Visualizing Predictive Policing in Los Angeles

Stop LAPD Spying Coalition*

June 7, 2017

Abstract

PREDPOL is an algorithm published in 2011 that analyzes crime reports and generates heat scores *on a block-by-block basis. LAPD uses PREDPOL by ensuring that the blocks with the highest heat scores have a disproportionately high police presence; these blocks are labeled* hotspots*. Data-driven, algorithmically-generated policies, like all LAPD policies, merit transparency and public scrutiny. To facilitate analysis of PREDPOL, the daily locations of hotspots generated between 2012 and 2015 are calculated here.*

I. Introduction

LAPD is by far the largest recipient of public funding from The City of Los Angeles[1]; even minor changes to LAPD policies have massive implications both for public health directly, and the operations of every other publicly-funded institution indirectly. Crime forecasting, has a very large effect on the day-to-day operations of LAPD. It is LAPD policy that officers spend a certain (WHAT PERCENTAGE) percentage of their off-mission time patrolling according to crime forecasts.

PREDPOL was first proposed in 2011 in [2]. The development of Predpol, the theories on which it's based, and its implementation in LAPD policy are described in detail in (CITE OUR OWN PUBLICATION).

PREDPOL assumes a statistical model of crime generation called ETAS (*Epidemic-Type Aftershock Sequence Model of Interacting Triggered Seismicity*). PREDPOL assigns a likelihood of a "crime" occurring on a given day, on a given city block.

The details of the PREDPOL algorithm are fully described in the original work [2], but were re-derived by K. Lum in [5] with a focus on transparency. K. Lum's implementation of PREDPOL precisely follows the details described in that work, and was graciously and supportively made available for use on this project.

The historical crime report data used as input to PREDPOL for this project was obtained through Los Angeles Open Data[2]. Hotspots generated for this work aim to display the true locations of hotspots for each of the dates analyzed, between 2012 and 2015.

II. Model Background

i. Broken Windows and Repeated Crimes

The core idea on which most predictive policing is built is that of Broken Windows Policing, developed by (XYZ PERSON), who (HAS XYZ CONNECTION TO PREDPOL's GENEOLOGY).

Part of this theory claims that certain offenses are, by their nature, likely to be repeated in the same location where they occurred. The

*A volunteer-led, grassroots community organizing network operating out of LA Community Action Network in Los Angeles.

[1] LAPD does not just receive more funding from the City of Los Angeles than any other institution, but actually received the *majority* of its $4.85 billion *"unrestricted"* budget in the 2016-2017 fiscal year. The 2016-2017 LA City Budget: http://cao.lacity.org/budget/summary/2016-17BudgetSummaryBooklet.pdf

[2] Los Angeles Open Data: https://data.lacity.org/

crimes that PREDPOL considers to adhere to this model are (1) "burglary", (2) "car theft", and (3) "in-car theft". The idea is that there are incentives to commit these crimes a second time, such as already having researched a target's vulnerabilities.

In practice, crime reports are labeled with a crime code, and only certain crime codes are considered as part of predictive policing methods. The specific crime codes used by PREDPOL are described in detail in Section III.

It may be worth noting that crimes may be chronically misclassified by the police who record them, as described in [10]. This was not considered as part of this work.

The idea that certain conditions may elevate the risk of a crime occurring is not necessarily objectionable. It is unclear, however, whether *policing* is a desirable solution to conditions of elevated risk. This work aims to provide tools to question that assumption.

ii. ETAS

ETAS (*Epidemic-Type Aftershock Sequence Model of Interacting Triggered Seismicity*) was first proposed to model the emergence of aftershocks following an earthquake.

The ETAS model assigns a likelihood of an event occurring at a given time and location. In the context of earthquakes, ETAS assigns a likelihood of an earthquake, which may or may not be an aftershock, occurring at a given location, at a given time. In the context of PREDPOL, an event is a crime, and an "*aftershock*" is by analogy a *follow-up* crime.

The ETAS model functions differently in different contexts, where the context is described by parameters that dictate the likelihood of occurrences. Specifically, ETAS considers a *base rate* of likelihood of an event occurring, as well as the likelihood of a *child* (follow-up) event. Several parameters also describe the decreasing likelihood of a follow-up event at increasing distances away from and lengths of time after *parent* events.

If all model parameters were known, such as in a simulation, then the likelihood of an event occurring at a given location and point in time, given all past events, is known.

iii. Expectation Maximization

In practice, model parameters are not known. Forecasting is estimating parameters that will generate accurate likelihoods of future events. A *forecast* is the likelihood of a future event according to a parameterized model.*Forecast* may also refer to the *most likely* event.

Which estimated parameters will produce the most accurate possible predictions of future events? There is no way to know. Assuming an estimate of parameters, the probability of an event can be determined according to a model. Without knowing the true parameters of a model, there is no way to know how unusual a real or hypothetical sample may be. The true parameters, and the true data generating process in general, are unknown.

The "best" estimate is often the parameterization under which the data the has already occurred would be least unusual. Finding such an estimate is called *expectation maximization*.

iv. Forecasting in ETAS: EM Cycling

The relevant parameters of ETAS are the *base rate* and the *probability of a follow-up event*. To determine how unusual historical data is according to some estimate, one could compare the estimated likelihood of a follow-up event to, say, the proportion of events that had follow-up events. However, this, too is typically not known, because it is not known which events are follow-up events!

A standard procedure for estimation within a model like ETAS is called *EM Cycling*. This method makes an initial guess of which events followed up from which other events; given that guess, the base rate can be estimated as the average number of non-follow-up events that occur in a given area and length of time, and the likelihood of a follow-up event can be estimated as the percentage of events labeled as follow-up events (M-Step). Then, given those estimates, the labels of *parent* and *child* are exchanged among the events so that the likeli-

hood of the entire sequence of events is most probable (E-Step). These two steps ("E" and "M") are iterated many times. Under certain conditions, EM Cycling is guaranteed to converge to a good estimate of the true model parameters.

In the context of EM Cycling, the labels of *parent* and *child* on a dataset are considered parameters. The "best guess" of those labels are the labels under which the data is least unusual.

v. PREDPOL

The PREDPOL algorithm aims to estimate the parameters of the ETAS model and forecast the likelihood of new events.

PREDPOL uses historical crime report data to determine which parameters would make that data least unusual. It accomplishes this through EM cycling. With the parameters in place, PREDPOL provides the likelihood of a new event, which is assigned by ETAS and can be called a *forecast*.

PREDPOL considers a variant of ETAS where crimes can only be considered *follow up* crimes if they occur on the same city block. A block is defined as a 500 by 500 foot *cell*. The ETAS parameter denoting the "probability of a follow-up event occurring a given time after and distance away from an event" is not just a number, but a distribution that decays in time and space, called a *kernel*. In PREDPOL, the kernel of follow-up events is zero outside of a block, and only depends on time within a block.

vi. Heat Scores and Hotspots

Finally, PREDPOL's forecast of the likelihoods of follow-up events (crimes following up from previous reported crimes) occurring at a given location are determined for all blocks (cells) across a division. Each division seems to generate its own hotspots, as explained in (CITE OUR OWN WORK. CITE POLICE TWEETS. CITE K. LUM'S SOURCE).

Given the likelihood of a follow-up event at a given location, the expected number of follow-up events occurring at that location during a given window is called a *heat score*. Usually it is, of course, most likely that no crimes will be committed on a given block within a given day.

The blocks with the highest heat scores in a given division are considered *hotspots*.

There is evidence that LAPD labels 10-20 hotspots per division, according to (CITE OUR OWN WORK), [3], and [6].

III. Data

PREDPOL purports to use historical crime data to forecast future crimes. But what is crime data?

The *crime data* used by PREDPOL is the record of *crime reports* having occurred before the algorithm is executed. In PREDPOL, "each [crime] is associated with a reported time window over which it could have occurred, often a few hour span" [2]., as well as a location.

The crime data for this project was obtained through Los Angeles Open Data[3]. Each crime report is listed with a *crime code*, a *crime code name*, an address or cross-street, geocoded coordinates, an LAPD division, information about the conditions of the report, and information about the result of the police intervention.

For this work, the date of occurrence, geocoded coodinates, LAPD division, and crime code determined whether and how a crime was inputted into PREDPOL.

i. Crime Codes

PREDPOL does not assert that *all* crime is generated according to the pattern of the ETAS model. The crimes considered are restricted to three types, which are (1) *burglary*, (2) *car theft*, and (3) *in-car theft* [3].

[3] Los Angeles Open Data hosted a dataset called "Crime Data 2012-2015" that was used for this project. In April 2017, this dataset was updated to include more recent reports. The data can currently be found in a dataset called "Crime Data from 2010 to Present", which is updated approximately once per week, and contains data through the previous month.

By compiling the unique crime code names from the dataset, however, we were able to manually identify which crime codes are likely to be used as input to PREDPOL. The crime reports used for this work are listed in Table 1, with the number of occurrences between 2012 and 2015 listed for reference.

ii. LAPD Division

Each reported crime in the dataset lists the LAPD division associated with the crime. The dataset does not specify whether the division listed is the division that processed the crime report, or the division in whose geographic jurisdiction the reported crime reportedly occurred.

During this work, it was noted that several crimes were listed with coordinates far outside of the jurisdiction of the division listed. Those crime reports often occurred on intersections whose street names are shared by intersections in other divisions. For example, many crimes listed with "Central Division" at 4th St. and Hill St. (in Downtown LA) were listed with coordinates at 4th St. and Hill in Santa Monica. This is a clear geocoding error, and it proves that divisions are associated with crimes at the time of recording, and not retroactively based on location.

For the visualizations in this work, LAPD division outlines were obtained from the LA Times [9].

IV. Methods

i. Binning LA

As described in Section II, PREDPOL only considers *follow-up* events that occur on the same city block, or 500 by 500 foot cell. Within a cell, all events are considered to occur at the same location.

As such, the raw coordinates in the LA Open Data crime report dataset had to be grouped according to which 500 by 500 foot cell they fell in.

For this work, the city of Los Angeles was divided evenly into 500 by 500 foot cells. Crimes were then assigned to the cell in which their coordinates fell.

ii. PREDPOL Implementation

The implementation of PREDPOL used for this work was written in Python by K. Lum as part of work related to [4] and [5].

The input to PREDPOL is the collection of all crime reports whose crime codes correspond to the types of crimes reportedly considered by PREDPOL (see Section III or Table 1). Each crime report is associated with a bin number. Each bin corresponds to a 500 by 500 foot cell, as described in Section IV-i.

The output generated by PREDPOL is the heat scores for each bin on each day. The heat score of a bin on a given day is the expected number of crimes to occur that day within that bin, as described in Section II-vi.

iii. Rankings and Hotspots

On each day, the bins in each division were sorted by their heat scores. The bins with the highest heat scores were labeled hotspots. The appropriate number of hotspots in a given day is discussed in VI-i.b.

For this work, 12 hotspots were labeled in each division on each day.

iv. Statistics, Visualizations, and Additional Datasets

Once hotspots were generated for this work, their context within and effect on the City of Los Angeles could be studied.

Hotspots generated for dates between 2012 and 2015 were plotted over a map of Los Angeles in ArcGIS. Every day, hotspot locations within a division change. An ArcGIS feature enables visualization of all hotspots having occurred during a time window.

By bringing in data regarding LAPD arrests, also from data.lacity.gov, it was possible to measure the correlation between arrests and hotspot labeling. This work hypothesizes that hotspots lead to an increase in overall arrests, as well as a disproportionate increase in what

Table 1: *Crime Reports Considered*

Code	Code Name	Occurrences 2012-2015
310	"BURGLARY"	56369
320	"BURGLARLY; ATTEMPTED"	4838
330	"BURGLARY FROM VEHICLE"	57291
331	"THEFT FROM MOTOR VEHICLE - GRAND ($400 AND OVER)"	11067
410	"BURGLARY FROM VEHICLE; ATTEMPTED"	1117
420	"THEFT FROM MOTOR VEHICLE ($950.01 & OVER)"	
	"THEFT FROM MOTOR VEHICLE - PETTY (UNDER $400)"	31760
421	"THEFT FROM MOTOR VEHICLE - ATTEMPT"	468
510	"BURGLARY"	56369

could be called "quality of life" crimes. For more information about this analysis, see (CITE OUR OWN WORK).

It is also possible to compare hotspot locations to demographic data regarding those locations, from the US Census. This work hypothesizes that hotspots are disproportionately labeled in communities of color that are adjacent to commercial areas.

V. Results

i. Hotspot List

The most basic way to navigate the output of PREDPOL was to list the 12 blocks in each division on each day with the highest heat scores. Each block, also called a *bin* or a *cell*, has a number associated with it.

Each bin was associated with an address by randomly selecting from amongst the addresses listed with each crime falling in that bin. Then, the hotspots on each day were listed with each bin's address.

ii. Visualization

The Southwest corner coordinates of each bin were recorded, and each hotspot on each day was plotted with a marker at this location in ArcGIS.

iii. Statistics Cross-Referencing other Datasets

iii.a LAPD Arrest and Citation Data

The record of every arrest and citation made in 2015 was loaded data.lacity.gov. The locations of each arrest were compared to the location of the closest hotspot on that day in that division.

Each arrest is listed with a "descent code" denoting the ethnicity or race of the individual arrested. The vast majority of descent codes were listed as "W" (White), "B" (Black), "H" (Hispanic/Latino/a), and "O" (Other); all other descent codes were processed as "O" for this work.

From these distances, the following statistics were calculated for all divisions:

- Average distance of arrest/citation to nearest hotspot in a given division
- Average distance of arrest/citation to nearest hotspot in a given division within a group denoted by descent code
- Percentage of arrests/citations made within distance d of a hotspot in a division, for $d = 0.25$km, 0.5km, and 0.75km
- Percentage of arrests/citations made within distance d of a hotspot in a division within a group denoted by descent code

iii.b Census Data

Document our work with census data here!

VI. Discussion

i. Unknowns in LAPD PREDPOL

i.a Crime Codes

The crime codes used by LAPD seem to conform to the same pattern as the *Uniform Crime Report* (*UCR*) system [8], but LAPD's crime codes do not correspond individually to the crime codes in any publicly-available UCR-related resource found during this work.

i.b Number of Hotspots per Division (K)

There is evidence that LAPD labels 10-20 hotspots per division, according to (CITE OUR OWN WORK), [3], and [6]. However, the true number is not publicly available. For this work, 12 hotspots were labeled in each division, on each day.

The LAPD Pacific Division's Twitter account tweeted the hotspots in their division on a daily basis several times in October of 2014 (CITE TWITTER??); LAPD Pacific Division posted hotspots on Facebook during that same month.

In October of 2014, a (bizarre) progression of seemingly experimental hotspot reporting strategies was attempted through social media, such as **this tweet**, followed by "Predictive Policing Preventable Crime Trends" like **this**, and "Crime Alert Updates" such as **this**. From these tweets, which seem to list all the hotspots on a given day, it seems that 12 hotspots were, in fact, generated in each division, on each day.

i.c Binning LA

There is no way to know whether the bins used by LAPD's PREDPOL implementation match those used in this work. This certainly *can* affect the resultant hotspots.

As an example, supposing many crimes were reported on both opposite corners of a block, that block could be labeled a hotspot. If adjacent blocks had very few reported crimes, this would be the only hotspot in the area. If, however, a binning of the region divided the two corners of that block into separate 500 by 500 foot cells, and given that adjacent blocks had very few crime reports, the number of crime reports in each of the two cells might be too low to be labeled a hotspot. This could be considered a "dilution" of what might be measured as a cell's high rate of crime reporting.

It may be likely that crime reporting happens due to conditions that span more than a single block; it may be unlikely for adjacent blocks to have highly different rates of reporting. That would make it less likely that differences in geographic binning would affect which areas have hotspots. The situation described above may be unlikely.

On the other hand, it may be very likely that crime reporting happens due to conditions that span less than a single block. There may be individual alleyways, overpasses, or unlit paths for which there are disproportionate numbers of crime reports. If some such region was divided into separate cells, the crime reports associated with it would indeed be "diluted" in those cells, which could affect the labeling of a hotspot. This region associated with high numbers of crime reports would have to be small enough to fit within a single cell in a given binning, and yet large enough to be split into separate cells in a different binning.

It is unclear how much of an effect the chosen binning of Los Angeles has had on the labeling of hotspots in this work.

i.d Separation of Crime Reports by Code

Based on LAPD's tweets regarding predictive policing (e.g. "burglarly alert"), it may be the case that LAPD produces a separate heat score for each of the categories of crime considered by PREDPOL (described in Section III-i).

The logic behind hotspot policing is that certain types of crimes lend themselves to repeat offenses. This logic would break down in a hotspot-generating scheme that *does not* separately process different types of crime. LAPD policy is unkown, but it may be the case that crimes are processed as such.

This work processed all crime reports together.

ii. Critique of PREDPOL

"Highly clustered event sequences are observed in certain types of crime data, such as burglary and gang violence, due to crime-specific patterns of criminal behavior" [2].

References

[1] Helmstatter, A, Sornette, D. Predictability in the Epidemic-Type Aftershock Sequence Model of Interacting Triggered Seismicity *J. Geophys. Res., 108(B10), 2482, DOI: 10.1029/2003JB002485, 2003*.

[2] 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, DOI: 10.1198/jasa.2011.ap09546*.

[3] Mohler, G.O., Short, M.B., Malinowski, S, Johnson, M, Tita, G.E., Bertozzi, A.L., Brantingham, P.J. Randomized Controlled Field Trials of Predictive Policing *Journal of the American Statistical Association, 110:512, 1399-1411, DOI: 10.1080/01621459.2015.1077710*.

[4] Lum, K., Isaac, W. (2016). To predict and serve? *Significance*.

[5] Lum, K. (2016). Derivation of the Expectation-Maximization Algorithm for the ETAS Model *Human Rights Data Analysis Group*.

[6] Levine, E.S., Tisch, J., Tasso, A., Joy, M. (2017). The New York City Police Department's Domain Awareness System *Interfaces*.

[7] Kleinberg, J., Mullainathan, S., Raghavan, M. (2017). Inherent Trade-Offs in the Fair Determination of Risk Scores. *Cornell University, arXiv:1609.05807v2*.

[8] U.S. Department of Justice, Federal Bureau of Investigation (2004). "UCR. Uniform Crime Reporting Handbook."

[9] "LAPD Divisions." *LA TIMES*. 2017. http://boundaries.latimes.com/set/lapd-divisions/

[10] Poston, B., and Rubin, J. "LAPD Misclassified Nearly 1,200 Violent Crimes as Minor Offenses." Los Angeles Times. 9 August, 2014.