). A formula below shows how we implemented this algorithm in an Excel spreadsheet.

fx: IF((1-POISSON.DIST(0, AVERAGE($Number\ of\ Crimes_{i,t-12}, \ldots, Number\ of\ Crimes_{i,t-1}$), TRUE))> 0.5, IF($Number\ of\ Crimes_{i,t-1}$> 0, 1, 0), 0)

i = is the grid cell id
t = the forecasted month

For each grid cell in each month, this formula returns either 1 or 0. By averaging these dichotomous values over all months in the study period, we obtained an average true positive value for each grid cell. Using these true positive values, we sorted the grid cells from the most predictable to the least. This is the population heterogeneity part.

For the state dependence part, we used the number of crimes in the just previous month to assess the elevated risk on the grid cell toward forecasted month. Thus, after we sorted the grid cells by their true positive values, then we sorted by the number of crimes in the prior month from the highest to the lowest. If there is a tie in true positive values among the grid cells, we placed more weight on the grid cells with more crimes in the previous month. To meet the NIJ's criteria of the total forecasted hot spot area, we select 83 grid cells from the top row in Excel.[2]

After we applied our algorithm using Excel, we re-plotted all CFS cases on a map. Figure 1 shows forecasted crime hot spots for all CFS. The map also shows the density of crime for the entire study period in a color ramp from blue (low density) to red (high density). Boundaries of each police district are also provided in a separate layer with district number.

[2] NIJ mandated the total forecasted area of hot spots must be in the range between 0.25 mi^2 and 0.75 mi^2.

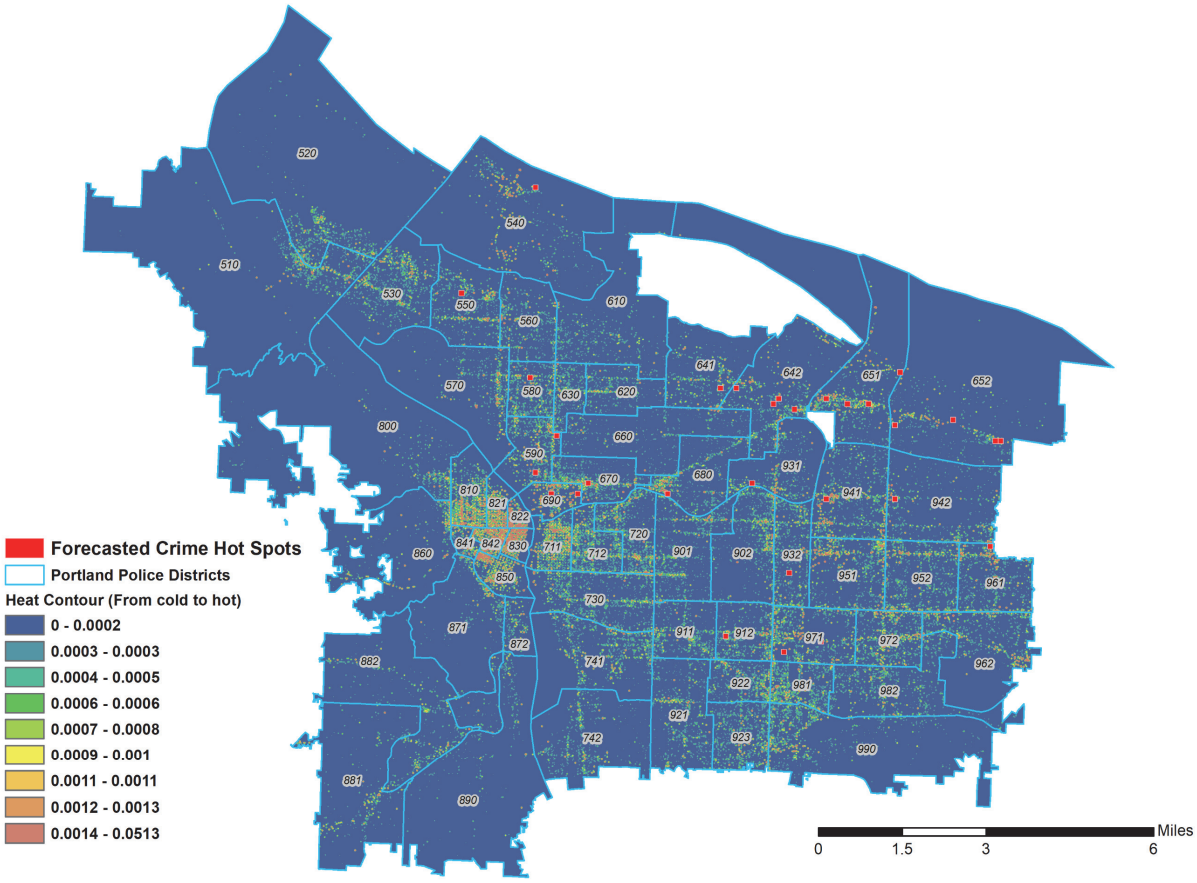

Fig 1. Map of forecasted crime hot spots for all CFS and crime density

This section provides numeric evidence of the accuracy and efficiency of our hot spot forecasting algorithm. To evaluate the efficiency of our algorithm we use the Prediction Efficiency Index (PEI). The PEI measures how well a forecasting algorithm does compared to how well it could have done (Hunt, 2016). Simply put, it is a ratio of the number of crimes in forecasted hot spots to the number of crimes in actual hot spots. The formula for the PEI is provided as follows.

$$PEI = \frac{\frac{n}{N}/\frac{a}{A}}{\frac{n^*}{N}/\frac{a}{A}} = \frac{\frac{n}{N}}{\frac{n^*}{N}} = \frac{n}{n^*}$$

Where,
n = is the number of calls-for-service forecasted
N = the total number of calls-for-service
a = the forecasted area
A = the total area
n^* = the maximum obtainable number of calls-for-service for the amount of area forecasted (a).
The PEI can be expressed as a proportion (varies from 0 to 1) or a percent (varies from 0 to 100).

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Table 1. Accuracy and prediction efficiency indices for crime hot spot forecasting algorithm (%)

Type of Crime	Forecasting Period		
	1 month	2 months	3 months
All calls-for-service	91.2	92.5	92.0
Hot spots	72.9	71.8	74.1
Burglary	20.9	19.5	29.5
Hot spots	23.5	8.2	15.3
Street crime	82.2	85.9	87.9
Hot spots	63.5	61.2	65.9
Theft of auto	37.5	47.7	59.3
Hot spots	22.4	43.5	44.7

Table 1 shows the performance of our forecasting algorithm. Each column under 'Forecasting Period' shows the duration of the forecast. For '2 months', for example, we forecasted hot spot locations and the number of crimes in hot spots in the following 2 months. The first row in each cell shows how many crimes are in the forecasted hot spots relative to crimes in actual hot spots. This is equivalent to the PEI. The second row shows how accurate the algorithm is at forecasting the locations of hot spots.

Apparently, our forecasting algorithm performs best for all calls-for-service hot spots. We find that about 72 to 74 percent of hot spot locations are correctly forecasted, and it captures 91 to 92.5 percent of crimes in actual hot spots. Street crime forecasting also shows high accuracy of hot spot locations as well as high efficiency.

Theft of automobiles and burglary show relative low performance in accuracy and efficiency. However, considering the number of vehicle thefts and burglaries in the data set (see Appendix), our forecasting algorithm still provides sound performance. For example, theft of automobiles is a type of crime which has high mobility and seems to be randomly happening elsewhere. Our forecasting results, however, confirm that theft of auto does not happen by chance. It is a type of crime which occurs at particular places more frequently than other places, thus making theft of auto predictable by our algorithm[3]. We could apply the same explanation to burglary cases. One interesting finding is that long-term forecasting (here, 3 months forecasting) is more accurate and efficient for the other types of crime than it is for burglary.

In this study, we operationalize two different theories to forecast crime hot spots. We use population heterogeneity framework to select grid cells where hot spot forecasting is consistently successful over months. Among these grid cells, we employ state dependence model to assess an

[3] We were awarded $10,000 from NIJ for achieving the highest forecasting efficiency for theft of auto hot spots.

enhanced state of risk on a grid cell. Findings from experiments demonstrate that our theory-driven algorithm forecasts hot spots with high accuracy and efficiency.

We would like to emphasize the simplicity and convenience of our algorithm. Once all the data was organized, we only used Microsoft-Excel to calculate Poisson probabilities, evaluate the state of risk, and forecast hot spots. Many police officers and most crime analysts should be able to implement this procedure. It might be possible to pay for a marginally more accurate and efficient algorithm, but one has to ask whether the increase is worth both the implementation and ongoing costs involved. Further, if the public is concerned about how the forecasts are created, the algorithm can be opened to public examination.

References

Farrell, G. (2005). Progress and prospects in the prevention of repeat victimization. *Handbook of crime prevention and community safety*, 143-170.

Groff, E. R., & La Vigne, N. G. (2002). Forecasting the future of predictive crime mapping. In *Analysis for Crime Prevention, volume 13 of Crime Prevention Series.* Monsey, NY: Lynne Rienner Publishers.

Hunt, J. M. (2016). *Do crime hot spots move? Exploring the effects of the modifiable areal unit problem and modifiable temporal unit problem on crime hot spot stability* (Doctoral dissertation, American University).

Madensen, T. D., & Eck, J. E. (2008). Violence in bars: Exploring the impact of place manager decision-making. *Crime Prevention and Community Safety*, *10*(2), 111-125.

Nagin, D. S., & Land, K. C. (1993). Age, criminal careers, and population heterogeneity: Specification and estimation of a nonparametric, mixed Poisson model. *Criminology*, *31*(3), 327-362.

Nagin, D., & Paternoster, R. (2000). Population heterogeneity and state dependence: State of the evidence and directions for future research. *Journal of Quantitative Criminology*, *16*(2), 117-144.

Pease, K. (1998). Repeat victimisation: Taking stock (Crime detection and prevention series paper 90). *London: Home Office*.

Appendix. Descriptive statistics of all types of CFS

Type of Crime	Statistics	Per Month	Per Grid Cell	Per Month Per Grid Cell
Calls-for-Services	Mean	16,172.44	88.23	1.40
	Median	16,158	37	0
	Min	11,869	1	0
	Max	22,878	4,246	118
	St. Dev	2,214.07	183.75	3.32
	Total			1,018,864
Street Crime	Mean	2,875.06	19.92	0.32
	Median	2,780	7	0
	Min	1,988	1	0
	Max	3,665	1,331	41
	St. Dev	427.20	47.60	1.00
	Total			181,129
Theft of Auto	Mean	578.67	4.75	0.08
	Median	511	3	0
	Min	338	1	0
	Max	1,400	105	10
	St. Dev	193.76	5.69	0.31
	Total			36,456
Burglary	Mean	399.79	3.40	0.05
	Median	404	2	0
	Min	288	1	0
	Max	583	51	6
	St. Dev	49.89	3.25	0.25
	Total			25,187