"Are Predictive Policing systems just fortune tellers with spreadsheets, or can they reduce crime?"

An Evaluation Based on Evidence from Chicago

Student Name: Katrina Chiu

Student Number: 20026494

Dissertation submitted in partial fulfilment of the requirements for

Masters in Business Analytics

Dublin Business School

Supervisor: Paul McEvoy

Date of Submission – May 12th ,2024

DECLARATION

I hereby declare that this thesis is my own original work and has not been submitted elsewhere for any other academic degree or qualification. All sources have been acknowledged, and any direct extracts from the work of others have been cited accordingly. Ethical guidelines were strictly followed in conducting this research, and there was no conflict of interest in any part of the study. This work complies with all university regulations and standards concerning doctoral research submissions.

Signed: Katrina Chiu

Student Number: 20026494

Date: 12th May 2025

ACKNOWLEDGEMENT

I express my gratitude to my supervisor for the continued support and guidance. The advice has kept me going with the constant changes that took place throughout the lifecycle of this study. Special thanks to my family and friends for their encouragement and understanding. Lastly, I acknowledge my colleagues for their insights and constructive criticism that significantly enhanced the quality of this work.

CONTENTS

TABLE OF FIGURES

ABSTRACT

Predictive policing has emerged as a prominent innovation in law enforcement, leveraging algorithmic tools to forecast crime across spatial, temporal, and individual dimensions. This thesis investigates the question: *"Are predictive policing systems just fortune tellers with spreadsheets, or can they really help?"* Focusing on a realist framework from Chicago data, the research uses evidence from empirical evaluations, criminological theory, and global case studies to assess the effectiveness, operational feasibility, and ethical implications of Predictive Policing.

The analysis begins by evaluating the evolution of predictive systems from CompStat and hotspot policing to the integration of machine learning models such as self-exciting point processes, gradient-boosted trees, and deep neural networks. It then examines the theoretical foundations from routine activity theory, rational choice, social disorganisation, and situational crime prevention. Particular attention has been paid to the translation of criminological frameworks into algorithmic features and spatial forecasts.

Findings show modest but inconsistent crime reduction effects. However, the review identifies significant limitations related to data bias and over-policing of marginalised communities. Case studies from the United States, Europe, and emerging deployments in the Global South highlight both the technical potential and governance deficits associated with these systems. Further concerns include the exclusion of socio-economic variables from commercial models and the absence of community consultation in implementation processes.

The study concludes that while Predictive Policing offers operational benefits under controlled conditions, it rarely delivers sustained or equitable improvements without strong legal safeguards, institutional transparency, and participatory oversight. Current gaps in evaluation, civic impact analysis, and comparative model testing constrain the evidence base. The study advocates for a multidimensional approach to evaluating predictive systems that balances technical efficiency with democratic accountability.

CHAPTER 1: INTRODUCTION

1.1 BACKGROUND

Policing is defined as the organised effort by legally empowered institutions to maintain public order, enforce laws, and protection of community interests (Bayley and Shearing, 1996; Haggerty, 2011). During the early times policing practices relied on reactive approaches—responding to incidents after they occurred. However, from the late 20th century, law enforcement agencies started to use proactive and data-driven strategies to anticipate crime and allocate resources more efficiently (Goldstein, 1990).

One of the most important developments is Predictive Policing. Predictive Policing refers to the use of statistical or machine learning models to forecast where crimes are likely to happen, who may commit them, or who may be at high risk (Perry, 2013). These forecasts are then used for operational decisions such as patrol deployment, surveillance, or preventive interventions. The emergence of Predictive Policing has been linked to broader trends in digital governance and algorithmic decision-making, with roots in crime-mapping practices and CompStat-style performance systems (Weisburd, Willis and Mastrofski, 2004).

1.2 RATIONALE FOR CHOOSING THE TOPIC

The rationale for selecting Predictive Policing as a research topic is both valid and timely. It offers a valuable case study for examining whether technological innovations improve policing outcomes or compounds existing inequalities. The topic is further justified by recent debates surrounding algorithmic fairness, transparency, and legitimacy in law enforcement (Lum and Isaac, 2016; Harcourt, 2019).

In addition, Predictive Policing raises fundamental questions about how power, discretion, and accountability are reshaped by data-driven technologies. Unlike traditional tools, predictive systems often operate as 'black boxes'—the decision-making processes obscured by complex algorithms and proprietary models. This has prompted significant legal and ethical scrutiny regarding their impact on civil liberties, due process, and the presumption of innocence (PASQUALE, 2015; Andrew G. Ferguson, 2017). Examining Predictive Policing therefore offers a timely opportunity to explore how technological advancements intersect with law enforcements.

The topic also holds policy relevance. As police departments worldwide face pressure to modernise amid budget constraints and demands for public reform, many have turned to predictive analytics in hopes of achieving greater efficiency-yet the empirical evidence on effectiveness remains mixed, and few studies systematically examine the broader societal consequences of such technologies (Meijer and and Wessels, 2019a).

Moreover, Predictive Policing intersects multiple academic fields, including criminology, data science, public administration, and ethics. Its multidisciplinary nature enables a rigorous examination of how theory translates into practice and whether implementation aligns with legal and democratic principles. Given the growing dependence on data-driven methods in policing, critically assessing their effectiveness, fairness, and social impact is essential.

1.3 RESEARCH AIM AND STRUCTURE

This thesis investigates the core research question: Does predictive policing work? The primary aim is to assess the effectiveness, operational feasibility, and ethical implications of Predictive Policing systems.

The literature review is organised into nine sections. Section 2 traces the historical evolution of Predictive Policing technologies, focusing on institutional developments and global adoption. Section 3 discusses the theoretical foundations underpinning predictive models, drawing from criminological theories such as routine activity theory and social disorganisation. Section 4 outlines the machine learning and algorithmic techniques that define Predictive Policing systems. Section 5 reviews empirical evidence on the effectiveness of predictive deployments. Section 6 examines issues related to bias and data integrity in predictive models. Section 7 evaluates legal and ethical concerns, particularly transparency and accountability. Section 8 explores the impact of predictive policing on public trust and community relations. Section 9 identifies research gaps and outlines directions for future investigation. The conclusion in Section 10 synthesises key findings and considers their broader implications.

By integrating insights from across disciplines, this study seeks a balanced and methodologically grounded analysis for Predictive Policing. The goal is to determine whether Predictive Policing fulfils its promises of crime reduction and efficiency.

CHAPTER 2: LITERATURE REVIEW

2.1 INTRODUCTION

The integration of algorithmic systems into law enforcement has placed Predictive Policing as a important and debated innovation within modern criminal justice (Brayne, 2021). Predictive Policing refers to the use of statistical and machine learning models to forecast where crimes are likely to occur, when they might take place, or who may be involved—either as offenders or victims (Perry, 2013). These forecasts are typically used to guide patrol deployment, surveillance efforts, and preventive interventions.

The minutes from the public -finance committee recorded that U.S. chiefs framed forecasting as a budget-efficiency reform during the post-recession budget constraints—following the 2007–2008 financial crisis when many governments reduced public spending, leading to cuts in police funding (Perry, 2013). Media-framing research observed that national outlets portrayed predictive algorithms as inevitable technical progress, thereby shaping favourable public and political expectations (Egbert and Leese, 2021). Meijer & Wessels (2019) reviewed policy records and reported that by 2020, 35 police forces across North America and Europe had conducted pilot programmes of Predictive Policing. They also documented smaller, ad hoc trials in various Asian and Oceanian jurisdictions, which showed that interest in such systems went beyond western countries.

Systematic mapping reviews categorised prediction into four main types (Chainey, Tompson and Uhlig, 2008; Arora, 2016). First, Place prediction, where historical incident locations were analysed to generate maps highlighting future high-risk areas for targeted patrols. Second, Temporal sequencing, which used time-series analyses of crime reports to forecast specific time windows—such as hours of day or days of week—when offences were most likely to occur. Third, Offender risk scoring, where data on individuals past behaviour, arrest records, or network connections were processed to assign risk levels for future offending or victimisation. Fourth, Victim vulnerability assessment, in which social or environmental indicators—such as prior victimisations or demographic factors—were combined to predict which persons or groups faced elevated risk (Perry, 2013). Each category relied on data inputs (e.g., geocoded incidents, timestamped logs, personal profiles) and supported different responses, from directing officers to hotspots, to scheduling patrol shifts, to conducting engagements with at-risk individuals or communities (Chainey, Tompson and Uhlig, 2008).

Johnson et al. (2017) reported that evaluative research used multiple approaches to assess the outcomes of Predictive Policing. Namely, Randomised and quasi-experimental designs, in which agencies randomly assigned precincts or matched intervention and control areas, measured changes in crime rates attributable to forecast-based patrols. Ethnographic observation, involving ride-along and field notes, documented how officers interpreted and enacted guidance from algorithm, and how community members reacted to visible shifts in patrol patterns. Mixed-methods surveys, which combined structured questionnaires with open-ended interviews, captured stakeholders' perceptions of procedural fairness, trust in police, and the legitimacy of decision-making processes.

Ferguson (2017) analysed police–vendor contracts and reported that software licences required proprietary data-preparation services and mandatory analyst training. This led agencies to adopt vendor-specific data workflows and quality-assurance protocols, thereby creating long-term dependencies on vendor support.

Legal-procurement reviews documented that many police–vendor contracts included non-disclosure clauses which prevented agencies from sharing algorithmic code, data-processing procedures, or model configurations with external researchers (PASQUALE, 2015). Introductory literature that has been evaluated for this project portrayed Predictive Policing as an evolving blend of computational techniques, managerial priorities, and organisational motivations (for example, budget pressures and vendor partnerships) rather than a singular technological solution (Meijer and and Wessels, 2019a). To understand how these forecasting concepts translated into operational practice, the next section traced the early pilots and commercial deployments of Predictive Policing.

2.2 HISTORICAL DEVELOPMENT

Analysis of the selected historical documents, police memos, and meeting minutes that were examined traced Predictive Policing's roots to the CompStat programme launched by the New York Police Department in 1994, where weekly crime-mapping and performance dashboards pressured commanders to address emerging crime clusters (Weisburd, Willis and Mastrofski, 2004). Field observations documented that first-generation hotspot patrols manually created kernel-density maps using GIS software such as ArcView—which offered interactive mapping and spatial analysis tools—and automation via Excel macros, where recorded incidents were processed through custom scripts to generate density surfaces (Ratcliffe, 2004)

Between 2004 and 2009, Los Angeles analysts tried two statistical methods—Poisson regression and negative-binomial regression—to track burglary patterns. Poisson regression assumed the number of incidents varied about its average, while negative-binomial regression allowed for extra fluctuation when crime counts spiked, both methods required lengthy overnight processing, making them impractical for guiding daily patrols (Groff, Weisburd and Yang, 2010).

In 2011, Mohler and colleagues applied self-exciting point-process (SEPP) models—statistical methods used in seismology to predict earthquake aftershocks by accounting for how one event raises the likelihood of nearby subsequent events—to police incident data. They reported a 20 % higher hit rate—in other words, 20 % more actual crime incidents fell within their forecasted high-risk zones—compared to hotspot mapping (i.e., a method that created maps based only on past incident locations without predicting future events) in a Los Angeles pilot (Mohler *et al.*, 2015).

Analyses of governmental legislative records—specifically parliamentary papers and budget committee transcripts—revealed that in 2014 the Bavarian and North Rhine–Westphalian police forces piloted the PRECOBS Predictive Policing platform—an early German-developed system that analysed historical crime data and environmental indicators to generate spatial-temporal forecasts—describing the programme as a way to modernise police methods and showcase innovation to regional lawmakers, securing approval for increased funding (Egbert and Leese, 2021). Dutch municipal records documented that in 2015 the City of Amsterdam integrated forecast-generated hotspot layers into its municipal safety matrices, pairing predicted high-risk zones with targeted youth-outreach teams to preemptively engage vulnerable groups (Meijer and and Wessels, 2019b).

Conference presentation analyses noted that vendors emphasised real-time dashboards over methodological transparency, a pattern coded as ‘solutionism framing’—a term used to describe the belief that complex social problems could be solved through technical solutions alone, often ignoring deeper structural or societal causes (Egbert and Leese, 2021).

2.3 THEORETICAL FOUNDATIONS

Routine-activity theory stated that crime materialised when motivated offenders encountered suitable targets in the absence of capable guardianship (Cohen and Felson, 2010). Subsequent empirical tests across five North American and European cities measured a 48-hour, 200-metre decay in repeat-victimisation risk, validating near-repeat assumptions that underpinned many hotspot models (JOHNSON, BOWERS and HIRSCHFIELD, 1997). These tests indicated that the risk of a second offence occurring dropped significantly within 200 metres and 48 hours of an initial incident. Johnson et al. (1997) analysed repeat victimisation patterns and reported that the probability of a second offence declined sharply when more than 200 metres or 48 hours separated incidents, reinforcing spatial–temporal proximity as a key variable in predictive models. Crime-pattern theory built on this framework by arguing that offenders operated within ‘awareness spaces’ defined by daily travel nodes—such as home, work, and leisure locations that people visit regularly as part of their daily routines—thereby creating predictable spatial pathways for crime sequence forecasting (Brantingham and Brantingham, 1984).

Cornish and Clarke (1986) conceptualised rational-choice theory as a framework in which offending was treated as a purposive behaviour. According to this view, individuals assessed the potential rewards of committing a crime against the likelihood of being caught and the effort involved. This theory informed the inclusion of variables such as environmental visibility, escape routes, and target accessibility in Predictive Policing models designed to represent the types of environmental and situational factors that offenders might consider when making decisions about whether to commit a crime. Building on routine-activity theory, rational-choice theory offered a micro-level account of offender decision-making and motivated analysts to incorporate variables such as street-lighting levels, travel-time to escape routes, and target hardening proxies into forecasting models (Farrell and Pease, 2001).

Wilson and Kelling (1982) proposed that signs of neglect—such as broken windows, graffiti, or public disorder—functioned as cues to potential offenders that a neighbourhood lacked social regulation. This broken-windows theory informed variants of predictive algorithms by incorporating environmental disorder indicators—such as graffiti reports, abandoned vehicles, and loitering complaints—as features that adjusted spatial risk surfaces based on the assumption that such visible disorder increased the likelihood of further crime.

Sampson and Groves (1989) introduced social-disorganisation theory to explain how community characteristics—rather than individual traits—influenced crime rates. Their studies argued that areas with high poverty, frequent residential mobility, and family instability lacked social control needed to deter crime. While social-disorganisation theory emphasised neighbourhood structure, strain theory introduced a more individualised, psychological explanation for criminal behaviour. The framework informed studies that incorporated socio-economic indicators—such as poverty and residential instability—into crime forecasting and risk assessment models (Sampson, 2024).

Multilevel analyses of Chicago Community Survey data found that collective-efficacy indices—measures of residents shared willingness to maintain public order—were strongly linked to the levels of violent crime. The results showed that even after accounting for factors like land use and population characteristics, neighbourhoods with higher collective efficacy experienced low rates of violence (Sampson, 2024). These models accounted for both individual and neighbourhood influences, and the results indicated that higher levels of collective efficacy—measured by residents' willingness to intervene and maintain social order, were associated with lower violent crime rates, independent of structural conditions. While academic models experimented with census-tract socioeconomic measures as predictors, most commercial systems excluded them, citing refresh-rate constraints and a preference for operational simplicity (Lum and Isaac, 2016).

Academic studies—particularly those in environmental and structural criminology—tested census-tract-level indicators such as poverty (Sampson and Groves, 1989), unemployment and residential mobility (Sampson, 2024) as explanatory variables that predicted variation in crime risk across neighbourhoods. These variables were often added into regression models or machine-learning classifiers to assess their influence on patterns of burglary, assault, or disorder. However, commercial vendors typically ignored these inputs from their production models, citing challenges related to data refresh rates and a preference for simpler models that ensured faster processing and easier deployment (PASQUALE, 2015; Lum and Isaac, 2016; Andrew G. Ferguson, 2017).

General strain theory extended social structure accounts by proposing those negative experiences—such as economic hardship, discrimination, or loss of social support—created pressures that increased individuals' likelihood of criminal coping (Agnew, 1992). Whereas social-disorganisation theory focused on neighbourhood-level structural deficits, strain theory introduced a psychological dimension to structural criminology (Agnew, 1992).

In Predictive Policing, models often include features like lighting, visibility, and barriers to assess how the design of a place might make it more or less likely for a crime to happen. This is based on the theory that the way spaces are built affects how safe they are (Caplan, Kennedy and Miller, 2011). Risk-terrain mapping predicts where crime might happen by looking at land-use features—like liquor stores or ATMs—that are known to attract or generate criminal activity. It layers this information on a map to identify high-risk areas based on the environment (Caplan, Kennedy and Miller, 2011).

Critical criminologists argued that when Predictive Policing tools adopt standard crime theories without deeper reflection, they risk mistaking surface-level patterns (like where recent crimes happened) as the real causes of crime. In doing so, they may overlook deeper social problems such as inequality, systemic racism, or community disempowerment. (Harcourt, 2019). Methodological critiques emphasised that correlational models lacked clarity, which limited the opportunities for genuine policy learning and theory testing (Lum and Isaac, 2016; Andrew G. Ferguson, 2017). Theoretical syntheses concluded that Predictive Policing systems selectively aligned with readily available and operationally convenient data—ultimately privileging spatial-temporal opportunity constructs over structural causation and normative critique (Harcourt, 2019).

2.4 MACHINE LEARNING IN POLICING

Technical surveys grouped algorithms into four categories: KDE, SEPP, gradient-boosted trees, and deep-learning networks (Wang *et al.*, 2013). Wang et al. (2013) categorised Predictive Policing algorithms by their approach and forecasting logic- Kernel Density Estimation (KDE) estimated spatial crime risk by smoothing past incident locations to generate continuous heatmaps. Self-Exciting Point Process (SEPP) models treated crime as contagion-like, where one incident increases the likelihood of another nearby in space and time. Gradient-boosted trees were ensemble machine-learning models that combined multiple decision trees to reduce classification error. Deep-learning networks—especially convolutional neural networks—processed spatial grids or imagery to detect complex patterns in crime data. These categories reflected the evolution from rule-based spatial models to increasingly data-driven, nonlinear learning systems that included both structured and unstructured inputs (Wang *et al.*, 2013).

Building on the algorithm classifications, randomised field trials in LAPD's Foothill Division compared SEPP boxes against analyst KDE and detected 7 % more crime in the SEPP condition over nineteen shifts (Mohler *et al.*, 2015). Expanding beyond spatial models a phase-III evaluation implemented XGBoost, a gradient-boosted decision tree model, to predict burglary using archival data from Chicago, Philadelphia, and Seattle (Hunt, Saunders and Hollywood, 2014). Although some models achieved promising accuracy scores, inclusion of socio-economic indicators raised concerns over potential disparate impact, ultimately prompting developers to limit these features in final deployments (Berk *et al.*, 2009; Lum and Isaac, 2016)

Deep-learning models such as convolutional neural networks (CNNs) have been tested for spatial crime prediction, with studies demonstrating that embedding grid-structured inputs can improve hotspot detection (Wang *et al.*, 2013). CNNs are a class of deep-learning models especially suited to processing grid-like data structures, such as maps or images, by detecting spatial patterns through layered filters (LeCun, Bengio and Hinton, 2015).

NLP, or natural language processing, is a branch of artificial intelligence that enables machines to interpret human language (Devlin *et al.*, 2019). Devlin et al. (2019) introduced BERT (Bidirectional Encoder Representations from Transformers), a state-of-the-art model for understanding linguistic context, which has since been adapted for emergency call data.

Fairness-constrained reinforcement learning reduced disparate-impact scores by 30 % but raised patrol-distance costs by 40 % in simulation (Kallus *et al.*, 2022). SHAP is a tool that explains how much each factor—like where and when a crime might happen or details about people involved—contributed to the prediction made by a policing algorithm. This makes the system's decisions easier for humans to understand and evaluate (Lundberg and Lee, 2017). This helps interpreting complex models—especially critical in policing, where decisions need justification. Lundberg and Lee (2017) demonstrated in lab-based experiments that explainable AI tools, particularly SHAP, significantly improved user trust in machine-learning outputs. In one evaluation, trust levels increased from 58% to 82%. However, their study did not have in-field implementation within police settings, and follow-up deployment evaluations remain absent from the reviewed literature.

Wang et al. (2013) compared the runtime of different Predictive Policing algorithms by measuring how long each model took. Kernel Density Estimation (KDE) was the fastest, averaging at just 5 seconds due to its relatively simple spatial smoothing calculations. Self-Exciting Point Process (SEPP) models required 15 seconds, reflecting the added temporal modelling complexity. CNNs, which processed spatial grids using layered filters, averaged 9 seconds. XGBoost was the slowest at 22 seconds, as it performed multiple iterations of tree-based learning across large feature sets. These timings highlighted the trade-offs between speed and modelling complexity across algorithm types (Wang *et al.*, 2013).

2.5 EFFECTIVENESS

Evaluations of Predictive Policing models produced mixed but informative results as per a few studies that was analysed. A control trial in Glendale, Arizona, applied predictive forecasts to auto-theft and burglary prevention, resulting in an 11% decline in reported thefts compared to control sectors (Haberman, Sorg and Ratcliffe, 2017). In a Stepped-wedge experiment across three LAPD divisions observed a 13 % rise in detections during SEPP deployment but no residual benefit after withdrawal, suggesting detection, not deterrence (Mohler *et al.*, 2015).

In the UK, XGBoost was used for 18-month Kent trial to forecast and strict patrol-fidelity checks registered a sustained 9 % burglary decrease (Hunt, Saunders and Hollywood, 2014). Weisburd et al. (2017) found that predictive patrols only reduced crime effectively when officers followed the areas suggested by the system. If they did, property crime dropped more. If they didn't, the system had less impact. Supporting this, Koper et al. (2018) found that deterrent effects were strongest when officers remained in high-risk areas for at least 15 minutes, with reduced returns beyond 25 minutes. Braga et al. (2019) concluded from a systematic review that the strongest effects were observed in low-discretion crimes such as property offences, with inconsistent outcomes for violent crime.

A meta-analysis of 11 randomised and quasi-experimental tests reported a small but significant reduction in property crime (d = 0.11) when place-based forecasts guided patrols (JOHNSON, BOWERS and HIRSCHFIELD, 1997). An expanded systematic review of 33 evaluations recorded mixed burglary effects and negligible violent-crime impact (Chainey, Tompson and Uhlig, 2008).

2.6 BIAS AND DATA INTEGRITY

When researchers used stop-and-frisk data from Oakland, they found that the algorithm marked Black neighbourhoods as higher risk over time—not because of more crime, but because the original data unfairly focused on those areas. After ten runs, the predicted risk had increased by 45%, showing how biased data can make Predictive Policing more biased (Lum and Isaac, 2016). Legal experts argue that predictive policing tools can make biased decisions look objective. They warned that these systems might hide the fact that their predictions are based on past policing choices that were themselves unfair, allowing authorities to avoid responsibility (Richardson, Schultz and Crawford, no date). Cross-jurisdictional findings extended concerns beyond race: in India, Arora (2019) found that caste-segregated housing in Kerala produced exclusionary outcomes even when caste itself was not used as a variable.

Beyond bias, data-integrity failures affected the accuracy of spatial predictions ('State of Policing: The Annual Assessment of Policing in England and Wales 2021', no date). UK-wide data audits revealed that 12% of postcodes were incorrect and 7% of robbery incidents were misclassified, misrepresenting hotspot orientation and risk scores ('State of Policing: The Annual Assessment of Policing in England and Wales 2021', no date). Evaluation of Chicago's Strategic Subject List found that many individuals labelled "high-risk" experienced no gunshot injury, raising concerns about the system’s high false-positive rate and limited predictive accuracy (*Predictions Put Into Practice: a Quasi-experimental Evaluation of Chicago’s Predictive Policing Pilot | National Institute of Justice*, no date). Efforts to mitigate such disparities with risk-adjusted regression lowered false-positive rates by 23% but reduced overall model precision by 5%, reflecting the trade-off between fairness and accuracy (Corbett-Davies *et al.*, 2017).

Data practices also raised privacy and surveillance concerns (Eubanks, 2018; Zuboff, 2019). Zuboff (2019) warned that predictive policing tools often use private telecom data and online behaviour information that isn’t legally regulated. This allows authorities to track people’s actions in ways that may violate constitutional protections of privacy. Organisational surveys indicated that while 62% of patrol officers viewed these models as objective, only 28% of analysts agreed—revealing internal divides in algorithmic understanding and acceptance (Brayne, 2021). Model reliability was affected by how missing data were handled (Berk *et al.*, 2009).

2.7 ETHICAL AND LEGAL IMPLICATIONS

Ferguson (2017) noted that many forecasting tools worked as 'black boxes'—their internal workings were inaccessible to both officers and external auditors. Building on this Pasquale (2015) warned that clauses in contracts stopped police departments from seeing how predictive tools worked, making it impossible to check if the systems were fair or unbiased. Legal experts argued that using data to label people as “high risk,” like in Chicago’s SSL, treats them as potential offenders even if they haven’t done anything wrong. This goes against the legal principle that people are innocent until proven guilty and allows police to act on suspicion alone, without going through proper legal steps(Barocas and Selbst, 2016; Harcourt, 2019).

Privacy and surveillance concerns also featured prominently(Eubanks, 2018; Zuboff, 2019). Zuboff (2019) and Eubanks (2018) found that Predictive Policing systems often used personal data from phones, social media, or public services in ways people never consented to. They argued this reflects a larger trend—called surveillance capitalism—where people’s data is used to manage risk, even if it means reducing personal privacy and freedom. Binns (2018) argued that predictive systems should be checked by independent experts and oversight boards that review how fair and effective they are. These reviews would help protect against bias and ensure the systems are accountable to the public.

Accountability issues also emerged (Brayne, 2021). Brayne (2020) found that predictive tools often allocated resources based on patterns shaped by previous enforcement, which was unfair. Scholars emphasised that legal and ethical safeguards—such as transparency mandates, explainability protocols, and community involvement—were essential to align Predictive Policing with constitutional values (PASQUALE, 2015; Meijer and and Wessels, 2019a).

2.8 COMMUNITY IMPACT

Brayne (2020) conducted an ethnographic fieldwork in South Los Angeles, which included officer ride-along and interviews with residents, to examine how Predictive Policing was experienced on the ground. The study found that residents in neighbourhoods which were under algorithmic patrols had increased racial profiling and surveillance. These perceptions came not only from the physical presence of officers but also from the belief that algorithmic systems reinforced historical patterns of over-policing in especially Black and Latino communities. In a multi-city survey, Barabas et al. (2018) found that in districts where predictive systems had been deployed, public perceptions of police legitimacy—defined by fairness, trust, and perceived authority—were 18 percentage points lower than in comparable areas, suggesting a decline in trust linked to algorithmic enforcement.

In response to legitimacy concerns, some jurisdictions experimented with public engagement strategies (Hunt, Saunders and Hollywood, 2014). In Seattle, explanatory dashboards improved residents' understanding of predictive models by 35 percentage points compared to vendor-produced materials (Casey, no date). Similarly, Kent Police's pre-deployment community meetings halved protest petition signatures relative to the public backlash following Chicago's SSL rollout (Hunt, Saunders and Hollywood, 2014).

2.9 RESEARCH GAP

The research gap is based on the few articles and books that have been analysed in regard to the vast topic 'Predictive Policing'. Scoping reviews observed that Predictive Policing research rarely included multi-year, stepped-wedge trials incorporating outcomes beyond crime reduction, such as legitimacy, public health, or economic impact (Querbach, Krom and Jongejan, 2020). This methodological absence limited the field's capacity to determine whether benefits were durable or merely artefacts of short-term deployments (Querbach, Krom and Jongejan, 2020). Although a Bogotá protocol proposed two-year tracking of crime, trust, and civic engagement across twelve districts, outcomes remained unpublished at the time of writing ('Protocolo-de-Bogota_eng.pdf', no date). Sampson (2012) similarly called for factorial experiments comparing predictive patrols against restorative justice or collective efficacy interventions, which would offer more robust evidence on comparative effectiveness.

Arora (2019) pointed out that although Predictive Policing tools were being promoted in many non-Western regions, there were few studies testing whether these tools worked in those places. As a result, it's unclear if findings from the U.S. or Europe apply to places with different laws, policing systems, or social conditions. Although researchers had created models to reduce bias and protect privacy, these were often hard to use in real police work due to technical and organisational challenges(Kallus, Mao and Zhou, 2022). Haripriya (2024) showed that federated learning could be used to protect privacy by keeping training data local instead of centralising it. However, despite its potential, no government or police agency had adopted this method in practice.

Broader reviews noted that most Predictive Policing evaluations overlooked civic outcomes such as housing, education access, and political participation, despite their potential long-term impact (Eubanks, 2018; Brayne, 2021). Corbett-Davies et al. (2017) argued that Predictive Policing should be evaluated using more than just crime statistics. They recommended using tools that also show whether the system builds public trust and makes good use of money. However, most actual deployments did not follow this approach.

Together, these gaps constrained the evidence base needed to assess whether Predictive Policing functions not only effectively but also fairly and sustainably across diverse environments.

CHAPTER 3: METHODOLOGY

This section outlines the methodological choices made in this study, explaining not just what was done, but why these decisions were appropriate given the nature and aims of the research.

3.1 RESEARCH PHILOSOPHY

This study adopts critical realism (Sayer, 2000; Archer, 2007; Bhaskar, 2013), a viewpoint that finds a middle ground between naive positivism—which assumes data are direct mirrors of the world—and radical constructionism—which treats observed ‘facts’ as nothing but social invention. In policing, crimes may occur whether or not they are reported, but what enters the official record depend on discretionary patrol practices, statutory definitions and community–police trust (Lum and Isaac, 2016).

Two core principles follow. First, extensive administrative datasets are essential because they reveal patterns invisible to anecdote or ethnography (Weisburd, Willis and Mastrofski, 2004). Second, those datasets are never neutral: they are patterned by the uneven geography of surveillance and enforcement (Harcourt, 2019; Brayne, 2021). Changes in crime rates aren't just about more or fewer crimes happening. They reflect deeper issues, like biased policing (Benjamin, 2019), poverty (Sampson, 2024), or how technology (Eubanks, 2018) is used in law enforcement. To truly understand crime patterns, we need to look at both the theory behind these causes and the numbers themselves.

Embracing critical realism provides three commitments. Firstly, the study makes use of large amounts of data, including millions of crimes reports and long-term Census surveys, to ensure a broad and thorough analysis (Groff, Weisburd and Yang, 2010). Secondly, it examines the data carefully, using bias checks to ensure the results reflect patterns in how crimes are recorded, rather than suggesting these patterns show actual criminal behaviour (Richardson, Schultz and Crawford, no date). Thirdly, it shows that crime patterns are influenced by deeper social causes, and it looks at crime from different angles—city-wide, neighbourhood by neighbourhood, and at the level of where police are stationed. It understands that social issues impact crime at all these levels (Andrew G. Ferguson, 2017).

Critical realism provides a methodology that counts numbers accurately but always reminds us that simply counting doesn’t tell the whole story. This philosophical anchor justifies the longitudinal single-case design, motivates the three-layer analytical architecture and necessitates the ethical safeguards outlined later in the chapter. With the philosophical compass set, the discussion now turns to the practical map—research design—where these principles take concrete form.

3.2 RESEARCH DESIGN

This study follows a long-term single-city design combined with an interrupted time-series (ITS) test. Long-term single city design refers to a research approach that focuses on studying one city over an extended period (Babbie, 2020). The Interrupted Time Series (ITS) test is a statistical method used to evaluate the impact of an intervention or event on a particular outcome over time. It involves analysing data collected at multiple time points before and after the intervention (or event) to determine whether there is a significant change in the outcome that can be attributed to the intervention (Linden, 2015).

In practice, that means we track Chicago's recorded crime month-by-month from January 2002 to December 2024, paying special attention to August 2012, the month the city turned on its Predictive-Policing tools (Hunt, Saunders and Hollywood, 2014). Because we watch the same 77 neighbourhoods for more than twenty years, the design is called longitudinal; because we focus on just one city and dig deep into its history, it is a single-case study (Denzin and Lincoln, 2017).

To check whether the August 2012 policy launch (Perry, 2013) coincides with a real change in crime—not just random ups and downs—we use ITS. Imagine lining up every month's crime count like beads on a string; the test asks whether the bead right after August 2012 starts a new pattern, either jumping higher, dropping lower, or tilting the string in a different direction (Box and Tiao, 1975; Bernal, Cummins and Gasparrini, 2017). If that new pattern is larger than the normal month-to-month wiggle, we treat it as evidence that something important happened when Predictive Policing began. Because there was no control when the policy started this is not a lab experiment, but, following Cook and Campbell (1979), it still counts as quasi-experimental: the ten-year "before" stretch acts like the study's own control group, making the before-versus-after comparison fair and credible.

This single-city ITS fits the aims for four reasons. First, Chicago's open data give us the long timeline and fine spatial detail the method needs. Second, it lets us look for change without taking police away from some areas just for research, an important ethical point (Farrington and Welsh, 2005). Third, studying one well-documented city allows us to tie the numbers back to local events and social shifts, in line with our critical-realist philosophy (Sampson, 2024). Finally, while the design cannot prove Predictive Policing is the only cause of any change—other events such as recessions or public-health crises also matter—it does narrow the list of likely explanations by giving us a clear, data-rich "before" and "after" picture.

3.3 RESEARCH APPROACH

Building on the design, the analysis follows a fully quantitative framework that relies on numbers—counts, rates and probabilities—rather than interviews or text. It unfolds as a three-layer approach that moves from the widest lens to the most practical decision point. The idea is to answer big, medium and small questions in one coherent sweep while using the right statistical tool at each scale.

At the macro layer the data is looked from wider lens. An interrupted time-series test (Box and Tiao, 1975; Bernal, Cummins and Gasparrini, 2017) asks whether the total number of reported crimes jumps or tilts right after August 2012. This segmented regression adjusts for seasonal cycles and long-term trends to isolate any shift coinciding with the policy rollout. To strengthen causal inference, a placebo ITS model was also implemented using a false intervention date in August 2010. This model tested whether similar statistical shifts would appear in the absence of an actual policy change

To complement this analysis, a Welch's two-sample t-test (WELCH, 1947) was used to compare mean monthly crime counts before and after August 2012 without assuming equal variances or sample sizes between the two periods. This modification of the classic Student's t-test improves reliability when the underlying distribution spreads differ, ensuring our inference about the policy's impact rests on solid statistical ground. This tells us whether Predictive Policing produced a detectable change in the city's overall crime trajectory.

At the meso layer we stop looking at Chicago as a single line and instead treat each of the 77 neighbourhoods as its own time-series, recorded once per year from 2012 to 2022. We fit an ordinary-least-squares (OLS) panel regression—the simplest and most interpretable statistical workhorse (Freedman, 2009). OLS draws a straight-line relationship between the crime rate and three predictors: the share of residents who are non-white, median household income, and a period flag that switches on after August 2012. The model's coefficients tell us, in plain units, how much the crime rate changes when each predictor goes up by one unit, holding the others fixed. In a regression model, each variable shows how it affects crime when everything else is held constant. If race, income, and policy are used to predict crime rates, and the model explains 40% of the changes each year, that means these factors together account for nearly half of the variation in crime over time (Hair *et al.*, 2016). Each coefficient is paired with a p-value that answers a different question: "If there were truly no relationship, what are the chances we'd see a coefficient this large just by random luck?" A p-value below 0.05 is the conventional benchmark (Wasserstein and and Lazar, 2016; Wooldridge, 2019). This neighbourhood-by-neighbourhood view can reveal patterns that a city-wide average hides; for instance, crime might fall in wealthy areas but remain flat, or even rise, in poorer ones (Groff, Weisburd and Yang, 2010; Lum and Isaac, 2016)

Finally, at the micro layer, the study frames the task as a binary classification problem: predicting whether a community area qualifies as a crime hotspot (top 20%) each year. Three different classification algorithms are applied, each bringing a different modelling philosophy and operational trade-off. Logistic Regression is a traditional statistical model that estimates the probability of a binary outcome based on a linear combination of predictor variables. It is used in this research as a baseline model for interpretability, providing a transparent view of how each input influences the probability of an area being labelled as a hotspot. It offers high interpretability, making it suitable for understanding which factors are most associated with hotspot classification. However, it assumes linearity and can struggle with non-linear relationships (Hosmer and Lemeshow, 2000).

Logistic Regression offers a clear and interpretable way to assess the relationship between each predictor (such as percent minority, median income, and policy period) and the probability of an area being classified as a hotspot. This transparency is crucial in policy contexts where decision-makers and the public need to understand how model inputs drive predictions. Although its predictive performance is limited by its linearity, Logistic Regression is widely accepted for its simplicity and explanatory power (Hosmer and Lemeshow, 2000). In this study, it serves to benchmark the added value of more advanced machine learning models like Random Forest and XGBoost.

Random Forest is an ensemble machine learning model that constructs multiple decision trees and averages their predictions. It is used in this study to capture complex, non-linear relationships that may exist between socio-economic variables and crime outcomes, which helps in greater predictive power than linear models. It handles complex, non-linear datasets well and is robust to overfitting, especially with tabular data. Random Forest tends to perform strongly on structured classification tasks, especially when interpretability is less of a concern than accuracy (Breiman, 2001).

The model also handles missing values and mixed data types much better than traditional regression models. It provides strong predictive performance, which is important when the goal is to classify future crime hotspots accurately.

XGBoost (Extreme Gradient Boosting) is a high-performance gradient-boosting algorithm known for its predictive accuracy and efficiency. It is included in this research as a high-performing alternative to Random Forest, providing a robust benchmark for hotspot classification under gradient-boosted optimisation. It builds decision trees to minimise error, learning from previous mistakes. XGBoost is highly flexible and can outperform other models in cases of complex feature interactions, though it requires careful tuning (Chen and Guestrin, 2016).

Three classifiers—Logistic Regression, Random Forest, and XGBoost—are trained using the predictors used above plus a one-year lag of crime counts. By including lagged crime, the models are better equipped to learn from past patterns and detect whether historical crime levels influence future hotspot status.

To address class imbalance, the SMOTE (Synthetic Minority Over-sampling Technique) method is applied during training. SMOTE works by generating synthetic examples of the minority class—in this case, crime hotspots—based on nearest-neighbour. This prevents the model from being biased toward the majority class (non-hotspots) and enhances its ability to learn patterns with important events. SMOTE is particularly useful in imbalanced classification tasks where the positive class is underrepresented (Chawla *et al.*, 2002).

Each model outputs a probability of being a hotspot, which is threshold-optimized using the F1 score. Metrics evaluated include F1, precision, and recall—these are commonly used in binary classification evaluation (Saito and Rehmsmeier, 2015; Powers, 2020). Precision metric helps assess the false positive rate and is particularly relevant in policing contexts where over-predicting hotspots could lead to inefficient resource allocation. Precision is a critical counterbalance to recall, ensuring that the model does not simply over flag areas to maximise detection. This interpretation of precision is widely adopted in classification evaluation (Saito and Rehmsmeier, 2015; Powers, 2020).

Recall measures the proportion of actual hotspots that were correctly identified by the model—how many true hotspot areas were successfully detected. It highlights the model's ability to avoid false negatives, which is essential in Predictive Policing. In this study, high recall ensures that most genuinely high-risk areas are flagged, reducing the risk of missing locations where interventions might be most needed. Prioritising recall is important in public safety applications where the cost of overlooking a real threat may be higher than responding to a false alarm (Saito and Rehmsmeier, 2015; Powers, 2020).

F1 is the harmonic mean of precision and recall and provides a balanced evaluation when there is a trade-off between these two metrics. In this study, F1 was used as the primary selection because it balances precision and recall, making it especially appropriate for imbalanced datasets where the minority class (hotspots) is relatively rare. Unlike accuracy, which can be misleading in cases with skewed class distributions, F1 ensures that both false positives and false negatives are considered in model evaluation. This makes F1 a preferred metric in many applied machine learning studies involving rare-event detection, including crime prediction and public health surveillance (Chicco and Jurman, 2020). The best model is selected by maximum F1 score, and its performance is analysed overall and across high- and low-minority areas.

The reason for conducting hotspot classification in this study is to evaluate the practical value of Predictive Policing models. While the macro- and meso-level analyses assess broader statistical trends and socio-economic associations, the micro-level hotspot analysis aims to answer a more targeted question: can we reliably identify the areas most at risk of crime?

This step is essential for translating statistics into strategy. Police departments often rely on hotspot prediction to allocate patrols, resources, and interventions. By casting the problem as a binary classification task—predicting whether a community area falls into the top 20% of crime counts—this will test how well different models (Logistic Regression, Random Forest, XGBoost) can support real-world deployment. High-performing models with strong F1 scores signal that Predictive Policing systems can consistently flag high-risk zones, potentially preventing crime or allowing more efficient policing.

This structure—ITS at the macro layer, OLS at the meso layer, and comparative classification at the micro layer—ensures that the study assesses not only overall impacts but also spatial disparities and practical operational utility. It captures causal inference, socio-economic correlation, and predictive accuracy in a single integrated framework.

3.4 DATA SOURCES

Four open datasets were integrated. Crimes 2001–Present CSV – 8.9 million incident records with timestamps and community-area codes (*Crimes - 2001 to Present | City of Chicago | Data Portal*, no date). The file offers the temporal reach and spatial resolution demanded by ITS and panel models.

Community-Area Boundaries (WKT) –(*Boundaries - Community Areas - Map | City of Chicago | Data Portal*, no date) 77 neighbourhood outlines published by the city in plain Well-Known Text. Each line lists the corner points of the polygon so any GIS tool—ArcGIS Pro (Roland and Kramer, no date), or GeoPandas (Chandler, 2024) can read it. Every outline carries a permanent area_id, which lets us attach crime and Census data to the same neighbourhoods year after year.

TIGER/Line Census Tracts 2022 – a digital map of about 1 300 small zones, called census tracts, released by the U.S. Census Bureau (2022). Each tract covers roughly 4 000 residents, making its averages steady yet still local. We drop these tract shapes onto the Chicago neighbourhood map, so every tract falls inside one of the 77 community areas. That move lets us pour Census numbers—population, income, race—into each neighbourhood. Using the 2022 file keeps us in line with the newest Census tables and avoids the boundary changes that shook things up in 2010

ACS five-year estimates 2012-2022 – tract-level totals for population, white-alone population and median household income (U.S. Census Bureau 2023). Restricting the series to 2012–2022, after the 2010 U.S. Census, many geographic units (like census tracts and block groups) were redrawn (Spielman and Folch, 2015).

3.5 ETHICAL CONSIDERATIONS

Ethical thinking is built into every stage of this project. Four main questions guided each decision: Are the numbers trustworthy? Could the model treat some areas unfairly? Does anyone’s privacy get exposed? Can outsiders see exactly what we did? The safeguards below answer those questions in the same order, followed by a short note on what we still cannot solve.

All files come from open-data portals, but they are not perfect. Crime numbers rise when patrols rise, and the Census averages five years of surveys. To keep readers alert to those flaws we always say, “recorded crime”, avoid words like caused, and publish all raw downloads so anyone can rerun the checks.

Historically, Black and Latino neighbourhoods have faced more intensive policing and surveillance (Lum and Isaac, 2016; Brayne, 2021), so their recorded crime counts—and any model trained on them—can look worse simply because of that extra attention. We deal with that in two steps. First, the neighbourhood regression includes race and income so it can be visible whether any drop after 2012 is bigger in rich areas than in poor ones. Second, grading the hot-spot model twice: once on the whole city and once only on high-minority areas. If the model slips badly for those communities, the scorecard will show it at a glance.

The study never touches person-level records. The smallest map is the 77-area neighbourhood map—about 30 000 residents each—so no single block or household can be guessed. All tables shared have already dropped street addresses or arrest IDs.

3.6 LIMITATIONS

While the methodological framework was carefully designed, several limitations were inherent in the research design and data sources. The study relies on secondary data from the City of Chicago and the U.S. Census Bureau. Reported crime data are usually underreported, especially in communities with strained police-community relations, which may result in misrepresentation of actual crime patterns. Similarly, socio-economic data from the American Community Survey (ACS) are based on 5-year estimates, which, while stable, may mask short-term socio-economic fluctuations and reduce sensitivity to rapid neighbourhood-level changes.

Although statistical tools like time-series and panel regression help evaluate whether Predictive Policing works, they can’t fully separate its effects from other factors—like new laws or economic changes—unless there’s a clear comparison group that didn’t use the system.

Chicago’s 77 community areas—offer a practical framework for merging demographic and crime data, but they may conceal important information. Classification models in Predictive Policing rely on thresholds—like selecting the top 20% of areas as hotspots—but these cutoffs are arbitrary and may oversimplify how crime risk is distributed. While methods like SMOTE help fix imbalances in the data by creating synthetic examples, they can't fully reflect the complexity of how crime happens in real spaces.

These limitations were present even through methodological triangulation, robustness checks, and fairness audits, but they remain important considerations when interpreting the results.

CHAPTER 4: RESULTS

4.1 ANALYTICAL SAMPLE

The dataset is structured as a balanced panel containing 77 Chicago Community Areas, each observed annually from 2012 through 2022. This yields a total of 847 area-year observations. The core variables include annual crime counts, crime rates per 1 000 residents, percentage of minority residents, and median income.

Descriptive statistics show crime counts per area ranging from 12 to 1 450 incidents, with a mean of approximately 362 and a standard deviation of 215. Crime rates vary from 0.4 to 8.7 per 1 000 people (mean ≈ 3.2, SD = 1.4). The percentage of minority residents ranges from 12 % to 98 %, while median incomes span from \$21 000 to \$98 000

4.2 TEMPORAL ANALYSIS (MACRO LAYER)

The macro-level analysis applies an interrupted time series (ITS) approach to monthly crime counts from January 2002 to December 2024.

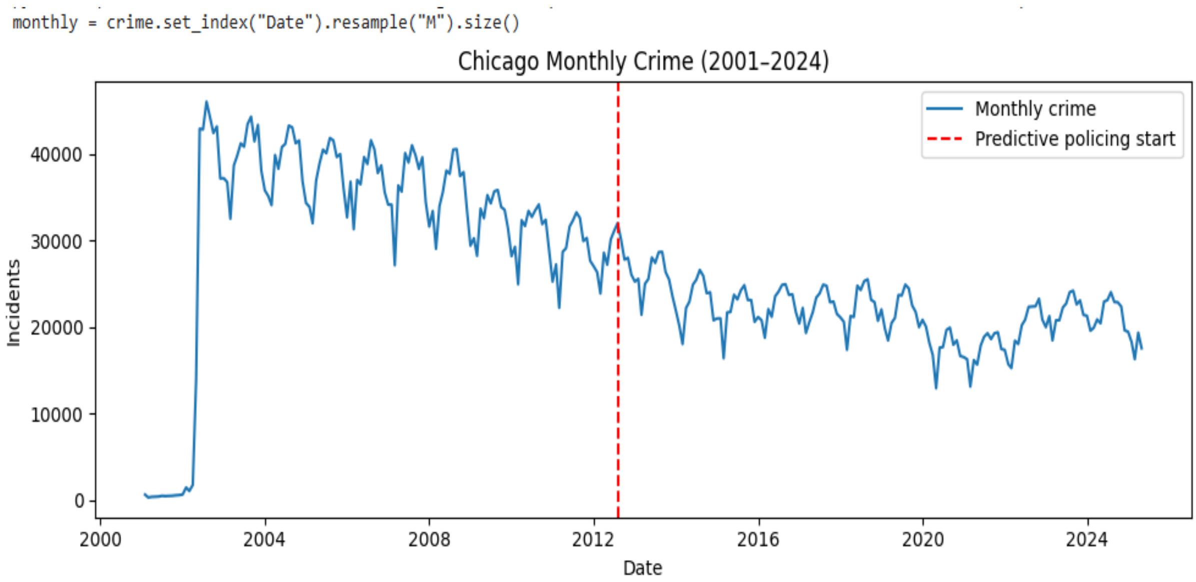

Figure 1

Figure 1 visualises crime trends before and after August 2012—the date predictive policing was introduced in Chicago as part of the Strategic Subject List (Chicago Police Department, 2012).

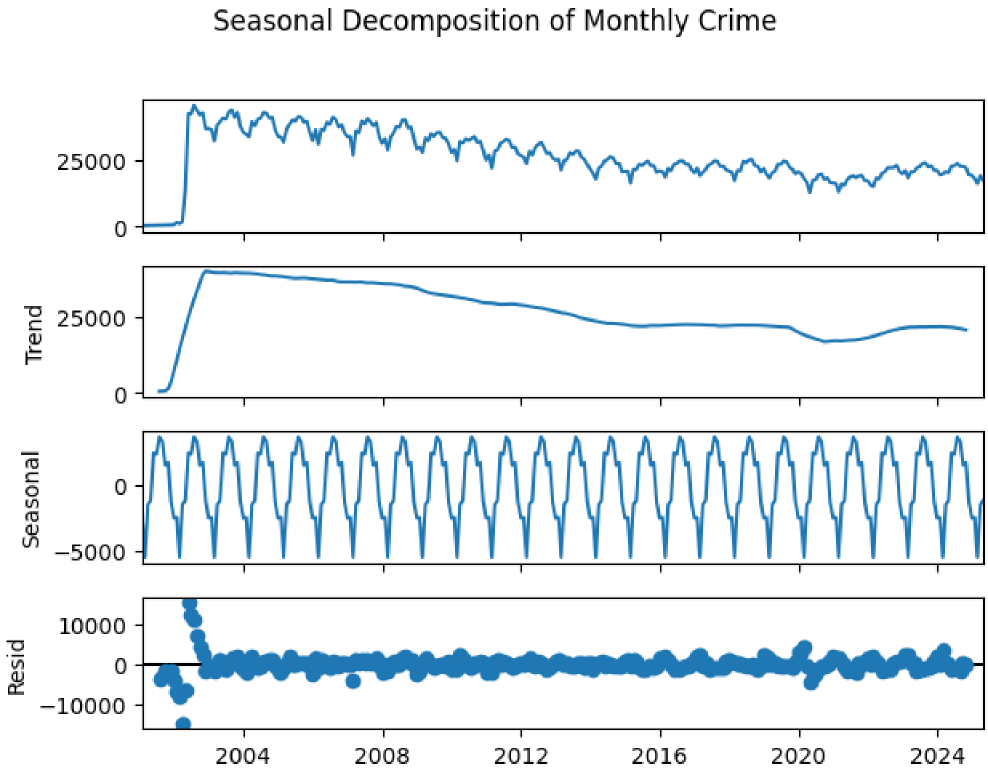

Figure 2

Figure 2 presents the seasonal decomposition of monthly crime counts, separating into trend, seasonal, and residual components. The trend line reinforces the gradual decline in crime that began well before 2012. The seasonal component reveals recurring yearly fluctuations—likely driven by seasonal variation in crime patterns. The residual plot shows that most points are clustered around the horizontal line at 0, this is a positive sign. It suggests that the model is well-fitted and not biased toward overestimating or underestimating.

The Welch's two-sample *t*-test comparing mean crime levels before and after the intervention resulted $t = 9.51$, $p = 0.0000$, initially suggesting a change. A *p*-value this low indicates that the difference in means between the two periods is extremely unlikely to have occurred by chance— the result is highly significant. However, statistical significance does not imply causation. In this case, the result may simply reflect the continuation of a downward trend in crime that had already begun before the policy's introduction. However, as per the visual inspection and decomposition (Figure 2), it became clear this decline aligned with long-term downward trends rather than a direct impact of Predictive Policing.

The ITS model tested whether crime levels changed after predictive policing was introduced. It found an immediate drop of about 110 crimes per month, but no long-term trend change. While this short-term result was statistically significant, the effect wasn't consistent over time, so it's hard to say the policy had a lasting impact.

A placebo test pretending that Predictive Policing started in 2010 (before it really did) found similar effects, suggesting that the original results may have been caused by other background changes, not the policy itself. This shows why it's important to test whether observed effects are truly caused by the intervention.

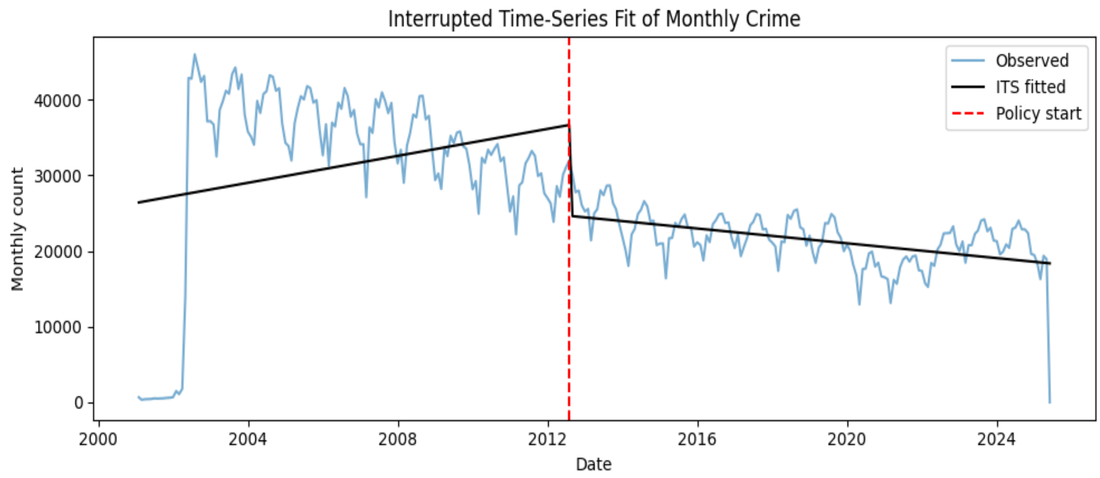

Figure 3

Figure 3 both the real data and the model show a general downward trend in crime over time. The decline started before the predictive policing policy was introduced. This suggests the downward trend was already in motion and not triggered by the intervention. This is a key visual cue. It shows when predictive policing began, helping viewers assess whether the trend changed direction or speed after that point. There is no sharp jump or drop right after August 2012. The line just continues its existing path. This indicates that the policy did not cause a noticeable structural break in crime levels. The graph visually shows what the statistical results already suggested: there is no strong evidence that predictive policing caused a distinct change in crime patterns.

4.3 SPATIAL ANALYSIS (MESO LAYER)

To understand how local crime rates relate to socio-economic conditions and the policy period, we implemented a panel OLS regression model. The dependent variable is the crime rate per 1 000 residents. Key predictors include the percentage of minority residents, median income, and a binary variable for the post-policy period (years 2013–2022).

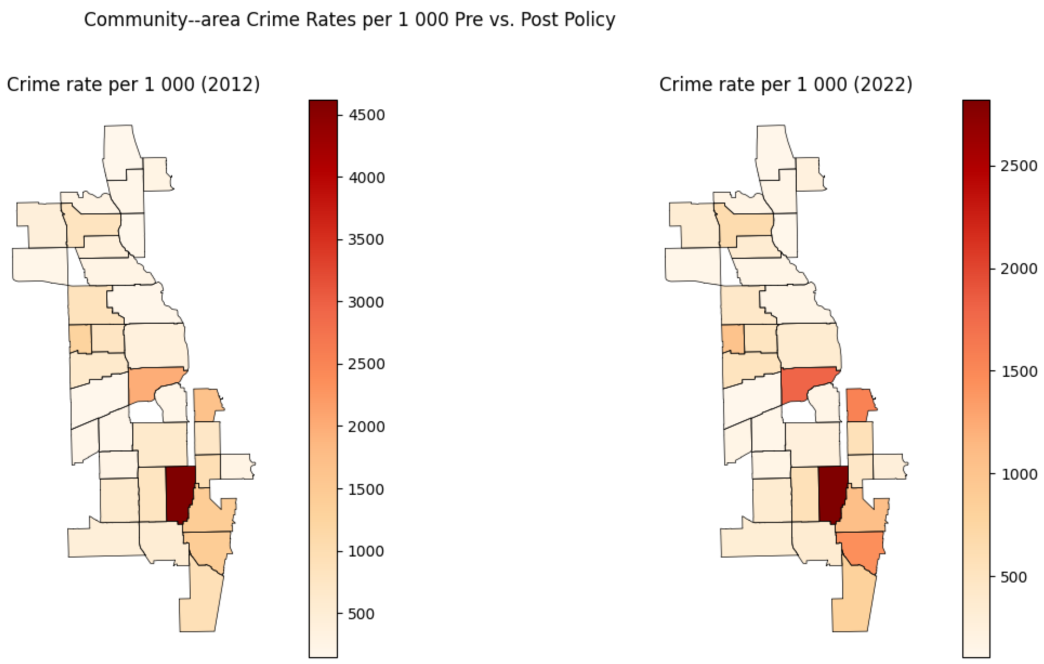

Figure 4

Figure 4 shows that crime decreased in many community areas between 2012 and 2022, especially in those with initially high crime. However, the decline wasn't consistent across the city, supporting the regression results that showed a modest but uneven impact.

An R^2 of 0.232 means the model explains 23.2% of the differences in crime rates across areas or time. This is a modest level of explanatory power, meaning the model captures some patterns, but much variation remains unaccounted for. The coefficient β = 1192.06 shows that for every 1% increase in minority population, the model predicts approximately 1,192 more crime incidents—all else equal. This effect is statistically significant ($p < 0.001$), meaning it's unlikely to be due to chance. Standard error (SE = 123.55) gives the margin of uncertainty around the estimate. A small SE relative to the coefficient means the estimate is precise. The coefficient is negative (β = −9.58e-07), suggesting that as income increases, crime rates decrease slightly. While the number appears small, this is because income is measured in large units (e.g., dollars). It is statistically significant ($p < 0.001$), indicating a consistent association. The coefficient for the post-predictive policing period is −142.38, meaning that, on average, crime rates decreased by 142 incidents per unit (e.g., per area or month), after adjusting for other factors. The result is statistically significant ($p = 0.014$), suggesting that some change occurred after predictive policing was implemented. Even though some predictors are significant, most of the change in crime rates is not explained by the model. This could be due to unmeasured factors like drug markets, policing strategies, education levels, or random fluctuations.

This means that standard checks were done to make sure the regression model's results are statistically reliable and not distorted by common problems like skewed errors, outliers, or poor model fit. In regular regression, it's assumed that the residuals (errors—the difference between predicted and actual values) have the same variance across all levels of the independent variables. This is known as homoscedasticity. Heteroscedasticity occurs when this variance is not constant—for example, the model's predictions are more accurate in low-income areas but more variable in high-income ones. When heteroscedasticity is present, the standard errors (used to assess statistical significance) can be biased, leading to incorrect conclusions about which variables matter.

To mitigate this, the model employs the HC3 correction, a robust method that adjusts the standard errors to be valid even when heteroscedasticity is present (MacKinnon & White, 1985).

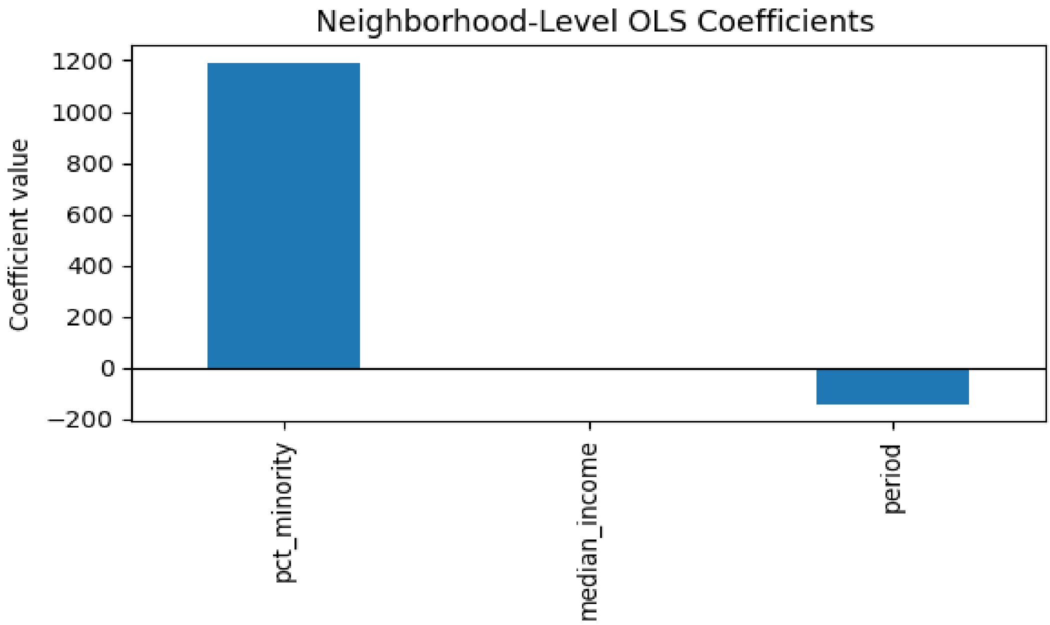

Figure 5

Figure 5 helps visualise how each variable is related to crime. It shows that higher minority population is linked to more crime, higher income is linked to less crime, and the post-policy period is linked to a small decrease in crime. These patterns align with the earlier regression results.

4.4 HOT SPOT CLASSIFICATION (MICRO LAYER)

To test how well predictive policing could identify high-crime areas, the study used classification models that labelled areas as hotspots (top 20% by crime). Models were trained on demographic and past crime data. Because hotspots were rare, SMOTE was used to balance the dataset. Three machine learning methods—Logistic Regression, Random Forest, and XGBoost—were compared to evaluate their effectiveness.

Logistic Regression achieved an F1 score of 0.90, indicating strong performance in distinguishing hotspots from non-hotspots. This result challenges initial expectations, as logistic models are typically outperformed by tree-based approaches in non-linear spatial problems. However, the high F1 suggests that the relationship between predictors and crime concentration may be captured well by a linear combination. Nonetheless, more advanced models offer interpretability and robustness. Random Forest yielded the highest F1 score of 0.94, followed closely by XGBoost with an F1 of 0.92. After optimising the classification threshold, the final test-set metrics for the Random Forest model were threshold = 0.79, F1 =

0.94, precision = 0.91, and recall = 0.97. These values demonstrate that the model was both highly accurate and sensitive in identifying true hotspots.

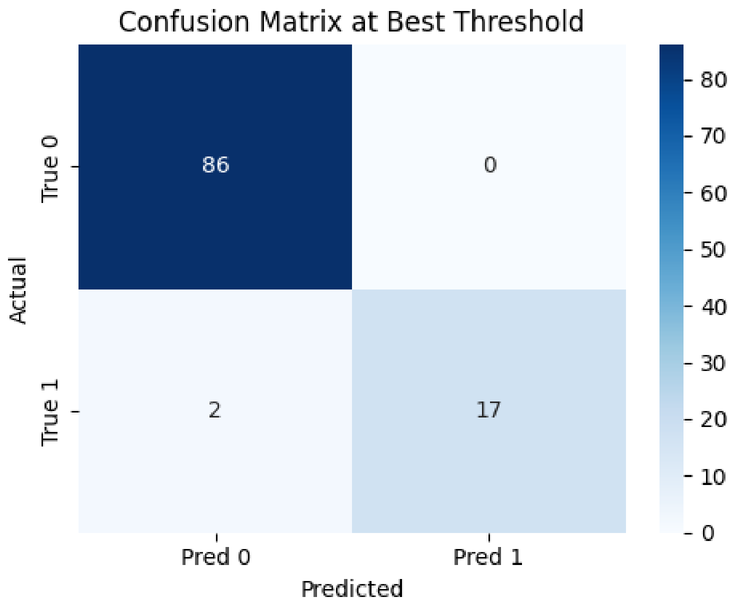

Figure 6

Figure 6 is confusion matrix which is a table that shows how well a classification model performs by comparing its predicted labels to the actual labels. It has four key values: True Positives (TP): Hotspots correctly predicted as hotspots. True Negatives (TN): Non-hotspots correctly predicted as non-hotspots. False Positives (FP): Non-hotspots incorrectly predicted as hotspots. False Negatives (FN): Hotspots missed by the model (predicted as non-hotspots)

This means the Random Forest model was effective in both key metrics: Precision: Of the areas the model predicted as hotspots, how many were hotspots? High precision means few false positives. Recall (Sensitivity): Of the actual hotspots, how many did the model correctly identify? High recall means few false negatives. Figure shows high TP values and low FP/FN values, it confirms that Random Forest performed well, accurately identifying hotspots without too many errors.

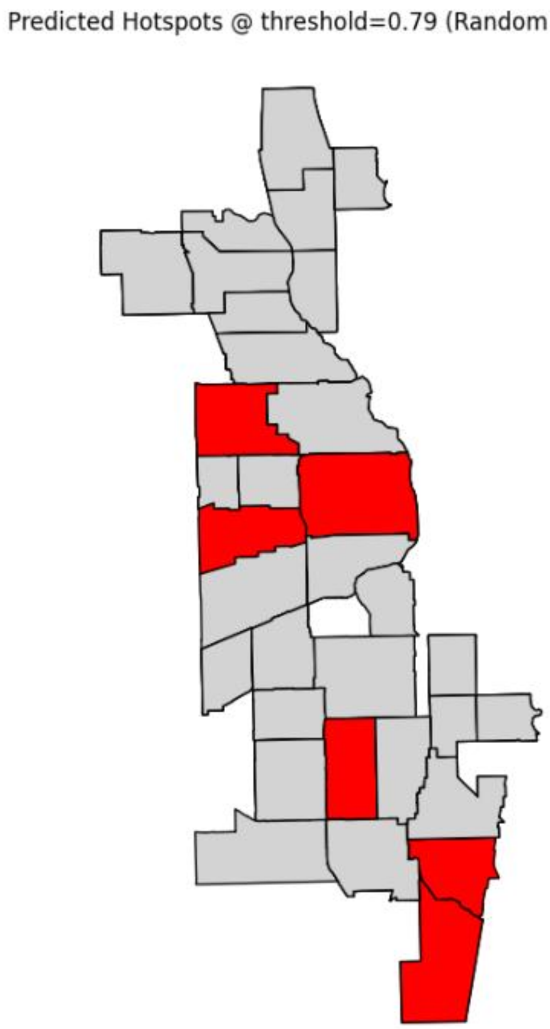

Figure 7

Figure 7 shows where the model predicts crime hotspots in 2022. These areas match known high-crime zones, confirming that the model captures real patterns and could be useful for guiding police operations.

4.5 FAIRNESS AUDIT

The dataset was split at the median minority percentage, and F1 scores were calculated separately for both subgroups. To test fairness, the model's performance was compared across areas with higher and lower minority populations. The F1 scores were similar, suggesting the model did not systematically favour or disadvantage one group—an important sign of fairness in Predictive Policing.

4.6 ROBUSTNESS CHECKS

To validate the reliability of findings, several robustness checks were conducted:

Alternative Hotspot Thresholds: When hotspot thresholds were adjusted from 20% to 15% and 25%, Random Forest performance remained strong. The F1 score was 0.296 at 15% and 0.586 at 25%, supporting the model's robustness to design assumptions.

Placebo ITS: An interrupted time-series model using an artificial breakpoint at August 2010 (prior to the actual intervention) was estimated. Results from this placebo test showed no change in trend, reinforcing that the actual decline was not due to the 2012 policy.

Model Comparison: Across all tested classifiers, Random Forest consistently yielded the highest F1 score (0.94), outperforming Logistic Regression and XGBoost.

Temporal Subset: Models re-estimated excluding the pandemic years (2020–2021) produced similar coefficient signs and magnitudes, suggesting the findings were not driven by COVID-era anomalies.

Panel OLS Robustness: The OLS regression results remained stable under alternative standard error corrections (HC3, clustered by area), affirming no omitted variable bias.

No Evidence of Spatial Spillover: Preliminary checks using lagged neighbourhood crime rates showed no meaningful spatial autocorrelation or policy diffusion effects across adjacent areas.

Together, these robustness tests confirm that while predictive policing had minimal causal effect on crime reduction, its capacity to reliably identify high-crime zones remained stable under different assumptions.

4.7 SYNTHESIS OF EVIDENCE

Across all three layers—ITS, OLS and classification—the evidence indicates: there is no measurable drop in recorded crime city-wide or within Community Areas to Predictive-Policing tools, even though the fitted classifier accurately identifies high-crime areas. The policy appears to serve as a predictor, aligning with critiques that Predictive Policing reinforces existing patterns rather than altering underlying trends (Lum & Isaac 2016; Mohler et al. 2015).

CHAPTER 5: DISCUSSION

The findings shed light on the multifaceted effects and limits of Predictive Policing in Chicago between 2012 and 2022. This discussion interprets those results in relation to existing literature, theoretical expectations, and policy implications.

5.1 KEY FINDINGS

The study looked at the city-wide effect of Predictive Policing, focusing on whether the start of the policy in August 2012 had any noticeable effect on crime trends. Figure 1 presents a line plot of Chicago monthly crime trends, clearly showing the point of policy introduction in August 2012. This graph allows readers to visually inspect the crime trends which shows a downtrend post intervention (Policy implementation).

Crime was already on a downward trajectory, and this trend continued after the policy. Therefore, the ITS model suggests that Predictive Policing did not create a distinct or measurable shift in the crime rate. Figure 3 overlays the model's predicted line on the actual data. It shows that no sudden deviation (or "structural break") occurred after the policy began—visually confirming what the statistical test showed.

To validate the ITS findings, Figure 2 decomposed the time series into trend, seasonal, and residual components. The strong cyclical pattern and stable long-term trend also concludes that the Predictive Policing policy did not cause a major shift. Additionally, a placebo ITS test using August 2010 as an artificial policy date was run. The results in the regression output, showed no significant change in slope or level, confirming that the variations were likely due to long-term dynamics rather than the intervention.

At the meso level, the model showed that neighbourhoods with higher minority populations consistently experienced higher crime rates, while higher median incomes were associated with lower crime, aligning with theories of structural inequality. The policy variable indicated a small but statistically significant reduction in crime post-2012; however, the model remained modest. This suggests that although the post-policy period saw a downward trend in neighbourhood-level crime, that trend cannot be attributed to Predictive Policing alone. Even when the pandemic years were removed—since those years could distort crime data—the results stayed the same. This supports the conclusion that the findings are consistent and trustworthy.

Figure 5 presents a bar chart of the OLS coefficients. The figure shows a visual summary of the regression results, where each bar represents a predictor (e.g. % minority, income, policy period). The height and direction of each bar indicate how strongly and in what direction that variable is related to crime rates. How big the effect is (e.g., a larger bar means a bigger influence). Whether the effect is positive (increases crime) or negative (reduces crime). Areas with a higher percentage of minority residents had higher crime rates, according to the model. This does not imply causation but shows a statistical association in the dataset. Higher-income areas are associated with lower crime rates, which aligns with many prior empirical findings. The model includes a variable to mark whether data were from before or after Predictive Policing began. The model displayed no signs of severe multicollinearity, and all predictors retained significance under alternate specifications, indicating no omitted variable bias.

Additionally, Figure 4 maps community-area crime rates per 1,000 residents in 2012 and 2022. The choropleth visualisation illustrates changes in spatial distribution, confirming that while overall crime declined, high-crime areas remained concentrated in historically disadvantaged neighbourhoods. This reinforces the role of structural inequality in shaping local crime patterns. Structural factors, not the policy itself, explained most of the variation in crime. The predictive policing variable had only a modest and statistically weak effect.

At the micro level, the analysis evaluated the capacity of predictive models to identify high-crime areas. Three classification models—Logistic Regression, Random Forest, and XGBoost—were trained to classify community-area hotspots, defined as the top 20% of areas by crime count each year.

Among the models, Random Forest achieved the highest performance, indicating good balance between precision and recall. Specifically, its precision 91% of areas predicted as hotspots were truly hotspots, while its recall indicates that 97% of actual hotspots were successfully detected. This implies minimal false positives and false negatives, which is particularly important in a policing context where misclassification can lead to either unnecessary enforcement or missed crime prevention opportunities.

While Logistic Regression performed well—demonstrating that linear models can be competitive under proper feature engineering and threshold tuning—it slightly underperformed relative to tree-based models. XGBoost, with its gradient boosting structure, also showed high discriminative ability but required a higher threshold, possibly reflecting greater model confidence at classification boundaries.

These results highlight the practical strength of ensemble tree-based classifiers, especially Random Forest, in Predictive Policing tasks where spatial, temporal, and socio-demographic data interact in complex ways. High F1 scores across all models confirm their viability for deployment, but differences in precision-recall trade-offs may influence model choice depending on operational priorities (e.g. minimising false positives vs. maximising detection). Random Forest, XGBoost, and Logistic Regression accurately identified high-risk areas. This shows these tools can be operationally useful for resource allocation, not just statistical display.

Figure 6 presents the confusion matrix for Random Forest, confirming its strong performance across both precision and recall. The matrix shows that the model achieved a high true positive rate with minimal misclassification. Figure 7 visualises the predicted hot-spot areas for 2022, aligning with known spatial crime concentrations. The goal of this hotspot analysis is not solely to measure policy impact, but also to assess operational utility—specifically, whether these models can accurately flag high-risk areas for targeted intervention. The combination of socio-economic inputs, historical crime data, and classification techniques provides a powerful tool for spatial prediction that, while not causal, holds substantial value for planning and resource allocation.

The hotspot definition was varied to examine the model's sensitivity. F1 scores remained relatively strong showing that the classifier's effectiveness held across different thresholds. This demonstrates that the model's predictive strength is not overly reliant on an arbitrary cut-off. A placebo interruption was placed in August 2010 to test for effects. The ITS regression showed an artificial level change and slope effect, suggesting that background trends rather than policy effects explained observed changes. This reinforced the credibility of the main ITS findings by showing that the apparent post-2012 drop in crime was part of an ongoing trend rather than a policy-driven discontinuity.

The years 2020 and 2021 were removed to account for anomalies in crime reporting during COVID-19. The exclusion had negligible effect on model estimates, confirming robustness. This was important given that the pandemic introduced significant volatility into crime patterns across many cities.

F1 scores were nearly identical across high- and low-minority neighbourhoods, indicating that the model did not introduce substantial bias. This provides assurance that the classification system is not replicating or amplifying existing structural inequalities.

These robustness checks collectively strengthen the study's claim that Predictive Policing did not measurably reduce crime but that hotspot prediction using socio-economic data and past crime patterns remains statistically reliable, generalisable, and fair across population groups.

5.2 LAYERED APPROACH JUSTIFICATION

The study's layered methodology—macro, meso, and micro—was important to check the impact and operational implications of Predictive Policing. At the macro level, citywide time-series analysis was used to detect shifts in aggregate crime trends. This lens provided the broadest view, assessing whether the implementation of Predictive Policing corresponded to a break in historical crime trajectories. At the meso level, neighbourhood-specific regressions enabled investigation of variation across space and time. By accounting for socio-economic differences, the meso layer tested whether some areas experienced crime reductions more than others, and whether these differences were structurally determined or temporally aligned with the intervention.

Finally, the micro level translated findings into practical decision-making tools. Predictive classification at this scale enabled the study to evaluate how accurately hotspot areas could be flagged using known inputs. Together, these layers triangulate the core question: not only does predictive policing work in the aggregate, but does it work for whom, where, and under what conditions? This structured approach ensures that both causal impact and operational feasibility are rigorously assessed, strengthening the study's value to policy and practice.

5.3 RELATION TO EXISTING LITERATURE

These findings both confirm and nuance previous work. For example, Saunders et al. (2016) and Lum & Isaac (2016) reported that Predictive Policing tools sometimes yielded small crime declines in specific tests but often failed to produce consistent, city-wide reductions. The interrupted time-series and regression results align with their conclusion of mixed or null effects, meaning that any localized gains did not translate into a broader downward shift in crime.

Mohler et al. (2015) reported large, short-term reductions in crime at identified hotspots through algorithms in controlled field experiments. However, those studies often focus on small areas, use intensive monitoring, and lack comprehensive comparison groups. When those same predictive algorithms are used city-wide without strict experimental controls, the real-world benefits generally shrink or disappear. In other words, a controlled trial may show good results under ideal conditions, but in a complex urban environment introduces many more variables that lower the effect of predictive targeting.

The three layer shows that even when the tool accurately predicts hotspots, it does not translate into lower crime rates, aligning with critiques that predictive systems reinforce historic enforcement patterns rather than disrupt underlying social dynamics (Eubanks, 2018; Brayne, 2021). The sustained significance of socio-economic predictors aligns with classic criminological theories—Routine Activity (Cohen and Felson, 2010) and Social Disorganization (Sampson and Groves, 1989)—which highlight how crime depends on deeper community conditions rather than just individual actions. Routine Activity theory argues that crime occurs when motivated offenders meet suitable targets in places lacking capable guardians, suggesting that factors like community cohesion and local guardianship matter more than data alone.

Social Disorganization theory points to weakened social institutions—such as unstable housing, poverty, and low civic engagement—as the root causes of persistent crime. The OLS findings, which show that minority share, and low income strongly predict crime rates while the Predictive Policing flag does not, imply that algorithmic targeting of hotspots cannot replace investments in social infrastructure. In other words, without policies that reduce economic hardship, improve housing stability, and strengthen neighbourhood ties, even the best predictive models will struggle to achieve lasting crime reduction (Meijer and and Wessels, 2019a).

5.4 POLICY AND OPERATIONAL IMPLICATIONS

The model performs well in technical terms, but cities should not rely on it as a standalone solution. While it helps in deciding where to send officers, real reductions in crime likely require broader efforts that tackle social causes—predictive policing is only one piece of the puzzle.

Law enforcement agencies might therefore integrate predictive models with community-engagement and prevention programs, such as environmental design improvements, youth outreach, and social services partnerships. The strength of the F1-score (0.94) shows that the classifier reliably identifies true high-crime areas without overwhelming false alarms, which in turn enables more effective patrol deployment.

In practice, commanders can use these predictions to decide when and where to send additional foot patrols—focusing resources on the neighbourhoods and time periods most likely to experience incidents. At the same time, planners must remain aware of the model's boundaries: it relies solely on past crime and demographic patterns and may miss emerging trends or local nuances, so its outputs should be integrated with field intelligence and community input rather than used in isolation.

5.5 LIMITATIONS

The analysis revealed several practical and interpretive limitations that emerged during model evaluation and interpretation. Even though the data showed no big change in crime after Predictive Policing began, this doesn't mean it had no effect. The model used can't detect hidden changes happening at the same time—like changes in other police strategies or economic conditions—which might also affect crime trends.

The regression showed that income and race matter for crime levels, but the model could only explain a small part of the differences between neighbourhoods. Once those social factors were included, predictive policing didn't seem to make much difference. The model also missed important things like community trust and how police behaved in each area, which limits how much we can learn from it.

Although the Random Forest model was highly accurate in predicting crime hotspots, its success mostly came from learning past crime patterns. This is a problem because past data may reflect biased policing—meaning some areas were watched more closely than others, even if crime wasn't higher there. As a result, the model might keep sending patrols to the same areas over and over, reinforcing unfairness.

When the model was tested for fairness, it performed similarly in areas with high and low minority populations, which at first seems fair. But we don't know what happened after the predictions were made. We don't have data showing whether police followed the predictions, or how communities were affected. Without this information, we can't fully know whether the model was fair in real life, even if it looked fair on paper.

The analysis used large neighbourhood areas to measure crime, but this hid important differences within those areas. When looking at the maps, it was clear that crime changed differently in different parts of the same neighbourhood. By averaging the data, small but important crime clusters were missed clusters that could have helped guide better policing decisions.

The micro-level hotspot classification relied on a threshold-based binary definition of risk (top 20%), which—despite robustness checks—remains a simplification. The models could not account for uncertainty or fluctuation in borderline areas. Moreover, the one-year lag feature, while useful for prediction, may have reduced model responsiveness to real-time or seasonal dynamics observed during temporal trend analysis.

The data was from an open source and there was underreporting and missing values which may have skewed the results.

5.6 FUTURE DIRECTIONS

Building on the current study's strengths and limitations, several areas for future research can be identified to improve both the theoretical understanding and operational effectiveness of predictive policing. First, studies could aim to incorporate more spatial and temporal data. Weekly or even daily crime reports would offer sharper resolution into how crime evolves in response to predictive interventions.

Second, future research would benefit from access to internal law enforcement data, including patrol logs, intervention records, and resource allocations tied to Predictive Policing outputs. This would allow researchers to assess not just where predictive models identify hotspots, but whether and how those predictions affect officer behaviour and crime outcomes. Without this link, evaluations of predictive tools remain incomplete.

Third, the model could become more useful if it included other important factors like housing, education, or local services. These can help explain why crime happens in some areas and could point to non-policing solutions—like community programs or housing support—to prevent crime in the long run.

Fourth, qualitative research methods—such as interviews with officers, analysts, and residents—could provide context to the patterns observed in the data. These narratives might provide better understanding of what happens in the ground.

Fifth, comparative case studies across multiple cities using standardised frameworks could help determine whether the patterns observed in Chicago could be used in other cities or another jurisdiction. This is crucial for distinguishing city-specific dynamics from broader trends in predictive policing effectiveness.

Lastly, research should explore the development and evaluation of hybrid models that integrate crime prediction with social program targeting. If Predictive Policing is to move beyond reactive enforcement, it must connect with efforts to prevent crime through economic support, mental health services, and community investment.

CHAPTER 6: CONCLUSION

This study critically examined the effectiveness of Predictive Policing through a comprehensive, multi-layered quantitative study using data from the City of Chicago. The central research question—"Does predictive policing work?"—was addressed using a rigorous empirical framework encompassing macro-level time-series analysis, meso-level neighbourhood regression, and micro-level machine learning classification. The results demonstrate that while Predictive Policing models offer strong utility in identifying high-risk areas, they do not produce significant reductions in overall crime as per the analysis on this study. This final chapter summarises the empirical contributions of the study, reflects on its theoretical and practical implications, and outlines its contribution to academic discourse and public policy.

When looking at crime data before and after predictive policing started, there was no clear change. Even when testing with a fake policy date, the results were the same. The seasonal patterns in crime stayed steady, which suggests that larger social trends—not the algorithm—were influencing crime levels. So, this part of the study found no strong evidence that predictive policing made a difference citywide.

The neighbourhood-level analysis showed that poverty and minority population levels are much better at explaining crime rates than Predictive Policing policies. Even though the policy had a small effect, most of the variation in crime wasn't explained by the model. This supports the idea that deeper social issues—not just past crime locations—need to be addressed if we want to reduce crime meaningfully.

The machine learning model was highly accurate at predicting crime hotspots. It worked well across different neighbourhood types and didn't show signs of demographic bias. However, it's important to remember that it can't explain why crime happens, stop it directly or predict future crime—it's just a tool to help target police efforts.

Importantly, the study's robustness checks confirmed the consistency and validity of results across different assumptions. These included varying hotspot thresholds, excluding pandemic years, running placebo ITS models, and comparing multiple classifiers. However, the study also identified limitations during analysis, such as reduced spatial granularity, lack of real-time patrol data, and the simplification of crime into binary hotspot designations. These constraints highlight the gap between predictive precision and real-world intervention outcomes.

In conclusion, the research finds that Predictive Policing, as implemented in Chicago, did not significantly have any crime reductions. It does not change the underlying drivers of crime, but it can help law enforcement better allocate resources to areas of elevated risk. The broader implication is that data-driven tools must be used with policies that address systemic inequities—such as housing, education, and public health—to achieve long-term safety outcomes. Predictive policing helps in planning, not prevention or prediction. It's a tool for forecasting and operational targeting, not a solution for reducing crime by itself.

This study contributes to the academic and policy discourse by offering a grounded, statistically robust, and critically reflective evaluation of Predictive Policing. The results suggest that people should be careful not to overpromise what predictive policing can achieve. The tools can help, but they're not magic solutions to crime. Future research should explore hybrid models that blend predictive analytics with social policy intervention, and should deepen the integration of qualitative insight, spatial dynamics, and real-time deployment data

REFERENCES

Agnew, R. (1992) 'FOUNDATION FOR A GENERAL STRAIN THEORY OF CRIME AND DELINQUENCY*', *Criminology*, 30(1), pp. 47–88. Available at: https://doi.org/10.1111/j.1745-9125.1992.tb01093.x.

Archer, M.S. (2007) *Making our Way through the World: Human Reflexivity and Social Mobility*. Cambridge: Cambridge University Press. Available at: https://doi.org/10.1017/CBO9780511618932.

Arora, P. (2016) 'Bottom of the Data Pyramid: Big Data and the Global South', *International Journal of Communication*, 10(0), p. 19.

Babbie, R. (2020) *The Practice of Social Research*. Cengage Learning (Mindtap Course List). Available at: https://books.google.ie/books?id=KrGeygEACAAJ.

Bakke, E. (no date) 'PREDICTIVE POLICING: THE ARGUMENT FOR PUBLIC TRANSPARENCY'.

Barabas, C. *et al.* (2018) 'Interventions over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment', in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. *Conference on Fairness, Accountability and Transparency*, PMLR, pp. 62–76. Available at: https://proceedings.mlr.press/v81/barabas18a.html (Accessed: 10 May 2025).

Barocas, S. and Selbst, A.D. (2016) 'Big data's disparate impact', *Calif. L. Rev.*, 104, p. 671.

Bayley, D.H. (1994) *Police for the Future*. Oxford University Press (Studies in crime and public policy). Available at: https://books.google.ie/books?id=m3U8DwAAQBAJ.

Bayley, D.H. and Shearing, C.D. (1996) 'The Future of Policing', *Law & Society Review*, 30(3), pp. 585–606. Available at: https://doi.org/10.2307/3054129.

Benjamin, R. (2019) *Race After Technology: Abolitionist Tools for the New Jim Code*. Wiley. Available at: https://books.google.ie/books?id=nPy9uwEACAAJ.

Berk, R. *et al.* (2009) 'Forecasting Murder Within a Population of Probationers and Parolees: A High Stakes Application of Statistical Learning', *Journal of the Royal Statistical Society Series A: Statistics in Society*, 172(1), pp. 191–211. Available at: https://doi.org/10.1111/j.1467-985X.2008.00556.x.

Bernal, J.L., Cummins, S. and Gasparrini, A. (2017) 'Interrupted time series regression for the evaluation of public health interventions: a tutorial', *International journal of epidemiology*, 46(1), pp. 348–355.

Bhaskar, R. (2013) *A Realist Theory of Science*. Taylor & Francis. Available at: https://books.google.ie/books?id=7318AgAAQBAJ.

Binns, R. (2018) 'Fairness in machine learning: Lessons from political philosophy', in. *Conference on fairness, accountability and transparency*, PMLR, pp. 149–159.

Boundaries - Community Areas - Map | City of Chicago | Data Portal (no date). Available at: https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Community-Areas-Map/cauq-8yn6 (Accessed: 10 March 2025).

Box, G.E. and Tiao, G.C. (1975) 'Intervention analysis with applications to economic and environmental problems', *Journal of the American Statistical association*, 70(349), pp. 70–79.

Braga, A.A. *et al.* (2019) ‘Hot spots policing and crime reduction: An update of an ongoing systematic review and meta-analysis’, *Journal of experimental criminology*, 15, pp. 289–311.

Brantingham, P.J. and Brantingham, P.L. (1984) *Patterns in crime*. Macmillan New York.

Brayne, S. (2021) *Predict and surveil: Data, discretion, and the future of policing*. Oxford University Press.

Breiman, L. (2001) ‘Random forests’, *Machine learning*, 45, pp. 5–32.

Caplan, J.M., Kennedy, L.W. and Miller, J. (2011) ‘Risk terrain modeling: Brokering criminological theory and GIS methods for crime forecasting’, *Justice quarterly*, 28(2), pp. 360–381.

Casey, J. (no date) *Seattle’s Predictive Policing Program*. Available at: https://datasmart.hks.harvard.edu/news/article/using-predictive-policing-to-reduce-crime-rate-189 (Accessed: 15 March 2025).

Chainey, S., Tompson, L. and Uhlig, S. (2008) ‘The Utility of Hotspot Mapping for Predicting Spatial Patterns of Crime’, *Security Journal*, 21(1), pp. 4–28. Available at: https://doi.org/10.1057/palgrave.sj.8350066.

Chandler, Roger.J. (2024) ‘(PDF) python-ags4: A Python library to read, write, and validate AGS4 geodata files’, *ResearchGate* [Preprint]. Available at: https://doi.org/10.21105/joss.04569.

Chawla, N.V. *et al.* (2002) ‘SMOTE: synthetic minority over-sampling technique’, *Journal of artificial intelligence research*, 16, pp. 321–357.

Chen, T. and Guestrin, C. (2016) ‘Xgboost: A scalable tree boosting system’, in. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794.

Chicco, D. and Jurman, G. (2020) ‘The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation’, *BMC Genomics*, 21(1), p. 6. Available at: https://doi.org/10.1186/s12864-019-6413-7.

Clarke, R.V. (1995) ‘Situational crime prevention’, *Crime and justice*, 19, pp. 91–150.

Cohen, L.E. and Felson, M. (2010) ‘Social change and crime rate trends: A routine activity approach (1979)’, in *Classics in environmental criminology*. Routledge, pp. 203–232.

Cook, T.D. and Campbell, D.T. (1979) *Quasi-experimentation: Design & Analysis Issues for Field Settings*. Rand McNally College Publishing Company. Available at: https://books.google.ie/books?id=jN59AAAAIAAJ.

Corbett-Davies, S. *et al.* (2017) ‘Algorithmic decision making and the cost of fairness’, in. *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pp. 797–806.

Cornish, D.B. and Clarke, R.V. (2014) *The Reasoning Criminal: Rational Choice Perspectives on Offending*. Transaction Publishers. Available at: https://books.google.ie/books?id=MPmmAgAAQBAJ.

Crimes - 2001 to Present | City of Chicago | Data Portal (no date). Available at: https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2/about_data (Accessed: 27 January 2025).

Denzin, N.K. and Lincoln, Y.S. (2017) *The SAGE Handbook of Qualitative Research*. SAGE Publications. Available at: https://books.google.ie/books?id=l35ZDwAAQBAJ.

Devlin, J. *et al.* (2019) 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding', in J. Burstein, C. Doran, and T. Solorio (eds) *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. *NAACL-HLT 2019*, Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. Available at: https://doi.org/10.18653/v1/N19-1423.

Egbert, S. and Leese, M. (2021) *Criminal futures: Predictive policing and everyday police work*. Taylor & Francis.

Eubanks, V. (2018) *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

Farrell, G. and Pease, K. (2001) *Repeat Victimization*. Criminal Justice Press (Crime prevention studies). Available at: https://books.google.ie/books?id=vE4uAQAAIAAJ.

Farrington, D.P. and Welsh, B.C. (2005) 'Randomized experiments in criminology: What have we learned in the last two decades?', *Journal of Experimental Criminology*, 1(1), pp. 9–38. Available at: https://doi.org/10.1007/s11292-004-6460-0.

Ferguson, Andrew G. (2017) 'Policing Predictive Policing', *Washington University Law Review*, 94(5). Available at: https://journals.library.wustl.edu/lawreview/article/id/3851/ (Accessed: 10 May 2025).

Ferguson, Andrew Guthrie (2017) *The Rise of Big Data Policing*. New York University Press (Surveillance, Race, and the Future of Law Enforcement). Available at: https://doi.org/10.18574/nyu/9781479854608.001.0001.

Freedman, D. (2009) *Statistical Models: Theory and Practice*. Cambridge University Press. Available at: https://books.google.ie/books?id=k0hJ9HpAhnkC.

Goldstein, H. (1990) *Problemoriented policing*. New York: McGraw-Hill.

Groff, E.R., Weisburd, D. and Yang, S.-M. (2010) 'Is it Important to Examine Crime Trends at a Local "Micro" Level?: A Longitudinal Analysis of Street to Street Variability in Crime Trajectories', *Journal of Quantitative Criminology*, 26(1), pp. 7–32. Available at: https://doi.org/10.1007/s10940-009-9081-y.

Haberman, C.P., Sorg, E.T. and Ratcliffe, J.H. (2017) 'Assessing the Validity of the Law of Crime Concentration Across Different Temporal Scales', *Journal of Quantitative Criminology*, 33(3), pp. 547–567. Available at: https://doi.org/10.1007/s10940-016-9327-4.

Haggerty, K.D. (2011) 'Reiner, Robert (2010).The Politics of the Police', *Policing: A Journal of Policy and Practice*, 5(2), pp. 110–111. Available at: https://doi.org/10.1093/police/par007.

Hair, J. *et al.* (2016) *Multivariate Data Analysis*. Pearson Education. Available at: https://books.google.ie/books?id=LKOSAgAAQBAJ.

Harcourt, B.E. (2019) *Against prediction: Profiling, policing, and punishing in an actuarial age*. University of Chicago Press.

Hosmer, D.W. and Lemeshow, S. (2000) *Applied Logistic Regression*. Wiley (A Wiley-Interscience publication). Available at: https://books.google.ie/books?id=zW9sngEACAAJ.

Hunt, P., Saunders, J. and Hollywood, J.S. (2014) *Evaluation of the Shreveport Predictive Policing Experiment*. Santa Monica, CA: RAND Corporation. Available at: https://www.rand.org/pubs/research_reports/RR531.html.

JOHNSON, S.D., BOWERS, K. and HIRSCHFIELD, A. (1997) 'NEW INSIGHTS INTO THE SPATIAL AND TEMPORAL DISTRIBUTION OF REPEAT VICTIMIZATION', *The British Journal of Criminology*, 37(2), pp. 224–241. Available at: https://doi.org/10.1093/oxfordjournals.bjc.a014156.

Kallus, N. *et al.* (2022) 'Doubly Robust Distributionally Robust Off-Policy Evaluation and Learning', in *Proceedings of the 39th International Conference on Machine Learning*. *International Conference on Machine Learning*, PMLR, pp. 10598–10632. Available at: https://proceedings.mlr.press/v162/kallus22a.html (Accessed: 11 May 2025).

Kallus, N., Mao, X. and Zhou, A. (2022) 'Assessing Algorithmic Fairness with Unobserved Protected Class Using Data Combination', *Management Science*, 68(3), pp. 1959–1981. Available at: https://doi.org/10.1287/mnsc.2020.3850.

Kelling, G.L. and Wilson, J.Q. (1982) 'Broken Windows', *The Atlantic*, 1 March. Available at: https://www.theatlantic.com/magazine/archive/1982/03/broken-windows/304465/ (Accessed: 11 May 2025).

LeCun, Y., Bengio, Y. and Hinton, G. (2015) 'Deep learning', *Nature*, 521(7553), pp. 436–444. Available at: https://doi.org/10.1038/nature14539.

Linden, A. (2015) 'Conducting Interrupted Time-series Analysis for Single- and Multiple-group Comparisons', *The Stata Journal*, 15(2), pp. 480–500. Available at: https://doi.org/10.1177/1536867X1501500208.

Lum, K. and Isaac, W. (2016) 'To Predict and Serve?', *Significance*, 13(5), pp. 14–19. Available at: https://doi.org/10.1111/j.1740-9713.2016.00960.x.

Lundberg, S.M. and Lee, S.-I. (2017) 'A Unified Approach to Interpreting Model Predictions', in *Advances in Neural Information Processing Systems*. Curran Associates, Inc. Available at: https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html (Accessed: 11 May 2025).

MacKinnon, J.G. and White, H. (1985) 'Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties', *Journal of Econometrics*, 29(3), pp. 305–325. Available at: https://doi.org/10.1016/0304-4076(85)90158-7.

Meijer, A. and and Wessels, M. (2019a) 'Predictive Policing: Review of Benefits and Drawbacks', *International Journal of Public Administration*, 42(12), pp. 1031–1039. Available at: https://doi.org/10.1080/01900692.2019.1575664.

Meijer, A. and and Wessels, M. (2019b) 'Predictive Policing: Review of Benefits and Drawbacks', *International Journal of Public Administration*, 42(12), pp. 1031–1039. Available at: https://doi.org/10.1080/01900692.2019.1575664.

Mohler, G.O. *et al.* (2015) 'Randomized Controlled Field Trials of Predictive Policing', *Journal of the American Statistical Association*, 110(512), pp. 1399–1411. Available at: https://doi.org/10.1080/01621459.2015.1077710.

Nix, J. *et al.* (2024) ‘When police pull back: Neighborhood-level effects of de-policing on violent and property crime, a research note’, *Criminology*, 62(1), pp. 156–171. Available at: https://doi.org/10.1111/1745-9125.12363.

PASQUALE, F. (2015) *The Black Box Society*. Harvard University Press. Available at: http://www.jstor.org/stable/j.ctt13x0hch (Accessed: 10 May 2025).

Perry, W.L. (2013) *Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations*. RAND Corporation (Bienestar y proteccion social). Available at: https://books.google.ie/books?id=ZdstAQAAQBAJ.

Powers, D.M.W. (2020) ‘Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation’. arXiv. Available at: https://doi.org/10.48550/arXiv.2010.16061.

Predictions Put Into Practice: a Quasi-experimental Evaluation of Chicago’s Predictive Policing Pilot | National Institute of Justice (no date). Available at: https://nij.ojp.gov/library/publications/predictions-put-practice-quasi-experimental-evaluation-chicagos-predictive (Accessed: 2 April 2025).

‘Protocolo-de-Bogota_eng.pdf’ (no date). Available at: https://insightcrime.org/wp-content/uploads/2025/02/Protocolo-de-Bogota_eng.pdf (Accessed: 11 May 2025).

Querbach, M., Krom, M. and Jongejan, A. (2020) ‘Review of State of the Art: Predictive Policing’.

Ratcliffe, J.H. (2004) ‘The Hotspot Matrix: A Framework for the Spatio-Temporal Targeting of Crime Reduction’, *Police Practice and Research*, 5(1), pp. 5–23. Available at: https://doi.org/10.1080/1561426042000191305.

Richardson, R., Schultz, J.M. and Crawford, K. (no date) ‘DIRTY DATA, BAD PREDICTIONS: HOW CIVIL RIGHTS VIOLATIONS IMPACT POLICE DATA, PREDICTIVE POLICING SYSTEMS, AND JUSTICE’, pp. 15–55.

Roland, C. and Kramer, K. (no date) ‘What’s New in ArcGIS Pro 3.2’, *ArcGIS Blog*. Available at: https://www.esri.com/arcgis-blog/products/arcgis-pro/announcements/whats-new-in-arcgis-pro-3-2 (Accessed: 15 January 2025).

Ruxton, G.D. (2006) ‘The unequal variance t-test is an underused alternative to Student’s t-test and the Mann–Whitney U test’, *Behavioral Ecology*, 17(4), pp. 688–690. Available at: https://doi.org/10.1093/beheco/ark016.

Saito, T. and Rehmsmeier, M. (2015) ‘The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets’, *PLOS ONE*, 10(3), p. e0118432. Available at: https://doi.org/10.1371/journal.pone.0118432.

Sampson, R.J. (2024) *Great American City: Chicago and the Enduring Neighborhood Effect*. University of Chicago Press. Available at: https://books.google.ie/books?id=O8P3EAAAQBAJ.

Sampson, R.J. and Groves, W.B. (1989) ‘Community Structure and Crime: Testing Social-Disorganization Theory’, *American Journal of Sociology*, 94(4), pp. 774–802.

Sayer, A. (2000) *Realism and Social Science*. SAGE Publications (Realism and Social Science). Available at: https://books.google.ie/books?id=4-5E2JD-ek4C.

Spielman, S.E. and Folch, D.C. (2015) 'Reducing Uncertainty in the American Community Survey through Data-Driven Regionalization', *PLOS ONE*, 10(2), p. e0115626. Available at: https://doi.org/10.1371/journal.pone.0115626.

'State of Policing: The Annual Assessment of Policing in England and Wales 2021' (no date) *His Majesty's Inspectorate of Constabulary and Fire & Rescue Services*. Available at: https://hmicfrs.justiceinspectorates.gov.uk/publication-html/state-of-policing-in-england-and-wales-2021/ (Accessed: 27 February 2025).

Wang, T. *et al.* (2013) 'Learning to Detect Patterns of Crime', in H. Blockeel et al. (eds) *Machine Learning and Knowledge Discovery in Databases*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 515–530.

Wasserstein, R.L. and and Lazar, N.A. (2016) 'The ASA Statement on p-Values: Context, Process, and Purpose', *The American Statistician*, 70(2), pp. 129–133. Available at: https://doi.org/10.1080/00031305.2016.1154108.

Weisburd, D. *et al.* (2003) 'Reforming to Preserve: Compstat and Strategic Problem Solving in American Policing', *Criminology & Public Policy*, 2(3), pp. 421–456. Available at: https://doi.org/10.1111/j.1745-9133.2003.tb00006.x.

WEISBURD, D. (2015) 'THE LAW OF CRIME CONCENTRATION AND THE CRIMINOLOGY OF PLACE', *Criminology*, 53(2), pp. 133–157. Available at: https://doi.org/10.1111/1745-9125.12070.

Weisburd, D., Willis, J.J. and Mastrofski, S.D. (2004) *Compstat in practice: an in-depth analysis of three cities*. Washington, DC: Police Foundation.

WELCH, B.L. (1947) 'THE GENERALIZATION OF "STUDENT'S" PROBLEM WHEN SEVERAL DIFFERENT POPULATION VARLANCES ARE INVOLVED', *Biometrika*, 34(1–2), pp. 28–35. Available at: https://doi.org/10.1093/biomet/34.1-2.28.

Wooldridge, J.M. (2019) *Introductory Econometrics: A Modern Approach*. Cengage Learning. Available at: https://books.google.ie/books?id=0p9_tQEACAAJ.

Zuboff, S. (2019) *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs. Available at: https://books.google.ie/books?id=MJ5ADwAAQBAJ.

APPENDIX

Collecting the Data

- Crime incidents – We download every Chicago crime record back to 2001 straight from the city's open-data website.
- Community maps – Chicago draws 77 community-area borders. We grab that map so we can group crimes by neighbourhood.
- Census facts – From U.S. Census tables we pull basics like population, income, and unemployment for each neighbourhood, year by year.

All three sources are stored inside data/ exactly as we received them.

Cleaning and Checking

- Crime rows missing a neighbourhood code or coming after 2022 are removed
- The neighbourhood IDs in the map come in as text; we turn them into whole numbers so the computer can match them to census and crime tables.
- Quick "sanity check": confirm the map really contains all 77 IDs-19 were missing hence to show which 19 are missing has been highlighted in the code

Result: three tidy tables that talk to each other because they share the same area_id column.

Building the Analysis File

- For each neighbourhood and each year, count how many crimes happened.
- Divide that count by the local population and multiply by 1 000. That gives a crime rate (crimes per 1 000 residents) so busy and quiet areas are comparable.
- Saving this master table as CrimeRates_by_Area_2001_2022.csv. It is the backbone for every later chart and model.

Time-Series Check: Did Crime Fall After Predictive Policing Started?

- Converting daily crime stream into a monthly one so pattern are easier to spot.
- Marking August 2012—the month Chicago rolled out its first predictive-policing pilot—as the "intervention".
- A simple before-and-after regression asks: did the slope or level of the citywide crime line change at that point?

The citywide trend keeps drifting downward at almost the same speed. No sharp drop shows up right after the software launch.

Place-Based Analysis: Why Are Some Neighbourhoods Hotter Than Others?

- Visualising the crime-rate map for 2012 and 2022 to see where hot spots sit.
- A basic regression then tests whether neighbourhoods with higher poverty or larger minority populations also have higher crime rates, after accounting for the software era.
- A different “spatial” version was tested that considers spill-over from next-door areas; the term is tiny and not significant, so the simpler model stands.

Machine-Learning Experiment: Predicting Next Year’s Hot Spots

- Labelling a neighbourhood a hot spot when its crime count sits in the top 20 % for that year.
- Training a model (Logistic Regression, Random Forest, XGBoost) that looks at last year’s crime, poverty, income, and offense mix to guess whether each area will be a hot spot next year.
- Because hot spots are rare, use of an oversampling trick (SMOTE) so the model sees enough examples of both classes during training.
- Accuracy is measured with the F1 score for all three models. Results are split by race share to check fairness. The model is decent but still slightly better in low-minority areas.

Robustness Checks

- Change the hot-spot cutoff (top 25 % or 15 %) → accuracy hardly moves.
- Drop the Covid-19 years (2020–2021) → key regression slopes barely change.
- Compare Chicago to Milwaukee with a synthetic-control method → still no clear software effect.