A Critical Study of Geospatial Algorithm Use in Crime Analysis and Predictive Policing

Katy Weathington
Marquette University

A CRITICAL STUDY OF GEOSPATIAL ALGORITHM USE IN CRIME ANALYSIS AND PREDICTIVE POLICING

by

Katherine Weathington

A Thesis submitted to the Faculty of the Graduate School,
Marquette University,
in Partial Fulfillment of the Requirements for
the Degree of Master of Science

Milwaukee, Wisconsin

May 2020

ABSTRACT
A CRITICAL STUDY OF GEOSPATIAL ALGORITHM USE IN
CRIME ANALYSIS AND PREDICTIVE POLICING

Katherine Weathington

Marquette University, 2020

We examine in detail two geospatial analysis algorithms commonly used in predictive policing. The k-means clustering algorithm is used to partition input data into k clusters, while Kernel Density Estimation algorithms convert geospatial data into a 2-dimensional probability distribution function. Both algorithms serve unique roles in predictive policing, helping to inform the allocation of limited police resources.

Through critical analysis of the k-means algorithm, we found that parameter choice can greatly impact how crime in a city is clustered, which therefor impacts how mental models of crime in the city are developed. Interviews with crime analysts who regularly used k-means revealed that parameters are overwhelmingly chosen arbitrarily. Similarly, KDE parameters greatly influence the resulting PDF, which are visualized in difficult to interpret heatmaps. A mixed method user study with participants of varying backgrounds revealed that those with backgrounds in law enforcement and/or criminal justice rarely actively chose the parameters used, in part due to not fully comprehending the meaning of less obvious parameters. It was also found that individuals with different backgrounds tended to interpret heatmaps and make resource distribution decisions differently.

There are several implications from these findings. Primarily, this implies that most would-be users lack the training and expertise to reliably implement and interpret geospatial crime analysis algorithms. Both within and without crime labs, critical thought is rarely given to parameter choice, especially for parameters without a clear, easily understandable explanation. These factors illuminate predictive policing being an inexact science, despite being taken as reliable and objective. These shortcomings and misconceptions, due to their pivotal role at the earliest part of the policing and criminal justice system, have long term consequences for denizens of any place being policed at behest of an algorithm.

ACKNOWLEDGEMENTS

Katherine Weathington

I would like to thank my colleagues who have worked with me on one or more of these projects (in chronological order): Marielle Billig, Christopher Supinger, Joe Chudzik, Md. Romael Haque, and Regina Vela-Mesta. Thanks to those who contributed to the early stages of these projects before my involvement: Griffin Berlstein and Justin Miller. Special thanks to my advisor, Dr. Shion Guha, and my other committee members, Dr. Michael Zimmer and Dr. Aleksandra Snowden.

TABLE OF CONTENTS

1 INTRODUCTION

The criminal justice system is the subject of many heated debates and has been at the center of many scandals, often with concerns raised over unfair treatment and systemic bias. As the impacts on those subjected to the criminal justice system are far reaching and entirely capable of ruining careers and personal relationships, small choices or overarching policies in the criminal justice system will have long lasting and hard to predict ripples. But the criminal justice system does not begin in a courthouse, rather, it begins with police actions and, by extension, policing decisions. While there are many well-known criticisms of both the attitudes and actions of police, there is a lesser known layer that influences the entire shape of policing in a city: predictive policing and crime analytics.

Predictive policing describes the practice of augmenting current knowledge of crime trends in an area through the implementation of data analysis tools such as clustering algorithms on crime locations and time series analysis of crime rates. In many American cities, these efforts are the result of a small crime lab of individuals with degrees in criminology, sociology, or public administration with at most a limited education in data science techniques.

Two of the most prevalent geospatial algorithms used in predictive policing are k-means clustering and kernel density estimations (KDEs). K-means partitions data into k descriptive clusters. KDEs process data points and outputs a probability density function. Ideally, these tools would allow analysts to visualize where crimes tend to happen in a given area, which would then inform the police on where to allocate car and foot patrols. There are several motivations for implementing predictive policing. By allowing policing to become more proactive, you can maximize police workhours in order to save cost. Furthermore, the grounding of decisions in data provides a facade of fairness and creates a point of accountability beyond individual officer choice. However, the criminology theory as well as common sense dictates that higher allocation of police to an area would lead to higher rates of arrests being made in that area. Therefore, misuse of predictive policing algorithms can easily create a self-reinforcing data loop leading to more, likely unnecessary and detrimental, police presence. This motivates our research questions: how are crime mapping tools being implemented, and how well are they being interpreted? This paper will combine the findings from two studies to attempt to answer these questions for the most popular geospatial crime analysis algorithms, utilizing a mixed methods approach to both identify trends and develop a meaningful understanding of the context surrounding the findings in the data.

2 RELATED WORKS

2.1 Algorithmic Crime Mapping

Algorithmic crime mapping is the use of modern information processing technology to combine GIS data, digital maps, and crime data to facilitate the understanding of spreading of crime. According to Mamalian et al., algorithmic crime mapping enables law enforcement agencies to analyze and correlate data sources to create a detailed snapshot of crime incidents and related factors within a community or other geographical area [22]. It is a versatile tool for crime investigation officers to understand the spreading of crime [43] and has already been applied to different crime types, including drug incidents [66], environmental crimes [18], burglary [18], gang violence [48], burglary repeat victimization [46], residential burglaries [2], and serial robberies [42].

The Bureau of Justice Statistics' Law Enforcement Management and Administrative Statistics (LEMAS) surveys of 1997 and 1999 indicate that crime mapping technology was adopted and used by law enforcement agencies after 1999. Following LEMAS's survey of 1997, the national survey conducted by the Crime Mapping Research Center (CMRC) of the National Institute of Justice was distributed to determine which agencies used GIS, the purpose of usage, and reasons for refusing it [58]. A pilot study was conducted to directly examine the adoption of algorithmic crime mapping in police agencies by choosing a random sample of 125 police agencies from the LEMAS 1999 survey of departments with 100 or more police officers [90]. Two additional important findings are, firstly, the existence of a direct link between the use of algorithmic crime mapping and hot spot approaches in policing, and second, both basic and applied research about crime places and hot spots played an important role in the process of diffusion of algorithmic crime mapping.

The early adoption of algorithmic crime mapping also happened in several countries outside the United States. A browser-based mapping application Map-based Analytical Policing System (MAPS) was released on the New Zealand Police network in late 2000 [6]. In Rio de Janeiro, Brazil, the space-time monitoring of geographical cells Monitora Espacio Temporal (CEMET) was applied across the entire state by using ArcGIS and digital maps to identify crime patterns [69]. In addition, Victoria Police department in Victoria, Australia, developed a tool to simplify the use of MapInfo GIS software by introducing Geographic Intelligence Unit (GIU) and implement crime mapping at many locations across the state [59].

In practice, the majority of hotspot and place-based predictive policing algorithms focus not on arrests, but on crimes predominantly reported to police by the public (e.g., robbery, burglary, assault) [3, 61]. Spatial clustering has been investigated to detect where crimes concentrate in space and time, e.g. to detect hotspots, or to predict future crime location [18, 93]. Spatiotemporal correlations over longer time periods have been investigated to further enhance hotspot detection [85]. The most common methods are spatial ellipses, thematic mapping of geographical areas, grid thematic mapping and Kernel Density Estimation (KDE) [19]. KDE is one of the most popular techniques and has proved itself to be very effective in terms of precision and prediction [19]. The technique is also known for producing smooth and precise maps [18, 28]. Several other approaches include a new crime hotspot mapping tool - Hotspot Optimization Tool (HOT), an application of spatial data mining to the field of hotspot mapping. The key component of HOT is the Geospatial Discriminative Patterns

(GDPatterns) concept, which can capture the differences between two classes in a spatial dataset. The pros and cons of utilizing related factors in hotspot mapping are discussed. [87]. Other research shows computational co-offending network analysis is an effective means for extracting information about criminal organizations from large real-life crime datasets, specifically police-reported crime data which is virtually impossible to obtain such information by using traditional crime analysis methods [84].

However, several concerns have been associated with such tools and methodologies. Research has demonstrated that a racial bias exists in policing, including the racial profiling of vehicles [11, 13, 44, 88], pedestrian stops [1, 31, 36], traffic tickets [27], drug enforcement and arrests [12, 49, 54], use of force [17, 52, 65], and even in the decision to shoot white or black criminal suspects while in a training simulator [70]. While the mechanisms driving these observed patterns of racial disparity (i.e., racial profiling, stereotyping/cognitive bias, deployment, racial animus/prejudice) remain difficult to disentangle [89], there is little doubt that racial disparities in policing outcomes do exist. Racial bias of predictive policing algorithms has been the focus in recent research articles [15, 45]. With regards to place-based predictive policing methods that forecast a time and location where a crime may occur, the concern is that racially biased police practices may be directed toward some areas rather than others. Knowing that they are in a prediction area may heighten the awareness of police officers in ways that amplify biases. That is, a minority individual observed in a prediction area may be more likely to be subject to biased police actions than the same individual observed outside of a prediction area [32]. There are also significant privacy concerns with hot spot policing [9, 50]. The dissemination of spatial crime data can be problematic when the locations of crimes can be linked to specific addresses and individuals as police reports are public record, and a number of police departments offer online crime mapping tools [9]. Lastly, some studies investigated factors [55, 73, 74] that affect the use of algorithmic crime technology. This research suggests algorithmic crime mapping should be used by law enforcement agencies to focus on increasing number of full-time paid employees, providing academy training, assigning patrol officers to specific areas/beats, and updating technology frequently to support the analysis of community problems [55].

2.2 Human Algorithm Interaction

For over 20 years, the academic community has proposed numerous guidelines and recommendations for how to design for effective human interaction with AI-infused systems [5]. Early supervised machine learning algorithms have relied on reliable expert labels to build predictive models. However, the gates of data generation have recently been opened to a wider base of users who started participating increasingly with casual labeling, rating, annotating, etc [64]. The increased online presence and participation of humans has led not only to a democratization of unchecked inputs to algorithms, but also to a wide democratization of the consumption of machine learning algorithms' outputs by general users. Hence, these algorithms, which are essential building blocks of recommender systems and other information filters, started interacting with users at unprecedented rates [75]. The result is machine learning algorithms that consume more and more data that is unchecked, or at the very least, not fitting conventional assumptions made by various machine learning algorithms [20].

Recently, there are findings that highlight opportunities and challenges in designing human-centered algorithmic work assignment, information, and evaluation and the importance of supporting social sense making around the algorithmic system [51]. The potential of rich human computer collaboration via on-the-spot interactions is a promising direction for machine learning systems and users to collaboratively share intelligence [62,83]. Several methods incorporating study participants into the process of data analysis and interpretation have been proposed recently [10,24,29,92]. Algorithmic interfaces in socio-technical systems rarely include a clear feedback mechanism for users to understand the effects of their own actions on the system. The increasing prevalence of these opaque algorithms coupled with their growing power raises questions about how knowledgeable users are versus how knowledgeable they should be about these algorithms' "existence", "operation", and the "biases" they might introduce to users' experiences [29].

More recent influences on user studies in interactive cartography include the related areas of human computer interaction (HCI) and usability engineering (UE) [80]. Scientists working in HCI have produced a range of technology-driven research on interaction design that is broadly applicable to the cartographic context [57,80]. Furthermore, cartographers have borrowed empirical methods commonly used in HCI such as interaction logging, task analyses, and think aloud studies to supplement psychology and geography-based approaches when digital interactivity is provided. Scholars in HCI are increasingly turning their attention to interactive maps [23,37,39,67,72,79], pointing to an increased mutual influence as maps become interactive and move online or to mobile devices.

For many years, interactivity has been touted as the primary way to support visual thinking in the context of geographic visualization, with the goal of generating new hypotheses in unknown datasets to support scientific exploration [56,60]. However, the ubiquity of interactive maps presents emerging opportunities to study interaction design beyond exploratory spatial data analysis [34]. Geo-visual analytic and “big data” science is one important use case [77]. Future research also needs to approach interaction design for a general audience, in which the interactive maps and visualizations serve the purpose of communication, personalization, and even entertainment. These very different use and user contexts present different methodological opportunities and challenges regarding participants, materials, and procedures, and the degree to which insights regarding exploratory visualization can be transferred to these different contexts currently remains unclear [78].

Moreover, questions derived from critical science and technology studies are also needed to inform qualitative research on interactive maps and visualizations, particularly to understand how interactivity empowers and potentially misleads or marginalizes its users [41]. For instance, how does interactivity deferentially impact user access to or trust in the information behind the map? [33]. Do interactive maps and visualizations that reach marginalized populations disproportionately serve as propaganda or surveillance? [21] Do they compromise our privacy, or change the ways we construct and negotiate public space? [91] Additional research must be adapted to approach such critical questions about new use cases for interactive maps and visualizations.

2.3 Transparency & Criminal Justice Algorithms

Algorithms, complex mathematical formulas and procedures through which computers process information and solve tasks, have an increasing impact on people's lives [4]. Algorithms are replacing or augmenting human decision making in crucial ways. People have become accustomed to algorithms making all manner of recommendations, from products to buy, to songs to listen to, to social network connections. However, algorithms are not just recommending, they are also being used to make big decisions about people's lives, such as who gets loans, whose resumes are reviewed by humans for possible employment, and the length of prison terms [82].

As artificial intelligence and algorithmic prediction come quickly to penetrate local governance, it would be desirable for the public to know what policy judgments the algorithms reflect and how well they perform in achieving the objectives set for them. It will be possible to assess a predictive algorithm's politics, performance, fairness, and relationship to governance only with significant transparency about how the algorithm works. One such use context of algorithms in which people's lives depend on the outcome is criminal justice algorithms.

Criminal justice algorithms, sometimes called "risk assessments" or "evidence-based methods," are controversial tools that purport to predict future behavior of defendants and incarcerated persons [7]. These proprietary techniques are used to set bail, determine sentences, and even contribute to determinations about guilt or innocence. Yet the innerworkings of these tools are largely hidden from public view. As criminal justice algorithms have come into greater use at the federal and state levels, they have also come under greater scrutiny. Many criminal justice experts have denounced "risk assessment" tools as opaque, unreliable, and even unconstitutional [16,25]. As many "risk assessment" algorithms take into account personal

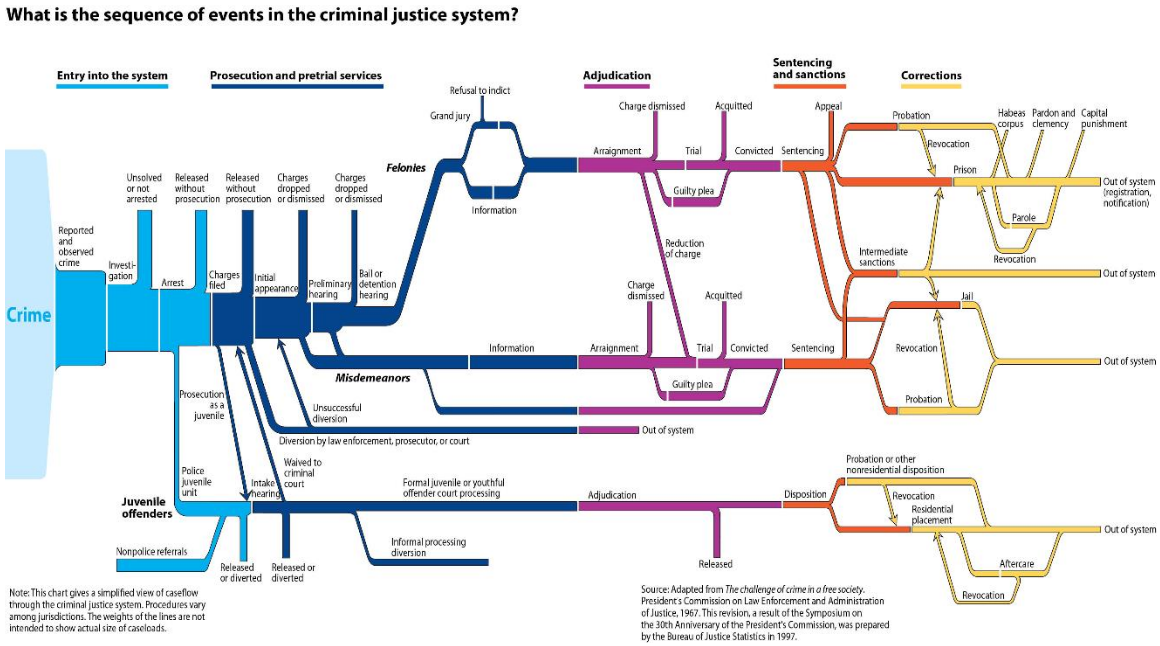

Fig. 1. **The sequence of events in the criminal justice system.**

characteristics like age, sex, geography, family background, and employment status, two people accused of the same crime may receive sharply different bail or sentencing outcomes based on inputs that are beyond their control and have no avenue to assess or challenge the results [47]. In May 2016, ProPublica released an in-depth report about COMPAS, suggesting that it was both racially biased and inaccurate. According to ProPublica's analysis, the scores not only proved "remarkably unreliable" in forecasting violent crime, but they also contained significant racial disparities—even though the formula does not officially take race into account [47].

However, even though in Figure 1 we can see "reported and observed crime", "investigation" and "arrest" (sky-blue region) are top in the chain of criminal justice system events which are usually fed into the future crime prediction algorithm and algorithmic crime mapping tools [3,61], very few works [25,50,74] usually focus on the hidden bias that might be involved in the tools or the data [1,13,36,44,49,54,88] that fed into it. Because the ways in which these systems reach their conclusions may reflect or amplify existing biases, or may not offer explanations that satisfy our accustomed social and judicial expectations, there is growing concern that the traditional frameworks for implementing transparency and accountability may not suffice as mechanisms of governance.

3 STUDY ONE: K-MEANS CLUSTERING ALGORITHM

The k-means algorithm partitions data into precisely k clusters based on distance. For geospatial data, it is often implemented based on latitude and longitude, with several possible formulas for measuring distance. K-means, if implemented correctly, can be an effective method of revealing natural clusters of events in your data. Our study sought to critically examine the inner workings of k-means and its parameters as well as gain insight into how it is used by real world crime analysts.

3.1 Methods

We started by interviewing two professional crime analysts to get initial insights into algorithmic crime mapping practices. We used publicly available arrest data about the city of Milwaukee for 12 years (2005-2016) as an empirical lens of investigation. We focused on the k-means algorithm because it is both a commonly

ALGORITHM 1: Potential Bias Index

```
Input: G: geodesic cluster
Input: E: list of unique euclidean clusters in G
Output: I: Potential Bias Index
numGeodesicPoints ← getPointCount(G)
minorityRatio ← getMinorityRatio(G)
clusterScore ← 0
// for each euclidean cluster found in G
foreach e_i ∈ E do
    // for each point in geodesic cluster
    euclideanPoints ← 0
    matches ← 0
    foreach p_j ∈ e_i do
        // if euclidean point is in geodesic cluster
        if p_j ∈ G then
            matches ← matches + 1
        end
        numEuclideanPoints ← numEuclideanPoints + 1
    end
    score ← matches/numEuclideanPoints
    weight ← matches/numGeodesicPoints
    index ← score ∗ weight
    clusterScore ← clusterScore + index
end
dissimilarity ← 1 − clusterScore
potentialBiasIndex ← dissimilarity ∗ minorityRatio
return potentialBiasIndex
```

Fig. 2. **Algorithm for Calculating Potential Bias Index**

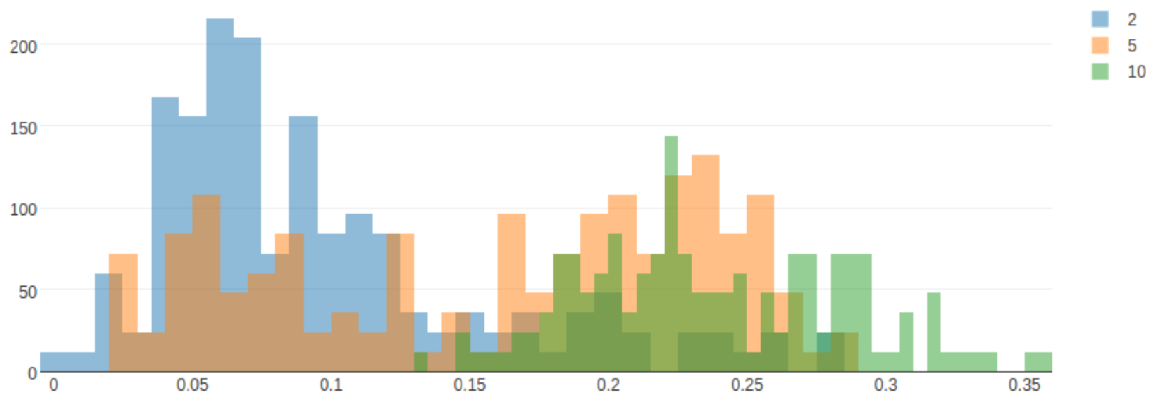

Fig. 3. **Frequency of Potential Bias Values Based on Number of Clusters**

used crime mapping tool and its flaws are intuitive to understand for the layperson. We restricted our analysis to four common crimes: robbery, simple assault, theft, and motor vehicle theft, which are commonly mapped by analysts. We created visualizations of potential bias and used publicly available demographic information to create a Potential Bias Index (PBI) (Figure 2) that we used as visual aids in the next round of interviews.

Then, we conducted follow-up interviews of 17 people. Eleven of them were professional crime analysts working in the greater Milwaukee and Chicago metropolitan area. Six participants were local community organizers working to improve opportunities and reduce crime in the inner city. We adopted a grounded theory perspective [94] to our work. After multiple iteration of thematic analysis, initial high-level themes have emerged from the qualitative data.

3.2 Results & Discussion

3.2.1 Deconstructing k-means for potential biases

Examining Lloyd's algorithm for k-means, we found two inflection points for potential human bias [95] to enter the model: (a) the initial selection of clusters and (b) the choice of the distance metric. Note that these are the two parameters required by the k-means algorithm. Considering the choice of number of clusters, shown in Figure 3, the potential bias values for both theft and motor vehicle theft ranged from 0 to a high of 0.36. The average potential bias for a given k ranged between 0.069 and 0.17 for theft and between 0.063 and 0.1706 for motor vehicle theft. In general, values of k greater than 4 produced an average bias value greater than or equal to .14, while values of k less than 4 produced values less than 0.1.

For theft, the gold standard of 5 clusters produced a low potential bias value of 0.0315 and a high value of 0.3099 with a mean of 0.1442 and standard deviation of 0.0562. Motor Vehicle Theft had a larger range with a low of 0.0180, a high 0.3495, a mean of 0.1457, and a standard deviation 0.0665. Theft exhibited lower standard deviation than motor vehicle theft, likely due to the higher number of data points (900 vs 400). But between both, when high potential bias values are produced, the associated clustering typically featured two different configurations of the city center, while the clusters in the northern and southern ends of the city tended to be similar. This is likely due to the sparser nature of points on the city periphery, while the density of points toward the center of the city created more "unstable" initializations that result in high potential bias scores.

Considering the parameter of distance metric and looking at a given geodesic cluster, dissimilarity can increase in two ways. First, dissimilarity will increase when the number of unique Euclidean clusters present increases. Geodesic cluster purity will decrease dissimilarity. Second, dissimilarity will increase if a small ratio of Euclidean points is found inside the geodesic cluster compared to the number of points in the Euclidean cluster. This dissimilarity score can be between 0 and 1. Zero means a

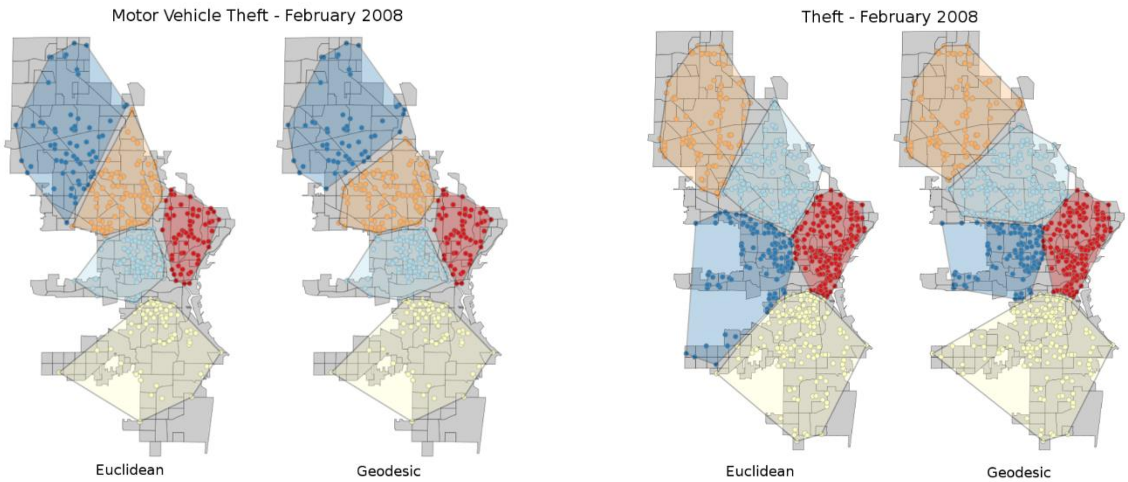

Fig. 4. **Comparison of Euclidean and Geodesic Distance Metrics on Two Sets of Crime Data**

geodesic cluster matches perfectly with a Euclidean cluster. If a geodesic cluster contains small fractions of many different Euclidean clusters, its score will approach 1. A visualization of this effect is presented in Figure 4.

3.2.2 Default behavior of crime analysts

One of the main findings from our interviews is that, in general, crime analysts were unclear about the theoretical design and inner workings of the algorithms that they were using. Decisions made during data analysis were mostly supplemented with prior knowledge and existing mental models of the city.

All the analysts we interviewed had master's degrees in criminology, crime analysis, sociology or public administration and some had taken a few courses in applied statistics, like Matthew. Some participants such as Jill reported complete unfamiliarity with statistical distance metrics after we explained how k-means worked and displayed our visualizations:

> "I didn't know what these distance things [metrics] are...I understand the Euclidean that...the calculation of the straight line because we

learnt it in high school but I didn't know that there were other ways to calculate distance. I just point and click [on the GUI based crime analysis software that they use developed by a private third party]."
-Jill

In this case, Jill does not change the default distance metric (Euclidean) that is provided in the software even though other options are present. Others point to a lack of transparency and clarity within the choices provided by the software that they use and a confusion in selecting appropriate options. This leads them to select default options. For instance, 37-year-old male crime analyst John stated:

"When I go to run the clusters [referring to k-means or other clustering methods], there are many other options on the menu but I don't know most of them so I just go with the default options on the menu... we were taught a basic idea of clustering but I didn't know that we could have so many different options." -John

This refers to a general lack of transparency in how this third-party software designs and implements the algorithms. When faced with a variegated menu of choices, the analysts select the one that is most familiar i.e. the default option. Taken together, this type of analysis is rule-based and path-bound [96]. It is natural to be paralyzed by a suite of potential options and then choose the most familiar one, however incorrect it might be under the given circumstances. However, when asked about how they decide to select the initial number of clusters, some participants responded that they depended on existing institutional knowledge about crime in Milwaukee. For instance, when asked about city-level clustering, Kevin referred to extant institutional knowledge that is likely already biased:

> "When I started the job, I was told that we always divide the city into five main divisions. There is the downtown cluster, the northshore cluster where all the rich folks live...you have the northwestern and southside clusters where there is a lot of gang activity and then the west side near the suburbs where a lot of people commute from." - Kevin

Any subsequent analysis depends on this initial categorization that is dependent on institutional knowledge. Therefore, this type of analysis is based on situated decision making [96]. We observe here that while domain knowledge is very important, when combined with what we learnt about the statistical (in)appropriateness of the actual process, there is a lot of potential for misclassification and untoward policy making. Relatively few people request to switch from the default regardless of what the default is. Clearly, the default selected by policymakers has important implications.

4 STUDY TWO: KERNEL DENSITY ESTIMATIONS

Kernel density estimation is an algorithm which converts discrete data points into a probability distribution, the shape and size of which depends on user input parameters. These calculate probability distributions are summed then normalized in order to define a singular probability distribution function which can be used to predict areas likely to have future events occur. For geospatial crime analysis, KDEs are generalized from the common univariate implementation to a 2-dimensional bivariate analysis. The output of a bivariate implementation is often visualized as a heatmap. Our study seeks to identify trends in how users with various backgrounds choose parameter values, interpret heatmaps, and use this knowledge to make resource allocation decisions in the specific context of Milwaukee crime.

4.1 Methods

We conducted a three month long empirical study with participants in Milwaukee, USA. A total of 60 participants were individually interviewed to discover how people from different educational and professional backgrounds interact with crime-mapping algorithms. We developed an online, interactive web application which displays a heatmap output of a Bivariate Kernel Density Estimation (also referred to as KDE here) based on data parameters and user chosen KDE parameters. We were able to analyze the interaction and ability to interpret the crime-mapping algorithm through an interactive activity where participants were asked to choose different parameters for the KDE algorithm and then to identify hotspots on the map. We were able to find participants' values and needs for such algorithms through a one-on-one interview. The interview helped to supply necessary context by providing participants' current thoughts and concerns on crime analysis tools. This study was approved by the Institutional Review Board (IRB) of Marquette University where this study took place.

4.1.1 Application design

Data used in our application was taken from the Wisconsin Incident-Based Reporting System (WIBRS), a publicly accessible database organized by the Wisconsin Department of Justice [8]. Data had to be accessed by time and was collected starting with 2017 (the last full year before this project began) and then worked backwards. We were able to cultivate a dataset of crimes in the City of Milwaukee from 2012 to 2017. We chose to focus on quality of life crimes, such as robbery, larceny, and motor theft, because they are common and tend to have the most effect on an average person's life. Moreover, most people would also be familiar with or have encountered such crime in their life. In contrast, crimes such as rape, murder, terrorism, etc. are relatively rare events in the grand scheme of things [53]. Street addresses provided in the dataset were then cleaned using a manually assembled regex library then geocoded using Google's geocoding service. Data was organized into individual months for each unique type of crime.

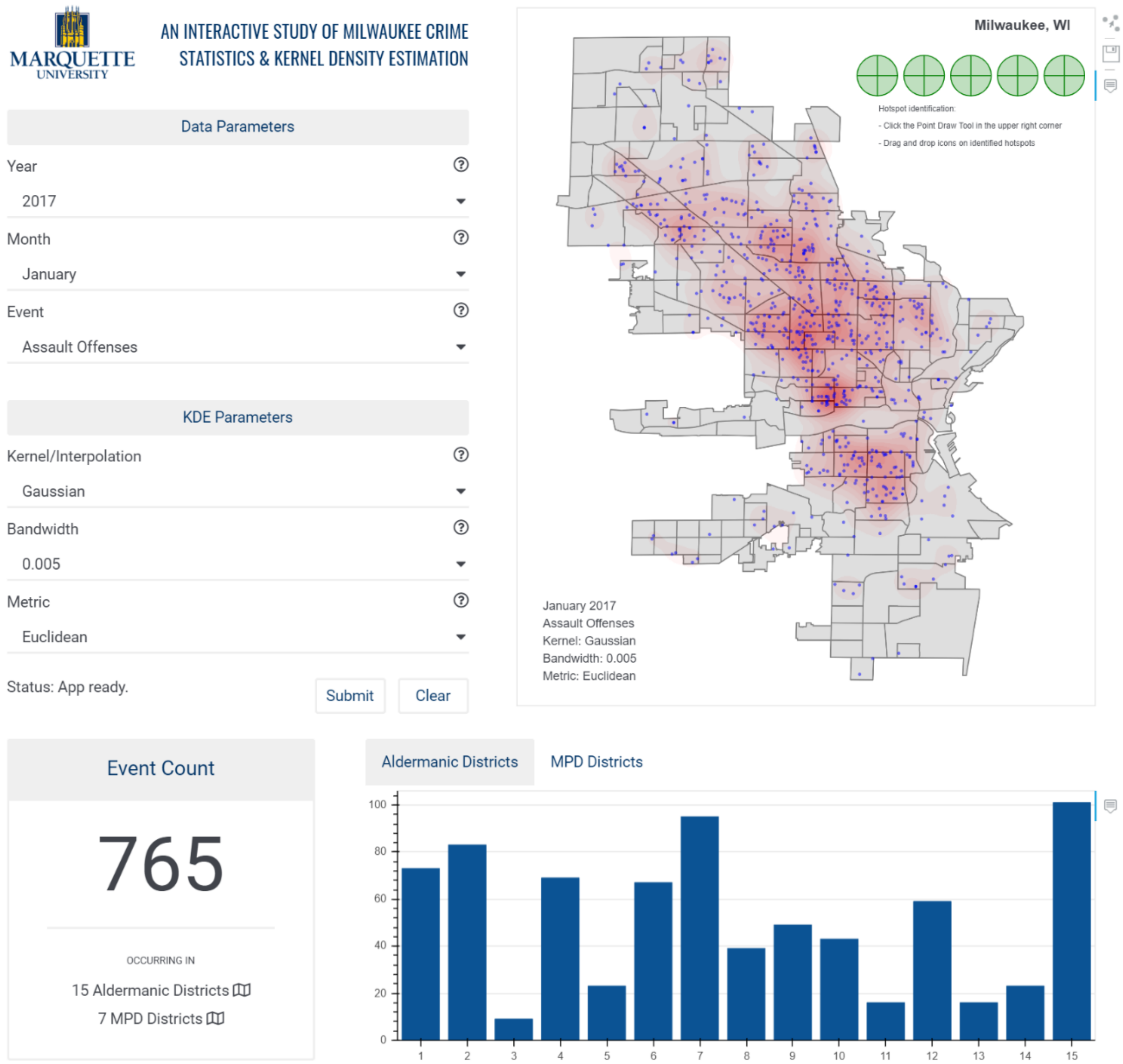

Fig. 5. **Screenshot of Application Used in Our Study**

To facilitate our study, we built an online, interactive application, shown in Figure 5, which prominently displays the heatmap output of a Bivariate Kernel Density Estimation (KDE) based on specified data and KDE parameters. We chose to present KDE outputs for two reasons: firstly, we know from prior knowledge of common practices by crime labs that KDEs are one of, if not the most, commonly used crime mapping algorithm [38], and secondly that we wished to understand the predictive capabilities of individuals which are fostered by the PDF outputs of a KDE heatmap.

KDE parameters	Default Options	Other Available Options
Kernel	Gaussian	Tophat, Epanechnikov, Linear
Bandwidth	0.005	0.0025, 0.0075, 0.01
Distance Metric	Euclidean	Manhattan, Chebyshev

Table 1. **List of Parameters, Default Setting, and Other Options**

Each parameter dropdown list had a question mark tooltip in order to answer some general questions about parameters. The data parameters allow users to choose one type of crime from a specific month and year from 2012 to 2017 in the City of Milwaukee. KDE parameters and options are outlined in Table 1.

KDEs attempt to find and represent the underlying probability density function that a set of data was taken from. By providing a PDF, the heatmap allows users to predict high and low probability areas for future events. KDEs accomplish this by smoothing each discrete data point into a two-dimensional probability distribution function (PDF) with the original point at the mean, then aggregating the PDFs into a singular heatmap for the entire area. The shape of the distribution that data points are smoothed into is defined by the kernel parameter, also called interpolation method. The shapes provided by these are equivalent to common probability distributions those found within standard statistics. The bandwidth, or smoothing, parameter controls the width of each distribution. In statistics, this would be analogous to the variance of a symmetric distribution. For example, a gaussian kernel with a bandwidth of one will result in the standard normal distribution, though you generally want a much smaller bandwidth in order to produce meaningful results. Higher values for bandwidth result in much smoother outputs, which would lose power but reduce bias from overfitting. Our final parameter is distance metric, which controls how distance between points is

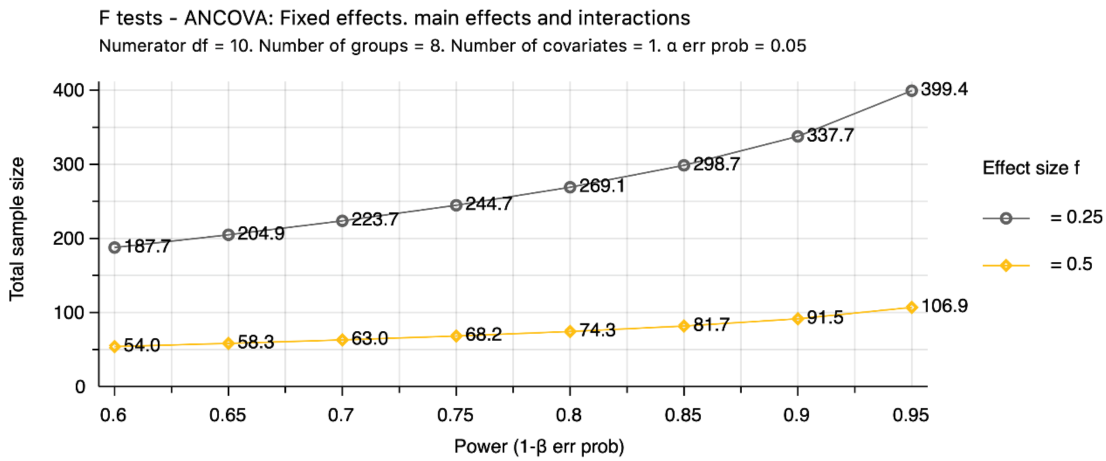

Fig. 6. ***A Priori* Power Analysis to Determine Number of Participants**

measured. The most commonly used distance metric is Euclidean, or merely a straight line between two points on a plane, which is the default setting for distance metric.

After submitting an initial set of parameters, users are shown the calculated heatmap with several interactive features, as well as a bar chart displaying the number of events in either each police district or each aldermanic district. Hovering over data points in the heatmap will show a pop-up with information about the date, districts, and location of the event.

4.1.2 *A Priori* Statistical Power Analysis

In order to determine how many participants should be recruited, we performed a standard *a priori* statistical power analysis, depicted in Figure 6. Based on our power analysis we set a minimum goal of 60 participants, ensuring 0.6 power for low effect size. Achieving this number affords us a reasonable ability to discover trends in our collected data.

4.1.3 Participants

We decided to recruit participants from the Milwaukee area as many studies today have been crowdsourced through Amazon Mechanical Turk. Having Milwaukee residents allows us to obtain community feedback from people who already have knowledge about the city they live in. Recruitment was performed in the form of advertisement flyers posted around the city of Milwaukee in coffee shops, community centers, and public libraries as well as various online forums and groups. If an individual found interest in the study, they would contact us via email and we would then confirm that the individual was at least 18 years of age. Next, we would propose specific meeting times for an in-person experiment. If the participant was unavailable for any of the initially proposed times, we would propose different session times. After a time was accepted by the participant, we would send a confirmation email to the participant with the accepted time, the location of the experiment, and mark down the time in a master schedule for study organization.

The participants were individuals (n=60) from the Milwaukee area with varying backgrounds with regards to algorithmic crime analysis. We assigned subjects to one of three groups depending on their self-reported background and experience. Group 1 (n_1=39) consists of those with no background relevant to algorithmic crime analysis. That is, they have no background in either data analysis nor in criminal justice and law enforcement. Group 2 (n_2=14) have a technical background involving programming or algorithmic analysis. Group 3 (n_3=7) consists of law enforcement officers or those otherwise professionally involved in criminal justice and civil peacekeeping. A small number of members of Group 3 also had technical backgrounds akin to Group 2. Our goal of separating participants into three different groups was to analyze the ability to

Demographic Criteria	Participant Description	No. of Participants
Gender	Male	23
	Female	36
	Transgender Male	1
Age Range (in Years)	18 -21	29
	22-30	20
	31-40	4
	40+	7
Education	High School Diploma or GED	2
	Undergraduate (Enrolled)	32
	Bachelors	20
	Masters	5
	Doctorate	1
Job Type	Full-Time	17
	Part-Time	14
	Unemployed	4
	Self Employed	3
	Student	21
	Retired	1

Table 2. **Demographic Overview of Study Participants**

interact and interpret crime analysis methods, specifically KDEs, through their different uses and needs for the algorithm. An overview of the participants' demographics can be found in Table 2. Each participant was given a specific participant identification number such that we may be able to use their data and interview for analysis without revealing personally identifiable information.

4.1.4 Experiment Overview

Once the participant arrived for the experiment session at the confirmed time, the experiment began by providing the participant with an informed consent form to obtain data. After signing the consent form, the experiment would continue by administering an interactive activity with our developed web application, moving to a two-part online survey consisting of the NASA-TLX scale to review workload of the activity and a demographic questionnaire, and concluding the experiment with a recorded, oral interview. The developed web application, surveys, and the oral interview are further described in detail in the following sections. After finishing the oral interview, participants would receive compensation and the experiment would conclude. Figure 7 displays a flowchart of the complete study process.

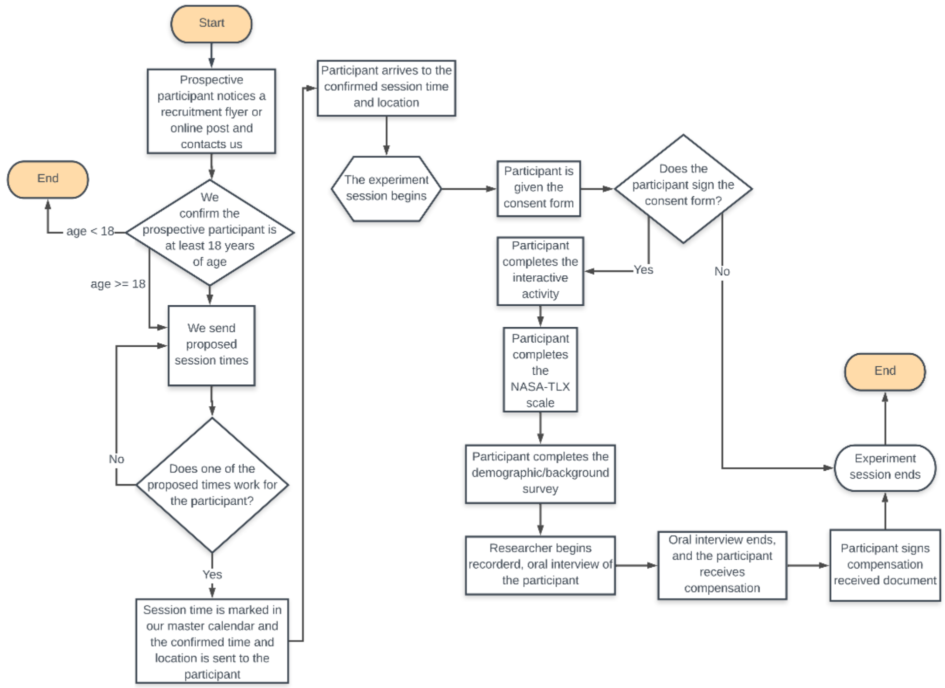

Fig. 7. **Flowchart of Experiment Process**

In order to gauge their ability to understand, interpret, and actively use implemented algorithmic crime analysis, participants were asked to perform a series of activities on heatmaps of various complexities. While participants were able to change parameters of the KDE to one of several options, data parameters were provided and consistent across maps. Therefore, complexity of the underlying data patterns rather than a specific heatmap output was considered when choosing exemplar data parameters for each complexity. To determine which maps would be shown, we went through many different heatmaps as a team and selected those with clearly different levels of complexity ranging from distinct hotspots scattered across the map, to hotspots that blurred together and were not easy to pinpoint. Map A is the least complex, with data points naturally forming relatively distinct clusters. Map B was

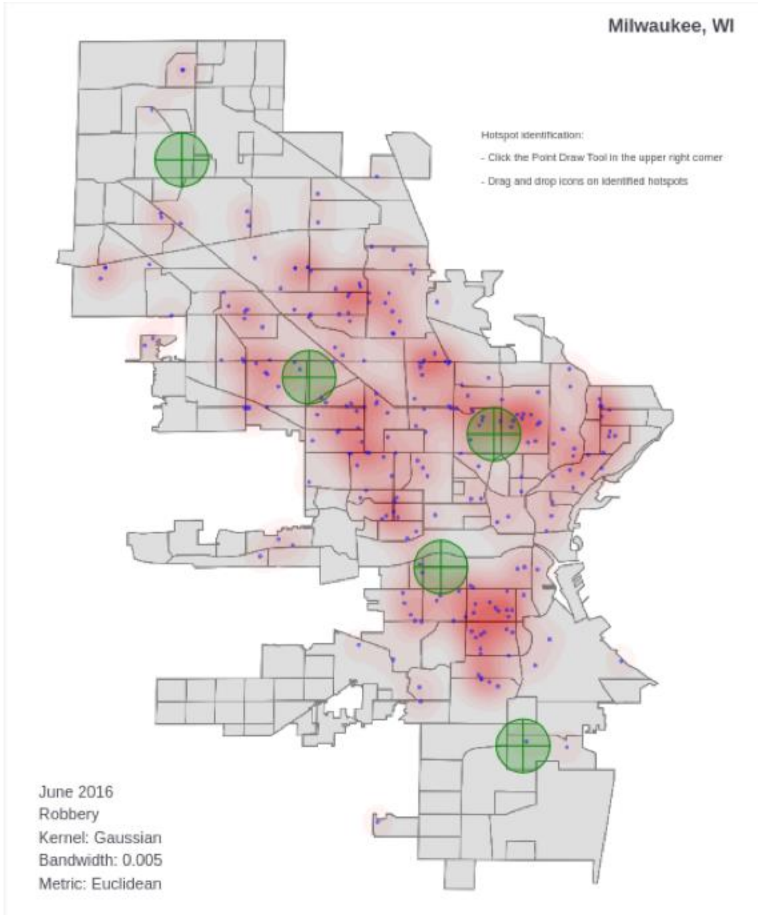

Fig. 8. **Example of Heatmap With User Allocated Circles**

more complex, with more blurry edges around natural clusters. Map C was the most complex, with many of the clusters bleeding together yet still maintaining slight variations in density. For each of these maps, we identified and agreed on four values to be collected.

For each map presented, participants were asked to estimate a minimum and maximum number of hotspots they see in the heatmap. Next, participants were prompted to imagine that the green circles with crosshairs represented the area that one police patrol unit could most effectively patrol. Then they were asked to, keeping in mind what they represent, provide a minimum and maximum number of green circles needed to effectively address crime based on this heatmap. Finally, participants were instructed to place 5 of the green circles in locations they believed would be the best allocation of resources. A picture of the heatmap with circle allocation and chosen parameters was saved, an example of which is depicted in Figure 8.

After the interactive activity, participants were assessed with NASA-TLX [40] for workload of the application and activities. The TLX is a measure of perceived workload. Workload, like usability, is a complex construct but essentially means the amount of effort people must exert both mentally and physically to use the interface, measured using six dimensions: Mental, Physical, and Temporal Demands, Frustration, Effort, and Performance. They then answered a survey with questions about their prior level of familiarity with algorithmic crime analysis, their understanding of the application, feelings towards law enforcement and government, and general demographics. Finally, participants were interviewed with a series of pre-written questions, though conversation was fostered and often moved away from the prepared questions. The conversation began by asking if the participant has ever used or seen a system of heatmaps such as in the activity to gauge the participant's background and knowledge of the algorithm. We asked questions about their views on the legality, ethicality, and fairness of data, about their willingness to offer personal data to law enforcement for algorithm training, and their level of concern about algorithmic crime mapping and the factors of which they would consider important to know. Other recurring, but not explicitly written, topics included interactions with law enforcement and relevant background with either law enforcement or algorithmic data analysis. Interviews were transcribed as they were being recorded with an application [68], then were saved and later review by multiple primary investigators. This ensured accuracy and consistency in qualitative analysis.

4.1.5 Analysis

In order to statistically confirm the significance of trends found in our descriptive figures and our intuitions regarding RQ1, we performed two types of tests. Because our data values are discrete and cannot be assumed to come from any well-defined distribution, we have used non-parametric, or distribution free analogues to the Student's T-Test and Levene's Test of Variance, the Mann-Whitney U test and Fligner's test respectively. After consideration, the Kruskal-Wallis test, a non-parametric equivalent to ANOVA, was found to be inappropriate for the low effect size commonly seen in these studies, especially given our sample size [76].

The Mann-Whitney U test is signed rank test [86] that can be used as a non-parametric alternative to the Student's T-test. By ranking the observations based on value, and maintaining their sign, it can indicate whether the median of the population the samples were taken from are equal or not.

Fligner's test is a squared rank test which tests whether the populations two samples are drawn from have equal variance. It acts much like a non-parametric analogue to Levene's Test. Even where measure of central tendency may not be significantly different, a difference in variance of the samples and therefore populations indicate some level of different capability or motivations. Lower variance indicates more precision, even if the overall accuracy, or measure of central tendency, happens to be the same. Fligner's test can use either mean or median as the measure of central tendency to use to calculate variance, we have provided results for both.

To address RQ2, we compute the probability distribution of sticking with the default parameters to understand the default behavior among the groups. We have also administrated the NASA TLX survey to assess their mental workload of the tasks performed. Administering the TLX involves two steps. First, a participant reflects on the task they are being asked to perform and looks at each paired combination of the six dimensions to decide which is more related to their personal definition of workload as related to the task. This results in a user considering 15 paired comparisons. For example, a given participant needs to decide whether Performance or Frustration represents the more important contributor to the workload for the specific tasks performed. The second step involves participants rating each of the six dimensions on scales from Low to High or from Good to Poor. The raw score for each of the six items is multiplied by the weight from step 1 to generate the overall workload score per task.

Our qualitative interviews were transcribed in an online platform - Otter [68]. Both the audio recordings and the transcriptions were initially stored in that platform. They were then downloaded to safe storage where two principal investigators checked the automated transcription by hand to eliminate any discrepancies. The transcripts were then analyzed using thematic analysis [14]. After reading through the transcripts carefully, we conducted several rounds of iterative coding to identify patterns and converge them into appropriate themes. This analysis was re-examined and confirmed by a PI with extensive domain knowledge in criminology. Low-level themes were created by synthesizing the findings of these steps. Finally, high level themes were developed through cross referencing. We adopted Grounded Theory [35] approach to qualitative data analysis.

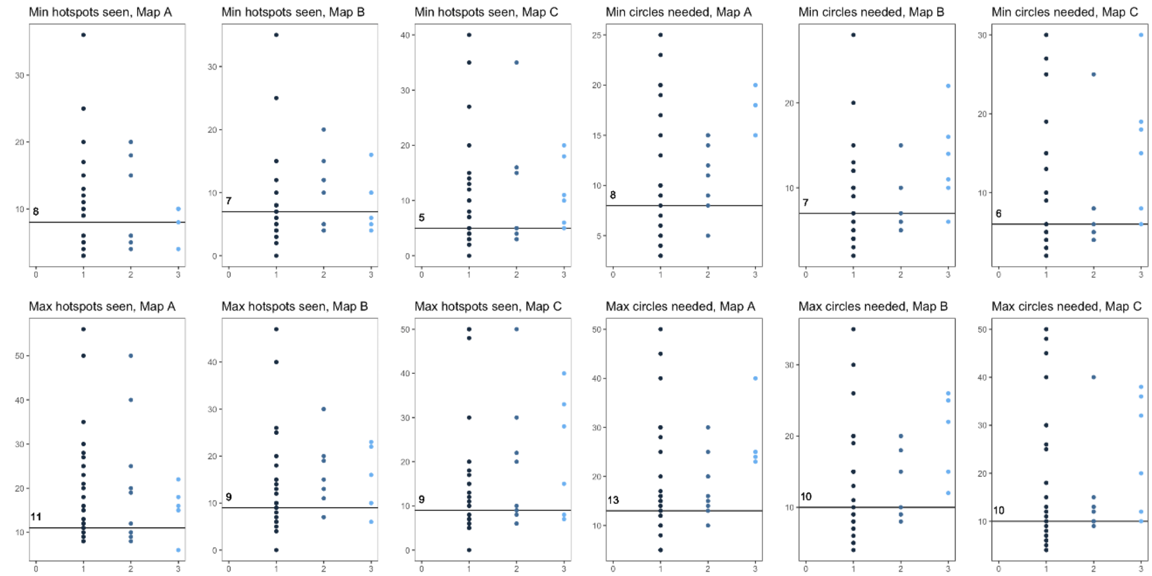

Fig. 9. **Distribution of Hotspot and Circle Estimations Compared to Gold Standard (Solid Line)**

4.2 Results

We have divided our results in three segments. In our findings, we elaborate on how people across different background interpret the crime mapping algorithm, how they interact with it, and lastly, what the values and needs required for crime mapping algorithm are.

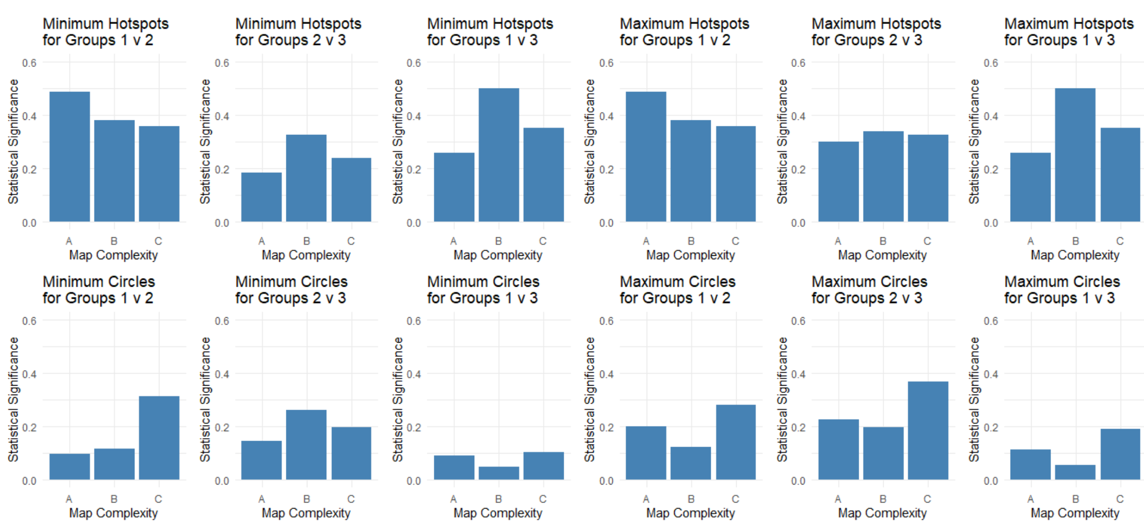

Fig. 10. **Statistical Significance of Mann-Whitney U Tests for Difference in Median Values by Group**

4.2.1 Interpretation of crime mapping algorithm

Trends Found Through Quantitative Analysis. Figure 9 shows the distributions of minimum and maximum hotspots and circles grouped by background. For our statistical analysis, we compared these values to a pre-determined gold standard value for each map complexity, which has been included in the graphs in order to see which distributions tended to over and underestimate. We see some variations both between groups and between the same groups in different complexities.

We ran Mann-Whitney U tests to determine if there is a difference in interpretation between individuals with different backgrounds and whether complexities of the heatmap factors in. The results of these tests (Figure 10) indicate that medians of each sample are different when estimating both minimum and maximum circles across different backgrounds. On average, members of Group 1 estimated a minimum patrol circles needed to be 15.82, an average of 8.82 greater than the relevant gold standards, and a maximum of 24.61, 13.61 more than the gold standard. This is much less than the averages of Groups 2 and 3, which estimated a low of 22.60 and 20.43 respectively and a high of 32.17 and 29.29. These values are higher than the relevant gold standards on average by 15.60, 13.43, 21.17, and 18.29 respectively.

Using Fligner's test to compare Group 1 to Group 2 for Map B and C, tests for all four measurements were found to be significant (Figure 11) , confirming the descriptive statistics that Group 2's estimates tend to be more precise than Group 1's, especially with more complex maps. Similarly, the statistically significant results of Groups 1 vs 3 at estimating hotspots for each map shows a difference in variance, confirming the higher precision of group 3 seen in the descriptive statistics. The lack of significance for the tests of estimating circles between Groups 1 and 3, however, indicates that there is equal precision at estimating circles. As a group, participants of different backgrounds tended to estimate the same median values. However, the different groups tended to have significantly different variances, especially in estimating hotspots. Groups 2 and 3 had much lower variance than Group 1, indicating individuals from Group 1 having less individual capability to identify hotspots and, to a lesser extent, estimate circles.

Members of Group 2 and especially Group 3 were much closer to the gold standard than members of Group1 when estimating hotspots. On average, Group 3 were 2.9048 and 7.7143 from the gold standard for minimum and maximum hotspots

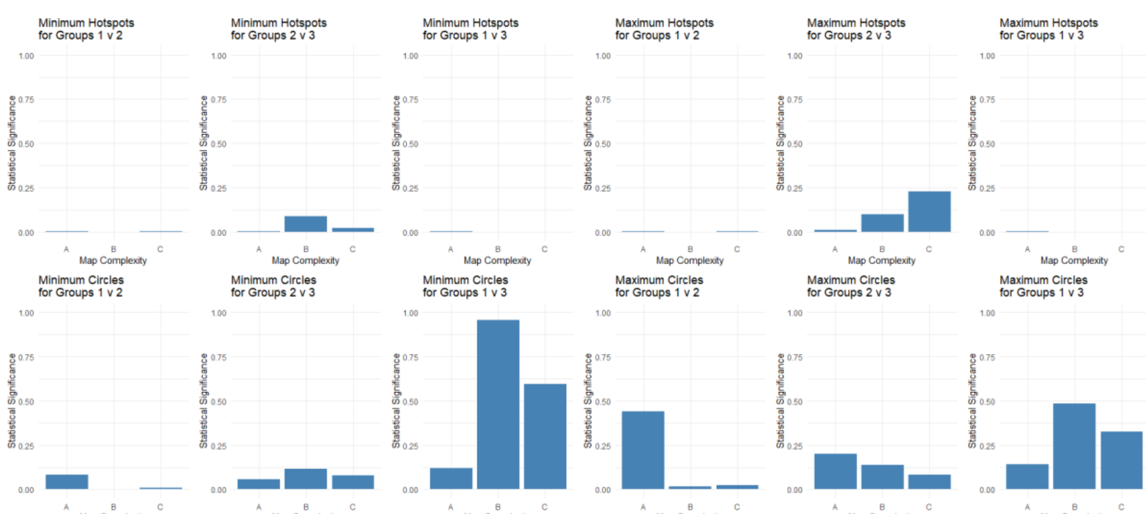

Fig. 11. **Statistical Significance for Results of Fligner's Test for Variance**

respectively. Similarly, Group 2 estimated 6.9048 and 13.3571 more hotspots than the gold standard. Group 1 estimated an average of 24.4017 minimum hotspots and 37.3248 maximum hotspots, which were on average 17.7350 and 27.6581 hotspots more than the gold standard. This indicates that members of Group 1 were worse at estimating hotspots than members of Group 2 and Group 3, grossly overestimating the number of hotspots present.

We have noticed several key factors that influences interpretation of heatmaps and how users estimated hotspots and patrol needs. Common factors are familiarity or lack thereof with crime in Milwaukee and familiarity or lack thereof with crime analysis algorithms.

Familiarity with Crime in Milwaukee. As all participants reside in or around Milwaukee, they entered the experiment with some general knowledge or stereotypes of the city, which may have impacted their choices. Some members of Group 3 used their background knowledge of general high crime areas to inform their decisions. We know this was the case for P16, a member of Group 3, who has detailed knowledge about the rates and places of drug crimes from her work as part of the judicial system. She somewhat ignored the hotspots of the map and placed circles based on her prior field knowledge.

> "... So I'm very familiar with drug crimes, the rates of drug crimes, the areas of drug crimes. If you were to ask me, such as about, like, robberies and stuff, if they're related to drug crimes, I would know roughly [the] rates but not so much other crimes outside of that.." - P16 LEA

On the other hand some LEAs combined their knowledge and the existing crime map they are seeing. Using both field experience and exploratory analysis in the crime map, they thoroughly made their decisions. Crime maps are especially useful for depicting where crime is, but LEAs must have knowledge of the city and its layout in order to properly put an end to the crime. P60 carefully described how he interpreted the map and how his prior knowledge influences his decisions on where to allocate resources.

> "With property crimes … burglaries … [for] things like that, I have to take a look at the way the map is as well. I know, for example, if like, Fond Du Lac Avenue on northside is a main thoroughfare. If there's robberies that are occurring, for example, in downtown, and we hear the broadcast, that's one thing that I'm really looking at is what's the quickest way to get out of downtown. A lot of our robbery offenders do tend to reside either on the north side or south side. And I look at the most direct route that would leave the downtown area. That, along with pawn shops, where they located or retail stores or strip malls, where are they located? Why am I seeing more dots here? Is it because of its prior to police work for traffic stops field interviews? Or is it calls for service for retail theft? " -P60, LEA

In this particular case, the participant is combining both his observation of the heatmap with knowledge about local roadways and thematic crime hotspots in order to place officers both in high crime hotspots and with ample avenues to address outlying crimes quickly. These observations pointed to the fact that LEAs have significant field experience which can be combined with the existing analysis to provide better and more efficient outcomes.

Familiarity with Crime Analysis Algorithms. Participants with technical backgrounds use their familiarity with the crime analysis algorithms in interpretation. This familiarity influences their decision-making process. Most of the participants who have seen crime maps before were used to static outputs of the algorithm instead of being able to choose their own parameters. While the interactive process was pretty new to people across different backgrounds and most, if not all, have little experience in interacting with the algorithm, people from technical backgrounds were a little surprised with the interactive session such, as P34 expressed.

> “Actually, yeah, not quite where it’s so interactive, and you can like, pinpoint things. I’ve seen a little reverse where, like, you have a map, and you can click on it, and it’ll tell you the demographics. Like, there was a robbery here like June 2016. Okay. And so a little more reverse, not quite, where you get to choose what you [want]" - P34, Technical

This shows some experience in the general topic of crime mapping software, though used in a somewhat different manner from our experiment. However, LEAs might have little knowledge with interacting the algorithm, but they were able decode the map with their familiarity of how crime analysis can be done. They used their experience in crime analysis to described why they are seeing such a result in the map. Using her experience in this domain, P61 explained the definition of a hotspot as a current area of where a patrol officer already is and used that in interpreting the map.

> “So the people, the officers who are the administrators who would be doing data analysis or be looking at crime maps, might give directive

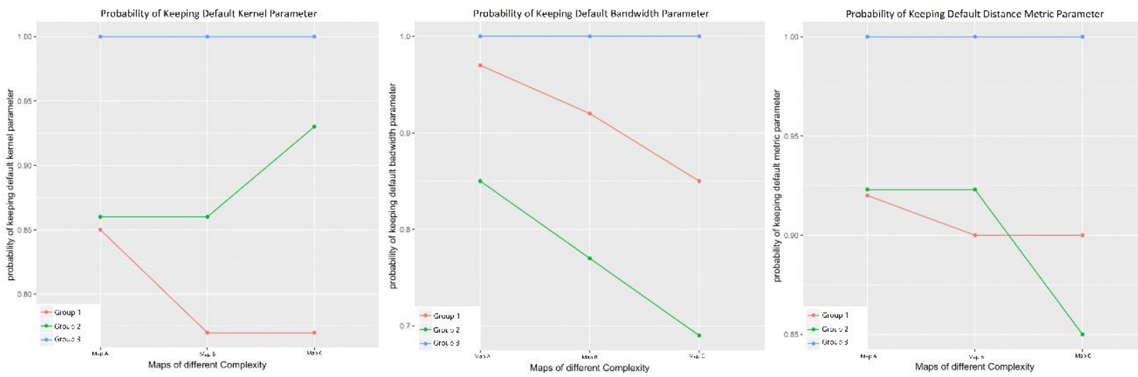

Fig. 12. **Interaction Graphs of Rates of Each Group to Change Default Parameter Values**

> of having a specialized patrol zone, and to make sure that officers were in that patrol zone. So even, even if we got a call for service, and there was no other officer available, that call for service, depending on his priority, my weight in order to keep an officer a patrol zone, that was like a hotspot" -P61, LEA

So, the interaction with the crime map was influenced by their familiarity with the map for participants with different backgrounds and their existing domain knowledge for participants with technical and law enforcement backgrounds. This provides some insights on how the decision of law enforcement allocation may be motivated by having human in the loop in crime mapping algorithms. The algorithm can provide meaningful insights to where crime is happening, but having a human in the loop to decipher the output allows for added contextual knowledge that the algorithm may not have.

4.2.2 Interaction with crime mapping algorithm

Each participant was given the option to choose any of three different parameters to tinker with. Participants had the option to adjust the bandwidth, kernel, and distance metric to different levels within the web application or to continue to use the web application with the default parameters. Based on the participant's selection, we extracted how people across different backgrounds tend to interact with the algorithm.

Interaction With Parameter Values: Accepting Defaults. The interpretation plots in Figure 12 depict the changing probability of changing parameter defaults for the different maps depending on their background. The solid line describes Group 2, the dotted line describes Group 3, and the hyphenated line Group 1.

As the map complexity shifts from maps A to B, the probability that the law enforcement agents will keep the default kernel parameter remains constant, whereas the non-technical participants are less likely to keep the default kernel parameter. As the complexity increases from maps B to C, technical participants are more likely to keep the default kernel parameter, whereas the non-technical participants show a constant probability.

As the maps increase in complexity, the law enforcement agents have a constant probability; these participants are most likely to not change the bandwidth parameter. Furthermore, as the map complexity changes, both the technical and non-technical participants show a decreasing probability that these participants will keep the bandwidth parameter as the default. As the complexity increases, both the technical and non-technical participants are more likely to change the bandwidth parameter from the default. Finally, the law enforcement agents show another constant probability that these participants will not change the default distance metric parameter. Technical participants begin with having a constant probability of keeping the default parameter from maps A to B, but lose probability of keeping the default metric parameter from maps B to C. The technical participants are more likely to change the default metric parameter as the map complexity increases from medium to hard. Non-technical participants show a slight decrease in probability of keeping the default parameter from maps A to B; it is only slightly probable the non-technical participants will change the default distance metric parameter, where they next have a constant probability of keeping the default parameter from maps B to C.

The reasons behind this behavior can be explained through the qualitative analysis of the interviews with our participants. Our qualitative analysis indicates that users of different backgrounds tended to have different motivations for changing defaults and choosing which parameters to change to. For those with a technical background, we see some level of familiarity and a level of curiosity. In cases where a participant was somewhat familiar with one of the parameters or kernel options, they would try that specific parameter.

> "So I noticed after, when I did it the first time, it was the default parameters, and I wanted to see, basically, what sort of difference it would make using the other sets. I don't really know the names. I've heard of a few of the names of these parameters, but as far as what they do is kind of blurry to me. So I figured the best way would just be to run it and see what actually happens. So yeah, that's basically why I changed it to see what would actually change on the map." -P70, Technical

For some participants they were becoming accustomed to the application as they were playing more with it. Some of them developed better understanding while hopping from map to map - for some it was confusing. P47 wanted to go back to the first map she was shown to change her answers after becoming more familiar with the application and seeing the second map.

> “I think I forget the details, but after I saw the next map, the first one, to me, looked … like there were more hotspots than I thought based on comparing it to the second map. I was trying to figure out, like, what I would consider a hotspot whether it was just anything that was slightly pink or if a hotspot would be the deeper, red color.” -P47, Non-technical.

People with no relevant background appeared to pick parameters randomly, with no real understanding or justification for their choices. Not a single person in the law enforcement group changed a default. This may be because they have been trained or become accustomed to accepting the path of least resistance.

Assessing Mental Workload of Tasks. During interaction, it was reasonable to find how challenging the task (interaction with our algorithm) was for participants. To assess participant workload, we adopted the NASA-TLX survey. Among the six factors recorded, we have elected to examine only four which would best allow us to evaluate psychological demand of interacting with the KDE algorithm in our application. They are mental demand, temporal demand, performance, and frustration level. Descriptive statistics of the weighted scores for each group is presented in Table 3.

Participants	NASA TLX Score (Weighted)			
	Mental Demand	Temporal Demand	Performance	Frustration Level
Group 1	Mean = 33.15, Median = 32, Max = 75, Min = 2, SD = 16.65	Mean = 8.49, Median = 6, Max = 50, Min = 0, SD = 10.81	Mean = 25.18, Median = 21, Max = 80, Min = 4, SD = 18.99	Mean = 11.17, Median = 3, Max = 70, Min = 0, SD = 18.22
Group 2	Mean = 36.21, Median = 33.5, Max = 75, Min = 15, SD = 16.78	Mean = 8.29, Median = 8.5, Max = 20, Min = 0, SD = 6.76	Mean = 45.57, Median = 46, Max = 90, Min = 24, SD = 16.73	Mean = 8.07, Median = 3, Max = 54, Min = 0, SD = 14.14
Group 3	Mean = 39.14, Median = 25, Max = 85, Min = 6, SD = 30.23	Mean = 11.57, Median = 4, Max = 40, Min = 0, SD = 16.37	Mean = 33.56, Median = 34, Max = 75, Min = 18, SD = 20.73	Mean = 5.71, Median = 6, Max = 10, Min = 2, SD = 3.15

Table 3. **Overview of Weighted NASA-TLX Scores**

To determine if the tasks our participants performed required significant mental effort or not, we examined the mental demand, which was a relatively high value [71] for all three groups (with mean values of 33.15, 36.21, and 39.14), indicating participants had to think critically about the tasks. The second factor considered is temporal demand, which evaluates the time pressure felt during the tasks. As participants had to decide on what options to choose on the spot, it was important to evaluate whether or not they felt any stress due to time limitations. Temporal demand was found to be somewhat low for all three groups (mean values of 8.49, 8.29, and 11.57) indicating participants did not feel rushed. The third factor we considered is performance which indicates to what degree a participant feels they performed their job successfully. Group 1 averaged much lower (25.18) than Groups 2 (45.57) and 3 (33.56), indicating a higher feeling of failure for non-technical users and, to a lesser degree, LEAs. The fourth and final factor we considered is frustration, in order to determine to what degree users found these tasks to be frustrating or mentally demeaning. Frustration was found to be fairly low (range 0-9 out of 100).

Through this analysis we can fairly say building an interaction model of crime mapping algorithm will not put that much of a pressure mentally, which implies it is feasible to build such a mental model. During our experiment, our researchers noted some of the interactions. Based on such jotted transcriptions, the task was somewhat mentally demanding as most participants did not have any knowledge about different parameters and settings. Overall, this ignites an important effect in terms of feasibility in building an interaction with crime mapping algorithm.

4.2.3 Values and needs for crime mapping algorithms

While analyzing the interviews, we have found distinct patterns of perceptions regarding the explainable requirement of algorithmic crime mapping. Depending on their background, participants had different values and needs regarding the existing system such as the ethical considerations of data collection, what kind of data has been used including the potential for flaws within it, and requirements of the current system. All the ideas raised by participants represent existing concerns and needs that may be necessary to improve an efficient, interactive crime mapping algorithm.

Tensions around the ethical basis of data collection. Law Enforcement Agents sided with non-technical participants on the need for the ability to explain issues of how the data has been collected. Law Enforcement Agents briefly mentioned their concerns of how crime has been forested, how to forecast that crime, and what kind of aspects they consider. Like the members of Group 1, Law Enforcement Agents also voiced concerns of their family. Moreover, not only that the explanation will release the public frustration around this hidden methodology, they also mentioned sometimes knowing what kind of data that was fed into the system might help the analyst or the developers of the algorithm with some improvements. Overall, incorporating field knowledge will help the algorithm to be robust. P61 gave insight to some of the data being entered and explained how sometimes information can be coded wrongly in the reports when the officer meant something else. To these LEAs, it is very important to understand how this data has been fed into the system and whether the crime analysis algorithm can decipher it.

> “I think the number of crimes or incidents in crime events, the type of event that is being reported, the number of them, but I think it’s also important to know, not just how that the event is being reported, but the actual outcome of it. So, you know, someone may report a robbery, but it’s not actually a robbery, but it could be coded as a robbery as an initial call. So I think it’s really important to make sure that there is attention to the classification of an event and making sure that it was what it was actually founded or found to be, as opposed to how it was reported." -P61, LEA

We observed every LEA in our interview wants the data to be collected ethically; few of them mentioned real case situations where the information has been collected unethically, yet still proved useful and provided safety. LEA participants also mentioned proper ethical steps must be included in a document around how far they can go with this kind of information collected. LEAs described how unethically obtained may be useful but raised the point that there may be need for limitations on what kind of data should be collected and used. P13 gave us some insight to how data may be collected and used unethically with the steps to avoid it.

> “Remember, a few years ago, they were having those problems with, … mobs of … 15 year old … running. They hit … State Fair, and there’s a park in River West, or I think, maybe down at the lakefront, but they use that … cell phone data to kind of more or less like hack whoever these kids were, and then the officers… visited their parents and stuff. And sounds like that seems pretty effective, but of course there needs to be, at some point, there needs to be some kind of a judge involved. If if they’re going to be tapping into people’s personal data like that.”
> - P13, LEA

However, technical participants showed the idea that all data should be collected, yet question whether the data has been sourced ethically or not. Sometimes "true data" can point to accurate predictions, proving to be a helpful and correct standpoint where everything else might result into something incorrect. P42 talked about the effectiveness and usefulness of such algorithms. The data might be collected unethically, but ultimately it would be used to achieve some greater good.

> “… you can see two ways from it, and that’s why I’m a little confused because, obviously, if there’s a corrupt precinct and they’re laying out charges left and right for minor infractions or made up infractions, obviously, that’s not accurate. But at the same time, the data is still predicting it accurately. But at the same time, the data is still predicting it accurately because even if it’s still a precinct, even if it’s still a corrupt precinct, if you’re in that corrupt precinct, you would still have that same chance of being incorrectly charged with something." - P42, Technical

Technical participants have pointed out the fact that data needs to be collected ethically but that will only happen in a perfect world; sometimes sacrifices have to be made in a big picture, real life case. The technical participants display more trust in the algorithms. Most of their opinions reflect the fact that, to make the algorithm robust and to make it work, it does not matter how the data has been collected or that it needs to be explained or not - as long as the end result provides "accuracy" or it fits well with some gold standard model. Although an algorithm may be highly accurate for a given dataset, the algorithm should not be used solely by itself. Data may be recorded for a specific person, but that person may not always act the same way in some instances. LEAs who work in the field mentioned how important it is to be in touch with the algorithm for added "human behavior" when analyzing the output.

> “You can take data from people, but you can’t expect people to always behave in a set way. So you have to have an understanding of human nature on top of how they interact with the public and how cops interact with the public." -P59, LEA

Finally, we see a tension between the LEAs and technical participants who work with the algorithm. LEAs in the field are the ones who are using the algorithm and they are the ones who continuously mention the data needs to be collected ethically. Participants with technical background mentioned however the data has been collected should not be a concern if it leads to a good outcome. If the effectiveness of the algorithm is more important, the algorithm might prove to be useful if the rest of the system fails.

Needs for getting rid of false alarms and bias. Even though people with Law Enforcement backgrounds have concerns around the ethical basis of data collection, we found other prominent concerns and opinions regarding the existing state of algorithmic crime-mapping system. LEAs are in the field and are the ones who actually use the information from crime maps the most to maintain the safety of that neighborhood. Through their field experience, several LEAs expressed concerns that it is becoming very inefficient. From our interviews we found that the crime analysis sometimes gives LEAs false alarms or reports. LEAs are being kept far from this analysis with little room to learn how the algorithm operates and are instructed to work solely on the information that is given to them. Because of this resource constraint, it's very difficult for them if the system gives wrong reports.

> "I never knew the algorithm that was used - that was never shared. There were a couple of occasions where we, me and myself, and my team would be given a packet of information about a specific crime that was occurring in our area of responsibility, and sometimes the predictive information that was given to us wasn't consistent with what we knew as law enforcement officers just being in the neighborhood. So we would work with our intelligence fusion center, and let them know that something was off. And then I think they were fine, their algorithm." -P53, LEA

Due to limited resources, LEAs expressed that they cannot afford truncated information from the system. The algorithm plays a major role in determining necessary patrol areas and resource allocation for observation. Because of some faults in the system, there can be major problems planning and managing resources. P60 mentioned the operations and the importance of contextual knowledge of the situation. The algorithm may have information about the crime, location, and the time of the incident to dispatch LEAs to the area, but the algorithm may not know the most efficient way on how to deploy the patrol officers in that neighborhood.

> “For example, we’ve had one person, in particular for burglaries, two occasions, two separate areas. He would specifically target detached garages, specifically look at the siding, going through the siding, which is normally concealed, but either by tree or other type of inanimate object and then burglarized garage, open up the overhead door and then load up this car and take off. Not many people would see what was going on at all. And we would plot all those locations out trying to determine the time of day, that is when this is occurring, and what districts are impacted and provide that information, like a one pager to district commander so they can make the right choice or make a better informed decision as to where to deploy their resources." -P60, LEA

LEAs have to deal with day to day operations around the statistics they are given. If they were given results which can be biased either racially or through another mean, they are still the ones who must deal with the results of the algorithm. LEAs felt it’s really important that this algorithm is properly scrutinized and free of these kinds of bias even couple of times they mentioned they want to know some important aspects of the algorithms how those suspicious red flags have been generated so that they can decide on what they should act and on what they shouldn’t - which we will discuss in next section. P50 shared such a story:

> "Well, I think it’s important to know how it all works together, just because it is really tricky. I mean, don’t want to get into things like

> how much of this is racially motivated? Or how are these calls coming in? A big part of it for me is how was the call actually generated? Did somebody call in and go, hey, there's this dude over here who's suspicious. And then when you ask them, why they're suspicious, they can't give you a reason, which usually means walking while black, which for PD means you still have to send officers. We have to send somebody no matter what. But then it makes the makes the department look racist, because we're going to check it out black dudes, because people are calling in and telling us they're suspicious."
> - P50, LEA

LEAs are excluded from all information regarding the algorithm's function. From not understanding how reports are generated to dealing with false alarms and racially biased results caused by the system's errors, LEAs have found themselves in a very difficult situation. According to them, it is a struggle to manage constrained resources, properly plan, and maintain a possible public image. Several occurrences of incorrect reports have shed a negative light on those who are on the field by simply following the works of the algorithm. LEAs have shown that a predictive crime analysis algorithm is very useful, but the LEAs are not able to fully rely on the algorithm's output for allocating resources.

Needs for algorithmic interpretation and interaction. With the growing concern of false alarms and becoming inefficient day by day, Law Enforcement Agents want to be able to explain themselves how this algorithm works. They want to know exactly how the system operates because many LEAs spoke about how crimes are connected. There may be an entry in the system about one specific crime but that can be linked with another. If crime happens in a specific area, there could be a related follow-up crime in that same area. A certain crime may be reported in order to observe a specific crime itself, but that area could potentially be prone to other types of similar crime. This is where Law Enforcement Agents felt like the system greatly lagged. As the system is static and there is no feedback loop in place (and not to mention the LEAs have very limited to no understanding of the system), the system's path to efficiency is greatly hindered. From the LEA prospective, it is very clear that they want to know, at least on a basic level, the workings of the algorithm. They felt a feedback system should be put in motion in order to make the system more efficient. LEAs think that information coupled with the knowledge of the officer who is familiar with the area might be very effective in deployment. P60 explained how knowing some of the facts behind the algorithm would be very useful in planning and decision making.

> "It influences my decision-making process as to where I'm going to spend more of my time researching crimes in this area versus another. Whether it's homicides or shootings, and why are they concentrated along, let's say Center Street, for example. Why are we seeing such a high increase in crime in gangs? Is it because it's a border for two districts? Is it because we're not allocating enough resources to that area and it's allowing crime to thrive? Is it the socioeconomic background of the citizens that reside there? What is the real root cause? " -P60, LEA

LEAs (in most cases) do not have any technical background. It is not expected that they will understand all the technicalities of an algorithm, but from their interviews it is very clear that LEAs want to know a few key aspects so that they can relate that information with their field experience, ultimately making the algorithm more efficient and effective. LEAs think it is important to know the basics of how to read the output. LEAs who have little experience to see how this algorithm works also expressed thoughts around the requirement of feedback system. They also mentioned how important the field reports are - how algorithms have to go beyond simple statistics to be more efficient.

> “Without having well trained investigators drilling down to find out why this person was shot. Was it because it was a drug deal? Even though some people will not cooperate and say, I don’t know, I just was randomly walking, then I got shot. Wow. That’s another example of a walkie sniper, you know, taking another shot at somebody. And you get those excuses quite frequently and a lot of cases and until we can find out a different way to interview the person or those around him to figure out what the real sources, I think you’re still going to continue to see the same areas pop up as having those serious crimes, whether it’s non-fatal shootings or sites occurring." -P60, LEA

Moreover, LEAs did mention how simple statistics or blindly data mining with a given statistic will not help the system in becoming efficient; rather it is very necessary to give more information on how the analysis is being done. Most importantly, the information must be presented in such a way so that LEAs can interpret. P53 provided some information regarding this issue such as how reports without proper information for interpretation proved to be unhelpful.

> “It’s important to know how to interpret, you know, the information that you receive from the analysis. Simply giving numbers or, I’ve

> experienced this previously, names of people with no context, no interpretation, Is that helpful? So, you know, going beyond just simply mining the data and giving raw numbers, put some interpretation that goes along with that is helpful." -P53, LEA

LEAs also pointed out some specific facts on how engaging police departments with the algorithmic analysis might improve the system. One of the key aspects they mentioned is how someone who builds the algorithm might have zero experience of what happens in the field. People with strong technical background might know mathematics and logic of an algorithm, but LEAs mentioned numbers cannot tell the whole story. Technical people with practical experience must be associated with the whole process of algorithmic crime analysis. P59 mentioned about such gap, which is concerning.

> “If they didn’t have any background, in working with a police department, they can have just the degree in you know, computer science or data or whatever, because, I mean, you can learn a lot of your crime stuff on the job. And by talking to people that if they just, you know, came out of university with zero experience with people and you know, police departments and said, Hey, we’re gonna, we’re going to do things by the numbers now, I’d be much more concerned."
> - P59, LEA

Lastly, while interviewing the LEA participants, they also provided some key examples of how the system can be improved. The following quote from P60 summarizes all of the concerns, thoughts, values, and needs of a crime mapping algorithm: the necessity of a feedback loop, key information for interpretation of the analysis, explainability of the algorithm so that LEA can understand deeply, and an algorithm that has been built beyond mining numbers and statistic. Squads are officers are being deployed in high crime areas, they're going there because there is, For example, a heatmap that that particular district commander might have looked at and made that informed decision. And then once they're done with that area in several hours, they want to try and disrupt the criminal activity that might be in another hotspot and chip resources there. Okay, that's kind of where it's not that important, because their observations when they make those stops there, everything else is all predicated on reasonable suspicion or probable cause. And often that background idea the only shows up in the police report is a intro paragraph. And if they do need to testify to that portion, they can always refer back to any product that was produced that particular commander. Taken together, it seemed like there should be reasonable intervention in removing the gap between the algorithm and people in the field.

5 DISCUSSION

Above results suggest some broader implications around interpretability and explainability of an algorithmic decision-making process, how prior mental model influences while interacting with an algorithm and lastly several strategies for making such systems more accountable to human actors and the challenges associated with it. Some of the implications includes the need to define "user control" while building an interactive model of crime mapping algorithm, behavioral patterns such as default behavior among the users and need for creating contestability in such a system.

5.1 Challenges of Integrating Control in Explainable AI

The goal for explainable artificial intelligence is that every user will be able to understand how a machine works. Besides, the machines will come with a high level of transparency and accountability. Every machine should be able to explain why certain actions need to be taken to its users. It should also explain why that is the best option and why other alternatives may not work out for a particular situation. Explainable artificial intelligence also aims at making it obvious to users when a particular machine has failed on a particular task and when it has succeeded in the task. It will make the users be in total control and not the machines. Users should also be able to know when to trust the actions of these machines and when to verify further. And if an error is noticed, users should be able to fix the error without the intervention of any developer. The goal of this human understanding of artificial intelligence is to put humans in total control so that no action is taken without human endorsement. Thus, before any machine takes any action, regardless of believing it the best course of action, the machine still has to seek the user’s permission.

While putting this discussion in the front, our model tried to integrate human actors in the system where the users can manipulate the algorithm by choosing different parameters and put them in the front seat of the decision-making process. But due to the lack of statistical knowledge and how that works, most of the users make their decisions based on the pre-notion, self-deterministic judgement, and prior mental model. While this discussion of implementation of control in algorithmic systems seemed very rational but we must also discuss what this control means. In our analysis we saw, merely giving them control over the algorithm was not enough. In such a case, where the human actors have such great domain knowledge in their respective field but little knowledge around complex mathematics of algorithms, what controls over the algorithm is being needed need to be discussed. In front of such restraints, researches need to be more and more focused on the identification of ways to deal with the interpretability of models.

Although integrating control can facilitate generalization and transfer of ideas across fields, the interleaving of human interaction and machine learning algorithms makes reductive study of design elements difficult. for example, inappropriately attributing success or failure to individual attributes of interactive machine learning solutions can be misleading. Therefore, new techniques regarding integrating control in algorithmic systems may be necessary to appropriately gauge the effectiveness of new interactive machine learning systems. In addition, as our case studies illustrated, some interaction techniques may be appropriate for certain scenarios of use but not others. Evaluations should therefore be careful not to overgeneralize successes or failures of specific interaction techniques. Rather, the scenarios and contexts of use should be generalized to better understand when to apply certain techniques over others.

For many years, priority has been given to the performance over the interpretability leading to huge advancements. However, the crucial questions driven by the reluctance to accept AI-based decisions may lead to a whole new dynamic where integrating control in such systems may be among the key measures for more accurate models but the challenges associated with the question - "what control means" and how they are different in diverse range of systems is crucial. Implementing control in such a way where the user can contribute from their domain knowledge would be very impactful contribution in the conversation of explainable machine learning systems.

5.2 Behavioral Patterns in Interactive Systems

In a context where advances in algorithms are reaching critical areas such as criminal justice systems, we have seen cases such as racial biases and false reports tending to be more frequent. In fact, there is a growing concern around the acceptance of AI agents and trust issues due to their lack of explainability. But at the same time it is also important to study the behavioral patterns and issues while implementing a interactive system. In an interactive system, users have control over the decision making process. In our case, we have extracted several important patterns.

One of the prominent patterns is that users kept with default parameters. Despite being presented with a diverse array of parameters, participants kept leaving the default parameters. Even though we saw some learning behavior around our non-technical participants, LEAs prefer to choose the defaults. This could be for two reasons. 1) They are trained to choose defaults, 2) seeing the complex unknown parameters of the algorithms, they chose not to tinker with it. The concept of an algorithm and its limitations can be difficult to convey to non-experts and are likely to rely on simplified explanations. Users without experience in statistics are unlikely to comprehend the implications of working with a given model. Furthermore, this decision-making process should be a co-adaptive process in that both the user and model will respond to the behavior of the other [26]. Establishing the right level of understanding among users and framing the task appropriately is critical and non-trivial.

Another behavioral pattern that can be drawn from the analysis is that, even though most users were unaware of the complex mathematics of the algorithms and they had to choose the decisive parameters while sitting in a close environment in an informal time-limit situation, in terms of mental pressure, they act really well. In our analysis, we found out they didn't feel mentally stressed while interacting with the algorithm. Emotions like feeling humiliated or neglected weren't noticed. This brings us to an important discussion how this interactive model can be built with the users. The placement of a user into an immersive environment can also facilitate more mechanisms of discovery and learning. Studies on Interactive Machine Learning [26] have shown the potential of full-bodied interaction as part of an algorithmic decision making process. A further ancillary benefit is that an enjoyable immersive experience may also improve user engagement. The construction of intuitive and informative multi-dimensional data representations can be significantly impactful.

5.3 Design for Contestability

Contestability fosters engagement rather than passivity, questioning rather than acquiescence. As such, contestability is a particularly important system quality where the goal is for predictive algorithms to enhance and support human reasoning, such as decision support systems. Contestability is one way "to enable responsibility in knowing" as the production of knowledge is spread across humans and machines [81]. Contestability can support critical, generative, and responsible engagement between users and algorithms, users and system designers, and ideally between users and those subject to decisions (when they are not the users), as well as the public. Efforts to make algorithmic systems knowable respond to the individual need to understand the tools one uses, as well as the social need to ensure that new tools are fit for purpose. Contestability is a design intervention that can contribute to both [30].

However, our focus here is on its potential contribution to the creation of governance models that support epistemically responsible behavior and support shared reasoning [63]. Contestability, the ability to contest decisions, is at the heart of legal rights that afford individuals access to personal data and insight into the decision-making processes used to classify them, [30] and it is one of the interests that transparency serves. Our model in algorithmic crime mapping invokes several discussions around design implications and contestability issues.

First, from our discussions with the Law Enforcement Agents, it came out several times that it is not reasonable for them to work with information that has been presented to them when there’s no explanation on how that has been generated. LEAs are the human actors on the ground who, while working with information, receive the blame from false or racially biased outcomes. By excluding their voice from such decisions, these algorithms are putting them in a complicated situation. Second, from our analysis there are several major points around how LEAs as human actors can contribute to the system. For example, they may know how a decision has been generated from prior contextual knowledge and familiarity with ancillary facets of the data rather than a purely mathematical explanation. They have the information in detail how the reports are being generated and it has been mentioned several times how they are not the proper representation of whole picture. By adding these explanations from the LEAs, it is possible to generate more accurate information which will help the algorithm to lead proper and reasonable decisions. Lastly, LEAs have been facing a moral dilemma regarding the privacy concerns and ethical basis around these decisions. As there is no explanations on how the decisions have been generated, whether or not they were developed without violating anyone's privacy and in a proper manner is very important to them. These issues around withholding explainability from the street level bureaucrats of such algorithms have been very obstructive and bringing contestability to such algorithms seems like a reasonable discussion to have.

We know, contestability as a design goal, however, is more ambitious and far-reaching. But a system designed for contestability would protect the ability to contest a specific outcome, consistent with privacy and consumer protection law. It would also facilitate generative engagement between humans and algorithms throughout the use of the algorithmic decision making system and support the interests and rights of a broader range of stakeholders, designers, as well as decision subjects in shaping its performance.

5.4 Education and Training in Algorithm Use

The role of parameters in an algorithm and the possible choices are rarely intuitive. Interviews with LEAs and crime analysts in both studies highlighted a concerning trend: practically all professionals who may actually implement algorithms for predictive policing purposes lack an educational background that would provide a deep understanding of the statistical models crime mapping software is dependent on or a good intuition of the roles and effects of parameter choice. Furthermore, during on job training, institutional knowledge is transmitted regardless of actual quality rather than attempting to develop a meaningful understanding of the tools used.

Our second study provided brief explanations of how KDEs work and the roles of their parameters in pop-up tooltips. Despite this, many participants still lacked confidence in their comprehension of the algorithm and its parameters. This shows that brief descriptions are not sufficient to promote understanding, even when examples are available. A surface level description and graphic does not replace rigorous training, proper education, or an immediate intuition.

Future work will have to address this dearth of knowledge and develop a training regimen for promoting the skills and background knowledge necessary to implement fair and transparent policing algorithms for those who may have a grounding in criminological theory but lack any technical skills or statistical education. This will involve walking the tightrope of simplifying enough to reduce the barrier to understanding while maintaining enough details to empower end users to make responsible choices. This breakdown could and should be created and trialed for every algorithm, geospatial or otherwise, that a crime lab would want to implement to help increase the fairness, accuracy, and accountability of predictive policing initiatives.

6 CONCLUSION

Limitations. There are several important notes with regards to the limitations of the studies performed and therefore the conclusions drawn from them. We used different datasets in our k-means tool and KDE tool. We chose to use arrest data for k-means in order to better emulate how crime labs implemented k-means. We chose to switch to crime data as court cases for the KDE experiment so that we would be able to utilize the judicial outcome (dismissed, guilty, innocent, etc.). Though that did not become relevant within the study and analysis contained in this paper, we intend to incorporate it in future work with this tool and thus committed to the newer dataset. Thus, it is important to note that the timeframe of our datasets are different and how crime is recorded has changed within their joint timeframes.

Our tools incorporated data on several different types of crime, but by no means did we use all types of crimes. We chose to focus on what we described as quality of life crimes. These were chosen for being both directly harming a person's quality of life, that is, warranting preemptive policing practices, while also being prolific enough that they formed a suitably large enough dataset to enable geospatial analysis. The specific subsets of data shown to participants in Study Two were chosen in order to create a variety of complexity in heatmap outputs rather than for some inherent value of the data.

Finally, our participants were significantly limited. All participants were from Milwaukee or nearby Chicagoland. Those from Group 3 of Study 2 were not necessarily professional crime analysts, we only required that they have a background pertaining to law enforcement or criminal justice. Both studies had regrettably few people who had a background in policing or criminal justice. Therefore, these results should be blindly applied to crime labs at large, and must be critically re-examined outside the specific context of Milwaukee, Southeast Wisconsin, and the Chicago area.

We have examined two geospatial algorithms commonly chosen to visualize where crime happens. Through two mixed method studies, we found points of entry for bias in the selection of parameter values. Furthermore, we uncovered a lack of education in the meaning of parameter choice and a severe lack of critical reflection when choosing parameters or leaving the default. We identified a number of trends in user behavior, such as often leaving default values when they are provided or by blindly following preexisting mental models or institutional knowledge when defaults do not exist. These factors lead to an underlying trend of unintentional bias in predictive policing initiatives which rely on geospatial algorithms. This reinforces more explicit biases already extant in the American criminal justice system.

We found a lack of education and training leading to choosing arbitrary parameters or even refusing to make a decision at all. This leaves power in the hands of software builders rather than end users. Because of the proprietary black box nature of crime analysis software, we cannot affirm that developers have good intentions, have put in checks and balances to preserve fairness, or even that they have basic domain knowledge inform their design. What options are given to the crime analysts are predefined and still limited in understandability. Thus, even when presented with options, institutional knowledge, and any baggage associated with it, becomes procedure without analysts having the toolset to challenge it.

This may be suitable if the goal of predictive policing is merely to optimize police workload. However, it drastically undermines the entire assumption of an unbiased, perfectly accountable process. Fairness, transparency, and accountability are not the focus of these procedures. Because predictive policing practices informs where police officers will be and where they will find crime, the oversights we found cast doubts upon the validity of police procedure and arrest patterns found in many large American cities. The institution of predictive policing warrants a re-examination and critical adjustments if it is to continue in a fair and equitable society. Crime analysis software developers need to design with transparency and provide thorough yet accessible definitions. End users need to be empowered to proactively make ethical decisions and have the capability to justify their actions to impacted community members, not just their superiors. Even with all these changes, predictive policing would still be compromised by using pre-existing crime data, which carries with it decades of societal biases. Predictive policing cannot be perfected. No matter the fixes suggested, the procedure must be performed critically, without the assumption of fairness. Fairness must be proven rather than taken for granted.

BIBLIOGRAPHY

[1] *Down, Out, and Under Arrest.* https://www.press.uchicago.edu/ucp/books/book/chicago/D/bo23530208.html

[2] GIS in Law Enforcement: Implementation Issues and Case Studies. https://www.crcpress.com/GIS-in-Law-Enforcement-Implementation-Issues-and-Case-Studies/Leipnik-Albert/p/book/9780415286107

[3] The Rise of Big Data Policing. https://nyupress.org/9781479892822/the-rise-of-big-data-policing

[4] 2015. Privacy expert argues “algorithmic transparency” is crucial for online freedoms at UNESCO knowledge café. (Dec. 2015). https://en.unesco.org/news/privacy-expert-argues-algorithmic-transparency-crucial-online-freedoms-unesco-knowledge-cafe

[5] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI ’19).* ACM, New York, NY, USA, Article 3, 13 pages. DOI:http://dx.doi.org/10.1145/3290605.3300233

[6] Uğur Argun and Murat Dağlar. 2016. Crime Mapping and Geographical Information Systems in Crime Analysis. *Journal of Human Sciences* 13 (04 2016), 2208. DOI:http://dx.doi.org/10.14687/ijhs.v13i1.3736

[7] Anonymous Author. Algorithms in the Criminal Justice System. (????). Retrieved August 29, 2019 from https://epic.org/algorithmic-transparency/crim-justice/

[8] Anonymous Author. WIBR Crime Data. Retrieved November 1, 2019 from https://data.milwaukee.gov/dataset/wibr

[9] Unknown Author. CRIME MAPPING (police). Retrieved August 29, 2019 from http://what-when-how.com/police-science/crime-mappingpolice/

[10] Eric P.S. Baumer, Xiaotong Xu, Christine Chu, Shion Guha, and Geri K. Gay. 2017. When Subjects Interpret the Data: Social Media Non-use as a Case for Adapting the Delphi Method to CSCW. ACM Press, 1527–1543. DOI:http://dx.doi.org/10.1145/2998181.2998182

[11] Frank R. Baumgartner, Derek A. Epp, Kelsey Shoub, and Bayard Love. 2017. Targeting young men of color for search and arrest during traffic stops: evidence from North Carolina, 2002–2013. *Politics, Groups, and Identities* 5, 1 (2017), 107–131. DOI:http://dx.doi.org/10.1080/21565503.2016.1160413

[12] Katherine Beckett, Kris Nyrop, and Lori Pfingst. 2006. Race, drugs, and policing: Understanding disparities in drug delivery arrests. *Criminology* 44, 1 (2006), 105–137.

[13] Benjamin Bowling. 2018. Pulled over: how police stops define race and citizenship. *Policing and Society* 28, 5 (2018), 626–629. DOI:http://dx.doi.org/10.1080/10439463.2018.1473978

[14] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[15] Sarah Brayne. 2017. Big data surveillance: The case of policing. *American Sociological Review* 82, 5 (2017), 977–1008.

[16] Tim Brennan, William Dieterich, and Beate Ehret. 2009. Evaluating the Predictive Validity of the Compas Risk and Needs Assessment System. *Criminal Justice and Behavior* 36, 1 (2009), 21–40. DOI:http://dx.doi.org/10.1177/0093854808326545

[17] James W Buehler. 2017. Racial/ethnic disparities in the use of lethal force by US police, 2010–2014. *American journal of public health* 107, 2 (2017), 295–297.

[18] Spencer Chainey and Jerry Ratcliffe. 2013. *GIS and crime mapping*. John Wiley & Sons.

[19] Spencer Chainey, Lisa Tompson, and Sebastian Uhlig. 2008. The Utility of Hotspot Mapping for Predicting Spatial Patterns of Crime. *Security Journal* 21, 1-2 (Feb. 2008), 4–28. DOI:http://dx.doi.org/10.1057/palgrave.sj.8350066

[20] Jeremy Crampton. 2014a. Collect it All: National Security, Big Data and Governance. *SSRN Electronic Journal* (01 2014). DOI:http://dx.doi.org/10.2139/ssrn.2500221

[21] Jeremy Crampton. 2014b. Collect it All: National Security, Big Data and Governance. *SSRN Electronic Journal* (01 2014). DOI:http://dx.doi.org/10.2139/ssrn.2500221

[22] Nancy G. LaVigne Cynthia A. Mamalian and the staff of the Crime Mapping Research Center. *The Use of Computerized Crime Mapping by Law Enforcement: Survey Results.* Technical Report. U.S. Department of Justice.

[23] Clare Davies. 1998. Analysing 'work' in complex system tasks: An exploratory study with GIS. *Behaviour & Information Technology* 17, 4 (1998), 218–230. DOI:http://dx.doi.org/10.1080/014492998119427

[24] Munmun De Choudhury and Michael Massimi. 2015. "She said yes" – Liminality and Engagement Announcements on Twitter. (March 2015). https://www.ideals.illinois.edu/handle/2142/73658

[25] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science Advances* 4, 1 (Jan. 2018). DOI:http://dx.doi.org/10.1126/sciadv.aao5580

[26] John J. Dudley and Per Ola Kristensson. 2018. A Review of User Interface Design for Interactive Machine Learning. *ACM Trans. Interact. Intell. Syst.* 8, 2, Article 8 (June 2018), 37 pages. DOI:http://dx.doi.org/10.1145/3185517

[27] Ronnie A Dunn. 2009. Measuring racial disparities in traffic ticketing within large urban jurisdictions. *Public Performance & Management Review* 32, 4 (2009), 537–561.

[28] J. Eck, S. Chainey, J. Cameron, and R. Wilson. 2005. *Mapping crime: Understanding Hotspots*. Report. National Institute of Justice, Washington DC. 1–71 pages. http://discovery.ucl.ac.uk/11291/

[29] Motahhare Eslami. 2017a. Understanding and Designing around Users' Interaction with Hidden Algorithms in Sociotechnical Systems. ACM Press, 57–60. DOI:http://dx.doi.org/10.1145/3022198.3024947

[30] Motahhare Eslami. 2017b. Understanding and Designing Around Users' Interaction with Hidden Algorithms in Sociotechnical Systems. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17 Companion)*. ACM, New York, NY, USA, 57–60. DOI:http://dx.doi.org/10.1145/3022198.3024947

[31] Jeffrey Fagan, Anthony A Braga, Rod K Brunson, and April Pattavina. 2016. Stops and stares: Street stops, surveillance, and race in the new policing. *Fordham Urb. LJ* 43 (2016), 539.

[32] Andrew Guthrie Ferguson. 2012. Predictive policing and reasonable suspicion. *Emory LJ* 62 (2012), 259.

[33] Andrew J. Flanagin and Miriam J. Metzger. 2008. The credibility of volunteered geographic information. *GeoJournal* 72 (2008), 137–148.

[34] Georg Gartner, David A. Bennett, and Takashi Morita. 2007. Towards Ubiquitous Cartography. *Cartography and Geographic Information Science* 34, 4 (2007), 247–257. DOI:http://dx.doi.org/10.1559/152304007782382963

[35] Barney Glaser and Anselm Strauss. 1998. *Grounded theory*. Strategien qualitativer Forschung. Bern: Huber.

[36] Sharad Goel, Justin M. Rao, and Ravi Shroff. 2016. Precinct or prejudice? Understanding racial disparities in New York City's stop-and-frisk policy. *The Annals of Applied Statistics* 10, 1 (March 2016), 365–394. DOI:http://dx.doi.org/10.1214/15-AOAS897

[37] Amy Griffin. 2004. Understanding how scientists use data-display devices for interactive visual computing with geographical models /. (01 2004).

[38] MD Romael Haque, Katherine Weathington, and Shion Guha. 2019. Exploring the Impact of (Not) Changing Default Settings in Algorithmic Crime Mapping-A Case Study of Milwaukee, Wisconsin. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*. 206–210.

[39] Mark Harrower and Benjamin Sheesley. 2005. Designing Better Map Interfaces: A Framework for Panning and Zooming. *Transactions in GIS* 9, 2 (2005), 77–89. DOI:http://dx.doi.org/10.1111/j.1467-9671.2005.00207.x

[40] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[41] David Hess. 2001. Ethnography and the Development of Science and Technology Studies. *Etnograficeskoe Obozrenie* (01 2001).

[42] Bryan Hill. 2002. Operationalizing GIS to investigate serial robberies in Phoenix, Arizona.

[43] Alexander Hirschfield and Kate Bowers. 2001. *Mapping and Analysing Crime Data: Lessons from Research and Practice.*

[44] William C. Horrace and Shawn M. Rohlin. 2016. How Dark Is Dark? Bright Lights, Big City, Racial Profiling. *The Review of Economics and Statistics* 98, 2 (2016), 226–232. DOI:http://dx.doi.org/10.1162/REST_a_00543

[45] Brian Jordan Jefferson. 2018. Predictable policing: Predictive crime mapping and geographies of policing and race. *Annals of the American Association of Geographers* 108, 1 (2018), 1–16.

[46] S. D. Johnson, K. Bowers, and A. Hirschfield. 1997. NEW INSIGHTS INTO THE SPATIAL AND TEMPORAL DISTRIBUTION OF REPEAT VICTIMIZATION. *The British Journal of Criminology* 37, 2 (1997), 224–241. http://www.jstor.org/stable/23638645

[47] Jeff Larson Julia Angwin. 2016. Machine Bias. (May 2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[48] David M. Kennedy, Anthony A. Braga, Anne M. Piehl, and David M. Kennedy Et Al. 1997. The (un)known universe: Mapping gangs and gang violence in Boston. https://citeseerx.ist.psu.edu/viewdoc/similar?doi=10.1.1.483.6248&type=ab

[49] Tammy Rinehart Kochel, David B Wilson, and Stephen D Mastrofski. 2011. EFFECT OF SUSPECT RACE ON OFFICERS'ARREST DECISIONS. *Criminology* 49, 2 (2011), 473–512.

[50] Ourania Kounadi, Kate Bowers, and Michael Leitner. 2014. Crime Mapping Online: Public Perception of Privacy Issues. *European Journal on Criminal Policy and Research* 21 (06 2014), 167–190. DOI:http://dx.doi.org/10.1007/s10610-014-9248-4

[51] Min Kyung Lee, Daniel Kusbit, Evan Metsky, and Laura Dabbish. 2015. Working with Machines: The Impact of Algorithmic and Data-Driven Management on Human Workers. ACM Press, 1603–1612. DOI:http://dx.doi.org/10.1145/2702123.2702548

[52] Joscha Legewie. 2016. Racial profiling and use of force in police stops: How local events trigger periods of increased discrimination. *Amer. J. Sociology* 122, 2 (2016), 379–424.

[53] Debra Livingston. 1997. Police discretion and the quality of life in public places: Courts, communities, and the new policing. *Columbia Law Review* (1997), 551–672.

[54] Mona Lynch, Marisa Omori, Aaron Roussell, and Matthew Valasik. 2013. Policing the 'progressive'city: The racialized geography of drug law enforcement. *Theoretical criminology* 17, 3 (2013), 335–357.

[55] Jia Ma. 2013. *How the use of computerized crime mapping was affected by characteristics of law enforcement agencies in 2007.* Ph.D. Dissertation. https://lib.dr.iastate.edu/etd/13616.

[56] ALAN M. Maceachren. 1994. Chapter 1 - Visualization in Modern Cartography: Setting the Agenda. In *Modern Cartography Series*, ALAN M. Maceachren and D. R. FRASER Taylor (Eds.). Visualization in Modern Cartography, Vol. 2. Academic Press, 1–12. DOI:http://dx.doi.org/10.1016/B978-0-08-042415-6.50008-9

[57] I. Scott MacKenzie. 1992. Fitts' Law As a Research and Design Tool in Human-computer Interaction. *Hum.-Comput. Interact.* 7, 1 (March 1992), 91–139. DOI:http://dx.doi.org/10.1207/s15327051hci0701_3

[58] LaVigne Nancy G. Mamalian, Cynthia A. and Elizabeth Groff. 2008. *Use of Computerized Crime Mapping by Law Enforcement in the United States, 1997-1998.* Technical Report.

[59] Timothy Mashford. 2008. *Methods for Implementing Crime Mapping within a Large Law Enforcement Agency: Experiences from Victoria, Australia.* 19–26. DOI:http://dx.doi.org/10.1002/9780470987193.ch3

[60] CAROL Mcguinness. 1994. Chapter 10 - Expert/Novice Use of Visualization Tools. In *Modern Cartography Series*, ALAN M. Maceachren and D. R. FRASER Taylor (Eds.). Visualization in Modern Cartography, Vol. 2. Academic Press, 185–199. DOI:http://dx.doi.org/10.1016/B978-0-08-042415-6.50017-X

[61] George O Mohler, Martin B Short, Sean Malinowski, Mark Johnson, George E Tita, Andrea L Bertozzi, and P Jeffrey Brantingham. 2015. Randomized controlled field trials of predictive policing. *Journal of the American statistical association* 110, 512 (2015), 1399–1411.

[62] Nazneen Fatema Rajani and Raymond J. Mooney. 2017. Using Explanations to Improve Ensembling of Visual Question Answering Systems. (2017). http://www.cs.utexas.edu/users/ai-lab/pub-view.php?PubID=127647

[63] Deirdre K. Mulligan, Daniel Kluttz, and Nitin Kohli. 2019. *Shaping Our Tools: Contestability as a Means to Promote Responsible Algorithmic Decision Making in the Professions.* SSRN Scholarly Paper ID 3311894. Social Science Research Network, Rochester, NY. https://papers.ssrn.com/abstract=3311894

[64] Olfa Nasraoui and Patrick Shafto. 2016. Human-Algorithm Interaction Biases in the Big Data Cycle: A Markov Chain Iterated Learning Framework. *CoRR* abs/1608.07895 (2016). http://arxiv.org/abs/1608.07895

[65] Justin Nix, Bradley A Campbell, Edward H Byers, and Geoffrey P Alpert. 2017. A bird's eye view of civilians killed by police in 2015: Further evidence of implicit bias. *Criminology & Public Policy* 16, 1 (2017), 309–340.

[66] A M Olligschlaeger. Artificial Neural Networks and Crime Mapping. *Criminal Justice Press/Willow Tree Press, pages* 313 - 347, 20531 (1997).

[67] Kristien Ooms, Arzu Coltekin, Philippe De Maeyer, Lien Dupont, Sara Fabrikant, Annelies Incoul, Matthias Kuhn, Hendrik Slabbinck, Pieter Vansteenkiste, and Lise Van der Haegen. 2014. Combining user logging with eye tracking for interactive and dynamic applications. *Behavior Research Methods* 47 (12 2014). DOI:http://dx.doi.org/10.3758/s13428-014-0542-3

[68] otter. Otter for meeting notes. (2019). https://otter.ai

[69] Ana Paula Mendes de Miranda and Marcus Ferreira. 2008. *An Analytical Technique for Addressing Geographical Referencing Difficulties and Monitoring Crimes in Rio de Janeiro, Brazil.* 9–18. DOI:http://dx.doi.org/10.1002/9780470987193.ch2

[70] E Ashby Plant and B Michelle Peruche. 2005. The consequences of race for police officers' responses to criminal suspects. *Psychological Science* 16, 3 (2005), 180–183.

[71] Atyanti Prabaswari, Chancard Basumerda, and Bagus Utomo. 2019. The Mental Workload Analysis of Staff in Study Program of Private Educational Organization. *IOP Conference Series: Materials Science and Engineering* 528 (06 2019), 012018. DOI:http://dx.doi.org/10.1088/1757-899X/528/1/012018

[72] Roman Rädle, Hans-Christian Jetter, Simon Butscher, and Harald Reiterer. 2013. The Effect of Egocentric Body Movements on Users' Navigation Performance and Spatial Memory in Zoomable User Interfaces. In *Proceedings of the 2013 ACM International Conference on Interactive Tabletops and Surfaces (ITS '13).* ACM, New York, NY, USA, 23–32. DOI:http://dx.doi.org/10.1145/2512349.2512811

[73] Jerry H. Ratcliffe. 2004. Crime Mapping and the Training Needs of Law Enforcement. *European Journal on Criminal Policy and Research* 10, 1 (March 2004), 65–83. DOI:http://dx.doi.org/10.1023/B:CRIM.0000037550.40559.1c

[74] J. H. Ratcliffe and M. J. McCullagh. 2001. Chasing Ghosts?: Police Perception of High Crime Areas. *The British Journal of Criminology* 41, 2 (03 2001), 330–341. DOI:http://dx.doi.org/10.1093/bjc/41.2.330

[75] Paul Resnick and Hal R. Varian. 1997. Recommender Systems. *Commun. ACM* 40, 3 (March 1997), 56–58. DOI:http://dx.doi.org/10.1145/245108.245121

[76] Judy Robertson and Maurits Kaptein. 2016. An introduction to modern statistical methods in HCI. In *Modern Statistical Methods for HCI.* Springer, 1–14.

[77] Anthony C. Robinson, Urška Demšar, Antoni B. Moore, Aileen Buckley, Bin Jiang, Kenneth Field, Menno-Jan Kraak, Silvana P. Camboim, and Claudia R. Sluter. 2017. Geospatial big data and cartography: research challenges and opportunities for making maps that matter. *International Journal of Cartography* 3, sup1 (2017), 32–60. DOI:http://dx.doi.org/10.1080/23729333.2016.1278151

[78] Robert E. Roth, Arzu Çöltekin, Luciene Delazari, Homero Fonseca Filho, Amy Griffin, Andreas Hall, Jari Korpi, Ismini Lokka, André Mendonça, Kristien Ooms, and Corné P.J.M. van Elzakker. 2017. User studies in cartography: opportunities for empirical research on interactive maps and visualizations. *International Journal of Cartography* 3, sup1 (2017), 61–89. DOI:http://dx.doi.org/10.1080/23729333.2017.1288534

[79] Johannes Schöning, Brent Hecht, Martin Raubal, Antonio Krüger, Meredith Marsh, and Michael Rohs. 2008. Improving Interaction with Virtual Globes Through Spatial Thinking: Helping Users Ask "Why?". In *Proceedings of the 13th International Conference on Intelligent User Interfaces (IUI '08).* ACM, New York, NY, USA, 129–138. DOI:http://dx.doi.org/10.1145/1378773.1378790

[80] Ben Shneiderman, Catherine Plaisant, Maxine Cohen, Steven Jacobs, Niklas Elmqvist, and Nicholas Diakopoulos. 2016. *Designing the User Interface: Strategies for Effective Human-Computer Interaction* (6th ed.). Pearson.

[81] Judith Simon. 2015. *Distributed Epistemic Responsibility in a Hyperconnected Era.* Springer International Publishing, Cham, 145–159. DOI:http://dx.doi.org/10.1007/978-3-319-04093-6_17

[82] Stuart S. Shapiro Jonathan M. Simson Garfinkel Smith, Jeanna Matthews. Toward Algorithmic Transparency and Accountability. https://cacm.acm.org/magazines/2017/9/220423-toward-algorithmic-transparency-and-accountability/fulltext

[83] Simone Stumpf, Vidya Rajaram, Lida Li, Weng-Keen Wong, Margaret Burnett, Thomas Dietterich, Erin Sullivan, and Jonathan Herlocker. 2009. Interacting meaningfully with machine learning systems: Three experiments. *International Journal of Human-Computer Studies* 67, 8 (Aug. 2009), 639–662. DOI:http://dx.doi.org/10.1016/j.ijhcs.2009.03.004

[84] M. A. Tayebi and U. Glasser. 2012. Investigating Organized Crime Groups: A Social Network Analysis Perspective. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 565–572. DOI:http://dx.doi.org/10.1109/ASONAM.2012.96

[85] Jameson L. Toole, Nathan Eagle, and Joshua B. Plotkin. 2011. Spatiotemporal Correlations in Criminal Offense Records. *ACM Trans. Intell. Syst. Technol.* 2, 4, Article 38 (July 2011), 18 pages. DOI:http://dx.doi.org/10.1145/1989734.1989742

[86] Graham Upton and Ian Cook. 2014. *A dictionary of statistics 3e.* Oxford university press.

[87] Dawei Wang, Wei Ding, Henry Lo, Tomasz Stepinski, Josue Salazar, and Melissa Morabito. 2013. Crime Hotspot Mapping Using the Crime Related Factors–a Spatial Data Mining Approach. *Applied Intelligence* 39, 4 (Dec. 2013), 772–781. DOI:http://dx.doi.org/10.1007/s10489-012-0400-x

[88] PATRICIA WARREN, DONALD TOMASKOVIC-DEVEY, WILLIAM SMITH, MATTHEW ZINGRAFF, and MARCINDA MASON. 2006. DRIVING WHILE

BLACK: BIAS PROCESSES AND RACIAL DISPARITY IN POLICE STOPS*. *Criminology* 44, 3 (2006), 709–738. DOI:http://dx.doi.org/10.1111/j.1745-9125.2006.00061.x

[89] Patricia Warren, Donald Tomaskovic-Devey, William Smith, Matthew Zingraff, and Marcinda Mason. 2006. Driving while black: Bias processes and racial disparity in police stops. *Criminology* 44, 3 (2006), 709–738.

[90] David Weisburd and Cynthia Lum. 2005. The Diffusion of Computerized Crime Mapping in Policing: Linking Research and Practice. *Police Practice and Research* 6, 5 (2005), 419–434. DOI:http://dx.doi.org/10.1080/15614260500433004

[91] Matthew Wilson. 2012. Location-Based Services, Conspicuous Mobility, and the Location-Aware Future. *Geoforum* 43 (03 2012). DOI:http://dx.doi.org/10.1016/j.geoforum.2012.03.014

[92] Susan Wyche and Eric PS Baumer. 2017. Imagined Facebook: An exploratory study of non-users' perceptions of social media in Rural Zambia. *New Media & Society* 19, 7 (July 2017), 1092–1108. DOI:http://dx.doi.org/10.1177/1461444815625948

[93] Yifei Xue and D. E. Brown. 2003. A decision model for spatial site selection by criminals: a foundation for law enforcement decision support. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 33, 1 (Feb 2003), 78–85. DOI:http://dx.doi.org/10.1109/TSMCC.2003.809867

[94] Barney Glaser. 2017. *Discovery of Grounded Theory: Strategies for Qualitative Research*. Routledge. Google-Books-ID:GTMrDwAAQBAJ

[95] A. Hartigan and M. A. Wong. 1979. Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28, 1 (1979), 100–108. https://doi.org/10.2307/2346830

[96] Samir Passi, Steven Jackson, Phoebe Sengers, Almila Akdag Salah, Sally Wyatt, and Andrea Scharnhorst. 2017. Data Vision: Learning to See Through Algorithmic Abstraction.. In *CSCW*. 2436–2447.