THE DUBIOUS BENEFITS OF PREDICTIVE POLICING

An Investigation into the Accuracy and Fairness of `PredPol`

MICHAEL W. YANG

Bachelor's of Science (BSc)
Computational Sciences
Minerva Schools at KGI

March 2019

ABSTRACT

A well-known predictive policing algorithm does not necessarily do better, either on accuracy or fairness with respect to race, than a simple algorithm (a running count of which locations have historically had the most crime). We implement a popular predictive policing algorithm, PredPol, and assess PredPol for both accuracy and fairness. In the context of this paper, we define fairness using *equalized odds* (Hardt, Price, & Srebro, 2016): It should be the case that a black criminal and a white criminal are equally likely to be caught by the recommendations from a predictive policing algorithm. We find that PredPol, when used to predict the top 5% of probable crime locations, does not achieve significantly better accuracy than the simple heuristic of visiting the areas with the most historical crimes. We discuss the normative implications of our research for PredPol. We also propose a post-processing knapsack task with constraints that can make PredPol and other predictive policing algorithms fairer. The technique improves the fairness of PredPol without compromising on accuracy on the data tested.

ACKNOWLEDGMENTS

I have been aided in many ways by many faculty over the past year and a half. First and foremost, I'd like to thank Professor Fost for being an excellent advisor: always helpful and exceedingly fair in his criticism and praise. I'd also like to thank Professor Srebro (TTIC) for introducing me to fair machine learning; Professor Sarwate (Rutgers) for sharpening my intuitions on this topic; Professor Sterne for providing me with my first research experience; and Professor Wisor for many valuable conversations. Finally, I'd like to thank my friends, especially Abby, and family for supporting me during this time.

CONTENTS

1

INTRODUCTION

Algorithms play an increasingly large role in many applications, including policing and criminal justice. Consequently, it is important to scrutinize both how effectively these algorithms meet their stated goals and whether such algorithms produce unintended side effects. In this thesis, we make two contributions to the assessment of algorithms with societal implications: first, we show empirically that PredPol, a well-known predictive policing algorithm, does not significantly outperform a simpler, baseline method. Briefly, PredPol uses kernel density estimation to approximate a statistical model of crime that accounts for, borrowing from seismology, "aftershocks" in crime. On the other hand, the baseline method (simply counting which locations have had the largest number of crimes) is both intuitive and easily explainable to the layperson. Second, we contribute a novel post-processing modification to make predictive policing fairer (according to the definition that we will supply). This method exploits the fact that while individual grid cells may be unfair, combinations of grid cells may in the aggregate be fair.

We also discuss the normative implications of our computational results. If the purported benefits of predictive policing algorithms are overstated while simpler alternatives are available, perhaps police departments ought to use the simple alternatives instead. Algorithms are essentially policy decisions that affect the wider community, and as such, these policy decisions ought to be relatively transparent and free from negative side effects.

The remainder of this thesis will proceed as follows: first, in Chapter 2 and Chapter 3, we introduce fundamental concepts: the PredPol model of crime and different technical notions of fairness. In the latter, we also weigh the varying normative concerns that each fairness definition addresses and discuss their implications in the context of predictive policing. The novel contributions then follow. In Chapter 4, we describe a computational task for making the outputs of a predictive policing algorithm fairer. Then, in Chapter 5, we compare the accuracy and fairness of PredPol, the aforementioned baseline technique, and the fairness modification. Finally, we conclude with the implications of the present research and future research directions in Chapter 6.

1.1 AN INTRODUCTION TO PREDICTIVE POLICING

Predictive policing systems are intuitively attractive; instead of reacting to crime, police departments can proactively patrol regions that are likely to see criminal activity (Perry,

2013).[1] Compared to human expert-based methods, automated systems are less costly, more efficient in processing large amounts of data, and potentially less biased. These advantages have driven an uptick in the number of predictive systems used in policing departments across the country.

Predictive policing, even when deployed by actors with good intentions, may create new issues instead of resolving them. There are several common myths and pitfalls associated with predictive policing, as described by Perry (2013). First, predictive policing is generally over-hyped relative to the actual performance of these models. Commonly, the media cites the film *Minority Report* to introduce an audience to the technology of predictive policing, but such an introduction is highly misleading—algorithms are not omniscient. Rather, predictive policing algorithms are like other machine learning techniques that extrapolate from historical data to make guesses about the future. The undue optimism of predictive policing is embedded within the very name itself. As a matter of practice, it may be more accurate to describe such algorithms as "forecasting" or "data mining."

Closely related to this first myth is the notion that only a sophisticated computer model can allow policing departments to become more effective. In reality, simple heuristics often perform quite well on the same predictive policing tasks, as this thesis will endeavor to show. Police departments with smaller budgets may not be significantly disadvantaged by the lack of access to sophisticated models.

Moreover, even a highly accurate forecast is not a panacea for crime, a complex sociological phenomenon. Accurate predictions do not automatically provide tactical utility. Just knowing where crime might occur does not suggest the best strategy for preventing or reducing crime in that area. Thus, a prediction system must be combined effectively with the other elements of a police department, including data collection, interventions, and policy.

In addition to the issues of effectiveness discussed so far, predictive policing may carry significant ethical risks. First, when algorithms are deployed in a societal context, they may possibly retrench and reinforce existing biases in the data. For example, a predictive policing algorithm could create a policing feedback loop in which the patrols visit the same areas repeatedly. The computer models then update solely on the basis of data collected by said patrols (the next section, 1.2, will highlight some of the relevant computer science literature on this topic). Of more relevance to the present thesis is the question of transparency: an algorithmic decision-maker in the context of something as sensitive as policing is effectively a decision-maker. Unlike other policy-makers, an algorithm cannot clearly articulate its policy decisions. Moreover, many of the affected members of the public may be unable to participate effectively in the discussion around an algorithm because the details of the algorithm, even if they are nominally available for inspection, are hidden behind mathematics and jargon. Critics question whether the benefits of predictive policing are worth these ethical costs.

1 Apart from the location-based systems discussed in this report, a newer system of person-based systems is also emerging. Such systems attempt to predict specific individuals who are likely to commit a crime. The ethical implications of such systems are even more concerning. For more, see Robinson and Koepke (2016).

1.2 RELATED WORK

Other researchers have investigated the policing feedback loops that PredPol and other predictive policing algorithms are liable to create (Lum & Isaac, 2016; Ensign, Friedler, Neville, Scheidegger, & Venkatasubramanian, 2017; Ensign, Frielder, Neville, Scheidegger, & Venkatasubramanian, 2018). The latter two papers characterize theoretically the possibility of feedback loops and offer technical remedies. The risk for feedback loops also highlights the importance of effective and reasonable "end-to-end" policing practices, since the data collection process and not just the accuracy of predictions affects the end result. This thesis does not directly explore the possibility for feedback loops.

Reports on the effectiveness of PredPol are mixed (Robinson & Koepke, 2016). An earlier investigation by a Swiss researcher comes to similar conclusions as this report on the effectiveness of PredPol relative to simple baseline techniques (Benslimane, 2014) (French language only). Moreover, real-world evidence about potential reductions in crime from PredPol are not conclusive. The makers of PredPol have been laudably transparent in publishing their methods and results in established scientific journals, including results from randomized-controlled trials (Mohler, Short, Brantingham, Schoenberg, & Tita, 2011; Mohler, 2014; Mohler et al., 2015). However, said work has been criticized by other researchers who are dubious that the observed improvements in PredPol can be attributed to PredPol itself rather than natural variations (Saunders, Hunt, & Hollywood, 2016). To date, the author of this report is not aware of any third-party controlled studies of the effectiveness of PredPol.

Some initial work, both theoretical and empirical, has been done to try and make PredPol fairer. Mohler, Raje, Valasik, Carter, and Brantingham (2018) use a penalized likelihood function to learn a fair version of PredPol. That work differs from the present both in the method of accomplishing fairness and in the fairness criterion considered. The fairness definition considered there, demographic parity, is deficient for an important reason that we will discuss in Chapter 3. Moreover, because of the fairness definition chosen, there is a strong trade-off between accuracy and fairness, while the post-processing task that we suggest does not perform as poorly. Empirically, Brantingham, Valasik, and Mohler (2018) perform a randomized-control trial that shows no significant increase in minority arrests from usage of PredPol. To the extent that there are doubts about other randomized control trials of PredPol, there ought to be similar concerns about this study as well.

Finally, computer scientists have built up a rich literature for assessing and maintaining fairness in machine learning algorithms. While these solutions, like PredPol, are not panaceas, they are interesting computational problems in their own right and may be useful in certain real-world contexts. The research on *fair ranking* tasks is most similar to the post-processing modification introduced in this paper (Celis, Straszak, & Vishnoi, 2017; Zehlike et al., 2017; Singh & Joachims, 2018).

2

A PRIMER ON PREDPOL

As a location-based system, PredPol aims to predict the probability of criminal activity in each location on a map, and from those predictions, generate a ranked list of crime locations. The rankings then direct the activity of police patrols. The novelty of PredPol lies in its "aftershock" model of crime that accounts for the fact that crime re-occurs in roughly the same geographic area.

Introducing a bit of notation, let (x_i, y_i) for $i < N$ be the location of the center of each grid on a map, where there are N grid cells total. The goal of predictive policing is to create a function $f(x_i, y_i, t)$ that is proportional to the probability of crime in the ith cell at time t (note that $\sum_i f(x_i, y_i, t)$ does not have to equal 1). One can then sort the grid cells by f to prioritize which grid cells ought to be visited first. Since each policing department has finite resources, in practice only the top few percentage of grid cells are visited (discussed further in Chapter 5).

PredPol makes use of only three pieces of information about historical crime: the date and time of the crime, the location of the crime, and the type of crime.[1] Because PredPol does not explicitly consider race or other protected attributes to make decisions, PredPol might be considered "fair through unawareness," a frequently cited notion of algorithmic fairness that will be discussed more thoroughly in Chapter 3.

PredPol posits that an instance of crime falls into one of two categories: either the crime was a "background" event caused by underlying, environmental factors not represented in PredPol explicitly, or that the crime was an "aftershock" event that was partially triggered by a recent crime in a nearby location (Mohler et al., 2011; Mohler, 2014). The terms "background" and "aftershock" derive from the model's origin in seismology.

Sociological studies of crime lend credence to the presence of "aftershock" crimes because of three kinds of behaviors: repeat victimization, in which offenders return to the location of previously successful crimes; near-repeat victimization, where offenders tend to re-offend in locations close to the location of previously successful crimes; and local search, in which offenders rarely travel far from common locations like home or work (Mohler et al., 2011). All three of these behaviors are modeled by the aftershock component of PredPol.

1 The results of this paper omit analysis based on the type of crime. For a full discussion of the ramifications of this decision, see Section 5.4.

Equation 2.1 describes how PredPol computes f for the ith location at time t:

$$f(x_i, y_i, t) = \sum_{j|t_j<t} \left(\underbrace{\mu(x_j - x_i, y_j - y_i)}_{\text{background}} + \underbrace{g(x_j - x_i, y_j - y_i, t_j - t)}_{\text{aftershock}} \right) \tag{2.1}$$

In practice, PredPol works like a kernel-based smoothing method over the dataset of historical crimes. Each crime j observed before time t contributes some amount of predicted intensity to the ith grid cell. The contribution of each crime to the predicted intensity is described by the μ and g functions. The first accounts for the "background" rate of crime, while the second accounts for "aftershock" crimes.

$$\mu(\Delta x, \Delta y) = \mathcal{N}_{pdf}\left(\Delta x, \Delta y \mid \eta^2\right) \tag{2.2}$$

$$g(\Delta x, \Delta y, \Delta t) = \lambda_{pdf}(\Delta t \mid \omega) \times \mathcal{N}_{pdf}\left(\Delta x, \Delta y \mid \sigma^2\right) \tag{2.3}$$

$\mathcal{N}_{pdf}$ is the probability distribution function for a symmetrical two-dimensional Gaussian distribution centered at the origin with variance η^2 or σ^2, respectively. λ_{pdf} is the pdf for the exponential distribution with decay parameter ω.

Three parameters have to be estimated from the data with this procedure: η, ω, and σ. The first governs the size of the Gaussian kernel used to estimate the background rate. Larger values of η mean that each crime contributes more intensity to a wider range of cells. The second two govern the influence of aftershock events. ω controls how long an event temporarily raises the intensity of its neighboring grid cell, with larger values of ω corresponding to more temporary contributions. The role of σ is similar to the role of η, and controls how close nearby events must be in order to contribute to the aftershock intensity.

Following the procedure established in Mohler (2014), we estimate these three parameters using an expectation-maximization (EM) procedure. See Appendix C for further details.

3

A PRIMER ON FAIRNESS DEFINITIONS

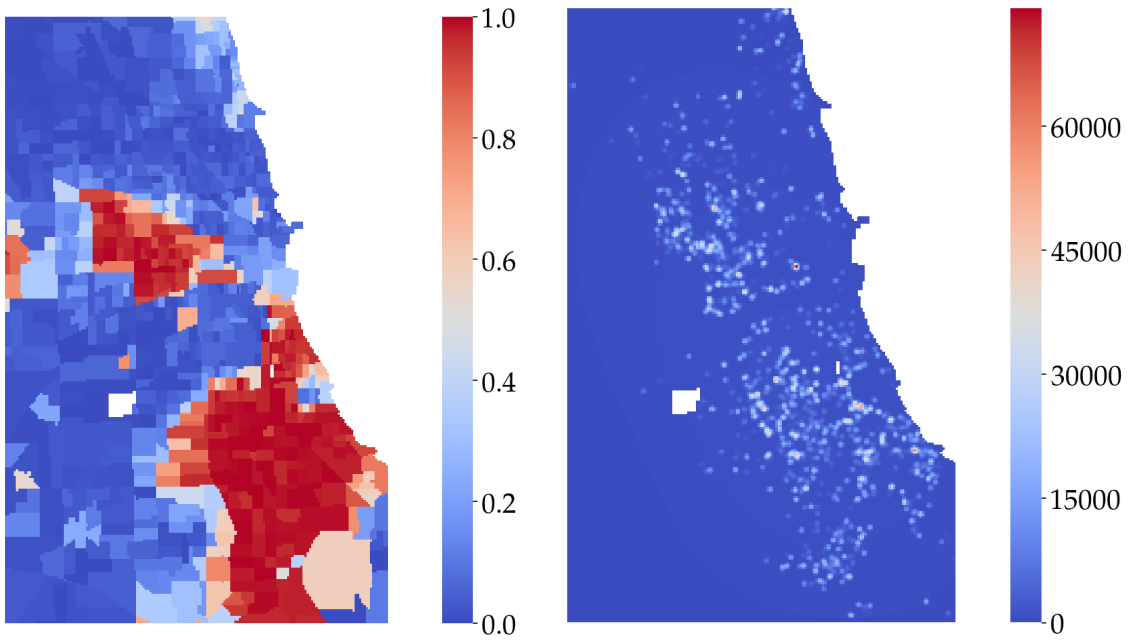
There are many mathematical definitions of fairness, each capturing different intuitive beliefs about equity. Some of the definitions of fairness turn out to be incompatible with one another in real-world scenarios. This chapter surveys common definitions of fairness and links the normative implications of each to predictive policing. Of the common notions of fairness, equalized odds is chosen as indicative of the viewpoint of an individual who cares about justice. That viewpoint will be set against the viewpoint of an individual whose main concern is accuracy in the assessment of PredPol.

3.1 FAIRNESS THROUGH UNAWARENESS

One of the most popular notions of fairness, "fairness through unawareness" is the assertion that a classifier or decision-maker who does not explicitly consider protected attributes such as race and gender is fair. To explicitly consider race and gender would constitute what the United States Civil Rights Act calls "disparate treatment," and would be impermissible in many settings, such as employment (Barocas & Selbst, 2016).

While intuitively satisfying to some, "fairness through unawareness" does not account for proxy discrimination. A classifier could still in practice be biased against a certain group if there is significant correlation between the protected attribute and another, seemingly innocuous variable. In the case of PredPol, which does not explicitly consider race or other protected attributes in its predictions, location acts as a proxy for race (Figure 3.1). The average PredPol predicted intensity and percentage black has a correlation coefficient of 0.32 over 39030 grid cells (a large enough sample size to detect moderate effect sizes with statistical significance). At the same time, the correlation coefficient between the actual number of crimes committed and percentage black is only 0.11. This evidence suggests that PredPol, which does not explicitly consider race, can still have different outcomes for different races (more detailed fairness analysis in Chapter 5). The point for the present discussion is that fairness through unawareness does not preclude disparate impact (also prohibited by United States civil rights law) (Barocas & Selbst, 2016).

Fairness through unawareness reflects a variety of intuitive beliefs about fairness. One intuition suggests that fairness through unawareness is the best that decision-makers or classifiers can hope for, because any other correlation between a protected attribute and the predictions might be explained as an unfortunate reflection of real-world correlation



(a) Black population as a percentage of total = black + white

(b) Average PredPol predicted values f in the year 2015

Figure 3.1: The visual correlation between race and PredPol

between the protected attribute and the target variable of prediction (in this case, criminality of a location). From this viewpoint, the results of PredPol ought to be taken as is without modification. Other definitions of fairness attempt to take the problem of proxy discrimination seriously and control for it.

3.2 THREE COMMON DEFINITIONS

Using conditional independence, three common definitions of fairness can be stated compactly as relationships between the ground truth Y, the prediction $\hat{Y}$, and the protected attribute A.[1] Each of these definitions is observational in the sense that one does not need to conduct an intervention or run an experiment to order to assess them (causal notions of fairness do require interventions) (Hardt et al., 2016). One of these definitions attempts to allow for the sort of proxy discrimination exhibited in this example up to the extent that is justified by correlation between the protected attribute and the true status. This definition also turns out to be relatively strong, insofar as it sometimes requires the decision-maker to make sacrifices to accuracy and even other notions of fairness.

The first definition, demographic parity ($\hat{Y} \perp A$), strengthens the "fairness through unawareness" condition, but ultimately to an unrealistic degree. Not only do the predictions

1 As a reminder of notation, the statement, "Random variable X is independent of random variable Y given the observation of random variable Z" is written as $X \perp Y \mid Z$.

have to be made without the protected attribute, the final predictions must also be statistically independent from the predicted attribute. This is a very stringent fairness requirement, since, as we have already seen, Y and A are related, and $\hat{Y}$ wants to track Y. Demographic parity further formalizes the notion of fairness that fairness through unawareness attempts to satisfy: in the ideal world, the true variables Y are independent from protected attributes A. Thus, in the ideal world, either fairness through unawareness or demographic parity are sufficiently strong notions of fairness.

The next definition, equalized odds ($\hat{Y} \perp A \mid Y$), can be seen as a relaxation of the previous criterion. After conditioning on the true status, the predictions ought to be independent from group status. This definition of fairness formalizes the amount of correlation that is permissible between the prediction $\hat{Y}$ and the protected group A: exactly as much as would be useful to predict true status.

Unlike the next and final fairness definition, which attempts to equalize accuracy across protected groups, equalized odds attempts to equalize false positive and negative rates across protected groups. The normative implications of this difference are significant, since false positives and negatives can have differing consequences on the individual and the community (Narayanan, 2018). In policing, a false positive might introduce an innocent individual into the criminal justice system, while a false negative might leave a criminal at large. In predictive policing, the individual units who are subjected to policing might be considered individual persons or the neighborhoods and locations which are recommended for patrol. If the individual units are locations, then the implication of a false positive is overpolicing a community unjustly while the implication of a false negative is underpolicing a community unduly. Equalized odds is a fairly stringent notion of fairness because satisfying it often requires sacrifices to accuracy (Hardt et al., 2016).

Finally, sufficiency ($Y \perp A \mid \hat{Y}$) requires that the algorithm's overall accuracy across different groups is the same. Put differently, the location of the grid cell ought to be sufficient information for prediction and further conditioning on race (in addition to the prediction $\hat{Y}$) would not add any accuracy. This is the least stringent fairness requirement, and many out-of-the-box ML algorithms already satisfy this criterion (Barocas, Hardt, & Narayanan, 2018). One might intuit sufficiency in the context of predictive policing as follows: if one race is responsible for p% of the overall crime, then that race ought to be policed p% of the time.[2]

Sufficiency and equalized odds are incompatible whenever the baseline rate of true status differs between groups (Kleinberg, Mullainathan, & Raghavan, 2016; Chouldechova, 2017).[3] In the real-world, one can usually only ask that either sufficiency or equalized odds are met.

We will assess PredPol on and make PredPol fairer with respect to equalized odds because equalized odds is a stronger notion of fairness than sufficiency without being

2 Out of the box, PredPol does not appear to satisfy this definition of fairness (Section B.1). The upshot is that the predicted intensities from PredPol have different interpretations based on the demographic makeup of the grid cell, a line of inquiry which is not pursued in this research.

3 Section B.2 offers an alternative proof of this incompatibility using the p% intuition for sufficiency suggested above.

obviously deficient like demographic parity. Sufficiency is too weak a notion of fairness because it ignores the consequential difference between false positives and negatives in many contexts. Some have argued that because of the impossibility of achieving both equalized odds and sufficiency, one ought to focus on deploying predictors which are as accurate as possible while also ensuring that the processes around the predictor are also just (Corbett-Davies, Pierson, Feller, Goel, & Huq, 2017). Essentially, this view bites the bullet on the impossibility theorem, and accepts the fact that equalized odds will not be met in real-world scenarios. While there is some merit to this view, especially since observational notions of fairness themselves do not guarantee that algorithms are used justly, there are likely still scenarios in which having an equalized odds predictor would be useful. Moreover, the aim of this thesis is to show how PredPol fails to impress either on accuracy or on a relatively stringent notion of fairness. We do not claim to have the final word on all normative concerns surrounding algorithmic fairness.

Finally, as a cautionary note regarding the usage of technical notions of fairness, ensuring equalized odds in predictions does not guarantee the equal distribution of risk or burden in reality. For example, in the context of predictive policing, individuals may still face heightened scrutiny for their racial features, regardless of which grid cells PredPol ultimately predicts. The task of equalizing burden in practice requires, as Perry (2013) suggested, best practices throughout the whole policing pipeline.

4

PREDPOL WITH EQUALIZED ODDS

In the previous chapter, we described various notions of fairness and settled upon equalized odds as the object of measurement for this study. In this chapter, we now focus on the computational task of ensuring equalized odds in PredPol (and theoretically, any predictive policing algorithm). The main theoretical contribution of this chapter is a post-processing task for approaching equalized odds. The task "post-processes" the predicted values of PredPol and does not modify the internal model of PredPol. Any predictive policing algorithm which also meets the description of the predictive policing task in Chapter 2 can also be modified using the following work.

The intuition for the post-processing modification is that if equalized odds is true, then the following quantity should be close to zero:

$$|\%\text{ black crime caught} - \%\text{ white crime caught}| \tag{4.1}$$

This measure captures the belief that grid cells with equal amounts of criminal activity ought to have similar predicted intensities, regardless of their demographic make-up. This measure also solves the problem that while observational notions of fairness are frequently deployed on binary variables, the variables in our setting (the predicted intensity per grid cell, the demographic make-up per grid cell, and the number of crimes in each grid cell) are all continuous.

Equation 4.1 is properly a measure of *un*-fairness, since smaller values of it correspond to fairer predictions and vice versa. One can turn Equation 4.1 into a fairness measure by subtracting it from 1.0.

The vague terms "black crime" and "white crime" accommodate both different interpretations of what fairness in policing requires and the possibilities afforded by different datasets. For example, the Chicago crime data used for the present research omits race as a feature of each observation. Instead, we operationalize "black crime" and "white crime" as "crime occurring in locations with larger black or white populations" respectively.[1] In other scenarios, researchers may choose to prioritize the race of either the perpetrator of a crime or the race of the victims; the approach used in this research attempts to consider both.

Given the goal of catching equal amounts of crime across protected groups, one ought to try and equalize the predicted intensity of crime across protected groups in the locations

1 To do so, we combined the Chicago crime data with other demographic data; see Appendix C for details.

visited. To formalize this notion for the ith grid cell, let f_i be the intensity, as predicted by PredPol, b_i the percentage of black individuals, and w_i the percentage of white individuals. The data are processed so that $\forall i$, $b_i + w_i = 1.0$. First, compute:

$$\hat{f}_i^{(\text{black})} = \frac{b_i f_i}{\sum_i b_i f_i} \tag{4.2}$$

$$\hat{f}_i^{(\text{white})} = \frac{w_i f_i}{\sum_i w_i f_i} \tag{4.3}$$

The $\hat{f}$ values indicate *racially-differentiated predictive value*.[2] If one is concerned about crime that disproportionately affects black communities, they should visit grid cells as ranked by the value of $\hat{f}_i^{(\text{black})}$, and vice versa for white communities. In a world where communities are fully integrated and there are no global disparities in demographics (so the demographics of each grid cell reflect the demographics over all locations), $\hat{f}_i^{(\text{black})} = \hat{f}_i^{(\text{white})}$ for all i. In this ideal setting, one could also achieve demographic parity.

Now, we can calculate the *predictive value gap* for each grid cell i by taking $\Delta_i = |\hat{f}_i^{(\text{black})} - \hat{f}_i^{(\text{white})}|$. Note that the values of $\hat{f}$ are in the unit interval $[0, 1]$. Thus, the difference $\hat{f}_i^{(\text{black})} - \hat{f}_i^{(\text{white})}$ will be in the range $[-1, 1]$, and Δ_i will be in the unit interval as well. When $\Delta_i = 0$ for a particular grid cell, it means that that grid cell captures an equal amount of crime affecting both races; in other words, that grid cell is "fair."

With the definitions of $\hat{f}$ in hand, we can define the post-processing fairness task, which takes its inspiration from the knapsack problem in theoretical computer science. Intuitively, even if no grid cell is fair individually, different combinations or subsets of grid cells might be fair (as indicated by their summed differences $\hat{f}_i^{(\text{black})} - \hat{f}_i^{(\text{white})}$.[3] This problem mirrors a knapsack task: we want to maximize the overall value of a subset of items from a larger collection while subject to certain constraints.

More formally, given a list of N items, each with an associated benefit f_i and two different costs $\hat{f}_i^{(\text{black})}$ and $\hat{f}_i^{(\text{white})}$, find a subset of at most k items such that: the total benefit is maximized, and the two total costs are approximately equivalent. If we let x_i be a binary indicator for whether an item is included in the knapsack, and give dummy weights of $w_i = 1$ for all i, we can state the problem as follows:

$$\max_x \quad \sum_i f_i x_i \tag{4.4}$$

$$\text{subject to } \sum_i x_i = k \tag{4.5}$$

$$\sum_i \hat{f}_i^{(\text{black})} x_i = \sum \hat{f}_i^{(\text{white})} x_i \tag{4.6}$$

2 If one had race information available for the perpetrator (or victim) of each crime, they could separately estimate intensities for each race $f^{(\text{black})}, f^{(\text{white})}$, and combine them in the following way: $\hat{f}_i^{(\text{black})} = \frac{f_i^{(\text{black})}}{\sum_i f_i^{(\text{black})}}$ and $\hat{f}_i^{(\text{white})} = \frac{f_i^{(\text{white})}}{\sum_i f_i^{(\text{white})}}$

3 We must drop the absolute value around this expression in order to allow different gaps from different cells to cancel out.

The tricky condition is the third line, which we can transform into two constraints:

$$\max_x \quad \sum_i f_i x_i \tag{4.7}$$

$$\text{subject to } \sum_i x_i = k \tag{4.8}$$

$$\sum_i \hat{f}_i^{(\text{black})} x_i = D \tag{4.9}$$

$$\sum_i \hat{f}_i^{(\text{white})} x_i = D \tag{4.10}$$

The value D is arbitrary—if the subset of grid cells selected polices the same percentage of black and white crime, then the subset is fair, by our standard. However, when implementing this problem in practice, we also relax all of the equalities to be less-thans. Then, D takes on the role of the *maximum tolerable gap* in fairness. Because the constraints become inequalities, the gap in fairness will be at most D.

We also observe that without the fairness constraints (alternatively, setting $D = 1$), this knapsack problem is trivial and amounts to taking the k cells with highest f_i value.

To provide further intuition for how this post-processing task might affect the predictions of an unconstrained procedure, we will consider two toy examples that illustrate the two possible ways in which post-processing task changes predictions. Each example considers a map of crime with four grid cells, and the resource-strapped police department can only visit two of the locations. Each map is represented by a two-by-two table, where the contents of each cell respectively indicate the predicted crime intensity f, the amount of black crime (as a percentage of all black crime) occurring in that cell $f^{(\text{black})}$, and the same quantity for whites $f^{(\text{white})}$. We further assume that we have a perfect predictor, so that $f = \hat{f}$.

5 / 0.4 / 0.2	4 / 0.2 / 0.2
4 / 0.2 / 0.2	1 / 0.2 / 0.4

Table 4.1: Example 1: A four-cell world

In example 1, the unconstrained algorithm would visit the top-left cell and either the cell to the immediate bottom or right to capture:

- 9 units of crime
- 60% of black crime
- 40% of white crime

However, the post-processing fairness modification with a sufficiently low tolerance threshold ($D < 0.2$) would choose both the top-right and bottom-left cell instead and thus capture:

- 8 units of crime
- 40% of black crime
- 40% of white crime

This example illustrates that the fair procedure will generally prefer cells which have smaller Δ_i values. Moreover, the accuracy of the fair procedure will degrade further as the "unfair" cells account for more of the overall crime (for example, let the 5 in the top-right cell be 50 or 500).

The fair procedure would also prefer combinations of cell that have a collectively small Δ_i value, as the second example will illustrate:

5 / 0.4 / 0.2	4 / 0.2 / 0.2
1 / 0.2 / 0.2	3 / 0.2 / 0.4

Table 4.2: Example 2: Another four-cell world

The unconstrained algorithm would choose the top row of cells and, as before, capture:

- 9 units of crime
- 60% of black crime
- 40% of white crime

The post-processing fairness modification with a sufficiently low tolerance threshold ($D < 0.2$) would instead choose the top-left and bottom-right cell and capture:

- 8 units of crime
- 40% of black crime
- 40% of white crime

In this situation, the fair procedure achieves greater fairness by "balancing" the unfairness of one cell with the unfairness of another cell. In this situation, the accuracy of the fair procedure degrades based on the difference between the second-best cell (the top-right cell in this example) and the "balancing" cell.

5

RESULTS

In this chapter, we present the major results from simulated tests on the Chicago crime data set. First, we find that the accuracy of PredPol does not exceed a baseline heuristic in the top 5% of cells predicted. Second, we find that PredPol, as assessed by the measure discussed in the previous chapter, is not fairer than the baseline heuristic. Third, we find that the fairness modifications presented in the previous chapter can work decently well at improving fairness while preserving the accuracy of PredPol. Taken together, these results show that the benefits of PredPol are overstated. We discuss this possibility as well as other implications in the next chapter.

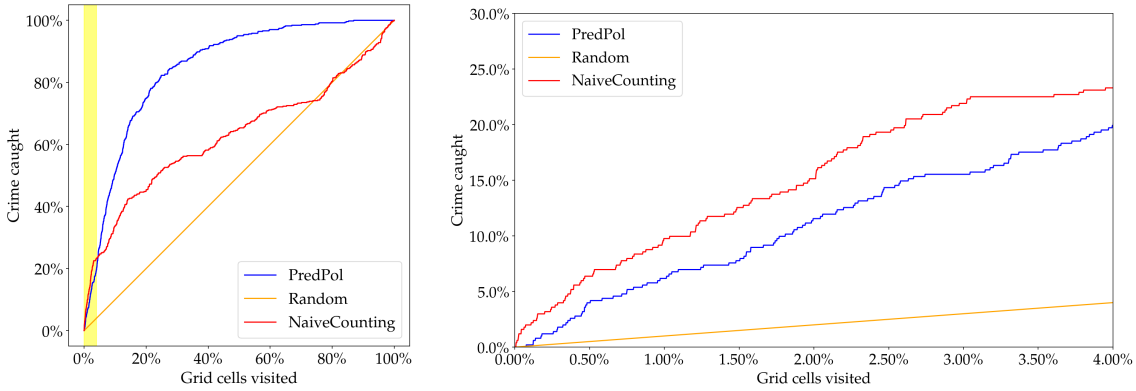
5.1 ACCURACY

Figure 5.1 shows the accuracy of PredPol compared to completely random prediction and the baseline measure, "naive counting." The baseline measure is comparable to a naive Bayes approach to predictive policing: rank each grid cell by the number of crimes that have taken place in that location over the whole observed dataset. In the language of Mohler (2014), the naive counting measure corresponds to the "chronic hotspots" method with no decay in time (observations further back in time are weighted equally as more recent ones). The results are shown as a function of the number of grid cells visited, since visiting more grid cells will tend to improve accuracy in general (these curves are similar to ROC curves in binary classification). The y-axis plots the percentage of crime caught, or the number of crimes that occur in the cells predicted as a percentage of the total crime occuring that day. Each curve is an average over 365 days of prediction in 2015.

Imagining that one had access to a perfect oracle, which knew in advance where each crime would occur, the results from such an oracle would plot a nearly perfect right angle at the upper left of Figure 5.1a. This reflects the fact that tiny percentage of grid cells each day are responsible for crime; knowing those locations in advance means that all of the crime occurring in a day could be captured by visiting just a few locations.

PredPol shows increased accuracy when considered over the whole range of grid cells. The naive counting measure breaks down at around 80% of grid cells predicted, when its performance becomes no better than randomly guessing.

However, visiting even 10% of the grid cells on a map is unreasonable for police departments to accomplish. Mohler (2014) states that "a range of 250–500 150m $\times$ 150m



(a) Overall, PredPol predicts well...

(b) ...but not in the top 5% of grid cells.

Figure 5.1: Accuracy curves comparing PredPol and other baselines

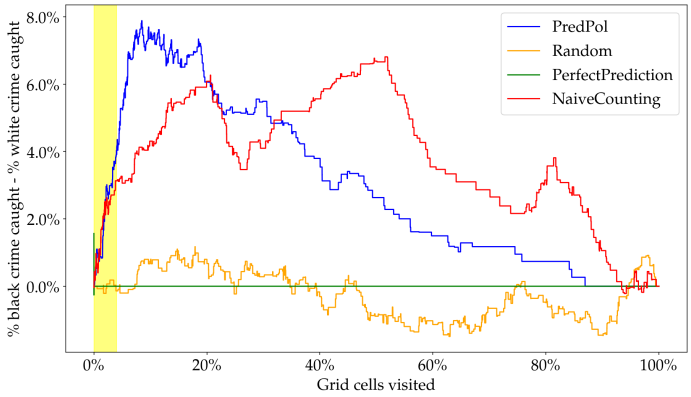
cells is a realistic number for a city the size of Chicago," which correspond on the above plots to a range of 0.5% to 1% of grid cells.

5.2 FAIRNESS

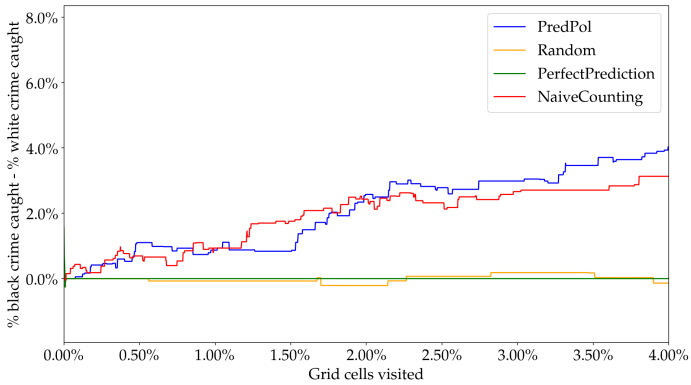
Figure 5.2 assesses the fairness of the same measures discussed in the previous section. Again, the x-axis is the percentage of grid cells visited/used by the algorithm. The y-axis computes the percentage of black crime captured (by the x% grid cells predicted by the algorithm) minus the percentage of white crime captured (by the x% grid cells predicted by the algorithm). When all grid cells have been predicted, 100% of crime of both races has been captured, so each of the curves in figure 5.2a meet at $(1.0, 0.0)$.

A completely fair algorithm, by this metric, would lie directly on the x-axis. Random prediction has this property (noise in the curve is due to the random variation, and on different runs, the shape of the curve changes completely). It is interesting to note that random prediction is fair with regard to equalized odds while at the same time qualifying as a "fair through unawareness" classifier. Of course, the accuracy of predicting randomly leaves much to be desired. Perfect prediction also lies on the x-axis, since perfect prediction captures all crime nearly instantaneously, thus also capturing equal percentages of both races' crimes. Curves above the x-axis indicate some amount of bias against blacks, while curves below the x-axis indicate some amount of bias against whites. However, because of how we operationalized race and fairness in Chapter 4, one could easily draw the opposite conclusions: capturing a larger percentage of crime in black districts *favors* blacks, and vice versa. We remain open to this alternative interpretation. The point is to try and achieve equality for both races.

The differences between PredPol and the baseline measure are less apparent on fairness. In the range of cells discussed in the previous section, PredPol hews closer to the x-axis than the baseline measure.



(a) PredPol is slightly fairer over all cells...



(b) ...but not especially so in the top 5% of grid cells.

Figure 5.2: Fairness curves comparing PredPol and other baselines

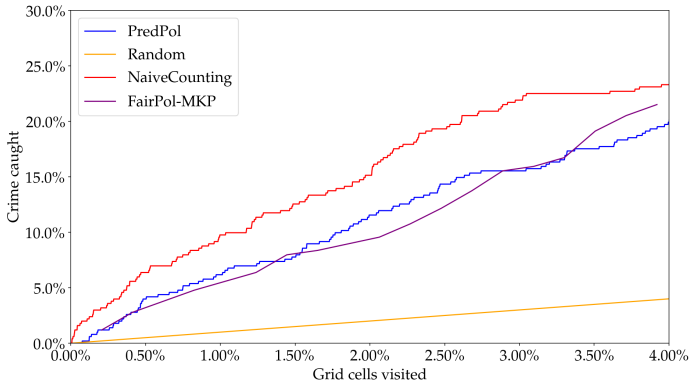
One interesting aspect is that the fairness curves for both PredPol and naive counting never go below the x-axis; in other words, both measures always capture a larger percentage of black crime than white crime. We suggest possible reasons for this in Section 5.5.

5.3 FAIRNESS MODIFICATION

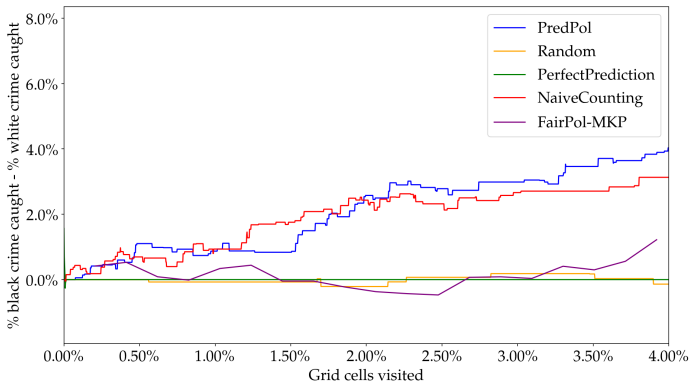
Finally, Figure 5.3 shows the result of the post-processing task suggested in Chapter 4. By seeking combinations of grid cells that together constitute fairer policing, these results show that achieving fairer prediction without comparable hits to accuracy is possible.

5.4 CAVEATS

Before proceeding to a full discussion of the results we would like to, in the interest of transparency and future research, highlight several limitations of the present research. We omit using different types of crime to predict one another, as done in Mohler (2014). That model uses the historical record of, say, burglaries, to assess increased likelihood of say,



(a) The modification has comparable accuracy...



(b) ...while being significantly fairer.

Figure 5.3: The accuracy and fairness of the post-processing modification (purple) ($D = 1.25\%$)

homicides. One direction for future research is to implement the full model described in Mohler (2014) and attempt to replicate the results of this research. Creating a plausible baseline method for this kind of prediction is complicated by the fact that different types of crime occur at very different rates than others; a naive Bayes method that treats all crime equally would be dominated by the most prevalent crime type.

Nevertheless, it still seems plausible that a baseline measure, such as the chronic hotspots method discussed in Mohler (2014) and similar to the naive counting measure used here, could achieve comparable performance to PredPol. In several places throughout the text, Mohler (2014) also acknowledges that chronic hotspots account for most of the crime in the dataset. If this is indeed the case, then the discussion of ethics in the following section is still likely to hold merit.

There are a few smaller differences that may have led to a difference in findings in the present research and Mohler (2014). First, the number of grid cells in the simulation differs. It seems that Mohler examined a smaller geographical region than in the present research, but Mohler did not make note of this detail in their paper. Second, the Chicago

dataset is known to change over time as the agency responsible for marinating the data re-randomizes the locations of crimes in order to protect privacy.

Finally, there may have been errors in our simulation, for which we take full responsibility. A more detailed explanation of methods along with replication code is available in Appendix C.

5.5 DISCUSSION

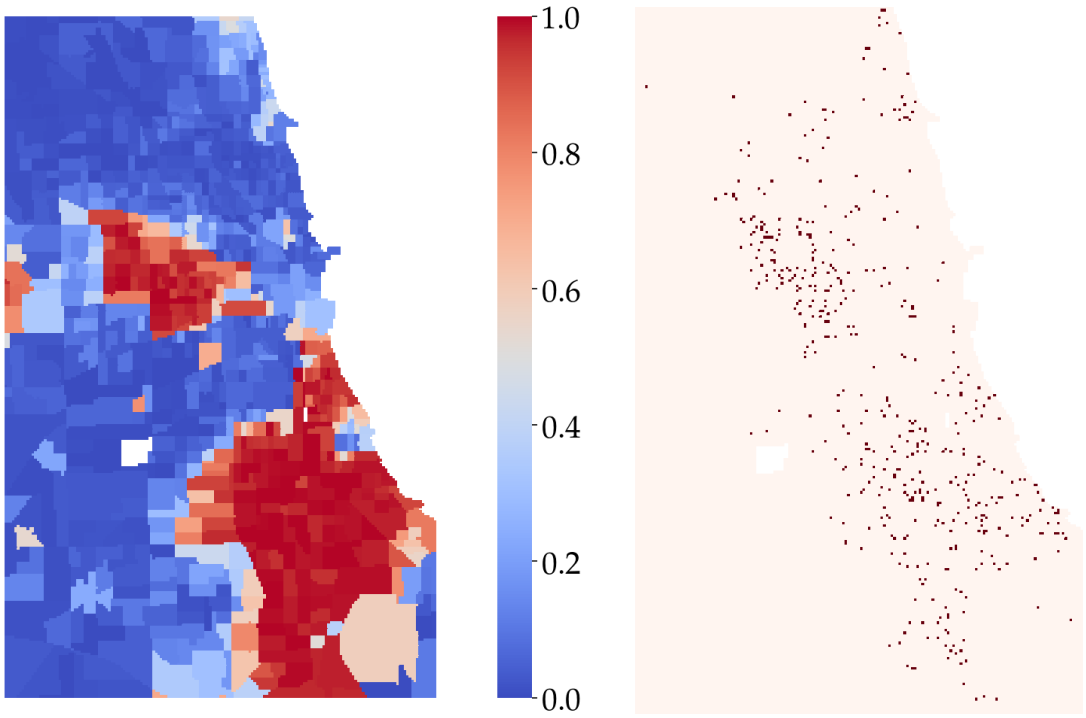
Barring the caveats mentioned in the previous section, it is unclear why PREDPOL does worse than the naive counting measure in the region specified. We hypothesize, for further testing, that some amount of constructive interference may be occurring: in the middle zone between two crime hotspots, PREDPOL predicts a high amount of crime, despite being exactly in the middle of where the crime would likely occur.

With regard to normative issues, Mohler (2014) states that their model results in a 17% relative improvement in accuracy from the measures they consider, and that, "[g]iven the high societal cost of homicide and serious gun crime (DeLisi et al., 2010), which is estimated to be billions of dollars a year in Chicago, even a few percent decrease in the homicide rate due to hotspot policing with a more accurate ranking method would be of significant societal benefit." Building off of the empirical results presented in this chapter, we posit that there may also be costs to fairness and transparency which should be weighed against the benefits of predictive policing. If a simpler method for prediction (i.e. the hotspot policing approach) is nearly as accurate, more equitable across protected groups, and easier to explain to the affected population, there are good reasons for preferring that simpler approach.

Both PREDPOL and naive counting police black crime more than white crime. We posit that this occurs because one group's crime is more concentrated in one area than another. Figure 5.5 show the distribution of $\hat{f}$ values in the Chicago region while Figure 5.4 shows the actual locations of crime. Since there is a weak relationship between race and number of crimes (as stated previously, Pearson's r of 0.11), we should expect crime in both whiter and blacker grid cells. However, the crimes in the whiter areas of the northern region are spread apart while the crime in blacker areas tends to be clustered closer together.

Because of PREDPOL's aftershock model, it would tend to exploit the clustering of black crime and predict those regions more frequently, even though a good predictor would also try and account for the sparser crime occurring in the north of the map. That explains how PREDPOL is initially more unfair than naive counting (in the range of 0-30% cells predicted) in Figure 5.2a. Nevertheless, Figure 5.2b does show that this difference in fairness is negligible in the range of grid cells that a police department could actually visit.

Finally, PREDPOL is likely to be less transparent to the police force and the community that they are responsible for than a measure like the naive counting measure. Patrol officers and officials are not likely to understand how and why the model generates its results, since the product is sold to police departments as a black box. Moreover, relying on a third-party vendor to handle predictions runs the risk of feedback loops, as suggested by

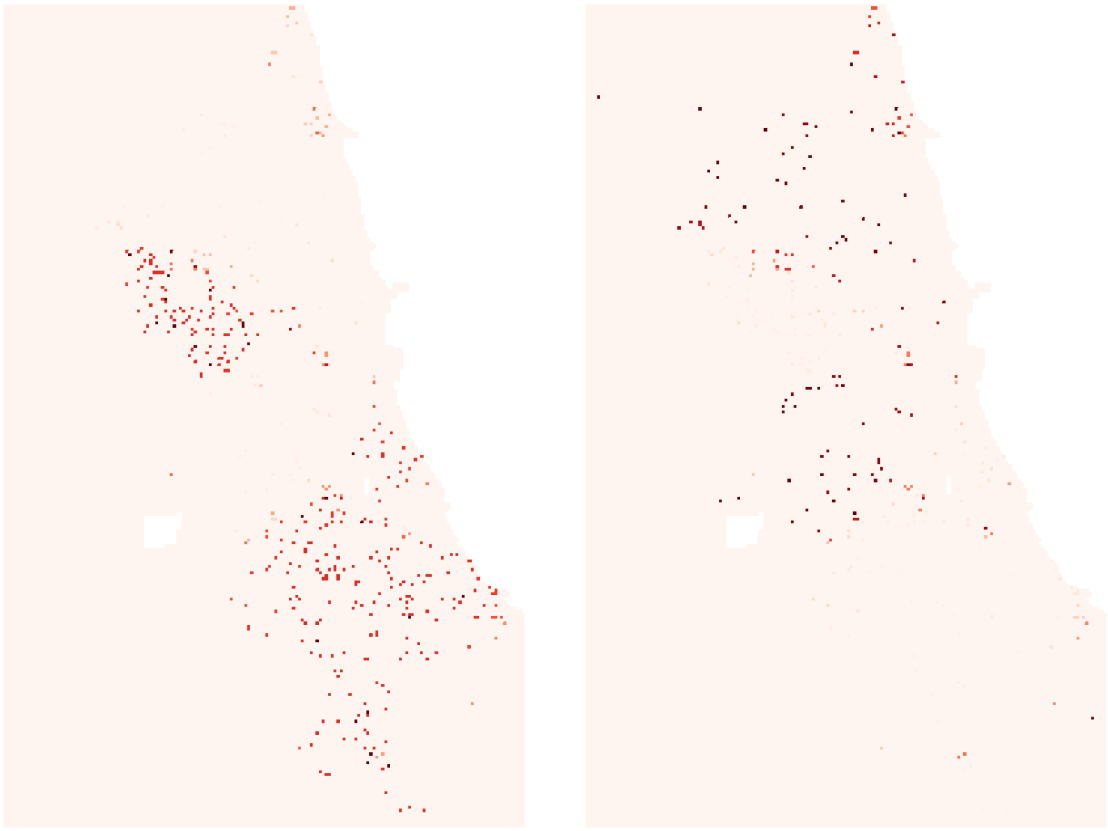


(a) Black population as a percentage of total = black + white

(b) Number of actual crimes

Figure 5.4: The visual relationship between race and actual number of crimes

other researchers (Lum & Isaac, 2016; Ensign et al., 2017; Ensign et al., 2018). The devil is in the details when using predictive policing software, and police department ought not treat the results or mechanisms of PREDPOL as infallible.



(a) Where crimes affecting black...

(b) ...and white people occur

Figure 5.5: Heatmaps of the "true" $\hat{f}$ values. These maps were produced using Equation 4.2 and Equation 4.3 but with the actual number of crimes in each region as the values of f.

6

CONCLUSION

To summarize, we implemented an existing predictive policing model and assessed it for accuracy and fairness relative to a baseline measure. We found that the method was not more accurate than the baseline method nor was it fairer. We proposed a post-processing fairness task that seems to improve the fairness of results while maintaining a good amount of accuracy. We also discussed some of the normative implications of this work on the use of predictive policing.

6.1 FURTHER WORK

The knapsack problem and many of its variants, including the multi-dimensional knapsack problem, are known to be NP-hard problems. It might be interesting to prove that the knapsack reduction discussed in Chapter 4 is also NP-hard, despite the fact that there is likely to be correlation between the value of each item and the costs associated with each item.

However, we would also like to suggest future lines of computer science research that more directly promote justice and equity. While there has been a heightened interest in assessing the social impacts of various computational techniques in society, much of the scholarly work from computer scientists has been, unsurprisingly, abstract and reductive. This work has been abstract in the sense that they detach themselves from real-world usage and the complexity of algorithmic applications and reductive in the sense that they boil down complex and ambiguous concepts such as "fairness" and "equity" into single and ultimately limited equations. Our hope is that this thesis shows that not all computer scientists have to be bound by the preferences of their field.

For instance, computer scientists can begin to model how communities respond to over-policing and how that might change the calculus of a police department (this work would be similar in nature to Liu, Dean, Rolf, Simchowitz, and Hardt (2018)). This work would interrogate whether some policing strategies are more strategy-proof than others.

Computer scientists could also consider different objective functions other than accuracy for predictive policing algorithm, such as the impact of policing on police-civilian relationships or sentiments of community safety.

Recent research on "dirty data" in predictive policing systems also complicates the results of this paper (Richardson, Schultz, & Crawford, 2019). The data in criminal justice

systems reflect not only incidences of crime, but also policing policy—and bias. Many crimes go unreported or unrecorded in the criminal justice system for a variety of reasons. At other times, false crimes are recorded or the number of crimes are over-reported. It may be worthwhile to study how different kinds of hypothetical systematic bias in data collection change the results of this paper. Essentially, one could conduct a robustness test by assuming that police data tend to systematically over-sample some regions and under-sample others based on demographics (as was the central thesis in Lum and Isaac (2016)).

Finally, it would be interesting to deploy causal notions of fairness (as opposed to the observational notions discussed in Chapter 3) in the context of predictive policing.

APPENDIX

A

HC AND LO USAGE

A.1 HCS

optimization (FA): Because the parameter space for PredPol is non-convex (see Chapter 2 for a discussion of each parameter), any given run of the EM algorithm is likely to return a local optimum. Thus, I added multiple restarts with random initial values and chose the set of parameters that produced the highest likelihood of generating the data.

simulation (FA): This thesis contains a combination of theory and simulation, with an emphasis on the latter. The post-processing modification task introduced in Chapter 4 may theoretically correspond to satisfying a notion of fairness, but that must be tested on actual data and prediction results. Technically speaking, one can only guarantee fairness at prediction time, in terms of the predicted intensity values. Simulation is required to verify these results in practice.

correlation (FA): In several places throughout this text (Section 3.1 and Section 5.5), Pearson's r is used as a cheap test for statistical independence, though the two terms are not synonymous. Pearson's r tests for a linear relationship among values and can be fooled by a variety of other types of relationships (such as a parabolic one). Statistical independence, on the other hand, is a more abstract notion that accounts for a variety of dependencies, including linear ones. I chose to use correlation as a indicator for dependence for three reasons. First, in the variables considered, non-linear dependencies would be quite odd and hard to explain (for instance, consider a parabolic relationship between proportion black/white in a grid cell and the number of crimes). Second, testing for conditional independence among continuous variables is a difficult problem. Moreover, the points involving Pearson's r are not the main results of this work (i.e. testing whether PredPol satisfies sufficiency and explaining the relationship between predicted intensities and demographics). Further work would be necessary to verify or refute the early results from Pearson's r, but either outcome would not undermine the central theses of this work.

modeling (EA): The word "model" has been overloaded; the aims of predictive and descriptive modeling are quite different. My results suggest that PredPol may function well as a descriptive model for criminologists and sociologists to examine,

but real-world constraints hampers its application to predictive modeling. It is possible to make abstract ML models that perform better than PredPol (e.g. Flaxman, Chirico, Pereira, and Loeffler (2018)), but those models have even less descriptive value than PredPol (which, all things considered, is not that sophisticated).

studyreplication (EA): The importance of reproducibility shows up in two places in my thesis. First, I took deliberate steps to make my implementation and simulation code as reproducible as possible. I included download and cleaning scripts for all the data used in this study, and I commented each of my functions as well as my scripts. This document also contains a prose description of the implementation (Appendix C). Second, in Section 5.4, I talk explicitly about the relationship of my work to previous studies on PredPol, and suggest ways in which my work might be flawed. Both of these are in the service of fostering open science and communication.

purpose (CS), ethicalframing (CS): Predictive policing algorithms like PredPol raise ethical conflicts because both proponents and detractors of their use claim to be concerned with the betterment of people's lives. Proponents will argue that preventing crime benefits all individuals, including and especially those who live in minority communities. Detractors, on the other hand, will argue that predictive policing algorithms run the risk of reinforcing existing bias and justifying discriminatory practices against minorities under the guise of algorithmic objectivity. In the interest of being charitable, we can say that there are good reasons on either side of the debate (even if in reality, it is probably the case that some individuals will be motivated by purely political or worse, racist views). Under this charitable view, where both viewpoints share the value of concern for others, we need a different approach to resolve the ethical dilemma. In this thesis, I took both points of view and demonstrated that PredPol, with regard to either set of motivations, fails to impress. This framing squares the circle of intuitive tensions between accuracy and fairness in a predictive policing context.

Had the results of my simulation been different (e.g. had PredPol performed substantially better on accuracy but much worse with regard to fairness, or had PredPol performed better on both metrics), the ethical dilemma would have been resolved differently. In the second case, one could likely recommend PredPol. The first case presents a trickier situation, and a new ethical framing would be necessary in order to resolve the debate.

multiplecauses (CS), sampling (FA): An unexplored fact of this work is the relationship between bias in crime data and my simulation results (I make mention of this issue in the conclusion). Computer scientists and practitioners have a tendency to treat data as the truth, when in reality, data collection practices are far from guaranteed to be a random sample from the population. In this paper, what has been used as and referred to as "crime data" is actually the result of a system of interactions, only one of which is the true occurrence of criminal activity. Other important aspects of the data generation process include the decision to report some crimes but not

others and historical police policies that may distort the picture of crime. There are issues with the Chicago data itself, due to a legacy of biased (both in the statistical and normative sense) policing practices and data entry, that might change how we should interpret the results of a predictive policing algorithm. Understanding the data in this paper as the result of this complex generation mechanism leads to different interpretations of a "predictive policing" algorithm. For example, one such alternative interpretation is that the results in this paper are actually predictive of police activity, and not of crimes themselves. That interpretation of the data suggests further reason to be wary of predictive policing algorithms in general, since the data they are trained on may not actually be usable. It would also be interesting to consider if the data are systematically biased (in terms of sampling) in such a way that undermines a central result of this paper (that PredPol does not necessarily predict more accurately than a simple baseline).

audience (MC): I intend for this paper to appeal to both technical and non-technical audiences. The latter group can just skip over any technical material or formulae, since most of the arguments and contributions of this paper can be explained in non-technical terms. As a point of differentiation from other computer science papers that I have read, I have explicitly included sections which talk about normative and ethical issues, such as the relationship between different notions of fairness (Chapter 3).

A.2 LOS

justice (AH164), technicalfairness (Custom), algorithmicimpact (Custom): I demonstrate my understanding of both philosophical justice and technical notions of fairness in Chapter 3. The kind of philosophical justice discussed here could be expanded in future work, since I restrict my attention mostly to procedural justice (is the allocation process fair?) rather than substantive justice (is the final standing of each person in society fair?). The lines between these two kinds of fairness do blur. One example of substantive fairness that differs from procedural fairness is to look at the number of people of different races that end up with a criminal record as a result of a predictive policing system. That goal is different from merely demographic parity (that there is no statistical relationship between the people predicted by an algorithm and the person's race). One could have demographic parity and still have unfairness (in the substantive sense) if the later decisions in the criminal justice process affected people of different races differently. Measuring and optimizing for that kind of fairness has not been discussed in this paper.

modelmetrics (CS156), technicalfairness (Custom): I demonstrated my grasp over different ML model metrics in three ways. First, I defined and implemented an appropriate measure of accuracy for PredPol and other predictive policing algorithms (ROC-like accuracy curves as a function of grid cells visited). Second, I discussed the trade-off between several kinds of classification errors in my section about fairness

definitions. Third, I defined my own custom measure for achieving equalized odds in a continuous setting (Equation 4.1) and improved PREDPOL with respect to that measure.

expectationmaximization (CS156): I had to implement and verify an expectation-maximization algorithm to train the internal parameters for PREDPOL; see the above discussion on optimization. While I did not prove convergence for the EM algorithm in this case (relying upon the work of other researchers), I did have to understand, implement, and debug a sparse description of the EM procedure in Mohler (2014).

probabilitytheory, graphicalmodels (CS146): I applied my knowledge of statistical models to understand and implement PREDPOL. While graphical models did not explicitly factor into this paper, I did draw graphical models for myself as I worked through the papers introducing PREDPOL.

novelapplication (CS110): I proposed how a fairness task could be transformed into a version of a different computational problem, the multi-dimensional knapsack problem (itself a variant of the knapsack problem). The conceptual work exemplifying this approach is in Chapter 4, where I explain the behavior that a fair predictive policing agent ought to have and show how that behavior maps onto a knapsack problem with particular values (the predicted intensity) and constraints (the racially differentiated predictive values). I also had to deal with the challenge that ideally, the fairness task has an equality constraint, but most formulations (and solvers) for the knapsack task expect inequalities.

B

SUPPLEMENTARY MATERIAL

B.1 VERIFYING PREDPOL FOR SUFFICIENCY

Sufficiency requires that conditional on the predicted status, the protected attribute is independent from the true status. In our setting, all three variables (protected attribute, prediction, and true status) are continuous, which complicates both statistical and graphical checks. Statistically, verifying conditional independence for continuous variables is a hard problem (Bergsma, 2004). Graphically, Barocas et al. (2018) introduce the approach of plotting accuracy curves for each predicted group separately. However, because the protected attribute in our case is continuous for each grid cell, this approach also does not work.

To provide initial evidence that PredPol satisfies sufficiency, we bin grid cells by predicted intensity and calculate the correlation between number of actual crimes and demographic (percentage black or white) for the grid cells in each bin. Within each bin of cells of similar predicted intensity, we should see no correlation (close to 0.0) between the number of actual crimes and the percentage black/white. We repeat this procedure for each day that was predicted in the test set, and plot each observed correlation as a separate point (so there may be more than one point for the same bin). Not every bin on each day has enough grid cells (more than two) to perform Pearson's r meaningfully; those points are omitted from the graph.

Figure B.1 shows the scatter plot of these results. If PredPol is calibrated, we would expect most of the y-values to be close to zero, which is what we observe. We see instead that for some days, there is indeed a strong correlation between demographics and number of crimes, even after conditioning on predicted intensity. Further research is necessary to verify these results and, if the results are verified, interrogate their implications.

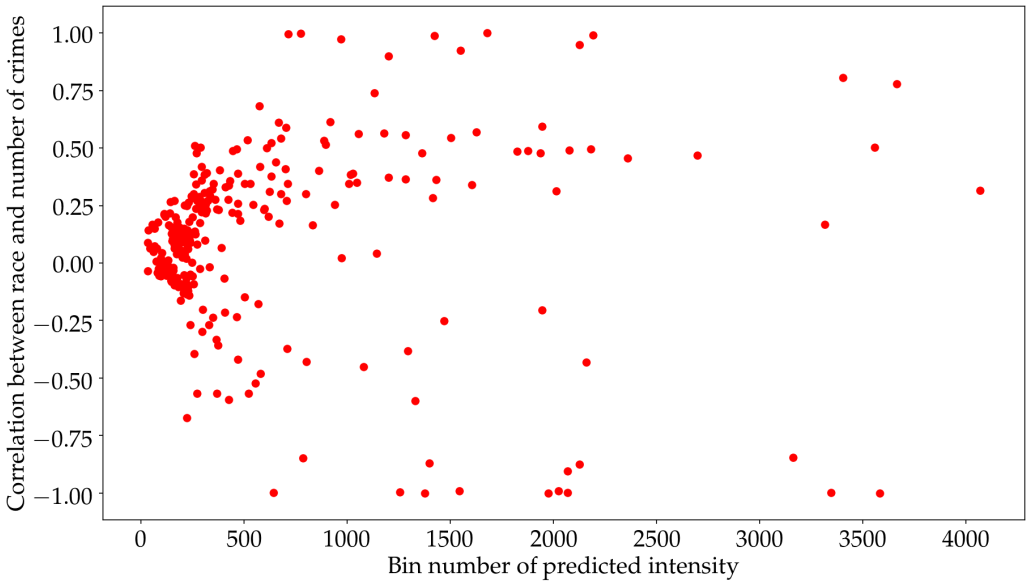


Figure B.1: Verifying that PREDPOL satisfies sufficiency

B.2 PROOF OF INCOMPATIBILITY FROM CHAPTER 3

The following proves the incompatibility of these two fairness notions in real-world situations:

Equalized Odds:

$$p(\hat{Y} \mid Y, A = 1) = p(\hat{Y} \mid Y, A = 0) \tag{B.1}$$

Equal Desert (Sufficiency):

$$p(A \mid Y = 1) = p(A \mid \hat{Y} = 1) \tag{B.2}$$
$$p(A \mid Y = 0) = p(A \mid \hat{Y} = 0) \tag{B.3}$$

Equation B.2 says, "the percentage of crime group A is responsible for equals the percentage of group A in the locations visited." The proof proceeds by showing that if all of the above equations are true, edge-case conditions must also be true (and so in general, not all of the above equations can be true).[1] Start by applying the law of total probability to the right-hand side of equation B.2.

$$p(A \mid \hat{Y} = 1) = \sum_{Y} p(A \mid Y, \hat{Y} = 1)\ p(Y \mid \hat{Y} = 1) \tag{B.4}$$
$$= \sum_{Y} p(A \mid Y)\ p(Y \mid \hat{Y} = 1) \tag{B.5}$$

1 The form of this proof closely resembles the proofs found in Barocas et al. (2018).

The second line holds from applying equation B.1, since the protected category A will be independent from predictions $\hat{Y}$ given the true status Y.

$$
\begin{aligned}
& p(A \mid \hat{Y} = 1) && \text{(B.6)} \\
& \quad = p(A \mid Y = 0)p(Y = 0 \mid \hat{Y} = 1) + p(A \mid Y = 1)p(Y = 1 \mid \hat{Y} = 1) && \text{(B.7)} \\
& \quad = p(A \mid Y = 0)p(Y = 0 \mid \hat{Y} = 1) + p(A \mid Y = 1)(1 - p(Y = 0 \mid \hat{Y} = 1)) && \text{(B.8)} \\
& \quad = p(A \mid \hat{Y} = 0)p(Y = 0 \mid \hat{Y} = 1) + p(A \mid \hat{Y} = 1)(1 - p(Y = 0 \mid \hat{Y} = 1)) && \text{(B.9)}
\end{aligned}
$$

Here, we make use of the assumptions in equations B.2 and B.3. Now dividing both sides by $p(A \mid \hat{Y} = 1)$:

$$
\begin{aligned}
1 &= \frac{p(A \mid \hat{Y} = 0)}{p(A \mid \hat{Y} = 1)} p(Y = 0 \mid \hat{Y} = 1) + 1 - p(Y = 0 \mid \hat{Y} = 1) && \text{(B.10)} \\
0 &= p(Y = 0 \mid \hat{Y} = 1) \left(\frac{p(A \mid \hat{Y} = 0)}{p(A \mid \hat{Y} = 1)} - 1 \right) && \text{(B.11)}
\end{aligned}
$$

So either $p(Y = 0 \mid \hat{Y} = 1) = 0$ or $p(A \mid \hat{Y} = 0) = p(A \mid \hat{Y} = 1)$. The first condition states that the algorithm has 100% precision. By conditioning on $\hat{Y} = 0$ instead of $\hat{Y} = 1$ in equation B.4, we also obtain the implication that $p(Y = 0 \mid \hat{Y} = 0) = 0$. In other words, this first condition states that the policing algorithm must be 100% accurate. The second condition, because of our assumptions in in equations B.2 and B.3, can also be rewritten as: $p(A \mid Y = 0) = p(A \mid Y = 1)$. It is rarely the case that the demographics of those who do commit crimes and those who do not are identical however.

Because Equation B.1, Equation B.2, and Equation B.3 together imply edge case conditions, we can conclude that in general all three cannot hold simultaneously.

C

METHODOLOGY

This appendix describes the data collection, data cleaning, and simulation process of the present research. It is meant to accompany the code provided at: https://github.com/MWYang/Capstone-FairPol, and offers a high-level description of the procedure specified in the README there.

With a fast Internet connection on a MacBook Pro (Retina, 13-inch, Early 2015, 2.7 GHz processor, and 16 GB of RAM), the whole experimental process (downloading the RAW data, processing the data, running the simulation, and collecting results) takes about 4 hours.

C.1 DATA COLLECTION

There are three sources of data for the project:

- Chicago crime data: https://data.cityofchicago.org/api/views/ijzp-q8t2/rows.csv?accessType=DOWNLOAD
- 2015 American Community Survey demographics by 2010 Census tracts: Accessed by the Python `CensusData` library (https://pypi.org/project/CensusData/)
- 2010 Census Tracts GeoJSON shapefile: https://github.com/uscensusbureau/citysdk/raw/master/v2/GeoJSON/500k/2015/17/tract.json

The first dataset is used to train PredPol and make forecasts, while the remaining two are used to associate each grid cell in the PredPol model with demographic information. Downloading demographic information makes use of the Python library `censusdata`, while the other two are available via public URLs.

C.2 DATA PROCESSING

The Chicago data are filtered to only include homicides in the years 2012-2015, inclusive. Rows with missing data in the 'X Coordinate', 'Y Coordinate', 'Latitude', 'Longitude', or 'Date' columns are dropped as these are all features required for simulation.

To prevent overflow errors in the PredPol simulation, x- and y-coordinate values are rescaled. The rescaling makes use of longitude and latitude calculations to ensure that 0.01

rescaled units corresponds to 150m, so that the final grid cell size in PredPol is 150m × 150m. Moreover, we also 0-1 normalize the dates in the data to also avoid overflow.

C.3 PREDPOL SIMULATION

As described in Chapter 2, PredPol simulation proceeds by iteration over the set of historical training data. First, PredPol learns its internal parameters (η, ω, and σ) via an expectation-maximization (EM) procedure. The EM equations are reproduced from Mohler (2014) below. In the E-step, compute two matrices $\mathbf{p}$ and $\mathbf{p^b}$ that indicate the probability of the ith event causing the jth event. If there are N crimes in the training data, then both these matrices will be $N \times N$ square matrices. The sum of all the values in $\mathbf{p} + \mathbf{p^b}$ also measures the likelihood of the data given the current parameters, which is used to assess the fitness of the parameters across random restarts.

$$p_{ij} = \frac{\omega \exp\left(-\omega(t_j - t_i)\right) \frac{1}{2\pi\omega^2} \exp\left(-\frac{(x_j - x_i)^2 + (y_j - y_i)^2}{2\sigma^2}\right)}{f(x_j, y_j, t_j)} \tag{C.1}$$

$$p^b_{ij} = \frac{\frac{1}{2\pi\eta^2} \exp\left(-\frac{(x_j - x_i)^2 + (y_j - y_i)^2}{2\eta^2}\right)}{T f(x_j, y_j, t_j)} \tag{C.2}$$

T is the length of the time window (the maximum t-value, if the minimum value is 0) in the training data. Given the values in the E-step, we update the values of the parameters in the M-step with the following equations:

$$\omega = \frac{\sum_i \sum_j p_{ij}}{\sum_i \sum_j p_{ij}(t_j - t_i) + \sum_i (T - t_i) \exp(-\omega(T - t_i))} \tag{C.3}$$

$$\sigma^2 = \frac{\sum_i \sum_j p_{ij} \left((x_j - x_i)^2 + (y_j - y_i)^2\right)}{2 \sum_i \sum_j p_{ij}} \tag{C.4}$$

$$\eta^2 = \frac{\sum_i \sum_j p^b_{ij} \left((x_j - x_i)^2 + (y_j - y_i)^2\right)}{2 \sum_i \sum_j p^b_{ij}} \tag{C.5}$$

EM proceeds by taking alternate E-step and M-step updates until the parameter values converge.

To make predictions, PredPol takes as an input all historical crime data observed up until the day desired for prediction. Predicted intensities are calculated for each grid cell using Equation 2.1 in Chapter 2. The grid cells can then be sorted by the predicted intensity to output a list of predictions. We do so for every day in 2015, using the data from years 2012-2014 and any previously seen days in 2015 to compute intensities. Because this process is fairly time-consuming, we store all of the predicted intensities for each day in 2015 to speed up our fairness modifications.

C.4 FAIRNESS MODIFICATIONS

Computing the post-processing modification task described in Chapter 4 is relatively straightforward. After computing the $\hat{f}$ values as described in that section, we pass those values as weights to a library for solving the multi-dimensional knapsack problem. To speed up the performance of the library, we restrict the number of items that the knapsack solver receives: if K items are desired ultimately, we take only the top $5K$ items, as ranked by the original PREDPOL intensity. We found experimentally that restricting items in this manner did not impact the ultimate accuracy of the knapsack solver, and instead resulted in a significant boost in runtime. Our function for computing the fairness modification takes in a parameter controlling this restriction, so researchers can experiment with this as well.

BIBLIOGRAPHY

Barocas, S., Hardt, M., & Narayanan, A. (2018). *Fairness and machine learning*. fairmlbook.org. Retrieved from http://www.fairmlbook.org

Barocas, S. & Selbst, A. D. (2016). Big data's disparate impact. *SSRN Electronic Journal*. doi:10.2139/ssrn.2477899

Benslimane, I. (2014, June 18). *Étude critique d'un système d'analyse prédictive appliqué à la criminalité : Predpol®*. Retrieved from https://cortecs.org/wp-content/uploads/2014/10/rapport_stage_Ismael_Benslimane.pdf

Bergsma, W. P. (2004). *Testing conditional independence for continuous random variables*. Eurandom.

Brantingham, P. J., Valasik, M., & Mohler, G. O. (2018, January 1). Does predictive policing lead to biased arrests: Results from a randomized controlled trial. *Statistics and Public Policy*, *5*(1), 1–6. doi:10.1080/2330443X.2018.1438940

Celis, L. E., Straszak, D., & Vishnoi, N. K. (2017, April 22). Ranking with fairness constraints. *arXiv:1704.06840 [cs]*. arXiv: 1704.06840. Retrieved December 15, 2018, from http://arxiv.org/abs/1704.06840

Chouldechova, A. (2017, June). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, *5*(2), 153–163. doi:10.1089/big.2016.0047

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017, January 27). Algorithmic decision making and the cost of fairness. *arXiv:1701.08230 [cs, stat]*. doi:10.1145/3097983.309809. arXiv: 1701.08230

Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., & Venkatasubramanian, S. (2017, June 29). Runaway feedback loops in predictive policing. *arXiv:1706.09847 [cs, stat]*. arXiv: 1706.09847. Retrieved September 18, 2018, from http://arxiv.org/abs/1706.09847

Ensign, D., Frielder, S. A., Neville, S., Scheidegger, C., & Venkatasubramanian, S. (2018). Decision making with limited feedback: Error bounds for predictive policing and recidivism prediction. *Proceedings of Machine Learning Research*, *83*, 9.

Flaxman, S., Chirico, M., Pereira, P., & Loeffler, C. (2018, January 9). Scalable high-resolution forecasting of sparse spatiotemporal events with kernel methods: A winning solution to the NIJ real-time crime forecasting challenge. *arXiv:1801.02858 [stat]*. arXiv: 1801.02858. Retrieved March 14, 2019, from http://arxiv.org/abs/1801.02858

Hardt, M., Price, E., & Srebro, N. (2016, October 7). Equality of opportunity in supervised learning. *arXiv:1610.02413 [cs]*. arXiv: 1610.02413. Retrieved May 30, 2018, from http://arxiv.org/abs/1610.02413

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016, September 19). Inherent trade-offs in the fair determination of risk scores. *arXiv:1609.05807 [cs, stat]*. arXiv: 1609.05807. Retrieved September 22, 2018, from http://arxiv.org/abs/1609.05807

Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., & Hardt, M. (2018, March 12). Delayed impact of fair machine learning. *arXiv:1803.04383 [cs, stat]*. arXiv: 1803.04383. Retrieved June 5, 2018, from http://arxiv.org/abs/1803.04383

Lum, K. & Isaac, W. (2016, October 1). To predict and serve? *Significance*, *13*(5), 14–19. doi:10.1111/j.1740-9713.2016.00960.x

Mohler, G. O. (2014, July 1). Marked point process hotspot maps for homicide and gun crime prediction in chicago. *International Journal of Forecasting*, *30*(3), 491–497. doi:10.1016/j.ijforecast.2014.01.004

Mohler, G. O., Raje, R., Valasik, M., Carter, J., & Brantingham, P. J. (2018). A penalized likelihood method for balancing accuracy and fairness in predictive policing. (p. 6). IEEE international conference on systems, man, and cybernetics (SMC2018).

Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., & Tita, G. E. (2011, March). Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, *106*(493), 100–108. doi:10.1198/jasa.2011.ap09546

Mohler, G. O., Short, M. B., Malinowski, S., Johnson, M., Tita, G. E., Bertozzi, A. L., & Brantingham, P. J. (2015, October 2). Randomized controlled field trials of predictive policing. *Journal of the American Statistical Association*, *110*(512), 1399–1411. doi:10.1080/01621459.2015.1077710

Narayanan, A. (2018, February 23). *21 fairness definitions and their politics*. Tutorial. Conference on Fairness, Accountability, and Transparency (FAT*), New York, NY, USA. Retrieved from https://www.youtube.com/watch?v=jIXIuYdnyyk

Perry, W. L. (2013). *Predictive policing: The role of crime forecasting in law enforcement operations*. Santa Monica, CA: RAND.

Richardson, R., Schultz, J., & Crawford, K. (2019, March 7). Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *New York University Law Review Online*, *Forthcoming*, 30. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3333423

Robinson, D. & Koepke, L. (2016, August). *Stuck in a pattern: Early evidence on "predictive policing"and civil rights*. Upturn. v1.0.1. Retrieved January 16, 2019, from https://www.upturn.org/static/reports/2016/stuck-in-a-pattern/files/Upturn_-_Stuck_In_a_Pattern_v.1.01.pdf

Saunders, J., Hunt, P., & Hollywood, J. S. (2016, September 1). Predictions put into practice: A quasi-experimental evaluation of chicago's predictive policing pilot. *Journal of Experimental Criminology*, *12*(3), 347–371. doi:10.1007/s11292-016-9272-0

Singh, A. & Joachims, T. (2018). Fairness of exposure in rankings. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining - KDD '18*, 2219–2228. doi:10.1145/3219819.3220088. arXiv: 1802.07281

Zehlike, M., Bonchi, F., Castillo, C., Hajian, S., Megahed, M., & Baeza-Yates, R. (2017). FA*IR: A fair top-k ranking algorithm. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management - CIKM '17*, 1569–1578. doi:10.1145/3132847.3132938. arXiv: 1706.06368