ABSTRACT

Title of Dissertation: FAIR URBAN CRIME PREDICTION WITH HUMAN MOBILITY BIG DATA

Jiahui Wu, Doctor of Philosophy, 2021

Dissertation directed by: Associate Professor Vanessa Frias-Martinez
College of Information Studies and UMIACS

Crime imposes significant costs on society. Reported crime data is important in quantifying the severity of crimes, based on which decision-makers would allocate resources for crime interventions. Human mobility big data has triggered the interest in various fields to study the relationship between urban crimes and mobility at a large scale, especially the predictive power of mobility for urban crimes. This research direction can enrich our understanding of crimes and better inform crime-related decision-making. One concern about reported crime data is the bias issue. The bias could be produced by different levels of residents' willingness to report potential crime incidents and police activity in neighborhoods. While lots of studies about crime prediction are aware of biases in reported crimes, few of them propose solutions to address or mitigate this issue or to evaluate how this issue would affect prediction models in terms of accuracy or fairness.

My dissertation research aims to explore the potential of human mobility big data for crime prediction. Specifically, my dissertation will advance the state-of-the-art by addressing three challenges in mobility-based crime prediction: 1) Constructing mobility features might be sensitive to different methodological choices. Without careful examination of these choices, there might be conflicting findings. One critical area of mobility analysis to predict crime is the identification of urban hotspots. Therefore, my work performs a systematic spatial sensitivity analysis on the impact of these choices and provides guidelines to identify the most stable ones. 2) Under-reporting generates biases in reported crime data. To address such bias, I develop a Bayesian model for long-term crime prediction that infers the unobserved true number of crime incidents. Comprehensive experiments show how the accuracy and fairness of long-term crime prediction would be affected by modeling the under-reporting of crimes. 3) Although empirical studies show promising results about the relationship between human mobility and long-term crime prediction, the effects of mobility features on short-term crime prediction have yet to be explored. Therefore, my work conducts a series of experiments to explore how incorporating mobility features into short-term crime prediction models affects their performance in terms of accuracy and fairness.

FAIR URBAN CRIME PREDICTION WITH HUMAN MOBILITY BIG DATA

by

Jiahui Wu

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2021

Advisory Committee:
Professor Vanessa Frias-Martinez, Chair
Professor Wei Ai
Professor Richard Marciano
Professor Taylor M. Oshan
Professor Kathleen Stewart

© Copyright by
Jiahui Wu
2021

To my family

Acknowledgments

My pursuit of the doctoral degree at the UMD iSchool has been full of joy, love and exploration. I am really grateful for all the people who have helped, supported and encouraged me.

First of all, I would like to thank Dr. Vanessa Frias-Martinez, who has been a great advisor, an inspiring mentor and a good friend. Over the past five years, I have been working with her on many interesting projects and have learned so much about research, data analytics and writing. Her guidance helped set me on the right course for the graduate life. Every time I have doubts and concerns, she is there for me. Without her support and encouragement, this dissertation would have been impossible. I hope one day I could become a good advisor like her.

I would also like to thank my committee: Dr. Kathleen Stewart, Dr. Richard Marciano, Dr. Wei Ai and Dr. Taylor M. Oshan for their time reviewing the manuscript. Thanks for their suggestions, guidance and support that helped me complete this work.

As my first time ever to study abroad, I have received a lot of help from my colleagues and friends. I am thankful for Jiqun Liu and Yufei Shen, who have talked me into pursuing the PhD degree. This journey has been more fulfilling than I ever imagined. I would like to thank Lingzi Hong, who has helped me adapted to the

life in the US smoothly. She also offered me the opportunity to work with her so that I could fit in the academic life easily. I really appreciate the fun, intense and inspiring discussions with Priya Kumar and Yuting Liao. I also learnt a lot through the chats with Myeong Lee, Joohee Choi and Jonathan Brier.

I send my love to my family who has always been proud of me, and has brought great strength for me to embrace challenges and keep fighting for my goal. I would like to thank my father, my mother, my father-in-law, my mother-in-law, my sister and my brother to accompany me through my graduate life. I also thank the family of my uncle, with whom I spent every Thanksgiving in the US.

And special thanks to my wife, Wenting Cheng. We have spent most of the past five years as a long-distance couple. It has been challenging but I am very grateful that we are going to see it through. It is your love, patience and understanding that support me to explore myself and find the most valuable things in my life.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Motivation

The popularization of mobile phones and the emergence of other information communication technologies (ICTs) has triggered research interests in urban computing from the perspective of human mobility. Human mobility data, such as Call Detail Records (CDRs), GPS logs and geotagged social media data, have shown high potentials in understanding many aspects of urban life and dynamics, such as analyzing daily patterns to understand the *pulse* of a city [3, 4]; investigating the correlation between human mobility patterns and land-use patterns as well as urban functions [5, 6] and assisting public emergency response during disasters [7]. Increasing attention has been paid to studying the relationship between human mobility and crimes. As suggested by the *routine activity theory* [8], crime is a complex social phenomenon emerging from the interaction of people and the surrounding environment. For example, there is evidence of a superlinear relationship between the presence of people and the number of property crimes, indicating that a disproportional number of property crimes occurs in regions where an increased flow of people takes place in the city [9].

Crime imposes high costs to society at the individual, community, and national

levels. In 2020, the US saw a significant crime rise across major cities[1]. A large body of literature from the fields such as criminology, geographic information science, urban planning and data science has been dedicated to understanding factors that cause criminal actions and, more importantly, measures to prevent crimes from happening. For example, environmental criminology [10] and crime opportunity theory [11] emphasize the importance of environmental factors in criminal actions and study the spatiotemporal patterns in crime incidents.

Countless reports, academic papers, books, news articles, social media discussions, and other materials about crimes rely on reported crime data [12, 13]. Reported crime data can be used to evaluate the efficiency of police force [14], predicting crime hotspots for patrol routes planning [15] and evaluate the effects of crime-related policies. Reported crime data is also the basis for crime prediction, which helps allocate resources more efficiently to prevent future crimes. There are individual-based crime prediction, such as the identification of offenders or victims, and place-based crime prediction, where places can be grids, census tracts, or large regions such as municipalities or cities, to predict the occurrence or volumes of crimes in a given place.

In this dissertation study, crimes mostly refer to traditional violent and property crimes, such as assaults and burglary, which are *direct-contact predatory violations* where the offenders and targets are in physical proximity when the violations happen [8], as opposed to modern crimes such as cybercrimes where the offenders are not spatially in contact with the victims. Crime prediction refers to place-based

[1]https://www.cnn.com/2021/04/03/us/us-crime-rate-rise-2020/index.html

crime prediction based on reported crime data. Depending on the time horizon of interest, place-based crime prediction can be distinguished as long-term crime prediction, *e.g.*, predicting the volumes of crimes in the next month or next year; and as short-term crime prediction, *e.g.*, predicting the occurrence of crimes in the next day or next week. Long-term crime prediction analysis allows us to understand how the environmental factors of places shape future crimes; and in turn, help us inform better urban planning that improves the urban environment potentially decreasing crime occurrence. On the other hand, short-term crime prediction analysis is generally used to better allocate policing resources to response to crimes more swiftly in a short period of time.

Traditionally, crime prediction utilizes the crimes' inherent spatiotemporal patterns [12, 16] and environmental factors such as spatial contexts, *e.g.*, land use and point-of-interest (POI) [17], and social contexts, *e.g.*, population diversity [18]. Recently, more and more studies leverage the relationship between human mobility and crimes. The general approach of mobility-based crime prediction is to first extract mobility features from human mobility data. These mobility features, such as number of identified visitors and footfall, *i.e.*, the number of people present in a given area and a time interval, can been considered as inputs for machine learning models to predict crimes. Various studies have shown that mobility features can be utilized along with other socioeconomic features to enhance long-term crime prediction in small spatial units such as grids and census tracts [18–21]. Although human mobility shows promising results in long-term crime prediction, at least three challenges have been identified in this dissertation research.

The first challenge is that the construction of mobility-based features might be sensitive to different methodological choices. For spatial data, there is a well-known issue called *modifiable areal unit problem* (MAUP) [22], *i.e.,* the areal units (zonal objects) used in many geographical studies are arbitrary and modifiable and could introduce statistical bias in the feature extracted. Due the MAUP issue, the extracted mobility-based features could be statistically biased and different, and therefore, different areal units of interest could generate conflicting findings about the descriptive relationship between human mobility and crimes or the predictive power of mobility for crime prediction. In addition to MAUP, differences in other methodological choices such as the delimitation of cities or urban areas and the specific definition of mobility features can also cause the findings such as the relationship between mobility and crimes to be less robust.

The second challenge is about the under-reporting issue in the reported crime data. As computational models are built upon reported crimes and have an increasing influence on future resource allocation, the bias encoded in reported crime data becomes more critical. One source of such bias is the under-reporting of crimes, where a crime has occurred but is not reported to the police, or the crime is reported but the police mark the incident as "no crime". Various factors contribute to the under-reporting of crimes, such as the crime being "too trivial/no loss [23] and lack of faith in the authorities [24]. The presence of female officers has also been shown to increase the willingness to report violent crimes, especially related to domestic violence [25]. As a result, reported crime data does not merely reflect the severity of crimes but also the community-police relationship and the bias towards

certain socially disadvantaged populations [10]. Without considering the potential bias of reported crime data, the crime prediction systems might relay the bias and provide unfair outcomes.

The third challenge is that the effects of human mobility data on the short-term crime prediction models are unclear. Traditional short-term crime prediction studies only utilize historical crimes and urban environment factors such as POI and focus solely on the accuracy of prediction. Although empirical studies have shown there exists a relationship between mobility-based features and long-term crime incidents, there exists a knowledge gap for the effectiveness of mobility features for short-term crime prediction in terms of accuracy and fairness. Studies have shown that short-term crime predictive systems could be reinforcing the data bias in the crime incident data [26]. With the increasing application of predictive models in policing, it is vital to understand not just accuracy but also the fairness of the prediction outcomes.

These challenges would impose obstacles on the mobility-based crime prediction research and applications in terms of *robustness*, as the extracted mobility features could be sensitive to and dependent on the underlying methodology choice; and *fairness*, as the under-reporting issue could generate and hide biases inside the reported crime data and lead to unfair crime prediction, both in the long-term and short-term.

1.2 Research Objectives and Questions

This study aims to improve the robustness, accuracy and fairness, and applicability of crime prediction using human mobility big data by addressing the three challenges identified above. Specifically, I will conduct three empirical studies to achieve the following three primary research objectives and address the corresponding research questions.

Research Objective 1*: Evaluate the sensitivity of human mobility features to different methodological choices for constructing such features.*

Urban spatial data are often heterogeneous, *e.g.*, these data are collected from different sources and have different spatial resolutions. There are different methodological choices to resolve the heterogeneity among different sources of urban spatial data. For example, one could extract human mobility features for grid-level spatial analysis or for Voronoi polygons. However, it is unclear how these methodological choices would impact the extracted mobility features and the subsequent analysis. In this study, I focus on one commonly used type of human mobility data, Call Detail Records, and a set of urban hotspots features that can characterize the spatial structure of cities. Although mobility-based hotspot analyses are extensive, there is no consensus among researchers about the process followed in computing them. Through an extensive literature review, I identify four kinds of methodological choices in constructing these mobility features: the definition of *city boundaries*, the *spatial units* of interest, the *interpolation method* to distribute individuals associated to a given cellular tower across grids or census tracts, and the *hotspot variables*

used to measure and characterize the computed hotspots. I quantify sensitivity as a stability index, and through the comprehensive spatial sensitivity analysis, I will answer the following research questions:

1A): Are mobility-based urban hotspot features sensitive to the methodology choices that construct such features?

1B): If sensitive, is there any combination of choices that provide the most stable result among all possible choices?

Research Objective 2*: Propose a model that can address the under-reporting issue in reported crime data and improve the fairness in mobility-based long-term crime prediction.*

Traditionally, reported crime data is considered as ground truth to train machine learning models for crime prediction. This approach ignores the under-reporting issue in reported crime data, which is a major concern about data bias in the reported crime data. Although the determinants for under-reporting behavior have been widely studied, none of the existing crime prediction models includes these insights into the modeling to address the under-reporting issue.

In this study, I focus on long term municipality-level crime prediction, i.e., the dependent variable is the annual number of crimes in a given municipality. The *reported crime data* is not assumed to be ground truth for the crime prediction model. Instead, it is considered as observed variables under a Bayesian framework. Two hidden variables are proposed to model the under-reporting process: 1) the *true crimes*, *i.e.*, all crimes that have happened whether or not they have been recorded in the reported crime data; 2) the *reporting rate*, *i.e.*, ratio of the number

of reported crimes to the number of true crimes. Existing literature has shown various factors that influence the under-reporting behavior. These studies provide domain knowledge about the under-reporting issue and can be used to model the under-reporting process. The mobility component of this model is to model the true crime generation process with mobility-based features. By explicitly modeling the true crime generation and the under-reporting process, this study will answer the following research questions:

2A): Does modeling the under-reporting process improve or hurt the performance of predicting the number of reported crimes?

2B): Does modeling the under-reporting process improve the fairness of crime prediction?

2C): What influence do the mobility-based features have on the true crime generating process?

2D): What influence do the determinants have on the reporting rate, *i.e.*, the under-reporting process?

Research Objective 3*: A comprehensive analysis on the effects of incorporating mobility features in short-term crime prediction in terms of accuracy and fairness.*

Existing empirical studies about the relationship between human mobility and crimes have been conducted in the context of long-term crime estimation or prediction analysis and no studies have incorporated mobility features in the short-term crime prediction. Also, studies proposing new short-term crime prediction focus solely on improving accuracy without taking fairness into account. As predictive

models become pervasive in daily policing decisions, I argue that it is important to evaluate new short-term crime prediction models in terms of both accuracy and fairness.

In this study, I focus on short-term census-tract-level crime prediction, *i.e.*, the dependent variable is whether there will be any crime incident occurring in a given census tract in the next day. Based on a publicly available human mobility dataset in the US, I will extract various mobility variables and evaluate the effects of incorporating mobility features on various state-of-the-art deep learning short-term crime prediction models. I consider three combinations of input (predictive) features for the crime prediction models: 1) historical crimes only, which serves as the baseline; 2) mobility features only; and 3) historical crimes and mobility features. In order to obtain robust results, the experiments and evaluation are conducted in four U.S. cities with diverse demographics, *i.e.*, Baltimore, Minneapolis, Chicago, and Austin, for two types of crimes, *i.e.*. property and violent crimes. The proposed approach requires large datasets with mobility data, which less-developed cities might not have access to. In fact, certain cities might not have the infrastructure or the mobile services might be relatively new, making the mobility data collected insufficient to properly train the mobility-based crime prediction model. Therefore, I also conduct a preliminary study on cross-city transfer learning to explore the effects of knowledge from data-abundant cities. The following research questions will be addressed in this study:

3A): Does incorporating mobility features improve or hurt the accuracy of short-term crime prediction?

3B): Does incorporating mobility features improve or hurt the fairness of short-term crime prediction?

3C): If incorporating mobility features makes short-term crime prediction less fair, what are the potential factors that contribute to the less fair prediction?

3D): How can the under-reporting process be modelled into short-term crime predictors and how will the accuracy and fairness be affected by it?

3E): Can transfer learning techniques help cities with limited mobility data improve mobility-based crime prediction by leveraging knowledge from cities with abundant mobility data?

1.3 Organization of the Dissertation

The organization of this proposal is as follows. Chapter 2 reviews related literature on human mobility, reported crime prediction and the under-reporting issue, and algorithmic fairness in the context of crime prediction. Chapter 3 introduces the framework and the results for the spatial sensitivity analysis on mobility-based urban hotspot features. Chapter 4 focuses on modeling the under-reporting issue to improve the fairness of mobility-based long-term crime prediction. Chapter 5 focuses on the effects of human mobility features on short-term crime prediction in terms of accuracy and fairness. Finally, conclusions and future direction of this dissertation are described in Chapter 6.

Chapter 2: Literature Review

2.1 Human Mobility Data in Urban Environments

Urban density and human mobility are essential to understand and characterize urban environments. For example, the population and population density are the foundations to define urban versus rural areas and measure the scale of urbanization, while the commuting flows are the key to classify metropolitan versus non-metropolitan [27]. Job density rather than residential population density is used to study the formation and functions of sub-centers in cities [28]. Traditionally, urban density is estimated by either census residential population density [29] or job/employment density, which is the population density during work hours [28]. Commuting flows and travel behavior can be collected through travel surveys [30]. These estimations are costly to collect and can only represent the snapshots of population density either during the night when people return to their homes or during work hours when people are at work.

The recent availability of pervasive technologies has triggered new ways of studying cities using the large scale human mobility data generated from mobile phones and other wearable devices as well as transportation systems with GPS sensors. With these technologies, we can collect human mobility data with low cost

and measure real-time fine-grained urban density and population dynamics in cities. The pervasive human mobility data has opened up many opportunities in a rising research field called *urban computing* [31].

On the individual level, various studies focus on modeling individual mobility behaviors and patterns. The most notable characteristics about individual mobility are the regularity and predictability. González et al.found that human trajectories show a high degree of temporal and spatial regularity, which is due the fact that people tend to spend most of their time in a limited number of places [32]. The regularity can be quantified as mobility entropy, where a low entropy means the individual's mobility is regular and vice versa [33]. Because most people's mobility is regular, various studies confirm their movements are also highly predictable [33, 34], and develop many next-location prediction methods [35, 36]. With human trajectories, we can also quantify the range of individuals' mobility by indices such as total mobility distance and radius of gyration [37]; identify individuals' meaningful location such as workplace and home locations [38].

On the city level, human mobility data provides a real-time sensing of the population's movement and gathering. One of the very basic applications is to estimate the density of population in the different regions covered by the dataset in fine-grained time resolution. The density of population is also called *footfall* and have been utilized in various studies [2, 9, 19, 39]. The estimation of population density provides a timely complement to the census survey which often takes places once every few years, especially for developing countries. In additional to population density estimation, studies have shown that mobility data such as CDR can further

model and evaluate the socio-economic characteristics of different regions, such as poverty and wealth [40, 41], and socio-economic development levels [42–44].

With the high spatio-temporal resolution, mobility data can also be used to study the built-in environment of cities by leveraging the dynamics of population movement. For example, the identified urban hotspots can be used to evaluate the dynamic spatial structure of cities, such as urban sprawl and compactness [2]. Urban spatial structure is an important topic in quantitative geography and urban economics as it is related to the transportation system in terms of energy consumption or air pollution [45, 46], and economic growth performance [47, 48]. Discovering functional regions or land use patterns is another important topic in applying mobility data for cities structure analysis. Ríos, Sebastián and Muñoz applied Latent Dirichlet Allocation (LDA) to mobile phone data to detect four land use patterns, office areas, residential areas, leisure-commerce pattern, and rush hour pattern [49]. Yuan et al proposed a topic modeling framework (titled DRoF) using point-of-interests data and GPS trajectories and discovered 9 types of functional region clusters. Geo-tagged social media can also be used to annotate functional areas in the city [50, 51].

Urban transportation systems can also benefit from the pervasive human mobility data. Various studies have been working on forecasting traffic in the city using mobile phone data, taxi trajectories, bikeshare usage, and so on [52–54]. The real-time mobility data can enhance government's response for emergency events, such as enhancing communication between the public and the local government during snowstorms [7], understanding population displacement due to earthquake [55], and

measuring inequality in community resilience to hurricanes and flooding events [56].

2.2 Reported Crimes Prediction

Crimes do not occur randomly. Environmental criminology has revealed various spatiotemporal patterns in different types of crimes [10]. Crimes are highly concentrated in space and cluster at a range of spatial scales. For example, At least half of the crimes took place in only approximately 5% of street segments in several cities [57]. Over short time ranges, near repeat victimization has been observed in different types of crimes over the world [58], *i.e.*, when a crime incident occurs at one location, there is a temporary increase in the probability that other crime incidents will occur nearby. Over long periods, the concentration of crimes is also stable. Based on the 14 years (1989-2002) of crime reports in Seattle, Weisburd et al. shows that the vast majority of street segments showed a remarkably stable pattern of crime [59]. That crimes stably cluster in both space and time is the basis of crime prediction using historical crimes. In the early efforts of crime prediction, Geographical Information System (GIS) enabled the generation of crime maps that assigned predictive risk scores to places, using techniques such as kernel density estimation based on historical crimes [16, 60]. A more sophisticated way to model spatiotemporal clustering is the self-exciting point process by making an analogy between near repeat victimization pattern and earthquake aftershocks [12, 61]. The advantage of this model is that it can be extended to model the relations between different types of crimes, such as gun crimes can provide information about risks

of future homicides [12]. Zhao and Tang observed that for a given region, with the increase of differences between two time-slots, the crime difference tends to increase [17]. Therefore they proposed an intra-region correlation to capture this temporal pattern of crimes in their prediction model. Neural networks have also been applied to modeling spatiotemporal patterns in historical crimes for future crime prediction [13, 62, 63]. In the neural networks, the spatial patterns of crimes are modeled by convolution layers, while the temporal patterns can either be modeled as multiple feature maps in the convolution layers [62] or modeled by the recurrent neural network layers such as LSTM [63].

Besides the spatiotemporal patterns within historical crimes themselves, the relationship between crimes and the environment is also very important in understanding crimes. The built environment characteristics of the neighborhoods, such as land use or point-of-interest, have a strong connection with the spatial distribution of crimes. For example, bars and liquor stores are associated with high levels of crime risks [64, 65], and spatial distribution of property crime is positively associated with non-residential land uses such as commercial uses and public transit stations [57]. Seasonality is also observed in crimes. Parks are more positively associated with property crime during spring and summer seasons [66]. Weather and climate change also influence the fluctuation of the number of crimes [66, 67]. In light of the connection between crimes and the environmental context, various contextual features, such as point-of-interest (POI), land use, meteorological data, 311 calls for service data, and census data, is used as to enhance the prediction accuracy [17, 18, 20].

In addition to the inherent spatio-temporal patterns of crime incidents, there are various theories about the relationship between human mobility and crime incidents; and Browning *et al.* provide a systematic review for the theoretical foundations at the intersection of place, neighborhood, crimes and human mobility. For example, the *routine activities* theory puts an emphasis on mobility and micro-places characteristics; the *social disorganization* theory has an implicit focus on mobility through the lens of neighborhood-level social interaction [68]; For example, the *routine activities theory* puts an emphasis on mobility and micro-places characteristics and argues that each successful violation involves at least one offender and at least one personal or property target and also requires the absence of any effective guardian capable of preventing its occurrence [8]; the *social disorganization theory* has an implicit focus on mobility through the lens of neighborhood-level social interaction [68]; while the *Opportunity makes the Thief* theory claims that the opportunity is the cause of crime [69] *i.e.,* the higher the presence of suitable *targets* such as people and property, the more crimes could happen.

Although various theories are suggesting the relationship between human mobility and crimes, only until recently empirical studies about what role and predictive power mobility might have on crimes at large scale and finer granularity become possible due to the increasing availability of human mobility data in the urban environment. Mobile phone data, geolocated social media, taxi pick-up/drop-off, and check-ins have been leverage to construct mobility features in order to predict crime incidents. One of the most common mobility feature used is *footfall*, defined as the number of individuals present in a given area at a given time span. Bogomolov *et al.*

estimate footfall and population diversity such as gender and age from mobile phone data and predict whether a regular grid cell will have a high or low level of crimes in the following month [19]. Caminha *et al.* showed that increased footfall in a particular area of the city was proportional to the increasing rate of property crimes happening in the region [9] Kadar and Pletikosa extracted footfall from check-ins, subway, and taxi data, along with other census and POI features, to predict the number of crimes for a given census tract using tree-based machine learning models [18]. Visitors for a given community can be identified using individual trajectories. Felson and Boivin show a strong correlation between visitors variables and crimes in all the census tracts studied [70]. De Nadai *et al.* proposed a spatially filtered Bayesian Negative Binomial model to study how social, built environment and footfall influence criminal activity [20]. In addition to footfall, Kadar *et al.* found out that the quantity of pass-through human flows are also helpful in modeling annual hourly profiles of crimes of census tracts [21].

In this dissertation study, I will explore the predictive power of urban hotspot features, which are extracted from mobility patterns, to characterize urban spatial structure for long-term crime prediction. Also, current mobility-based crime prediction models are based on traditional machine learning models such as random forest in the context of long-term crime prediction. It remains unknown how the mobility features could affect the short-term crime prediction. To fill in this gap, I will develop neural network models to exploit the large scale human mobility data to enhance mobility-based short-term crime prediction.

2.3 Under-reporting of Reported Crimes

Although reported crime statistics are widely used, the debate over the validity of reported crime statistics is almost as old as the reported crime statistics themselves [71]. However, all the crime prediction models mentioned in the above section treat reported crimes as ground truth and do not consider the under-reporting issue. Concerns about under-reporting in crime data are highly related to the production of the reports themselves where exists the systematic reporting bias. Although crime reporting systems around the world vary a lot, in a simplified way, before being a record in the reported crime database, a crime incident goes through two main phases: 1) being reported to the police, either by an individual or by the police officials on scenes; 2) being recorded to the database by the police.

In the first phase, crimes are generally reported by victims or witnesses, who report around 80% of the crimes, 6% by the police on scenes, and the rest by offenders, alarm systems, officials other than police [10, 72]. There are various reasons why the public might choose not to report a crime. The crime being "too trivial/no loss" used to be the most important reason, but recently "Police could do nothing" has come on top [23]. "Lack of faith in authorities" is another reason for under-reporting that highlights the importance of community-police relationships [24]. Increased presence of female officers can significantly increase the willingness to report violent crimes, especially domestic violence [25].

In the second phase, after incidents are reported, the police decide whether or not to record incidents as crimes in the database. Bottomley and Coleman showed

the percentage of incidents written off as "no crime" may be as high as 10 percent. Various factors can influence the police's decision, such as the perceived seriousness of the case, its detectability, and the officer's desire to minimize paperwork [73]. As a result, under-reporting in crimes is heavily impacted by social disparities. For example, in Kensington, middle-class crime complaints are more likely to be reported and accepted by the police (*i.e.,* high reporting rate and high recorded rate), while the reports from white working-class tend to be rejected (low recorded rate) and racially-mixed communities are less willing to report (low reporting rate) [10].

There are ongoing discussions about how to improve the efficiency of crime reporting systems, such as using digital technology to enhance anonymity [74], applying predictive policing to enable police officers to detect more crimes at scenes [15], and advocating for more timely national incident-based reporting system [75]. As we wait for these systems to be developed, it is critical to address the existing bias in reported crime data, so that crime predictions are fair across social groups. Although the reporting and recording of crime incidents are two different phases, in this study, I make no distinction between them as it is almost impossible to obtain such information from the local police force. Instead, we simplify and quantify the under-reporting issue of crimes as the reporting rate, which is the ratio of the number of reported crimes in the police database to the number of (unobserved) true crimes that have occurred. This simplification is common in the literature [25, 72].

The severity of under-reporting of crimes is often quantified by reporting rate, which is estimated by victimization surveys. These surveys, such as National Crime Victimization Survey in the U.S. [76], contain questions about whether respondents

experienced crimes in the past few months, details about when and where these crimes happened, and whether they called the police about those crimes. There are also a few studies that match call for service data, *i.e.*, 911 calls in the U.S., with reported crimes to estimated the reporting rate [77, 78]. The reporting rate for different types of crime in different countries varies. The average of reporting rate for property crimes in 2007 is around 50% in UK, 2007 [23] and 36% in US, 2000 [72]. While the reporting rate of all crimes in Mexico in 2010 is only 12% [79]. Studies have shown that the poverty rate [80] and unemployment rate [81] could decrease the likelihood of property crime incidents, such as burglaries, being reported. For violent crimes, on the other hand, gender, age, and marital status of the victims [72] as well as the percentage of female-headed households with children, poverty, and foreign-born population of census tracts [78] are shown to influence reporting behavior.

Instead of victimization surveys and call for service data, in this dissertation study I will develop mobility-based under-reporting-aware models, *i.e.*, modeling the true crime generation and under-reporting process to provide quantitative insights about under-reporting issue from the perspective of crime prediction.

2.4 Algorithmic Fairness in Crime Prediction

As there are systematic biases in the reported crime data, there is a risk of feedback loops in reported crime prediction if the systematic biases are not addressed and future resource allocation decisions are based on these prediction models. Feed-

back loops are situations where the trained machine learning model informs decisions that then affect the data collected for future iterations of the training process [82]. Therefore, it is necessary to evaluate and improve the fairness of the crime predictions.

The concern about biases in crime prediction based on reported crimes corresponds with the increasingly popular research topic, algorithmic fairness, due to the emergence of computational algorithms making decisions with high societal impact such as loan requests, crime prediction, and criminal sentencing. Algorithmic fairness, especially the most commonly used notion of group or statistical fairness, is based on the notion of protected or sensitive attributes, such as gender and race (minority and non-minority). A protected attribute usually represents a population sub-group that has historically suffered from discrimination, and therefore some form of (approximate) parity or non-discrimination regulation in the predictive algorithm is desired for these protected groups [82, 83].

Fairness is a complex concept, and there are different and sometimes conflicting definitions, and thus, a variety of fairness metrics. Verman and Rubin summarize three categories, five sub-categories, and 20 fairness metrics. Through an example experiment of credit score classification (whether a loan applicant has a good or bad credit score), they show the conflicts among these metrics, *i.e.*, the same predictions are considered as fair in some metrics but unfair in other metrics [84]. Adding fairness constraints or regularization in the modeling process is the most common and generic approach to improve fairness [85–87]. Although the definitions of fairness vary, it has been empirically shown that there is usually a trade-off between the

accuracy and fairness of prediction, *i.e.*, improvement in fairness is generally at the expense of the algorithmic accuracy [87–89].

In contrast with the large number of studies on building more accurate crime prediction models, only a few papers are focusing on the fairness or bias issue in place-based crime prediction. These papers focus on the potential feedback loop in predictive policing. As the increasing usage of predictive policing tools in the police system, Lum and Isaac conducted the seminal study analyzing the possible bias in the *PredPol*'s prediction, a crime prediction platform [26]. Through a simulation experiment, Lum and Isaac show that black people would be targeted by predictive policing at roughly twice the rate of whites in PredPol and argue that there could be a feedback loop as police officers may be incentivized to make more arrests to increase their productivity. Ensign et al. follow Lum and Isaac's work and propose a model that can reduce the potential feedback loop in predictive policing [90]. The creators of PredPol responded to Lum and Issac work by arguing the difference between reported crimes, the data that PredPol is trained on, and police arrests, the outcomes from police patrol. They analyzed the results from a randomized controlled trial in collaboration with the Los Angeles Police Department (LAPD) and show that there is no significant difference in the arrest rate for different races by introducing PredPol [91].

With the increasing application of short-term crime predictive systems in policing, although the effect of predictive policing on data bias and feedback loop is inconclusive, I posit that it is critical for any new proposed model to be properly evaluated both in terms of accuracy and fairness. However, short-term crime pre-

diction models have exclusively focused on the evaluation of prediction accuracy, without taking into account any fairness analysis [13, 92, 93].

The feedback loop from the bias in arrest to future model training is one of the potential unfair or biased issues that could be generated by place-based crime prediction. As mentioned by [91], reported crime data is different from the police's arrests. Only 7% of the reported crimes are reported by police on scenes, and about 80% of the crimes are reported by the community [10]. The study by Lum and Isaac clearly shows that the prediction from predictive policing tools can be biased toward the minority population due to the under-reporting of crime. Therefore, it is necessary to comprehensively evaluate the fairness in crime prediction models and improve the fairness of crime predictions by addressing the bias in reported crimes themselves.

In this dissertation study, I will focus on addressing one of the sources of reported crime data bias, that is, the under-reporting issue to improve the fairness of long-term and short-term crime prediction. Also, I will conduct comprehensive analysis on crime prediction models in terms of accuracy and fairness and explore the potential sources of unfairness in the crime prediction outcomes.

2.5 Cross-city Transfer Learning

Transfer learning is important when training data is insufficient. Generally speaking, transfer learning aims to extract the knowledge from one or more source settings (tasks) and to apply that knowledge to a target setting (task) [94]. In the

context of urban computing, cross-city transfer learning aims to transfer knowledge from source cities with abundant data resources to target cities where services and infrastructures are not ready or just in place, and where data resources are insufficient. Cross-city transfer learning often times is described as domain adaptation as in Pan and Yang's framework, where the tasks are the same for both source and target cities. Cross-city transfer learning has been applied to multiple areas in urban computing including POI recommendation in new cities[95, 96], mobility generation [97, 98], bike services distribution [99], crowd flow prediction [63, 100, 101].

The main challenge of cross-city transfer learning is that different cities have different road network infrastructure, public transportation systems, socioeconomic development levels and traditions and cultures. Therefore, a critical step in knowledge transfer for these tasks is spatial matching among source cities and target cities *i*.e., identify city regions that are similar and for which the transfer knowledge can be applied. Related work has sometimes solved the spatial matching problem by proposing to identify similar grids in source and target cities. Fan et al. construct a spatial matching matrix connecting all grids in the source city and the target city so that user trajectories in the source city can be explicitly transferred as trajectories in the target city [97]. The spatial matching matrix also enables visualizing how the knowledge in the source city is transferred to the target city. Wang et al. identify the most similar grid in the source city for each grid in the target city based on Pearson correlations between time series of crowd flow or other auxiliary data [100]. Instead of finding grid-to-grid matches, Yao et al. proposed a spatial memory mechanism to compute nonlinear similarity scores with clusters of different

grid-level spatiotemporal patterns [63].

Another approach for spatial matching is to learn a feature transformation function that can convert a city-dependent spatial distribution to a feature space that is not city-dependent and can be shared among source cities and target cities. For example, Liu et al. apply Factor Analysis over the extracted features of two cities jointly and obtain latent features for city grids [99]. Instead of pairing one source city and one target city, He et al. trained an adaptation function on multiple source cities so that this function can transform features from city-dependent spatial contexts to a mobility intention space that can be shared among all source cities, and which is assumed to also be shareable with the target cities [98].

In terms of spatial units, a large part of the related work divides cities into grids. Nevertheless, there are other types of spatial units considered in the literature depending on the tasks of interest. In fact, cities can also be constructed as graphs based on the road networks. Using a graph partition method, Mallick et al. divide a large road network into multiple local networks so that a graph convolution network can utilize knowledge learned from seen network structures to predict traffic in unseen networks [101]; and Lin et al. propose a Clustering-based Transfer Model for Prediction (CTPM) based on dynamic time wrapping similarity between each pair of road segments from the source cities and the target cities. The prediction of traffic speed for a road segment in a target city would be based on the k-nearest road segments in the source city. [102].

The effectiveness of cross-city transfer learning in the mobility-based crime prediction has yet to be explored. In this dissertation study, I will conduct a pre-

liminary study on transfer learning to improve short-term crime prediction models in cities with data scarcity issue in human mobility data.

Chapter 3: Study 1: Spatial Sensitivity Analysis for Urban Hotspots using Cell Phone Traces

3.1 Introduction

A critical area in mobility behavior analysis is the identification of activity centers or dense regions *a.k.a.* hotspots, defined as regions with high concentration of individuals for a given period of time [2, 3, 103, 104]. Hotspot analyses using CDR data are generally carried out in two different scenarios (1) modeling, with a focus on analyzing the urban structure, such as the quantification of the urban sprawl or compactness of cities [2, 47]; or the analysis of the spatio-temporal evolution of popular locations for a given region [105, 106]; and (2) prediction, with a focus on the analysis of the predictive power of dense regions with respect to a given variable; for example, high footfall (number of estimated visits) in a region has been associated to high crime [107, 108], or large numbers of individuals at night or work times have been associated to the identification of home (residential) and work locations [109, 110]. These studies are often carried out at two different spatial scales: *intra-city*, where researchers focus on models or predictions for a given city [3, 5]; and *inter-city*, where researchers focus on comparing behaviors across cities [2, 4].

Although hotspot analyses using cell phone traces are extensive, there is no consensus among researchers about the process followed to compute them in terms of three important features: (i) *city boundaries* used to define the area under study *e.g.,* some researchers use metropolitan areas [2] that represent *cities* as labor market areas comprising commuting behaviors, while others use a smaller entity - the core municipality - which represents the physical boundary of a city rather than its economic activity, and which is generally contained within a metropolitan area together with other non-core municipalities [3, 111]; (ii) *spatial units* considered to compute the hotspots, which in the literature range from using Voronoi polygons that simulate cell phone coverage areas [103, 112, 113]; to uniformly distributed grids [2, 5, 109, 114–116]; or census tracts [6, 117], with the latter two approaches requiring the use of *interpolation methods* [4, 6, 118, 119] to distribute individuals associated to a given cellular tower across grids or census tracts; and (iii) *hotspot variables* used to measure and characterize the computed hotspots, such as the total number of hotspots for a region or hotspot compactness measures [2, 120]. The combination of these different features could produce significant differences in the hotspots identified, which could in turn provide conflicting findings.

In this study, I provide a spatial sensitivity analysis of the impact that the choice of a given set of city boundaries, spatial units and interpolation methods might have on the stability of the hotspot variables computed using cell phone traces (CDR). This spatial sensitivity analysis will answer the following questions:

1A) Are mobility-based urban hotspot features sensitive to the methodological choices that construct such features?

1B) If sensitive, is there any combination of choices that provide the most stable result among all possible choices?

The answers to these questions will provide guidelines for researchers looking to identify the most stable combination of parameters that will preserve the stability of the CDR-based hotspots independently of the city boundaries, spatial units and interpolation methods selected; and will also pinpoint into risky combinations of features that might produce non-stable, CDR-based hotspot measures. Additionally, these recommendations will also guide researchers into whether results across papers can be compared or not, based on the reported stability of certain combinations of features.

The systematic analysis will be carried out for two cases: inter-city and intra-city, where most of the related literature in CDR-based hotspot analyses has focused. Inter-city analyses will evaluate the stability of the city-rankings, based on a given hotspot variable, across different combinations of city boundaries, spatial units and interpolation methods; while intra-city analyses will assess the stability of a hotspot variable - computed hourly and represented as a 24 hour vector - across different combinations of city boundaries, spatial units and interpolation methods.

3.2 Methodology

Hotspot analyses using CDR data are critical to study city dynamics and the spatial structure of cities. To detect hotspots, researchers generally follow a set of common steps, although its implementation varies widely depend on research focus,

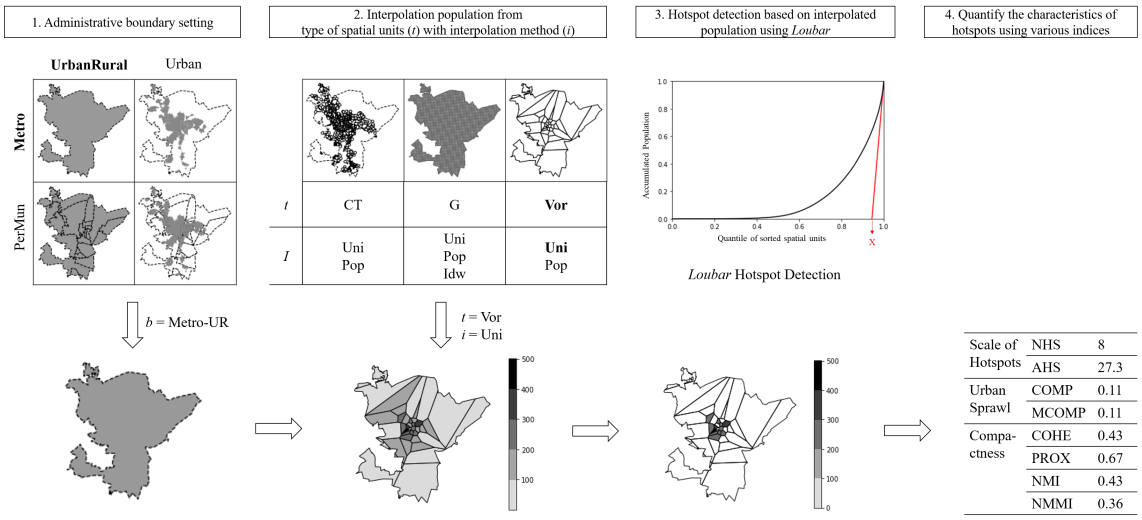


Figure 3.1: Hotspot Identification Process. The grey areas in Step 1 are the areas considered in each city boundary setting *e.g.,* the grey areas in Urban settings are urban areas, while the white areas are the rural areas. The outer boundary is the metro area boundary and the inner boundaries are the municipalities' boundaries.

application area or data availability. In this section, I will explain the different choices that researchers have in hand when computing hotspots, and I will describe the methodology I will use to assess the impact that the choice of varying feature combinations might have on the stability of a given hotspot measurement variable, both at the inter-city and intra-city levels.

Researchers generally follow these four steps to the identify the hotspots in a region: (i) define the city boundary of the city under study, (ii) define the type of spatial units used to compute hotspots, and estimate the population for each spatial unit, (iii) detect hotspots based on the estimated population, and (iv) compute hotspot indices to quantitatively characterize crowded regions in a city. Next, I explain each component in Figure 3.1 in detail.

3.2.1 City boundaries

The delimitation of cities or *urban* areas is in itself one of the traditional tasks in urban geography and planning [121]. Although not the focus of this study, it highlights the importance of understanding the impact that different city delimitations might have on hotspot analyses. Most related studies focused on the computation of CDR-based hotspots consider two different dimensions.

The first one is the definition of the physical city boundary. While some researchers define cities by their metropolitan area [2, 122], others only consider the urban core [3, 119, 123, 124]. Metropolitan areas are often defined as an aggregation of municipalities that share industry, infrastructure and housing, and that represent the *economic city* with a densely populated urban core area - that might span across multiple municipalities - and its surrounding rural, less-populated areas. On the other hand, municipalities are generally smaller spatial units embedded within a metropolitan area, with its urban core representing the physical boundary of the city and the region that has emerged historically as the most prominent in the metropolitan area [111]. Therefore, when the term *city* is used in current CDR-based hotspot analyses, it is important to understand whether it refers only to the densely populated areas within a metropolitan area [3, 123, 124]; or to the metropolitan area as a whole, including both the densely populated urban area and its less-populated, rural surrounding territories [2, 45, 122]. See *Urban* and *UrbanRural* columns in Figure 3.1.

The second dimension focuses on whether to treat the metropolitan area as

a whole unit to compute hotspots, or to consider each embedded municipality independent of each other, albeit connected by secondary population flows. Since metropolitan areas delimit the economic city, with mobility flows between its core urban area and other regions, it makes sense to identify hotspots at that scale, which would mostly characterize the commuting population [2, 45]. However, by computing hotspots at that scale, local characteristics or economic structures of individual municipalities might be ignored. For example, non-core municipalities within a metropolitan area might be sub-centers for jobs in the region [121]. As a result, the mobility patterns characterizing these municipalities might be more affected by its internal flows that by movements to and from other municipalities [122].

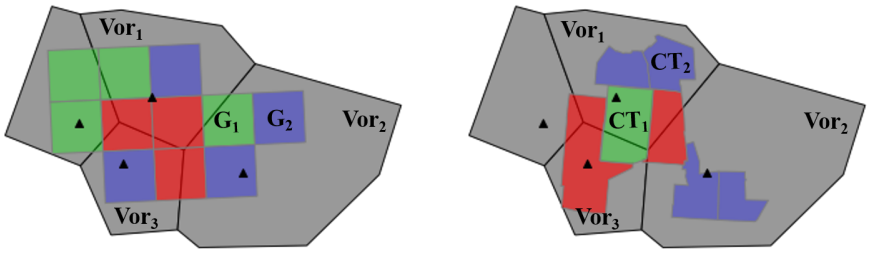
To carry out a comprehensive assessment of the different city boundary settings that are used by researchers when computing CDR-based hotspots, I propose to explore the following four settings: (i) *Metropolitan Area Urban-Rural* (Metro-UR), where hotspots are computed across the the whole metropolitan area that includes all urban and rural areas; (ii) *Metropolitan Area Urban* (Metro-U), where hotspots are computed across the whole metropolitan area which is defined exclusively by its urban areas; (iii) *Municipalities Urban-Rural* (PerMuni-UR), where hotspots are computed per individual municipality, and considering both urban and rural areas within the municipality; and (iv) *Municipalities Urban* (PerMuni-U), where hotspots are computed individually only for the urban areas within each municipality. The boundary setting is denoted as b with $b \in$ {Metro-UR, Metro-U, PerMuni-UR, PerMuni-U}. An example of the four different boundary types are shown in Step 1

in Figure 3.1.

3.2.2 Spatial units and interpolation methods

Voronoi tessellation is a common spatial unit of choice when using CDR data to understand population dynamics [103, 112, 113, 117]. For each cell tower c in the cell phone infrastructure, Voronoi tesselation is used to represent its spatial coverage or service area (see Figure 3.2). The assumption on which Voronoi tessellation is based is that users would always use the closest cell tower. In this way, researchers associate to a given Voronoi polygon v_c all the individuals that have been observed at that cell tower c. However, Voronoi polygons (Vor) are not the only type of relevant spatial unit. Some researchers have focused on spatial regularity and have chosen grids (G) [2, 109, 114] as the spatial units of interest, while others prefer to census tracts or blocks (CT) [6, 117] because these are the same geographic units as census data and can represent the boundaries of neighborhoods to some extent. In this study, I will denote the type of spatial unit as t with $t \in \{\text{CT, G, Vor}\}$.

Voronoi tessellation assigns a set of individuals to a given Voronoi polygon, and the number of individuals *i.e., the footfall*, is then used to compute hotspots. However, when using grids or census tracts, or when a Voronoi polygon needs to be clipped because it spreads outside the boundary of a city or a municipality, additional processing is required to assign the presence of individuals to a different spatial unit. Grid and census tracts polygons will overlap with Voronoi polygons, and as a result, interpolation methods that approximate the footfall in a given



(a) Grids (b) Census tracts

Figure 3.2: An example of grids intersecting with Voronoi polygons (the underlying grey polygons). The locations of cell towers are represented as black triangles. Grids or census tracts in red, green and blue intersect with three, two and one Voronoi polygons, respectively. For example, G_1 intersects with Vor_1 and Vor_2, G_2 intersects with Vor_2, CT_1 intersects with Vor_1 and Vor_3 and CT_2 intersects with Vor_1.

overlapping polygon area is required (see Figure 3.2 for an example). Similarly, clipped Voronoi polygons will require to approximate the footfall for any given sub-polygon. With that objective in mind, I explore three types of interpolation methods commonly present in the literature: uniform (Uni), population-based (Pop) and inverse-distance (Idw) [2, 6, 118]. In this study I will denote the interpolation methods as i with $i \in \{\text{Uni, Pop, Idw}\}$.

The most common interpolation method in the CDR literature is the uniform method (Uni) [2, 119, 125]. This method assumes that all individuals are located within a given polygon uniformly. Therefore, the number of individuals in any grid or census tract polygon overlapping with a Voronoi polygon will be proportional to its area. Let $\boldsymbol{v}$ be the set of Voronoi polygons intersected with spatial unit u, v_c be the Voronoi polygon and n_c be the footfall in cell tower c, the interpolated footfall $n_{u,i=\text{Uni}}$ using the Uniform method for u is computed as follows:

$$n_{u,i=\text{Uni}} = \sum_{v_c \in \boldsymbol{v}} \frac{\text{Area}\left(\text{Intersection}\left(u, v_c\right)\right)}{\text{Area}\left(v_c\right)} \cdot n_c \tag{3.1}$$

The limitation of the Uniform method is that people are unlikely to be distributed over a spatial unit uniformly, especially for vast rural areas where people are less likely to be present. Therefore, researchers have used population-based methods (Pop) that distribute the footfall of a Voronoi polygon over a spatial unit proportionally to the population density *e.g.,* urban areas in the spatial unit are assigned larger numbers of individuals than rural areas [6]. The population-based method requires information about census population. Let the given census population distribution be at the census tract level, y_x be the shape and p_x be the population of the x-th census tract. The census population is assumed to be normally distributed within y_x, as no finer-grained information about population is available. Let S be an arbitrary polygon *e.g.,* a spatial unit or the intersection between a spatial unit and a Voronoi polygon. If S itself is a census tract, the population of S is straightforward. That is one of the reasons why some researchers prefer to use census tract as spatial units. But if S is not a census tract, then S intersects with a set of census tracts denoted as $\boldsymbol{y}$. The population of S is computed as:

$$\text{Pop}(S) = \sum_{y_x \in \boldsymbol{y}} \frac{\text{Area}\left(\text{Intersection}\left(S, y_x\right)\right)}{\text{Area}\left(y_x\right)} \cdot p_x \tag{3.2}$$

Let $\boldsymbol{v}$ be the set of Voronoi polygons intersected with spatial unit u, v_c be the Voronoi polygon and n_c be the footfall for cell tower c, the population-based method interpolates the footfall $n_{u,i=\text{Pop}}$ at a given spatial unit u proportional to

the population, instead of to the total area:

$$n_{u,i=\text{Pop}} = \sum_{v_c \in \boldsymbol{v}} \frac{\text{Pop}\left(\text{Intersection}\left(u, v_c\right)\right)}{\text{Pop}\left(v_c\right)} \cdot n_c \tag{3.3}$$

Nevertheless, the population-based method has an important drawback since the population retrieved from the census will represent residential population rather that footfall, which might affect the way population dynamics in non residential areas are computed. Also, both uniform and population-based methods assume that the association of individuals to the closest cell tower location is always correct, which might not be the case specially for users who are at the boundaries of a given Voronoi polygon. Thus, researchers have used a third method to overcome this limitations, the inverse distance weighting (Idw) [4, 118] that determines that the number of individuals in a spatial unit is the weighted average of its neighbor cellular towers where the weights are inversely proportional to the distance. The distance between a spatial unit and a cell tower is computed using their centroids. It is important to clarify that this method has only been used by researchers in combination with grids, not census tracts [4, 118]. Given $\boldsymbol{v}$ as the set of neighbor Voronoi polygons of spatial unit u, n_c as the footfall for cell tower c, and $d(u, v_c)$ as the distance between the centroids of u and v_c, the interpolated population for u is computed as follows:

$$n^*_{u,i=\text{Idw}} = \frac{\sum_{v_c \in \boldsymbol{v}} \frac{1}{d(u, v_c)} n_c}{\sum_{v_c \in \boldsymbol{v}} \frac{1}{d(u, v_c)}} \tag{3.4}$$

Let $\boldsymbol{s}^{(a,b)}$ be the set of spatial units and $\boldsymbol{v}^{(a,b)}$ be the set of Voronoi polygons intersecting with all spatial units in city a under boundary setting b. Interpolating using inverse distance weighting does not guarantee the sum of all $n^*_{u,a=\text{Idw}}$ for $u \in \boldsymbol{s}^{(a,b)}$ to be the same as the sum of all n_c for $v_c \in \boldsymbol{v}^{(a,b)}$. Therefore, I re-scale $n^*_{u,a=\text{Idw}}$ as follows:

$$n_{u,i=\text{Idw}} = n^*_{u,i=\text{Idw}} \frac{\sum_{v_c \in \boldsymbol{v}^{(a,b)}} n_c}{\sum_{u \in \boldsymbol{s}^{(a,b)}} n^*_{u,i=\text{Idw}}} \tag{3.5}$$

In summary, I consider in this study the following combinations C of spatial units and interpolation methods: *(CT, Uni), (CT, Pop), (G, Uni), (G, Pop), (G, Idw), (Vor, Uni), (Vor, Pop)*, with grids of 500 x 500 meters, since this is one of the most common choices in the literature [2, 116, 123, 126]. This is not meant to be a complete list of combinations. There are different types of spatial units and interpolation methods of interest. For example, one could take into account the terrain or land cover information to assign different relative population density [39, 127]. Here I aim to analyze commonly used methods to shed light on potential stability issues. An example of the different spatial units and interpolation methods explored are shown in Step 2 in Figure 3.1.

3.2.3 Hotspot detection

To compute the hotspots of a city, I need to first identify the spatial units with a significant number of individuals. Hotspot detection is a binary classification problem, where the spatial units with a estimated number of people above a

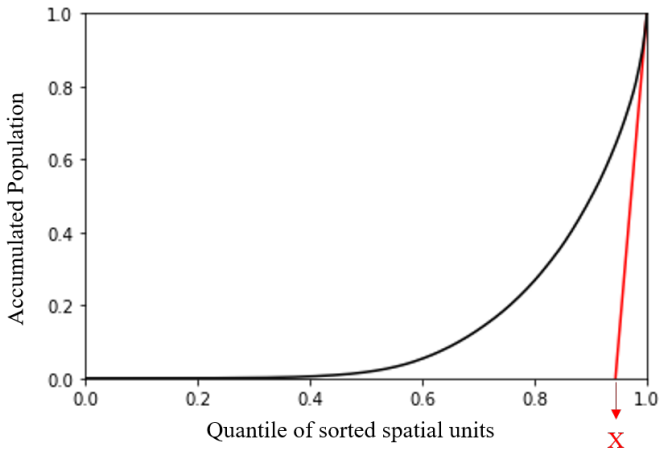


Figure 3.3: *Loubar* hotspot detection for a set of U spatial units with interpolated population. 1) the units are sorted in ascending order by the population; 2) draw the Lorenz curve of the accumulated population with x-axis being the ranking of units normalized by U; 3) compute the intersection of the tangent line at x=1.0 (the red line) and the x-axis. Let the intersection point be (X, 0); 4) The threshold δ is the population of the $U * (1 - X)$th spatial unit; 5) all spatial units with population $\geq \delta$ are the hotspots detected. The detailed explanation of *Loubar* method can be found in Louail et al. [2].

threshold value δ, *i.e.,* $n_u > \delta$, are considered as hotspots.

There exist different methods to determine the threshold δ [28, 104]. However, as previous work has shown, δ can be constrained within a lower and an upper bound [2]. Given the estimated footfall for a set of spatial units, the lower bound of δ is defined as the average of the set of footfall values. On the other hand, the upper bound *i.e.,* the strictest definition of hotspot, is computed using the *Loubar* method based on the Lorenz curve. The *Loubar* method is briefly explained in Figure 3.3. In this study, I will focus on the use of the upper bound, since it constitutes the strictest approach to measure the spatial structure of the most important places, and a result, strongest common denominator across different thresholds considered in the literature.

Given a CDR dataset, the *Loubar* method will be applied as follows. I will first aggregate the number of unique users for each cell tower c to obtain the average hourly number of unique users: $\overline{\boldsymbol{n}}_c = \{\overline{n}_{c,h}\}_{h=0}^{23}$, where $n_{c,h}$ represents the average of unique users between h:00:00 and h:59:59. Using equations (3.1) - (3.5), I can calculate the interpolated population with method i for a spatial unit u: $\overline{\boldsymbol{n}}_{u,i} = \{\overline{n}_{u,i,h}\}_{h=0}^{23}$.

Let $\boldsymbol{s}^{(a,b)}$ be all the spatial units in city a, $\boldsymbol{s}_m^{(a,b)}$ be the spatial units in a municipality m in city a in boundary setting b, and $\overline{\boldsymbol{N}}_{\boldsymbol{s},i,h} = \{\overline{n}_{u,i,h}\}_{u\in\boldsymbol{s}}$ be the interpolated population using method i for a set of spatial units $\boldsymbol{s}$ at hour h. I apply the *Loubar* hotspot detection method to $\overline{\boldsymbol{N}}_{\boldsymbol{s}^{(a,b)},i,h}$ if b is Metro-UR or Metro-U, or apply *Loubar* to $\overline{\boldsymbol{N}}_{\boldsymbol{s}_m^{(a,b)},i,h}$ for each municipality m in city a if b is PerMuni-UR or PerMuni-U to compute the threshold value δ and detect whether a spatial unit u is a hotspot at hour h. Finally, a spatial unit will be identified as a hotspot if it is permanent *i.e.,* it is considered a hotspot throughout the 24 hours of the day (*all-day*) $\mathbf{1}_{u,i,h} = 1$ for $0 \leq h \leq 23$ [2]. This binary decision can be denoted as:

$$\mathbf{1}_{u,i,h} = \begin{cases} 1, \text{if } \overline{n}_{u,i,h} \geq \delta \\ \\ 0, \text{if } \overline{n}_{u,i,h} < \delta \end{cases} \tag{3.6}$$

However, since population density [45] and employment density [28] are often used in quantitative geography, which roughly correspond to the presence of people during nighttime and daytime, in this study I will also explore *home-hour* and *work-hour* hotspots. These are formally defined as permanent hotspots during working

hours (9am-5pm) or home hours (10pm-5am) *i.e.,* $\mathbf{1}_{u,i,h} = 1$ for $h_s \leq h \leq h_e$ with $h_s = 9$ and $h_e = 17$ and with $h_s = 22$ and $h_e = 5$, respectively.

3.2.4 Hotspot measurement variables

In this study, I will explore three types of hotspot measurement variables or *indices* that have been traditionally used in related literature for hotspot analyses at inter-city and intra-city levels: (1) scale of the hotspots, (2) degree of urban sprawl and (3) urban compactness. The first type quantifies the number of hotspots detected and the geographical area covered by them. The last two types of indices focus on the quantification of urban structure [2, 30, 45, 46]. Research in quantitative geography and urban economics has shown the importance of studying urban structure, as it can shape people's mobility in terms of travel distance, model choice and car usage [30, 45], the transportation system in terms of energy consumption or air pollution [45, 46], and economic growth performance [47, 48]. Next, I explain each set of indices in detail.

(1) Hotspot Scale quantified in terms of number of grids in a city that are detected as hotspots (NHS) and the total geographical area covered by the hotspots detected (AHS).

(2) Urban sprawl characterizes a type of metropolitan decentralization or sub-urbanization where a large percentage of a city's residential and/or business activity takes place outside of its central location [128]. I use the following indices to quantify the degree of urban sprawl:

- Compacity coefficient (*COMP*) [2] measures the sprawl of the detected hotspots over a city, with smaller *COMP* values associated to less dispersed hotspots with respect to the size of the city. Let A be the geographic area of the city of interest, hs be the set of hotspots, $|hs|$ be the number of hotspots and $d_{j,k}$ the distance between the centroids of hotspot j and k.

$$\text{COMP} = \frac{\text{D}_{hs}}{\sqrt{A}},\ \text{D}_{hs} = \frac{\sum_{j=1}^{|hs|}\sum_{k=j+1}^{|hs|} d_{j,k}}{|hs|\,(|hs|-1)/2} \tag{3.7}$$

- Mass Compacity coefficient (*MCOMP*) [45] is a modified compacity coefficient that weights the distance between hotspots by the population of each grid, and measures the average distance between individuals located within the detected hotspots. The smaller MCOMP is, the less dispersed the hotspots are with respect to the size and population of the city. Let p_j be the population in grid j.

$$\text{MCOMP} = \frac{\text{MD}_{hs}}{\sqrt{A}},\ \text{MD}_{hs} = \frac{\sum_{j=1}^{|hs|}\sum_{k=j+1}^{|hs|} d_{j,k}\,p_j\,p_k}{\sum_{j=1}^{|hs|}\sum_{k=j+1}^{|hs|} p_j\,p_k} \tag{3.8}$$

(3) Urban compactness The major difference between urban compactness and urban sprawl indices is that sprawl is always measured with respect to the size of a city *e.g.,* the indices are normalized by the square root of the geographical area, while compactness is based on the assumption that the most compact form of a shape is a circle [120]. Therefore, compactness indices measure compactness in terms of geometrical properties, and are thus normalized by the reference circle, *e.g.,* an equal-area or equal-perimeter circle. Urban compactness indices range from

0 to 1, with 1 representing the exact continuous cirle. I consider the following four indices that are commonly used in hotspot measurement literature:

- Cohesion ($COHE$) [120] is the ratio of the average distance-squared among all points in the reference circle and the average distance-squared among all points in the hotspot areas. Large cohesion means people in hotspot areas are very close to each other. Let r be the points of hs in the rasterized format, $|r|$ be the number of points and $dr_{i,j}$ be the distance between the i- and j-th point.

$$\text{COHE} = \frac{\text{AHS}/\pi}{\frac{2}{|r|(|r|-1)} \sum_{j=1}^{|r|} \sum_{k=j+1}^{|r|} dr_{i,j}^2} \tag{3.9}$$

- Proximity ($PROX$) [120] is the ratio of the average distance from all points in the reference circle to its centre and the average distance to the geometry center of the hotspot areas. The proximity index focuses on the distance between points from the geometry center instead of the point-wise distance in the cohesion index. Let g be the center of gravity of hs and dg_i be the distance between the i-th point and the center g.

$$\text{PROX} = \frac{\frac{2}{3}\sqrt{\text{AHS}/\pi}}{\frac{1}{|p|} \sum_{i=1}^{|p|} dg_i} \tag{3.10}$$

- Normalized moment of inertia (NMI) [129] is based on the dispersion of points from the center of its shape. It involves the calculation of the second moment of an area about a point, also known as the moment of inertia (MI). The MI is then normalized by the MI of the reference circle, hence normalized moment

of inertia.

- Normalized mass moment of inertia (*NMMI*) [130] takes into account the mass distribution of a shape. The previous three compactness indices consider only the geometric shape *i.e.,* each point in the shape is equally important in the compactness. Nevertheless, each hotspot might have a different estimated population or mass, and they can still be compact - even though their geometry shape is not - by having the majority of the population concentrate around the mass center. The reference circle in *NMMI* is no longer an equal-area circle, but a circle with equal-effective-area. The mathematical derivation for the calculation of *NMI* and *NMMI* can be found in [129] and [130], respectively.

3.2.5 Hotspot index stability

Hotspot indices are often used as a lens to study various aspects of urban life, and have been used to compare cities (inter-city analyses) [2, 4, 48, 124], or to compare hourly hotspots within a city of interest (intra-city analyses) [3–5, 119]. For example, research has shown that high density cities in the UK share higher levels of public transportation use by low-income residents [29]; while other studies have revealed that hotspot compactness has a significant negative correlation with purchasing power parity in cities in Asia, US, Europe, Latin America and Australia [48].

Nevertheless, the choice of city boundaries, spatial units and interpolation methods prior to the computation of hotspots and hotspot indices could produce

significant differences in the hotspots identified, which could in turn provide conflicting findings. For example, a researcher interested in using hotspots to predict crime in a city, could find a strong correlation or no correlation at all, depending on the set of city boundaries, spatial units and interpolation methods as well as hotspot variables considered; or a researcher interested in comparing number of hotspots across cities, could identify largely different city-rankings depending on the sets of features used.

The main objective of this study is to provide a systematic analysis of the impact that the choice of a given set of city boundaries, types of spatial units and interpolation methods might have on the stability of the hotspot indices presented. The recommendations of these analyses will provide guidelines for researchers looking to identify the combination of parameters that will preserve the stability of the hotspot indices; and will pinpoint into risky combinations of features that might produce hotspot indices with little stability. Next, I explain my approach to measuring the stability of a hotspot index for both inter-city and intra-city scales.

Inter-city Index Stability. Inter-city analyses focus on comparing rankings of cities based on a given hotspot index, or on comparing rankings of cities based on correlations between a given hotspot index and another urban feature such as crime or economic growth. Thus, to measure the stability of a given inter-city index, I propose the following approach. For each combination of city boundary b, type of spatial unit t (Vor, G or CT) and interpolation method i (Uni, Pop or Idw), I compute the hotspots and hotspot indices described in the methodology section across all cities under study. Next, I conduct Spearman correlation for each pair

of city rankings resulting from different combinations of features, and compute the stability of an index for a given city boundary, as the average of all correlation coefficients across spatial units and interpolation methods. High average correlation coefficients across combinations of features will reveal that the hotspot index is stable *i.e.,* the ranking of the cities for a given index is similar independently of the spatial features used. Researchers could select any set of features since the rankings do not appear to change, and as result, any correlation analyses between hotspots and other features would also be robust. On the other hand, low average correlation coefficients will identify indices that should not be used since the rankings vary widely depending on the combination of features.

Let $\boldsymbol{C}$ = {*(CT,Uni), (CT, Pop), (G, Uni), (G, Pop), (G, Idw), (Vor, Uni), (Vor, Pop)*} be the list of combinations of type t spatial units and interpolation method i considered in this study, with $|\boldsymbol{C}|$ as the number of combinations and C_j as the j-th combination where $1 \leq j \leq |C|$. For each hotspot index $ind \in$ {NHS, AHS, COMP, MCOMP, COHE, PROX, NMI, NMMI}, city boundary b and combination $C_j = (t_j, i_j)$, I first compute the permanent hotspots (all-day/work-hour/home-hour) and then compute the index values for all cities under study. Each combination C_j will produce an array of index values, one per city, defined as $\boldsymbol{ind}_{b,C_j}$. Next, for each pair of combinations C_j and C_k, I compute the correlation coefficient for index ind and city boundary b as:

$$Coef_{ind,b,j,k} = \text{Spearman}(\boldsymbol{ind}_{b,C_j}, \boldsymbol{ind}_{b,C_k}) \tag{3.11}$$

The coefficient measures the similarity of rankings among cities. Then the stability for index ind and city boundary b is computed as:

$$Stability_{ind,b} = \frac{1}{|\boldsymbol{C}|\,(|\boldsymbol{C}|-\mathbf{1})} \sum_{j,k \leq |C|,\, j \neq k} Coef_{ind,b,j,k} \tag{3.12}$$

Intra-city Index Stability. Intra-city analyses generally focus on comparing hotspot rankings across time for a given city. Index stability at the intra-city level is helpful to identify the indices that provide similar hourly rankings independently of the spatial and interpolation features used. Stable indices can thus be used to robustly study the relationship between hotspots and urban growth or transportation efficiency, for example; while unstable indices will be discouraged from use given the variability of the rankings they provide. To measure the stability of an index at the intra-city level, given a city a defined using an city boundary setting b, I first compute the hotspots at each hour h using combination $C_j = (t_j, i_j)$. This yields a 24-hour vector $\boldsymbol{ind}_{b,C_j,a}$ for each combination C_j and city a. Pairs of combinations C_j and C_k are compared in terms of ranking similarity in city a via Spearman correlation.

$$Coef_{ind,b,a,j,k} = \text{Spearman}(\boldsymbol{ind}_{b,C_j,a}, \boldsymbol{ind}_{b,C_k,a}) \tag{3.13}$$

The intra-city level stability for index ind at city a using city boundary b is the computed as:

$$Stability_{ind,b,a} = \frac{1}{|\boldsymbol{C}|\,(|\boldsymbol{C}|-\mathbf{1})} \sum_{j,k \leq |C|,\, j \neq k} Coef_{ind,b,a,j,k} \tag{3.14}$$

Finally, to identify the stability of a given index for a certain city boundary and spatial combination, I average the stability measure across all the cities $a \in A$ under study:

$$Stability_{ind,b} = \frac{1}{A} \sum_{a=1}^{A} Stability_{ind,b,a} \tag{3.15}$$

Spearman correlation coefficients will be interpreted as follows [131]: a stability score in range of $[0.8, 1)$ is considered very strongly stable; in range of $[0.6, 0.8)$ is considered strongly stable; moderately stable in the $[0.4, 0.6)$ range; weakly stable in the range of $[0.2, 0.4)$ and unstable in the $[0.0, 0.2)$ range.

3.3 Results

3.3.1 Study areas and Dataset

To carry out the sensitivity analyses, I use pseudonymised CDR data from the 59 top metropolitan areas in Mexico (see Figure 3.4). The data covers cell phone activity from October 2009 to June 2010. No individual data has been used, only aggregated statistics at the cell tower level to quantify the number of unique users per hour. City boundaries have been defined using official shapefiles for metropolitan areas as defined by CONAPO, the National Population Council in Mexico [132]. Municipalities, census tracts (known as AGEBs in Mexico) and urban and rural

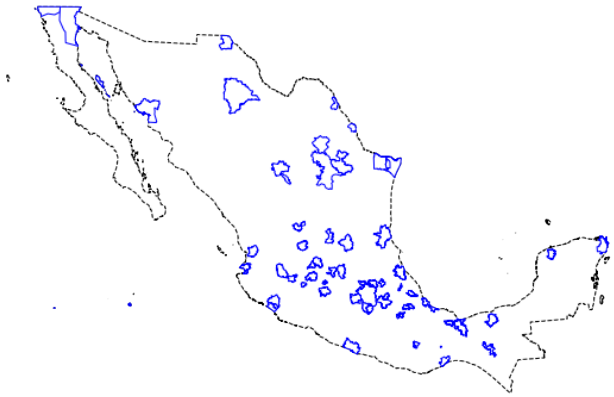


Figure 3.4: 59 metropolitan areas in Mexico

areas have been extracted from INEGI, the statistical department in Mexico [133], with data from 2010.

3.3.2 Inter-city level analysis

Figure 3.5 shows the stability scores for each hotspot index and city boundary, with each table representing all-day, work-hour and home-hour hotspots. Recall that each stability value is computed as the average of all Spearman correlations between all pairs of combinations of spatial units and interpolation methods. To measure that variance, each table also shows the standard deviation of the stability in parentheses. Next, I describe the main outcomes, followed by an in-depth discussion in the next section. There are the following observations based on the Figure:

1) All hotspot indices are the most stable when cities are defined by their urban municipalities only (PerMuni-U) with average correlations between different spatial unit and interpolation combinations ranging from 0.62 to 0.86 - very strongly correlated - across methods. Hotspot indices are least stable when cities are defined

	Metro UR	Metro U	PerMuni UR	PerMuni U
NHS	0.39 (0.31)	0.61 (0.17)	0.65 (0.16)	0.76 (0.09)
AHS	0.33 (0.24)	0.64 (0.17)	0.55 (0.17)	0.81 (0.06)
COMP	0.55 (0.17)	0.61 (0.16)	0.71 (0.17)	0.77 (0.08)
MCOMP	0.56 (0.16)	0.61 (0.16)	0.72 (0.11)	0.73 (0.10)
COHE	0.40 (0.19)	0.53 (0.10)	0.49 (0.17)	0.64 (0.11)
PROX	0.40 (0.19)	0.53 (0.13)	0.50 (0.17)	0.62 (0.11)
NMI	0.40 (0.19)	0.53 (0.10)	0.52 (0.15)	0.64 (0.11)
NMMI	0.41 (0.17)	0.53 (0.11)	0.57 (0.11)	0.62 (0.14)

(a) All-day

	Metro UR	Metro U	PerMuni UR	PerMuni U
NHS	0.52 (0.28)	0.74 (0.12)	0.65 (0.18)	0.80 (0.09)
AHS	0.46 (0.22)	0.76 (0.12)	0.58 (0.16)	0.86 (0.05)
COMP	0.57 (0.14)	0.68 (0.13)	0.70 (0.19)	0.78 (0.10)
MCOMP	0.59 (0.15)	0.63 (0.16)	0.73 (0.11)	0.73 (0.10)
COHE	0.41 (0.19)	0.55 (0.11)	0.44 (0.19)	0.67 (0.12)
PROX	0.41 (0.19)	0.53 (0.13)	0.45 (0.19)	0.65 (0.13)
NMI	0.41 (0.19)	0.55 (0.11)	0.47 (0.17)	0.67 (0.12)
NMMI	0.43 (0.19)	0.55 (0.13)	0.59 (0.12)	0.65 (0.15)

(b) Work-hour

	Metro UR	Metro U	PerMuni UR	PerMuni U
NHS	0.42 (0.30)	0.68 (0.14)	0.67 (0.15)	0.80 (0.08)
AHS	0.37 (0.24)	0.73 (0.14)	0.57 (0.16)	0.85 (0.05)
COMP	0.58 (0.15)	0.64 (0.15)	0.72 (0.16)	0.78 (0.08)
MCOMP	0.59 (0.14)	0.68 (0.12)	0.74 (0.11)	0.73 (0.13)
COHE	0.35 (0.22)	0.53 (0.11)	0.46 (0.19)	0.65 (0.11)
PROX	0.37 (0.22)	0.54 (0.13)	0.47 (0.18)	0.63 (0.11)
NMI	0.35 (0.22)	0.53 (0.11)	0.48 (0.16)	0.65 (0.11)
NMMI	0.42 (0.19)	0.55 (0.11)	0.56 (0.12)	0.64 (0.14)

(c) Home-hour

Figure 3.5: Stability (Standard deviation) of all indices in different boundary settings. The gradient background color is based on the stability score ranging from 0 to 1, the darker the orange color is, the closer it is to 1.

by their metropolitan area and considering both urban and rural regions (Metro-UR, with stability values from 0.33 to 0.59). Generally, it is fair to say that all indices tend to be more stable in settings that consider only urban areas and independent municipalities, rather than whole metropolitan areas. As a result, and whenever possible, city boundaries that consider only urban municipalities should be favored in inter-city analyses since the ranking of cities will likely remain stable and comparisons with other urban features - *e.g.,* crime - will be robust thus avoiding conflicting results.

2) Scale of hotspot indices (NHS, AHS) and urban sprawl indices (COMP, MCOMP) are the most stable. For 3 out of 4 city boundary settings, NHS, COMP and MCOMP are between strongly and very strongly stable. Therefore, compared to the compactness indices, researchers have more freedom to choose the boundary settings and interpolation methods for inter-city level comparisons that are based on scale of hotspots and degree of urban sprawl. This means that under the same city boundary setting, comparison among cities in terms of these two indices or correlation with other factors using them are less likely to produce conflicting findings

across combinations of types of spatial units and interpolation methods. Finding (1) revealed that PerMuni-U produces the most stable indices across types. However, the second most stable boundary setting for scale of hotspots and urban sprawl is different. Scale of hotspots' second best setting is Metro-U while for urban sprawl is PerMuni-UR. For urban sprawl, the difference between the stability in PerMuni-UR and in PerMuni-U is small. Therefore, as long as researchers are using PerMuni-based settings, whether or not to include rural areas does not have a large impact in terms of stability.

3) Compactness indices (COHE, PROX, NMI and NMMI) are the least stable indices. When making comparisons at the inter-city level in terms of compactness indices, researchers should first pay attention to the boundary settings because compactness indices are strongly stable only in the PerMuni-U setting and moderate to weakly stable in the other three settings. When PerMuni-U setting is undesired, *e.g.,* rural areas need to be incorporated, myproposed method allows researchers to explore in depth the relationship between spatial units, interpolation methods and stability for a selected city boundary; to then choose the most stable combination within the unstable setting. For example, Figure 3.6 shows the Spearman correlations among the different combinations for the PROX index in the Metro-UR boundary setting. In general, the stability across all combinations is low. However, if researchers need to use a Metro-based setting, the (G, Pop) combination tends to have the largest average correlations (0.49) and thus, the largest stability which would make it the top candidate combination to use when extracting hotspots at both urban and rural scales. Similar results are observed for COHE, NMI and

	CT Uni	CT Pop	G Idw	G Uni	G Pop	Vor Uni	Vor Pop	Mean
CT Uni	1.00	0.55	0.47	0.50	0.65	0.15	0.23	0.43
CT Pop	0.55	1.00	0.40	0.49	0.68	0.30	0.31	0.45
G Idw	0.47	0.40	1.00	0.31	0.53	0.32	0.40	0.41
G Uni	0.50	0.49	0.31	1.00	0.57	0.18	0.06	0.35
G Pop	0.65	0.68	0.53	0.57	1.00	0.20	0.31	0.49
Vor Uni	0.15	0.30	0.32	0.18	0.20	1.00	0.80	0.32
Vor Pop	0.23	0.31	0.40	0.06	0.31	0.80	1.00	0.35

Figure 3.6: Spearman correlation coefficient $Coef_{ind=PROX,b=Metro\text{-}UR,j,k}$ between each pair of combinations (C_j, C_k) for all-day permanent hotspots. The coefficient matrix is symmetric. The row mean is the average of coefficients in each row, excluding the values on the diagonal. The row mean of j-th row shows the average coefficients of combination C_j correlated with other combinations.

	Metro UR	Metro U	PerMuni UR	PerMuni U
NHS	0.38 (0.17)	0.38 (0.17)	0.31 (0.12)	0.35 (0.17)
AHS	0.43 (0.18)	0.41 (0.19)	0.36 (0.16)	0.39 (0.18)
COMP	0.52 (0.23)	0.52 (0.24)	0.30 (0.13)	0.33 (0.14)
MCOMP	0.49 (0.22)	0.49 (0.24)	0.36 (0.19)	0.40 (0.21)
COHE	0.25 (0.10)	0.29 (0.11)	0.29 (0.15)	0.35 (0.16)
PROX	0.26 (0.11)	0.30 (0.12)	0.29 (0.15)	0.35 (0.17)
NMI	0.25 (0.10)	0.29 (0.11)	0.29 (0.15)	0.35 (0.16)
NMMI	0.30 (0.11)	0.33 (0.14)	0.35 (0.15)	0.39 (0.15)

(a) 24-hour vector

	Metro UR	Metro U	PerMuni UR	PerMuni U
NHS	0.40 (0.14)	0.42 (0.15)	0.40 (0.16)	0.43 (0.18)
AHS	0.39 (0.15)	0.42 (0.16)	0.34 (0.14)	0.41 (0.17)
COMP	0.48 (0.21)	0.49 (0.23)	0.30 (0.12)	0.36 (0.15)
MCOMP	0.46 (0.22)	0.47 (0.23)	0.39 (0.18)	0.41 (0.22)
COHE	0.32 (0.14)	0.35 (0.14)	0.32 (0.16)	0.37 (0.16)
PROX	0.33 (0.15)	0.37 (0.15)	0.32 (0.17)	0.38 (0.16)
NMI	0.32 (0.14)	0.35 (0.14)	0.32 (0.16)	0.37 (0.16)
NMMI	0.37 (0.15)	0.39 (0.17)	0.39 (0.17)	0.44 (0.16)

(b) 4-hour-bin vector

Figure 3.7: Stability (Standard deviation) of all indices in different boundary settings. The coefficients in (a) are computed using 24-hour vector of indices and in (b) are computed using 4-hour-bin vector. The gradient background color is based on the stability score ranging from 0 to 1, the darker the orange color is, the closer it is to 1.

NMMI.

4) In most cases, the stability for different indices and boundary settings is similar among all-day, work-hour and home-hour permanent hotspots. But the home-hour stability in Metro-UR settings for COHE, PROX and NMI is consistently smaller than all-day and work-hour.

3.3.3 Intra-city level analysis

The stability scores at the intra-city level for each hotspot index and across boundary settings are shown in Figure 3.7(a). Standard deviation values for each stability score - representing the spread of the correlations among different combinations of spatial units and interpolation methods - are also shown. The Figure 3.7(a) shows that:

1) Hotspot scale and urban sprawl indices are more stable in Metro-based settings, while urban compactness indices highest stability is achieved when using municipalities to define city boundaries. Nevertheless, the impact of boundary settings is small for urban compactness indices. Based on these findings, researchers could favour one type of indices versus others depending on the boundary settings of interest, which would in turn produce more robust indices to measure hotspot rankings over time for a given city.

2) The highest stability score is achieved by urban sprawl indices (COMP and MCOMP) under Metro-UR and Metro-U boundary settings. This is different from the inter-city level stability analyses where the best score was achieved with PerMuni-based settings. However, even the highest stability is only moderately stable. AHS also has moderate stability under Metro-UR and Metro-U boundary settings. The rest indices in all settings are weakly stable, much lower than the stability scores at the inter-city level, indicating that a change in the boundary or spatial unit choice could produce widely different results.

3) The intra-city stability scores per index are computed extracting permanent

hourly hotspots *i.e.,* hotspots that are considered as such throughout the 24 hours. As a result, the low stability scores could be potentially due to the stringent definition of hotspot. To assess that, we grouped the 24 hours into 6 bins - each bin is a 4-hour-bin - and computed the hotspot indices again. In this case, the Spearman correlation was computed between two 6-bin vectors of coefficients - instead of the previous 24-hour vectors. Figure 3.7(b) shows that although the stability scores increase, except for the COMP index in the Metro-based setting, they are still mostly weakly stable.

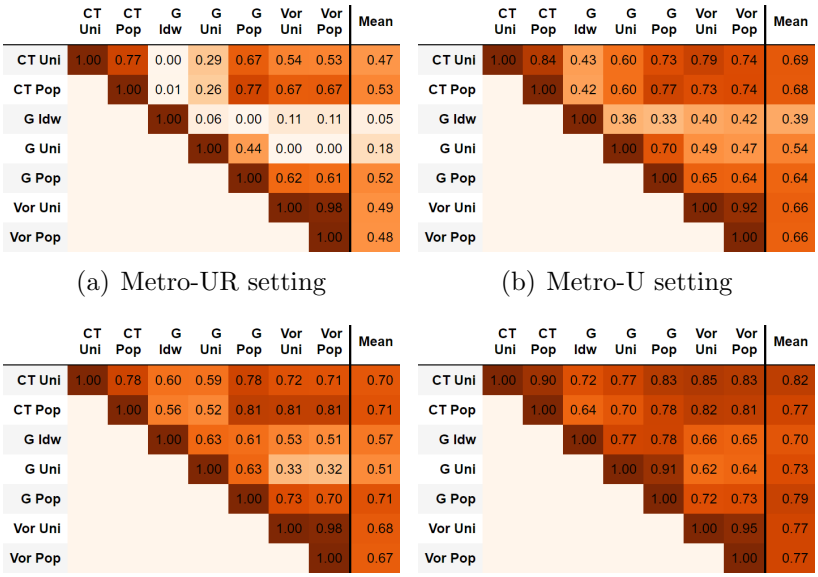
Finally, it is important to mention that although the index stability scores at the intra-city scale are on average low, some cities do have much higher stability scores than others. This finding might be indicating that, unlike inter-city scales, intra-city hotspots' stability might depend on other types of cultural or social trends not studied in this study.

3.4 Discussion

In this section, I explore potential reasons behind the stability findings described in the previous section.

3.4.1 Stability of hotspot scale indices (NHS and AHS) at the inter-city level

As explained in the previous section, hotspot scale indices (NHS and AHS) are more stable when city boundaries are defined using only urban municipalities.



(a) Metro-UR setting (b) Metro-U setting

(c) PerMuni-UR setting (d) PerMuni-U setting

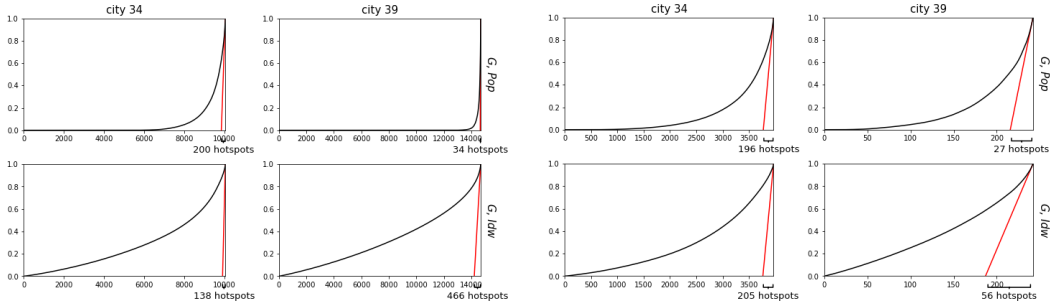
Figure 3.8: Spearman correlation coefficients $Coef_{ind=NHS,b,j,k}$ between each pair of combinations (C_j,C_k) under four different boundary settings. The coefficient matrix is symmetric whose lower triangular part is omitted.

I posit that this might be due to the fact that rural areas, which generally have smaller footfall, are dramatically changing the Lorenz curves and, as a result, the scale of the hotspots computed. I will now analyze in depth a few case examples that are representative of the global trends observed in the analyses. Figure 3.8(a) shows the Spearman correlation coefficients for the NHS index across all the combinations of spatial unit and interpolation methods per each of the four boundary settings. Combinations (*G, Uni*) and (*G, Idw*) are weakly or even not correlated with other methods in the *Metro-UR* causing the low stability score. Changing to other boundary settings, such as excluding the rural areas, all the coefficients involving methods (*G, Uni*) and (*G, Idw*) have a large increase and as a result NHS becomes more stable in *Metro-U* setting. Therefore, I will compare the NHS computed based on combinations *(G, Idw)* and *(G, Pop)* to explore what might cause the unstability

or dissimilarity in the cities ranking.

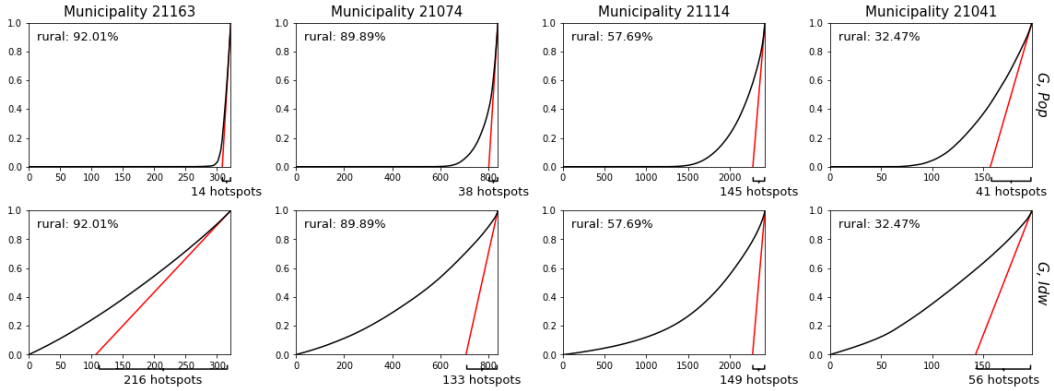
Figure 3.9(a) and 3.9(b) shows the comparison between Metro-UR and Metro-U settings for city 34 (Puebla-Tlaxcala Metropolitan Area) and city 39 (Rioverde-Ciudad Fernández Metropolitan Area). Each plot represents the Lorenz curve used to compute NHS_a based on a combination $C = (t, i)$ of type of spatial units t and interpolation method i in city a. The results show that the more rural areas a city has, the more heavily the shape of Lorenz curve is changed by different combinations C, which causes variations in the number of hotspots identified. For example, Figure 3.9(a) shows that under the Metro-UR setting, $\text{NHS}_{34} = 200$ and for $\text{NHS}_{39} = 34$ based on the combination *(G, Pop)*, which means that the ranking of city 34 is higher than 39. Changing to the combination *(G, Idw)*, NHS_{34} decreases to 138 but NHS_{39} increases to 466 showing that the ranking of cities 34 and 39 has reversed. This is caused by the fact that city 39 has more rural areas (99%) than city 34 (70%), and the Lorenz curve for city 39 is impacted more (much more closer to the diagonal) by changing from *(G, Pop)* to *(G, Idw)* than the curve for city 34.

On the other hand, the spatial units in the rural areas are not considered in the hotspots detection under the Metro-U setting. For example, the number of grids considered in city 39 in Metro-U setting is about 250, much smaller than in Metro-UR setting which is about $14,500$. The impact brought by the variation in percentage of rural areas is mitigated by focusing on urban areas only in the Metro-U setting *i.e.,* the change in the Lonrenz curve and change in NHS from *(G, Pop)* to *(G,Idw)* is smaller and more consistent in both cities. Changing from *(G, Pop)* to the *(G, Idw)*, NHS_{34} increases from 196 to 205 and NHS_{39} increases from 27 to

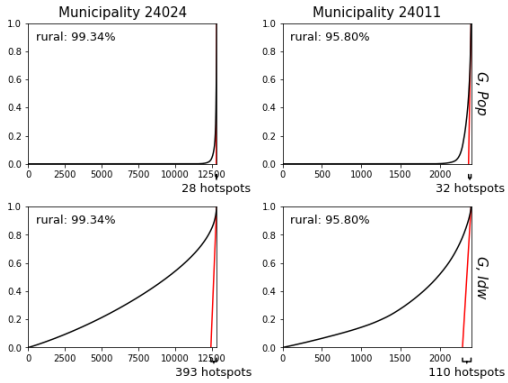


(a) Metro-UR

(b) Metro-U



(c) PerMuni-UR, municipaliites in city 34



(d) PerMuni-UR, municipaliites in city 39

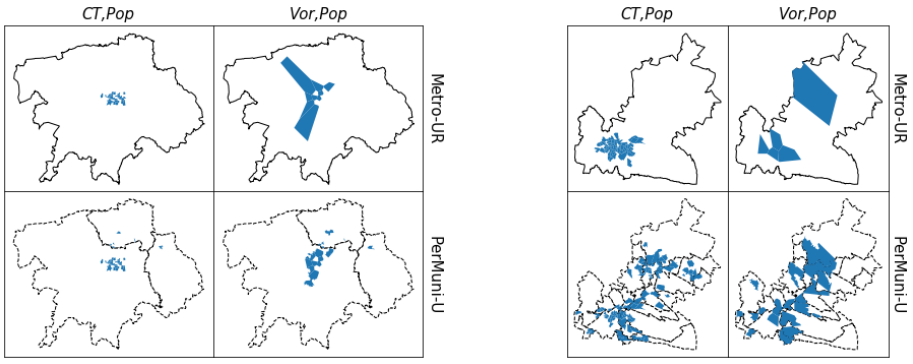
Figure 3.9: Lonrenz curves for loubar-based hotspots detection in city 34 and 39 in different boundary settings. The x-axis is rescaled to the number of spatial units to better explain the difference in NHS. Combination *(G, Pop)* and *(G, Idw)* are shown for comparison. City 34 has 70% of rural areas and 39 municipalities with various percentage of rural areas while city 39 has 99% of rural areas and 2 municipalities both with more than 95% of rural areas.

56. Therefore the ranking of both cities are better preserved between these two combinations.

Next, focusing on the Permuni-UR setting. In PerMuni-based settings hotspots are detected per municipality. Therefore, the Lorenz curves might be affected by the percentage of rural areas in each municipality. City 34 as a whole metropolitan area has 70% rural areas, but it has 39 municipalities with various percentages of rural areas from 1% to 92%. Four example municipalities are shown in Figure 3.9(c). City 39 has 2 municipalities, both of which have similar percentages of rural areas as city 39 as a whole, shown in Figure 3.9(d). It can be observed that for municipalities with high percentage of rural areas *e.g.,* municipality $22163, 21074, 24024, 24011$ in Figure 3.9(c) and 3.9(d), the Lorenz curves are also heavily impacted by changing from *(G, Pop)* to *(G, Idw)*. For municipalities with lower percentage of rural areas *e.g.,* municipality $21114, 21041$ in Figure 3.9(c), the Lorenz curves are less impacted. As a result, changing from *(G, Pop)* to *(G, Idw)*, the NHS_{34} changes from 905 to 3663 and the NHS_{39} changes from 60 to 503. Although city 34 still has a smaller change in NHS (increased 3 times) than city 39 (increased 7.4 times), the ranking between them is preserved. Therefore the stability of NHS in Permuni-UR is better than in Metro-UR.

3.4.2 Stability of urban sprawl indices (COMP and MCOMP) at the inter-city level

As discussed in the previous section, urban sprawl indices computed with PerMuni-based boundary settings appear to be more stable than Metro-based ones. One of the possible reasons might be the different ways in which the population of a metropolitan area can be distributed across municipalities. For example, some metropolitan areas have a dominant urban core with the majority of human activities, while in other metropolitan areas there might be municipalities acting as sub-centers with similar levels of activity as their urban cores.



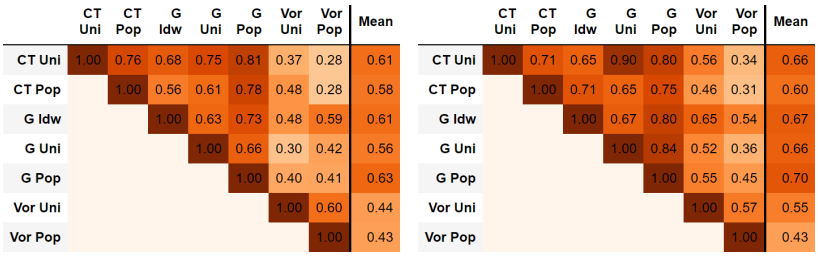
(a) city 25 (b) city 46

Figure 3.10: Permanent hotspots detected by combinations *(CT, Uni) (Vor, Pop)* for city 25 and 46.

Take city 25 (Morelia Metropolitan Area) and 46 (Tlaxcala-Apizaco Metropolitan Area) as examples (Figure 3.10). Both city 25 and 46 have multiple municipalities and each municipality has an urban area. But in city 25, the urban core is dominating, that is, the population of city 25 is mostly concentrated in one core urban region *e.g., (CT, Pop)* and *(Vor, Pop)* in Metro-UR setting in Figure 3.10(a).

While in city 46, the location of the hotspots varies. With the combination *(CT, Pop)* in Metro-UR, the permanent hotspots concentrate in the left-bottom corner (Figure 3.10(b)), just like city 25. But with the combination *(Vor, Pop)* in Metro-UR, permanent hotspots in other municipalities are detected thus increasing the value of the COMP index (Figure 3.10(b)). As a result, since different metropolitan areas have different distributions of population over multiple municipalities, hotspots indices computed over metropolitan-based settings are not as stable. On the other hand, detecting hotspots in the municipality-based settings is more stable because the permanent hotspots are local to each municipality and overall the hotspots spread over multiple municipalities (see the second row in Figure 3.10(a) and 3.10(b)). And because COMP is normalized by the square-root of city's geographical area, the distance between hotspots spreading over the city is normalized. Therefore variances in the population distribution bring less instability to COMP indices across spatial units and interpolation methods.

Rural areas also appear to play a role in index stability. Figure 3.11 shows the correlation coefficients for each pair of spatial unit and interpolation method combination across all boundary settings for the COMP index. Combinations *(Vor, Uni)* and *(Vor, Pop)* are the least correlated with other combinations, especially in settings including rural areas. When considering rural areas, the Voronoi polygons can cover large regions *e.g.,* the plot of *(Vor, Pop)* using the Metro-UR setting in Figure 3.10(a) and 3.10(b), as the cell towers tend to be sparse in rural areas. Since urban sprawl indices are computed based on the distance among centroids of polygons, these large Voronoi polygons drag the centroids away from dense population



	CT Uni	CT Pop	G Idw	G Uni	G Pop	Vor Uni	Vor Pop	Mean
CT Uni	1.00	0.76	0.68	0.75	0.81	0.37	0.28	0.61
CT Pop		1.00	0.56	0.61	0.78	0.48	0.28	0.58
G Idw			1.00	0.63	0.73	0.48	0.59	0.61
G Uni				1.00	0.66	0.30	0.42	0.56
G Pop					1.00	0.40	0.41	0.63
Vor Uni						1.00	0.60	0.44
Vor Pop							1.00	0.43

(a) Metro-UR setting

	CT Uni	CT Pop	G Idw	G Uni	G Pop	Vor Uni	Vor Pop	Mean
CT Uni	1.00	0.71	0.65	0.90	0.80	0.56	0.34	0.66
CT Pop		1.00	0.71	0.65	0.75	0.46	0.31	0.60
G Idw			1.00	0.67	0.80	0.65	0.54	0.67
G Uni				1.00	0.84	0.52	0.36	0.66
G Pop					1.00	0.55	0.45	0.70
Vor Uni						1.00	0.57	0.55
Vor Pop							1.00	0.43

(b) Metro-U setting

	CT Uni	CT Pop	G Idw	G Uni	G Pop	Vor Uni	Vor Pop	Mean
CT Uni	1.00	0.81	0.88	0.91	0.82	0.50	0.68	0.77
CT Pop		1.00	0.85	0.80	0.94	0.41	0.66	0.74
G Idw			1.00	0.90	0.88	0.47	0.67	0.78
G Uni				1.00	0.83	0.44	0.69	0.76
G Pop					1.00	0.45	0.67	0.76
Vor Uni						1.00	0.72	0.50
Vor Pop							1.00	0.68

(c) PerMuni-UR setting

	CT Uni	CT Pop	G Idw	G Uni	G Pop	Vor Uni	Vor Pop	Mean
CT Uni	1.00	0.75	0.81	0.92	0.74	0.72	0.65	0.77
CT Pop		1.00	0.81	0.78	0.86	0.59	0.72	0.75
G Idw			1.00	0.90	0.85	0.78	0.70	0.81
G Uni				1.00	0.85	0.79	0.72	0.83
G Pop					1.00	0.69	0.87	0.81
Vor Uni						1.00	0.74	0.72
Vor Pop							1.00	0.73

(d) PerMuni-U setting

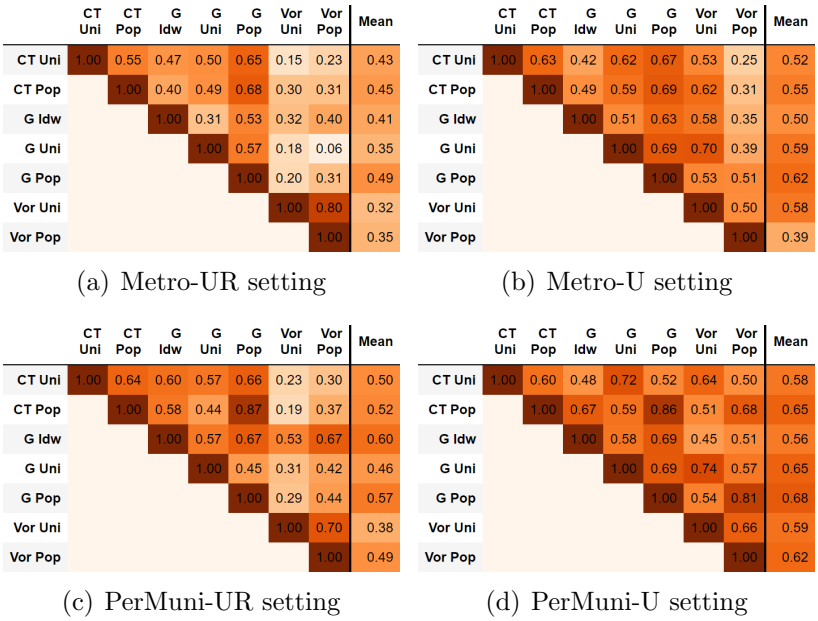
Figure 3.11: Spearman correlation coefficient $Coef_{ind=COMP,b,j,k}$ between each pair of between each pair of combinations (C_j,C_k) under four boundary settings. The coefficient matrix is symmetric whose lower triangular part is omitted.

areas; and since the sizes of the Voronoi polygons are not homogeneous across different metropolitan areas, the inclusion of rural areas brings in high instability to hotspot urban sprwal indices.

3.4.3 Stability of urban compactness indices (COHE, PROX, NMI and NMMI) at the inter-city level

Urban compactness indices, similarly to urban sprawl, are computed using the distance between the detected hotspots. Therefore, these indices are subject to the same instability issues due to the various ways in which population and footfall can be distributed across municipalities, and to the varying shapes that Voronoi polygons might have in rural areas. Comparing Figures 3.11 and 3.12 it can also be observed that the correlation between *Vor*-based and other spatial combinations

is much worse for urban compactness indices than for urban sprawl indices. This is because compactness indices measure the compactness of the shape of hotspots, and the varying nature of the Voronoi polygons causes the *Vor*-based combinations to be weakly to no correlated with other combinations when the rural areas are considered.



(a) Metro-UR setting (b) Metro-U setting

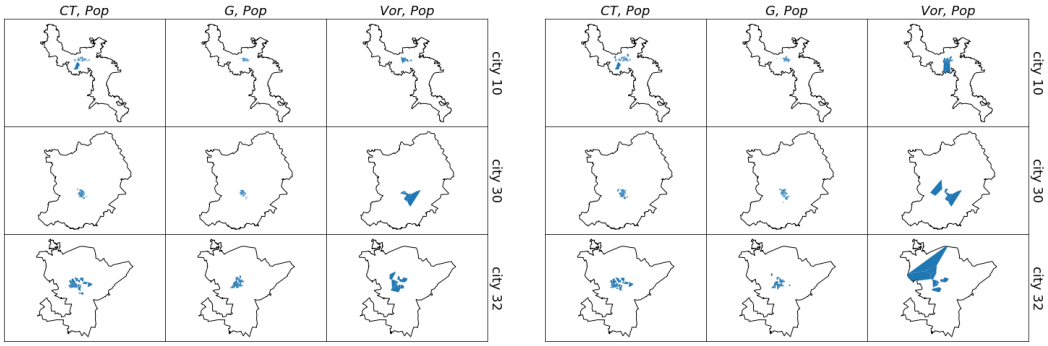
(c) PerMuni-UR setting (d) PerMuni-U setting

Figure 3.12: Spearman correlation coefficient $Coef_{ind=PROX,b,j,k}$ between each pair of between each pair of combinations (C_j,C_k) under four boundary settings. The coefficient matrix is symmetric whose lower triangular part is omitted.

3.4.4 Difference of stability between home-hour and work-hour permanent hotspots at inter-city level

There exists little difference in the stability of hotspot indices computed for home- and work-hour periods across different city boundary settings. Nevertheless, it is worth noting that the stability scores for COHE, PROX and NMI computed for home-hour periods are slightly more unstable than their work-hour counterpart,

especially in the Metro-UR setting. I argue that this might be due to the fact that for some cities home locations are more spatially scattered, possibly including the outskirts where CT and Vor polygons tend to be larger when compared to work locations in downtown areas that tend to have smaller CT and Voronoi polygons. Thus, these varying area sizes in home location polygons might be increasing the instability of the indices. See Figure 3.13 for an example of this setting with city 10 (Tuxtla Gutiérrez Metropolitan Area), city 30 (Tepic Metropolitan Area) and city 32 (Oaxaca Metropolitan Area). The home-hour permanent hotspots are more scattered than the work-hour permanent hotspots. Using *(G, Pop)*, a few small grids away from the core area in city 30 and 32 are considered as permanent hotspots. But using *(Vor, Pop)*, the Voronoi polygons away from the core area have large variation in size. The variation in size would have a huge impact on the covered geographic area and subsequently on the equal-area circle considered in the compactness indices. Therefore, COHE, PROX and NMI have slightly lower stability in home-hour period.



(a) Work-hour permanent hotspots (b) Home-hour permanent hotspots

Figure 3.13: Permanent hotspots detected by methods *(CT, Pop)*, *(G, Pop)*, *(Vor, Pop)* for city 10, city 30 and city 32 in work and home hours under boundary setting b=Metro-UR

Chapter 4: Study 2: Addressing Under-Reporting to Enhance Fairness and Accuracy in Mobility-based Long-term Crime Prediction

4.1 Introduction

Historical crime data is of great importance to understand the severity of crimes in society. Countless reports, academic papers, books and news articles rely on reported crime data [12, 13]. This data can be used to, for example, evaluate the effects of programs and policies designed to prevent crime in a city [14]. Crime prediction, on the other hand, is an important topic of research that uses reported crime data to predict future occurrences. For example, historical crime data has been used to predict hotspots so as to assist patrol route planning [15, 92] Traditionally, historical crimes and socioeconomic data have been used as input to build crime prediction models at various geographical levels *e.g.,* grids, cities, municipalities [12, 60, 134, 135]. Nevertheless, due to the increasing availability of mobility data such as geolocated social media and mobile phone data, a large number of studies have explored the predictive relationship between mobility patterns and reported crimes [17, 19, 136].

There are various theories about the relationship between mobility and crime in urban environments. For example, the *Opportunity makes the Thief* theory claims that the opportunity is the cause of crime [69] *i.e.,* the higher the presence of suitable targets such as people and property, the more crimes could happen; and empirical work has confirmed that theory, showing that there is a super-linear relation between the daily floating population (number of people that has been in a neighborhood) and incidence of property crimes [9]. Mobility patterns not only trace the movement of people, but can also characterize the dynamic spatial structure of the urban environment by detecting urban dense areas, *a.k.a.* hotspots [2]. Urban spatial structure is a critical and well understood concept in environmental criminology and urban quantitative geography, and it has been shown to be correlated to crime incidents [29, 137].

One of the major concerns of using reported crime data in crime prediction is data bias, especially when computational models - built upon such data - could influence future resource allocation *e.g.,* planning police patrol routes [92]. Data bias in this context can be framed under algorithmic fairness whereby crime predictive models can behave differently for disadvantaged groups such as low-income or minorities due to over- or under-representation in the historical crime data [138]. In fact, not all crimes are reported. Sometimes the public does not report crimes that are considered minor [23]; and low-income has been related to higher under-reporting for certain types of crimes [80]. In addition, not all reported crimes end up recorded in the official crime statistics, a decision mostly made by the police [24, 73]. Police may decide not record a report as a crime because of insufficient evidence and/or

individual biases [24]. Therefore, the reported crime data statistics that are used in research will naturally be biased, reflecting partial crime incidents mediated by community engagement, police resources and potential police biases towards disadvantaged populations. Although a few papers have looked into the identification of biases in predictive policing tools that exclusively use historical crimes [26], there is no work in the analysis of biases for mobility-based crime prediction models, nor in mitigation strategies to enhance fairness without sacrificing accuracy.

In this study, I propose a Bayesian hierarchical model to identify and mitigate under-reporting issues that could lead to biases and lack of fairness in mobility-based crime incident predictions [139, 140]. Specifically, the predictive model uses mobility-based features to infer the number of *true* crime, *i.e.,* the actual number of crimes that will occur regardless of whether they will be reported; and use domain knowledge of determinants for under-reporting (*e.g.,* poverty, unemployment rate) to model the reporting rate, *i.e.,* the ratio of the number of reported crimes to true crimes. By conducting experiments on different types of reported crimes, I aim to answer the following questions:

2A): Does modeling the under-reporting process improve or hurt the performance of predicting the number of reported crimes?

2B): Does modeling the under-reporting process improve the fairness of crime prediction?

2C): What influence do the mobility-based features have on the true crime generating process?

2D): What influence do the determinants have on the reporting rate, *i.e.*, the

under-reporting process?

4.2 Method

Mobility-based crime prediction can be framed as a regression problem: given a region of interest i, *e.g.,* a city, a set of mobility-based features $\boldsymbol{u_i}$ characterizing the dynamic spatial structure of i extracted from past mobility data and a set of determinants $\boldsymbol{s_i}$ that characterize under-reporting in i, predict the number of future crimes z_i in that region, *i.e.,*

$$z_i = F(\boldsymbol{u_i}, \boldsymbol{s_i}), \tag{4.1}$$

where F is the predictor to be trained. F can represent under-reporting-unaware models, *i.e.,* models that do not address under-reporting and use crime data as is, such as generic machine learning models; I hypothesize these models will make biased crime prediction due to the inherent bias in the reported crime data. F can also represent under-reporting-aware models, such as the proposed Bayesian model that explicitly models the under-reporting issue so as to mitigate bias.

In this section, I will introduce three major components of the proposed method: 1) The construction of mobility-based features $\boldsymbol{u_i}$ based on Call Detailed Records (CDR) data; 2) the proposed Bayesian hierarchical model for mobility-based crime prediction that addresses the under-reporting issue using a set of under-reporting determinants $\boldsymbol{s_i}$; 3) The process of fairness and accuracy evaluation for crime prediction.

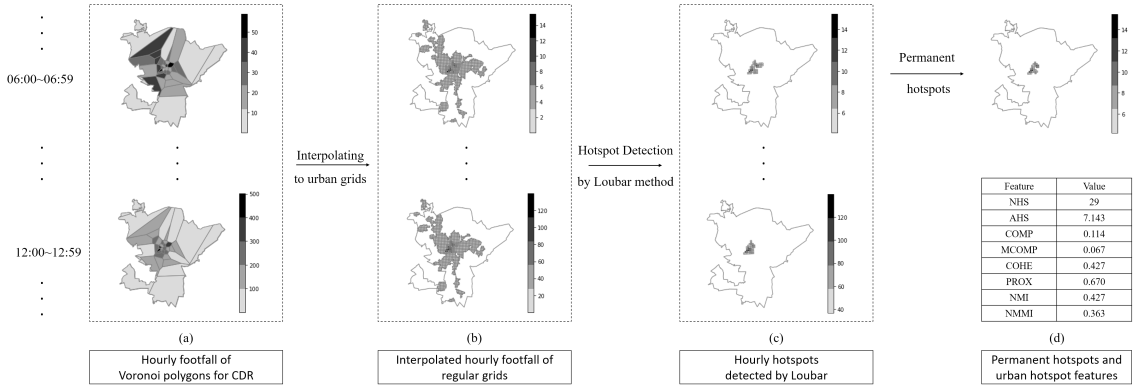


Figure 4.1: Extracting urban hotspot features from CDR data.

4.2.1 Mobility-based Hotspots Features

In this study, I will focus on mobility-based crime prediction models that exploit the predictive power of the dynamic hotspots and urban spatial structures in cities by analyzing the relationship between hotspots and crime incidents.

The mobility data used in this study are Call Detailed Records (CDR). CDR are a commonly used mobile phone data collected by telecommunication networks for billing purposes. CDR provide - among other features - spatio-temporal data about individual mobility behaviors. CDR locations are represented as the (latitude, longitude) pairs of the cellular towers that mobile phones are using when making phone calls or sending texts. The spatial coverage of cellular towers is often approximated via Voronoi tessellation. Hourly footfall is computed for each Voronoi polygon, defined as the average number of hourly unique users present at a given polygon (Figure 4.1(a)).

Due the irregularity of Voronoi tessellation, I interpolate footfall from Voronoi polygons to regular grids, with the assumption that footfall within a Voronoi polygon

is uniformly distributed over space. That is, the footfall for a grid within a Voronoi polygon is proportional to the overlap between grid and Voronoi polygon (see Figure 4.1(b)). In order to detect urban hotspots, I follow a similar approach to [2]: 1) for each hour of the day, I apply the Loubar method to the hourly footfall of each grid so as to detect the upper bound of the number of hourly hotspots (Figure 4.1(c)); 2) the grids that are detected as hotspots over the 24 hours of the day are identified as permanent hotspots (Figure 4.1(d)). The permanent hotspots represent the most important centers of dense activity in the urban environment and are the ones that I will use to predict crime incidents.

After detecting urban hotspots, I compute the three types of urban hotspots features as described in Section 3.2.4.

- Hotspot scales: number of hotspots (NHS) and total gegraphical area covered by hotspots (AHS);
- Urban sprawl: compacity coefficient (COMP) and mass compacity coeffcient (MCOMP);
- Urban compactness: cohesion (COHE), proximity (PROX), normalized moment of inertia (NMI), and normalized mass moment of inertia (NMMI).

4.2.2 Bayesian Model for Under-Reported Crimes (BURC)

As explained in previous sections, the problem of under-reporting in crime data is an important source of potential bias in mobility-based crime prediction algorithms that might affect protected groups. In this study, I develop a Bayesian

hierarchical model to mitigate under-reported crime incidents by inferring two variables (1) the unobserved "true" crime incidents *i.e.*, all the crimes that have occurred regardless of whether they have been reported and collected; and (2) the reporting rate, *i.e.*, the ratio of true crimes being reported in the crime data. The core of this Bayesian model is that the crime incidents are hypothesized to be generated following a Poisson distribution given the urban spatial structure features and the reporting rate is dependent on the determinants of the under-reporting issue through a logistic link function.

For a given city i, let y_i be the volume of true crime incidents (hidden variable), λ_i be the average incident occurring rate for true crime incidents for the Poisson distribution, z_i be the volume of reported crime incidents, and π_i be the reporting rate (hidden variable). I model the generative process of crime incidents as follow:

$$y_i|\lambda_i \sim \text{Poisson}(\lambda_i) \tag{4.2}$$

$$z_i|\lambda_i, \pi_i \sim \text{Poisson}(\pi_i\lambda_i) \tag{4.3}$$

$$log(\lambda_i) = \alpha_0 + \sum_{k=1}^{K} \alpha_k u_i^{(k)} \tag{4.4}$$

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{j=1}^{J} \beta_j s_i^{(j)} \tag{4.5}$$

The volume of true crimes y_i follows the Poisson distribution given the aver-

age occurring rate λ_i and the volume of reported crimes z_i also follows a Poisson distribution but the occurring rate is $\pi_i\lambda_i$, discounted by the reporting rate π_i. The λ_i is modeled by the logarithmic link function to ensure $\lambda_i > 0$ and the π_i is modeled by the logistics link function to ensure $\pi_i \in (0, 1)$. $\boldsymbol{u}_i = (u_i^{(1)}, ..., u_i^{(K)})^T$ is the feature vector for city i and $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, ..., \alpha_K)^T$ is the coefficients to model the true crimes occurring rate. $\boldsymbol{s}_i = (s_i^{(1)}, ..., s_i^{(J)})^T$ is the feature vector for city i and $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_J)^T$ is the coefficients to model the reporting rate of reported crimes. In this study, the feature vectors $\boldsymbol{u}_i$ for the true crimes occurring rate λ_i are the urban hotspots features that characterize the dynamic spatial structure of a city and which has been related to crime incidents [137]. The feature vectors $\boldsymbol{s}_i$ are determined by domain knowledge about the determinants for the reporting rate of the types of crimes of interest. For example, studies have shown that poverty rate [80] and unemployment rate [81] could decrease the likelihood of property crime incidents, such as burglaries, being reported. Therefore the feature vectors $\boldsymbol{s}_i$ for the under-reporting process for property crimes would contain poverty rate (PR) and unemployment rate (UR) for each municipality. For violent crimes, on the other hand, gender, age and marital status of the victims [72], which correspond to male to female ratio (M/F), adult rate (AR) and never marriage rate (NMR) for a city respectively, as well as the percentage of female-headed households with children (transformed as male-headed to female-headed household ratio in this study, M/FHH), poverty rate (PR) and foreign born population rate (FR) of census tracts [78] are shown to influence reporting behavior. Therefore these factors would be the $\boldsymbol{s}_i$ determinants in violent crimes modeling.

By treating the volume of reported crime incidents as observed variables and the volume of true crime incidents and reporting rate as hidden variables, this model manages to separate the bias in the crime reporting process from the volume of true crimes. In Section 4.3.2 I will show that this model can more accurately infer crime volumes while making more fair predictions.

4.2.3 Fairness and Accuracy Evaluation

As mentioned in the related work, fairness evaluation is often based on the notion of protected attributes such as gender, race or income levels. Although there are various definitions of fairness, its main objective is to achieve some form of (approximate) parity across the various groups defined by the protected attribute *e.g., female vs. male, low-income vs. high-income*. In this study I will consider two protected attributes that have been observed to receive unfair treatment and suffer from discrimination in the criminal justice system: income and race [141].

In this crime prediction problem setting, fairness is evaluated for a regression problem. A common choice of fairness metric for regression problems given a binary protected attribute is the mean difference, i.e., the difference between average prediction values in the positive group *e.g.* female, and the average prediction values in the negative group *e.g.,* male [85]. The mean difference is a real number with a value of zero signifying no attribute effect or dependency. The larger the absolute value of mean difference is, the less fair the predictions are for a given protected attribute. Given that the protected attributes are non-binary (income and race) *i.e.*

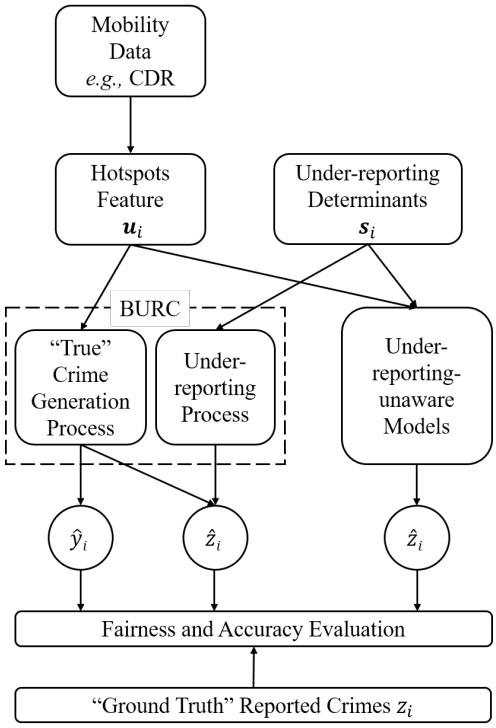


Figure 4.2: The framework of this study. BURC is the proposed Bayesian hierarchical model. $\hat{z}_i$ is the predicted number of reported crimes by different models and $\hat{y}_i$ is the predicted number of "true" crimes by BURC.

attributes defining more than two groups of population, I generalize the definition of mean difference for binary groups into multiple groups. I compute the mean difference for each group in the "1 vs all" setting, i.e., the MD_i for group i is computed as the difference between average prediction for group i and average prediction for other groups. In addition, I will compute the group error *i.e.,* the RMSE between predicted and ground truth reported crimes within each group, to show the fairness in terms of performance difference across different protected groups.

To assess the impact of addressing under-reporting in crime data, I will use the proposed BURC model to infer the volumes of true crimes; and evaluate the fairness and the accuracy of the BURC predictions via mean difference across protected groups (see Figure 4.2). Finally, these results will be compared against a set

Figure 4.3: Municipalities in Mexico studied in this study, colored in grey.

of baseline classifiers that use the reported crime data without any under-reporting treatment. Given that the volumes of true crimes have a different scale than the reported crimes, and given that the mean difference used to measure fairness is scale dependent, I will use the mean difference normalized by the average of reported crimes and true crimes to allow comparison between models. Similarly, since different protected groups will have different scales for average prediction, I will also normalize RMSE by the group average of the prediction to show the relative group error.

4.3 Experiments

To assess whether addressing under-reporting in crime data can improve the fairness and accuracy of crime prediction models, I focus on crime and mobility data from $1,379$ municipalities in Mexico as shown in Figure 4.3. 90% of the municipalities have population less than $80,000$ and geographic area less than $2,000$ km^2 while the largest population is $1,815,786$ and the largest area is $53,256$ km^2. In

this study, I consider two types of crime: property crimes and violent crimes across municipalities in Mexico.

4.3.1 Experiment Setting

4.3.1.1 Data

There are four types of data used in this study:

1. Mobility data is extracted from pseudonymised Call Detailed Records (CDR) from October 2009 to June 2010 across all $1,379$ municipalities in Mexico. No individual data has been used, only aggregated statistics at the cell tower level. As described in Section 4.2.1, CDR data is used to extract footfall and hotspot features.

2. Reported crime statistics are obtained from Mexico's *Secretary General of National Public Security* (SESNSP) [142]. I have retrieved property and violent crime data from 2011 for the $1,379$ municipalities under study. Property crimes in this study mainly include thefts, thefts from vehicles and burglaries, while violent crimes include robbery, sexual offense, homicide, battery, assault and kidnapping. These annual volumes of reported property crime or violent crime are used as the observed variables $z_i, i = 1, ..., 1379$ in the BURC model. The range of number of reported property (violent) crimes in these municipalities is [0, 17655] ([0, 28329]), the average is 265 (522) and the standard deviation is 1091 (1868). The volumes of reported property (violent) crime for 90% of the municipalities are less than 450 (900) Therefore there is a large variation in the number of crimes across municipalities.

3. Determinants of under-reporting in BURC include poverty rate (PR), unemployment rate (UR), adult rate (AR), the percentage of people who are never married (never married rate, NMR), male to female ratio (M/F), male-headed to female-headed household ratio (M/FHH) and the percentage of population born in other municipalities (foreign-born rate, FR). Poverty rate are obtained from Mexico's *National Council for the Evaluation of Social Development Policy* (CONEVAL) [143] and the other indicators are obtained from the 2010 Population Census [144]. PR and UR are used in the BURC model as the domain knowledge features $\boldsymbol{s}_i$ to characterize the reporting rate of property crimes ı.e., factors that affect the percentage of crime incidents being reported; while AR, NMR, M/F, M/FHH, FR and PR are used as $\boldsymbol{s}_i$ in the violent crime model.

4. Protected attributes for fairness evaluation include average income and statistics of indigenous population from CONEVAL [143]. The average income is a real-value attribute. I have divided income into quartiles of average income, and assign assign an income group label to each municipality: from *IcQ1* (lowest average income) to *IcQ4* (highest average income). On the other hand, the census identifies 4 types of municipalities determined by the presence of indigenous population (IP): *IP1* characterizes municipalities without indigenous population (there are 5 such municipalities in the dataset); *IP2* are municipalities with less than 40% of the population being indigenous and the indigenous population is less than 5000 (955 municipalities); *IP3* characterizes municipalities with less than 40% of the population being indigenous and the indigenous population is 5000 or more (213 municipalities); and *IP4* that represents municipalities with more than 40% of the

population being indigenous (206 municipalities). These four types of indigenous municipalities, from *IP1* to *IP4*, characterize the increasing presence of indigenous population in a municipality.

4.3.1.2 BURC settings.

The BURC model is implemented using NIMBLE in R [145] and the posterior distribution is inferred by Markov chain Monte Carlo (MCMC) sampling. The basis of MCMC sampling is that when the Markov chain converges, the samples generated by MCMC sampling are the joint posterior distribution of the Bayesian model. The burn-in period is $80,000$ iterations where samplings from MCMC are discarded before the Markov chains converge to the posterior distribution. After the burn-in period, another $80,000$ iterations are used to generate posterior samples with thinning intervals of 40. Four independent chains are used to sample and examine the convergence of the model.

The prior distribution for α and β in the BURC model is computed as follows: α_0 is defined by a normal distribution N(4,2) to make the model conservative to make large prediction of volumes of crimes, *i.e.,* the probability of $\alpha_0 > 8$ (number of true crimes $>$ 2981 given all $\boldsymbol{u}$ features equal to 1) is 2.5%. The prior distribution for β_0 is defined as N(-2, 0.5), because the national survey of victimization in Mexico (ENVIPE) suggests that the under-reporting rate of all crimes is around 88% in 2010 [79] and the inverse logit of -2 is 0.12. The prior distributions for other coefficients, α_k and β_j are defined with N(0,100) which are relatively non-informative priors.

After assessing the convergence of the MCMC sampler, I use the mean point estimate of the parameters to make predictions for each municipality i. Specifically, I compute: 1) the predicted volume of true crimes, which is the expected value of the Poisson distribution for the generation of true crimes, and which is estimated as $\hat{y}_i = \hat{\lambda}_i$; 2) the predicted reporting rate, which is estimated as $\hat{\pi}_i$; 3) the predicted volume of reported crimes, which is the expected value of the Poisson distribution for the generation of reported crimes, and which is estimated as $\hat{z}_i = \hat{\pi}_i \hat{\lambda}_i$.

4.3.1.3 Evaluation

The evaluation focuses on understanding if addressing the under-reporting issue in mobility-based crime predictors improves the fairness and accuracy of the predictive models. To achieve that, I will analyze fairness and accuracy of the proposed BURC model against a battery of three baselines, which are commonly used machine learning models for regression: Random Forest (RF), Bagging (BAG) and XGBoost (XGB). All baselines use random search hyperparameter tuning with a validation set from the training data to select the best hyperparameters. The feature vectors used to train the baselines are a concatenation of hotspot features ($\boldsymbol{u}_i$) and domain knowledge features *e.g.,* unemployment, poverty rates or gender ($\boldsymbol{s}_i$) so that baselines have access to the same information as the proposed BURC model. I used the implementations of RF, BAG and random search from scikit-learn [146] and the XGB implementation from [147].

In the experiment, I use 5-fold cross validation to split the data into training

and testing sets: the 1,379 municipalities are randomly split into 5 folds and in each experiment, 1 fold is used as testing set for evaluation and the 4 remaining folds are used for training models. Model performance and model fairness are reported as averages across all 5 runs. To evaluate the performance of the mobility-based crime prediction models, I use the Root Mean Square Error (RMSE), the Mean Absolute Error (MAE) and the correlation between predicted and *ground truth* volumes of reported crimes. Compared to MAE, RMSE penalizes models that have large errors. Fairness, on the other hand, is measured by the mean difference and group errors as described in Section 4.2.3. Next, I present the main results.

4.3.2 Results

4.3.2.1 Convergence of BURC

Assessing convergence of MCMC based models is important because only when MCMC models converge to the posterior distribution, the samples being drawn from the Markov chain are the samples from the desired posterior. I assess the convergence of the BURC model by autocorrelation tests and Gelman–Rubin convergence diagnostic [148]. Samples from MCMC samplers are not independent *i.e.*, the current sample being drawn is dependent on the previous sample, and thus there is autocorrelation among the posterior samples. Autocorrelation tests compute the autocorrelation with lag k, which is defined to be the correlation between the samples k steps apart. If the MCMC sampler has converged and reached the stationary distribution, the autocorrelation value should be small as k increases and 0 means

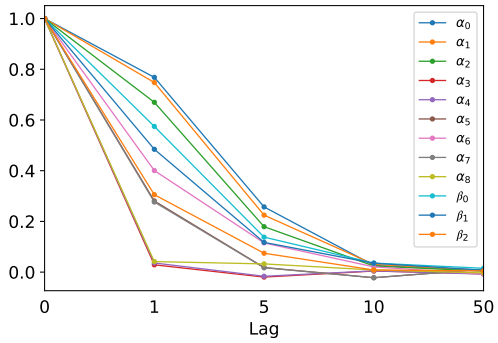


Figure 4.4: Lag-k autocorrelation for the coefficients in BURC for the reported property crime experiment. The autocorrelation for all coefficients drops to zero with lag larger than 10.

samples are independent with samples after k iterations [149]. In the reported property crime experiment, as shown in Figure 4.4, the autocorrelation drops as k increases and eventually converges around 0 after 50 iterations (Figure 4.4). Similar behavior was observed for the violent crime model. Gelman–Rubin convergence diagnostic requires multiple Markov chains with different starting points and assesses the convergence by computing the potential scale reduction factor (PSRF) based on between-chain and within-chain variance. If the MCMC sampler converges, the PSRF is close to 1 [148]. The PSRFs of all coefficients in BURC in both types of crime experiment are less than 1.01 suggesting the BURC model converges well in both experiment and the samples from this sampler can be used to estimate the posteriors.

4.3.2.2 Performance of Reported Crime Prediction

In this section, I compare the BURC model performance against the baselines. Table 4.1 summarizes the experimental results. For property crime prediction, the

	Metric	RF	BAG	XGB	BURC
Property Crimes	RMSE	763.1	803.8	810.8	**601.2**
	MAE	198.4	210.5	211.5	**180.4**
	Correlation	0.73	0.68	0.67	**0.82**
Violent Crimes	RMSE	1301.4	1294.5	1306.0	**1160.0**
	MAE	404.7	406.1	404.9	**346.1**
	Correlation	0.73	0.74	0.73	**0.81**

Table 4.1: Average cross validation performance for baselines and BURC model. BURC model has much lower error and higher correlations than the baselines.

best correlation between actual and predicted crime incidents for the baseline models is 0.73 (Random Forest) while that value increases to 0.82 for the proposed BURC model. As for violent crime prediction, the performance of three baselines is similar and the BURC model still has the highest correlation. This result shows the effectiveness of using urban hotspots features to predict future crime incidents; but more importantly, it also demonstrates that by explicitly modeling under-reporting in crime data, the BURC model can perform better than common machine learning models. In addition to higher correlation, BURC reduces the RMSE and MAE by 21.2% and 9% for property crimes prediction and by 10.4% and 14.4% for violent crimes.

4.3.2.3 Fairness: Mean Difference

In this section, I evaluate the fairness of the BURC and baseline models for two protected attributes, income and presence of indigenous groups, using the mean difference (MD) described in Section 4.2.3. Tables 4.2 and 4.3 summarize the normalized MD for both protected attributes in the property crime prediction, and

	IcQ1	IcQ2	IcQ3	IcQ4	AbsSum
$\boldsymbol{z}$	-1.30	-1.16	-0.89	3.45	6.80
$\hat{\boldsymbol{z}}_{RF}$	-1.20	-0.99	-0.60	2.84	5.63
$\hat{\boldsymbol{z}}_{BAG}$	-1.16	-0.98	-0.56	2.76	5.46
$\hat{\boldsymbol{z}}_{XGB}$	-1.21	-1.05	-0.66	2.98	5.90
$\hat{\boldsymbol{z}}_{BURC}$	-1.28	-1.14	-0.71	3.22	6.34
$\hat{\boldsymbol{y}}_{BURC}$	**-0.87**	**-0.62**	**-0.32**	**1.84**	**3.66**

Table 4.2: MD for protected attribute income group in property crime prediction. The predicted volumes of true crimes with under-reporting correction ($\hat{\boldsymbol{y}}_{BURC}$) are fairer than baselines across all and each of the income groups.

	IP1	IP2	IP3	IP4	AbsSum
$\boldsymbol{z}$	-0.60	-2.04	4.40	-1.07	8.10
$\hat{\boldsymbol{z}}_{RF}$	-0.56	-1.14	2.80	-0.94	5.44
$\hat{\boldsymbol{z}}_{BAG}$	-0.53	**-0.98**	2.51	-0.92	4.94
$\hat{\boldsymbol{z}}_{XGB}$	-0.54	-1.11	2.80	-0.99	5.44
$\hat{\boldsymbol{z}}_{BURC}$	**-0.47**	-1.36	3.27	-1.08	6.17
$\hat{\boldsymbol{y}}_{BURC}$	-0.52	-1.10	**2.37**	**-0.59**	**4.58**

Table 4.3: MD for protected attribute indigenous group in property crime prediction. The predicted volumes of true crimes with under-reporting correction ($\hat{\boldsymbol{y}}_{BURC}$) are fairer than baselines across all groups and are in favor of IP3 and IP4 which have more presence of indigenous population.

Table 4.4 and 4.5 summarize the normalized MD in violent crime prediction. I mostly discuss the MD results for property crime prediction, since results for violent crime follow a similar trend. The first 4 columns in each Table represent the mean difference MD_i between the average volume of crime incidents for group i and all other groups *e.g.,* column *IcQ1* represents the mean difference between the average volume of crimes for group *IcQ1* and the average of crime volumes in municipalities that are not in group *IcQ1*. The last column *AbsSum* is the sum of the absolute mean difference from all four columns, and I use it to evaluate the overall fairness across different groups. On the other hand, the row $\boldsymbol{z}$ represents the mean difference of the actual reported crimes in the testing set (ground truth)

	IcQ1	IcQ2	IcQ3	IcQ4	AbsSum
$\boldsymbol{z}$	-1.28	-1.09	-0.76	3.20	6.32
$\hat{\boldsymbol{z}}_{RF}$	-1.16	-0.96	-0.54	2.70	5.36
$\hat{\boldsymbol{z}}_{BAG}$	-1.11	-0.94	-0.54	2.65	5.24
$\hat{\boldsymbol{z}}_{XGB}$	-1.16	-0.99	-0.55	2.76	5.46
$\hat{\boldsymbol{z}}_{BURC}$	-1.22	-1.01	-0.69	2.98	5.90
$\hat{\boldsymbol{y}}_{BURC}$	**-0.77**	**-0.51**	**-0.19**	**1.47**	2.94

Table 4.4: MD for protected attribute income group in violent crime prediction. The predicted volumes of true crimes with under-reporting correction ($\hat{\boldsymbol{y}}_{BURC}$) are fairer than baselines across all and each of the income groups.

	IP1	IP2	IP3	IP4	AbsSum
$\boldsymbol{z}$	-0.60	-1.95	4.22	-1.02	7.79
$\hat{\boldsymbol{z}}_{RF}$	-0.45	-1.12	2.73	-0.90	5.20
$\hat{\boldsymbol{z}}_{BAG}$	**-0.43**	-1.13	2.69	-0.85	5.10
$\hat{\boldsymbol{z}}_{XGB}$	-0.49	-1.27	2.96	-0.89	5.60
$\hat{\boldsymbol{z}}_{BURC}$	-0.57	-1.61	3.62	-1.02	6.82
$\hat{\boldsymbol{y}}_{BURC}$	-0.46	**-0.85**	**1.84**	**-0.47**	**3.63**

Table 4.5: MD for protected attribute indigenous group in violent crime prediction. The predicted volumes of true crimes with under-reporting correction ($\hat{\boldsymbol{y}}_{BURC}$) are fairer than baselines across all groups and are in favor of IP3, IP3 and IP4 which have more presence of indigenous population.

and evaluates the fairness in the data itself; the row $\hat{\boldsymbol{z}}_{model}$ represents the MD of the predicted reported crimes for the three baseline models and for the proposed BURC model without under-reporting correction (model in formula 4.3, Section 4.2.2); and $\hat{\boldsymbol{y}}_{BURC}$ represents the MD of the predicted volumes of true crimes *i.e.,* volumes of crimes post under-reporting correction as computed by BURC (model in formula 4.2, Section 4.2.2).

Tables 4.2 and 4.3 show that the BURC model addressing under-reporting ($\hat{\boldsymbol{y}}_{BURC}$) has fairer crime predictions across all groups *i.e.,* AbsSum MD is the lowest for both income and presence of indigenous groups. These results highlight that by correcting the under-reporting, the BURC model does a better job at providing

fairer predictions across groups. Looking in depth into each protected attribute, I see that BURC provides the lowest mean difference (highest fairness) across all income groups: from low (IcQ1) to high (IcQ4) average income. However, although BURC has the lowest sum of absolute MD across all four types of indigenous population presence, I observe that the BURC model provides fairest predictions only for groups IP3 and IP4, which are the groups with the largest indigenous population, and those who have traditionally suffered more from biased predictions; while other models that do not correct for under-reporting provide slightly fairer predictions for groups IP1 and IP2, which are those with the lowest percentages of indigenous population and that represent groups that have been traditionally associated to lower biases by prediction models. Similar experiments with violent crimes revealed that BURC achieved the highest fairness for all IP groups except for IP1 (Table 4.5). These results also show that by correcting for under-reported crime rates, the model is slightly positively discriminating in favor of disadvantaged municipalities with mid to high volumes of indigenous people.

Based on results for the mean difference for both property and violent crimes, I have the following high-level observations for both protected attributes: 1) the ground truth reported crimes shows high bias both in terms of income and indigenous groups, as the sum of absolute MD is large; 2) the predictions from all the models have lower MD for each group than the ground truth, suggesting that using urban hotspot features for crime prediction decreases the bias (increases fairness) when compared to the *ground truth*; 3) although BURC's prediction for reported crimes ($\hat{\boldsymbol{z}}_{BURC}$) is less fair in terms of MD, the advantage of BURC is that it can

predict the true crime incidents including those failed to be recorded in the crime statistics. The inferred true crimes ($\hat{\boldsymbol{y}}_{BURC}$) reduce the AbsSum almost by half compared with the reported crimes and is much fairer than the baselines. This suggests that modeling the under-reporting issue can improve the prediction accuracy and fairness at the same time, because BURC disentangles the fairness burden from making accurate predictions. That is, BURC can make accurate predictions for the observed reported crimes ($\hat{\boldsymbol{z}}_{BURC}$) without adding fairness regularization/penalty that sacrifices accuracy [89] and at the same time improves the fairness by inferring the reporting rates ($\hat{\boldsymbol{\pi}}$ in formula 4.5 in Section 4.2.2), which a major source for data bias in reported crime data, and the volumes of true crimes ($\hat{\boldsymbol{y}}_{BURC}$).

4.3.2.4 Fairness: Group Error

In this section, I evaluate the fairness in terms of group error. With this metric, I aim to find the best model with balanced performance for each protected attribute. Together with the mean difference, these two metrics will allow us to identify the best model in terms of performance (lowest error) and fairness. Here I use RMSE to measure performance as the error of the predictions. However, as shown in Tables 4.2 and 4.3, different groups for a protected attribute have different scales for the number of reported crimes *e.g.,* the average of reported crimes in group *IcQ1* is different from the average in group *IcQ4*. Therefore, I calculate not only the absolute RMSE but also the relative RMSE - normalized by the group average of reported crimes - as shown in Tables 4.6 and 4.7. I only discuss property crime

	IcQ1	IcQ2	IcQ3	IcQ4
RF	50.0 (7.4)	96.1 (3.0)	197.0 (2.5)	1507.3 (1.6)
BAG	102.2 (14.8)	105.1 (3.5)	212.9 (2.8)	1579.2 (1.7)
XGB	69.7 (10.1)	87.1 (2.8)	**159.9 (2.0)**	1602.9 (1.7)
BURC	**26.4 (3.9)**	**73.3 (2.5)**	180.6 (2.3)	**1154.3 (1.2)**

Table 4.6: The absolute group RMSE (relative RMSE to the group average in the brackets) for income groups. BURC has more balance performance across all income groups and reduces relative errors in the low income group substantially.

	IP1	IP2	IP3	IP4
RF	**14.5 (5.2)**	**275.4 (2.9)**	1824.0 (1.5)	89.3 (4.0)
BAG	27.7 (10.7)	309.8 (3.3)	1901.3 (1.5)	151.9 (6.5)
XGB	21.5 (11.1)	337.4 (3.4)	1901.1 (1.6)	93.9 (4.3)
BURC	59.6 (19.5)	363.4 (3.6)	**1277.5 (1.0)**	**46.4 (2.0)**

Table 4.7: The absolute group RMSE (relative RMSE to the group average in the brackets) for indigenous groups. BURC reduces substantially the prediction errors for IP3 and IP4, *i.e.,* municipalities with large indigenous population.

prediction results. Violent crime analyses have a similar outcome and are presented in Tables 4.8 and 4.9.

Based on the results for group errors as well as the mean difference, I make the following observations: 1) In terms of absolute errors, BURC substantially reduces the large errors observed in the baselines *e.g.,* the RMSE for *IcQ4* is reduced from 1507.33 to 1154.32 or *IP3* is reduced from 1823.94 to 1277.39; this error reduction allows BURC to make more balanced predictions across different groups, thus increasing accuracy and fairness. This also explains why BURC decreases the RMSE by 21%, a much larger improvement than MAE, as mentioned in Section 4.3.2.2; 2) In terms of relative errors, the prediction errors are distributed more evenly over the income groups when compared to other baselines, and BURC substantially reduces the relative errors in the lowest income group; 3) Although BURC

	IcQ1	IcQ2	IcQ3	IcQ4
RF	102.0 (4.7)	**183.8 (2.3)**	390.0 (2.0)	2560.4 (1.4)
BAG	205.8 (9.6)	195.0 (2.5)	393.0 (2.1)	2547.7 (1.4)
XGB	168.6 (8.0)	226.2 (2.9)	378.6 (2.0)	2568.9 (1.4)
BURC	**98.2 (4.7)**	194.7 (2.4)	**331.9 (1.7)**	**2305.4 (1.4)**

Table 4.8: The absolute group RMSE (relative RMSE to the group average in the brackets) for income groups in violent crime prediction.

	IP1	IP2	IP3	IP4
RF	109.5 (33.9)	**502.5 (2.5)**	3128.0 (1.3)	189.7 (3.0)
BAG	127.5 (31.9)	522.1 (2.6)	3078.0 (1.3)	302.9 (4.7)
XGB	211.8 (93.6)	518.2 (2.6)	3131.6 (1.3)	255.3 (4.0)
BURC	**22.9 (9.8)**	518.4 (2.7)	**2719.8 (1.2)**	**120.2 (1.8)**

Table 4.9: The absolute group RMSE (relative RMSE to the group average in the brackets) for indigenous groups in violent crime prediction.

does not achieve the lowest group errors for all groups, BURC consistently makes good predictions for disadvantaged groups, such as municipalities with low income or municipalities with high percentages of indigenous population. This is meaningful because BURC provides higher confidence in that disadvantaged groups are not unfairly treated in the prediction; 4) BURC performs similarly both in terms of mean difference and group error *i.e.,* BURC has good scores for almost all income groups and for municipalities with large indigenous population, confirming that by addressing under-reporting both performance and fairness can be improved.

4.4 Insights about Crime Occurrence and Under-reporting

In this section, I aim to quantify the influence of different mobility and socio-economic features on the true crime occurring rates and reporting rate in BURC. This analysis will reveal insights that could be used to 1) understand better the

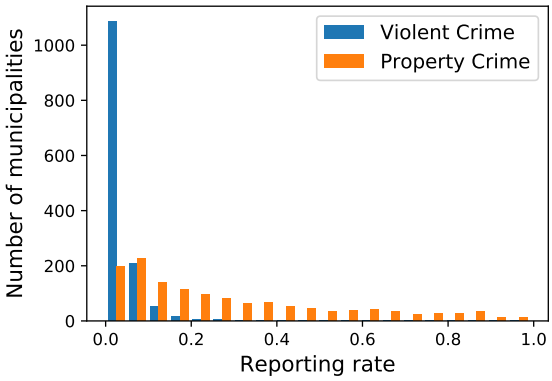


Figure 4.5: Distribution of the reporting rate for violent crimes and property crimes across all municipalities. Violent crimes have more serious under-reporting issue.

relationship between crimes and mobility patterns so as to improve safety in cities, and 2) evaluate the role that demographic and socio-economic data including poverty rate, unemployment or gender play in under-reporting so as to inform policies to encourage reporting.

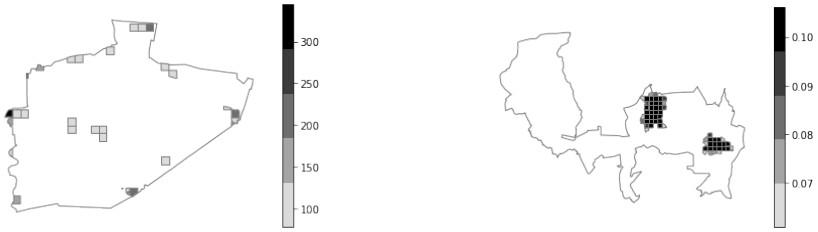
For that purpose, I fit the proposed BURC model with all the reported crime statistics for: (1) property crimes and (2) violent crimes for the $1,379$ municipalities. The distribution of the mean point estimate for the reporting rate for property and violent crimes across all municipalities, reveals a prevalent under-reporting issue with 94% of municipalities having less than 10% of violent crimes being reported (Figure 4.5). These results are consistent with the findings of the ENVIPE survey in Mexico where under-reporting rates were reported to be around 90% from 2010 to 2014 [79]. To understand the role that mobility and socio-economic features play on the true crime occurring rates and reporting rate in BURC, I compute the mean point estimate of the coefficients $\boldsymbol{\alpha}$ in the log link function and $\boldsymbol{\beta}$ in the logistics link function, respectively. Table 4.10 shows the coefficients for both property and violent crimes models.

Coefficient	Feature	Property Crime	Violent Crime
$\boldsymbol{\alpha}$	α_0 (intercept)	0.78 (0.03)	5.89 (0.03)
	log(NHS)	0.93 (0.01)	0.08 (0.01)
	log(AHS)	0.45 (0.01)	1.01 (0.01)
	log(COMP)	0.49 (0.02)	1.05 (0.02)
	log(MCOMP)	-1.49 (0.02)	-1.83 (0.02)
	log(COHE)	-3.89 (0.51)	-12.49 (0.39)
	log(PROX)	-0.71 (0.01)	-0.09 (0.01)
	log(NMI)	5.67 (0.51)	13.55 (0.39)
	log(NMMI)	-2.12 (0.01)	-1.58 (0.01)
$\boldsymbol{\beta}$	β_0 (intercept)	2.71 (0.04)	-27.69 (0.19)
	log(UR)	0.18 (0.01)	/
	log(PR)	-1.59 (0.01)	-0.28 (0.00)
	log(AR)	/	2.08 (0.04)
	log(NMR)	/	4.90 (0.03)
	log(M/F)	/	-0.61 (0.04)
	log(M/FHH)	/	-1.40 (0.01)
	log(FR)	/	0.47 (0.00)

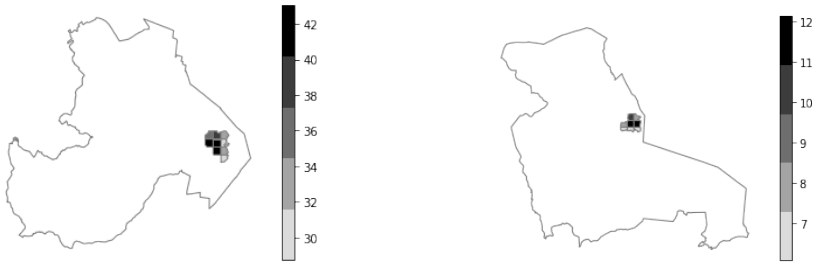
Table 4.10: Mean and standard deviation (Std) for posterior distribution of the coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in the link function for corresponding features.

4.4.1 True Crime Rates Analysis

For the true crime occurring rates, There are one intercept term, α_0, and eight coefficients corresponding to the eight urban hotspot features (expressed in log scale). $\alpha_0 = 0.78$ in the property crime model represents the setting when all urban hotspot features take the value of 1 and for which the true crime occurring rate is 2.18 (α_0 is in the log link function and $exp(0.78) = 2.18$). Similar interpretation applies to the violent crime model. Positive (Negative) coefficients mean that larger (smaller) feature values are associated to the larger (smaller) true crime occurring rates. The coefficients for NHS and AHS are positive, which means that the more hotspots detected *i.e.,* the more active people move around in the municipality, the



(a) large $\hat{y}$ and large $\hat{\pi}$ (b) large $\hat{y}$ and small $\hat{\pi}$



(c) small $\hat{y}$ and large $\hat{\pi}$ (d) small $\hat{y}$ and small $\hat{\pi}$

Figure 4.6: Permanent hotspot distribution in four sample municipalities to show the diverse spatial structure with one or multiple activity centers. The legends represent the footfall per hotspot. Varying levels of predicted volumes of true crimes $\hat{y}$ and reporting rate $\hat{\pi}$ per municipality are reported in Table 4.11.

more crimes there are. For the urban sprawl features, whether or not to weigh the distance between two hotspots by population density has different effects on the crime occurring rate. MCOMP has a negative coefficient and the scale is larger than the coefficient for COMP, suggesting that if the population is more spread out relative to the size of the municipality, the crime incident numbers will be smaller. In fact, having the population more spread out translated into low population density, which means that the potential targets for property crimes are sparse. Note that the maximum values for the urban compactness features (COHE, PROX, NMI and NMMI) is 1 representing the most compact form *i.e.,* the reference circle. The negative coefficient for COHE, PROX and NMMI suggests that the minimum crime occurring rate is achieved in the most compact form; as the hotspots become less

municipality	**(a)**	**(b)**	**(c)**	**(d)**
z	4579	46	9	20
$\hat{z}_{BURC}$	4563.41	65.45	30.96	6.01
$\hat{y}_{BURC}$	4738.33	1176.45	50.32	64.30
$\hat{\pi}$	0.96	0.06	0.62	0.09
NHS	27	60	9	8
AHS	5.42	13.39	1.92	1.43
COMP	0.75	1.20	0.55	0.32
MCOMP	0.79	1.18	0.53	0.31
COHE	0.03	0.13	0.80	0.88
PROX	0.18	0.36	0.92	0.95
NMI	0.03	0.13	0.80	0.88
NMMI	0.02	0.13	0.80	0.88
PR	0.85	41.47	5.30	25.49
UR	4.79	7.52	8.74	2.39

Table 4.11: The ground truth volumes of reported property crimes z, predicted reported crimes $\hat{z}_{BURC}$, predicted true crimes $\hat{y}_{BURC}$, predicted reporting rate $\hat{\pi}$, urban hotspots features and reporting rate determinants, poverty rate (PR) and unemployment rate (UR), for the examples in Figure 4.6.

compact with respect to the equal-area reference circle, the crime occurring rate increases.

As an example to delve into these relationships, Figure 4.6 and Table 4.11 show the detected permanent hotspots and the variables in the BURC property crime models for four example municipalities. These four municipalities have different levels of true crime occurring rate and reporting rate. Recall that the predicted volume of reported crimes is the expected value of the Poisson distribution and therefore could be smaller than the ground truth reported crime; and that the reporting rate is the ratio of predicted reported crimes $\hat{z}$ over predicted true crimes $\hat{y}$. Comparing Figure 4.6(a) and 4.6(b) with Figure 4.6(c) and 4.6(d), I observe that municipalities with high volumes of true crimes tend to have disperse spatial

structure, *i.e.*, have multiple activity centers, while municipalities with low volumes of true crimes tend to be more compact, *i.e.*, only one compact activity center is identified.

4.4.2 Reporting Rates

For the reporting rates, There are one intercept term, β_0, and two coefficients for property crimes and six for violent crimes corresponding to different socioeconomic determinants in log scale. $\beta_0 = 2.71$ in the property crime model reflects that when the PR and UR are 1%, the reporting rate is 93.8% (β_0 is in the logistics link function and inverse logit of 2.71 is 0.938). Similar interpretation applies to the violent crime model. In previous studies about under-reporting of property crimes, when studied independently, higher poverty and unemployment levels are associated to higher under-reporting [80, 81]. Here I model the PR and UR together and the results show that the scale and direction of influence on the reporting rate is different. In the BURC model, poverty rate has a much larger influence on the reporting rate than the unemployment rate. The larger the poverty rate is, the smaller the reporting rate is, as reflected previously in the literature [80]. On the other hand, unemployment rate has a small and positive coefficient, meaning that controlling the influence from the poverty rate, unemployment rate only has a small effect on reporting rate with larger unemployment rate corresponding to slightly higher reporting rate. However, this coefficient is extremely small to draw any conclusions. For violent crimes, the direction of influence of these determinants are mostly con-

sistent with the findings in the literature [72, 78]. Going back to the examples in Figure 4.6 and Table 4.11, when we compare columns (a), (c) with (b), (d) in Table 4.6, we can observe that the poverty rate is much lower when the reporting rate is large than when it is small.

Chapter 5: Study 3: Enhancing Short-term Crime Prediction with Human Mobility Flows: An Analysis of Accuracy and Fairness

5.1 Introduction

Crimes negatively impact the wellbeing of individuals and society as a whole. In 2020, the US saw a significant crime rise across major cities[1]. Researchers from various fields such as criminology, geographic information science, urban planning and data science, have conducted studies about the patterns of urban crimes. These studies help us better understand when and why certain crimes might happen and, more importantly, provide insights into the design of interventions to reduce the volumes of crimes. One critical research direction of such efforts is place-based crime prediction that focuses on predicting the number of crime incidents or crime occurrence for a given location. Environmental criminology provides theoretical foundations to study crimes from the perspective of places [57, 150]. Places with different urban functions can be viewed as crime attractors and crime generators [151]. Through the lens of place-based crime prediction, we can study the com-

[1]https://www.cnn.com/2021/04/03/us/us-crime-rate-rise-2020/index.html

plex relationship between future crimes and historical crimes, built environment and social interactions in different places. Place-based crime predictions are typically carried out using either long-term or short-term approaches. Long-term crime prediction analysis, such as monthly or annual crime prediction, allows us to understand how the environmental factors of places shape future crimes; and in turn, help us inform better urban planning that improves the urban environment potentially decreasing crime occurrence. On the other hand, short-term crime prediction analysis focuses on next-day crime prediction *i.e.,* the identification of places where there will be crimes the next day. Short-term crime prediction is generally used to better allocate policing resources to response to crimes more swiftly.

In this study, I focus on short-term crime prediction analysis. Various models have been developed to tackle this problem. Kernel density estimation - which was very common in the early efforts of crime prediction - uses the estimated density of historical crimes as a measure of risk for future crime areas. [16]. Epidemiological models have also been used to explain crime; for example, Mohler *et al.* proposed an epidemic-type aftershock sequence model to utilize the near repetition patterns of historical crimes [92], whereby the spatio-temporal patterns of crimes in one location increase the probability of other incidents occurring at nearby locations [58]. In addition, the popularity of deep learning has recently brought in several deep learning approaches to model the non-linear spatio-temporal patterns of crime and the built environment [13, 93, 152]. However, although the *Crime opportunity theory* suggests that human mobility is a key factor in crime generation *i.e.*, the higher the presence of suitable *targets* such as people or property, the more

crimes could happen, no studies have incorporated human mobility into short-term crime prediction. In fact, all the empirical studies about the relationship between human mobility and crimes have been conducted in the context of long-term crime estimation or prediction analysis. For example, researchers have shown that the number of people present at a given place can be predictive of crime occurrence in the long term (months) [19]; and human mobility features such as the presence of people, the urban spatial structure or pass-through flows have also been shown to be predictive of the future number of crimes in the annual horizon [18, 21, 153]. To fill the existing knowledge gap, in this study I will explore the effectiveness of using mobility features in short-term crime prediction models that predict the presence of next-day crimes. The main objective is to understand if mobility features can effectively enhance next-day crime prediction when compared to models that do not use mobility data.

With the increasing application of predictive modeling in high stakes social impact settings, algorithmic fairness has become a critical component of predictive systems. Algorithmic fairness, which focuses on understanding and correcting bias in data and algorithms, is especially important for short-term crime prediction models as these models might influence the allocation of public resources such as police patrol scheduling. The debate over data bias issues in crime incident datasets is almost as old as the crime datasets themselves [71]. Quantitative work has shown that bias might be present in crime data due to under-reporting and under-recording issues. For example, low-income and female-headed households are related to crime under-reporting [78]; and research has revealed police under-recording of crimes

associated to certain demographics [153]. As a result, crime predictive algorithms have been shown to replicate and sometimes exacerbate the bias present in the training crime data during the prediction stage. For example, Lum and Issac's work suggested that *PredPol*, a widely used predictive policing system, could be reinforcing the bias already existing in the crime data (feedback loop) by targeting black people roughly at twice the rate of white people for drug abuses [26]. In the second part of this study, I evaluate the algorithmic fairness of the short-term crime prediction models proposed in the first part. I analyze the impact of adding mobility features on the algorithmic fairness of the models designed, when compared to short-term prediction models that do not use such information. Following the steps of prior work [26], I focus the algorithmic fairness analysis on whether short-term crime predictive models achieve similar performance across majority and minority race and ethnicity groups; and I explore the role that both data and algorithmic bias might play in crime prediction models.

The second part of this study will show the presence of pervasive unfairness in short-term crime prediction, which I identify to be partially due to bias in the crime datasets used for the prediction. In fact, as my previous work has shown, crime datasets suffer from under-reporting. To mitigate the data bias caused by under-reporting of crimes, I propose and evaluate a convolutional gate mechanism to model the crime (under-)reporting process in the context of short-term crime prediction. In addition, I also propose a decision-making framework to look at the trade-offs between fairness and accuracy, providing insights into how to balance the two for decision making in policing, and extending current state of the art in short-term

crime prediction that exclusively focuses on accuracy [13, 92, 152].

The proposed experiment in this study requires large datasets with mobility data, which less-developed cities might not have access to. In fact, certain cities might not have the infrastructure or the mobile services might be relatively new, making the mobility data collected insufficient to properly train the mobility-based crime prediction model. This issue can be addressed by transfer learning, a technique used when training data is insufficient. Generally speaking, transfer learning aims to extract the knowledge from one or more source domains (tasks) and to apply that knowledge to a target domains (task) [94]. In the context of urban computing, cross-city transfer learning aims to transfer knowledge from source cities with abundant data resources to target cities where services and infrastructures are not ready or just in place, and where data resources are insufficient. Therefore, I finalize this chapter conducting a preliminary study on cross-city transfer learning to explore the effects of re-utilizing knowledge (models) from data-abundant cities on data-scarce cities by measuring the accuracy of the short-term crime prediction models on the data-scarce cities.

To carry out the design, testing and fairness evaluation of short-term crime prediction models, I use publicly available fine-grained human mobility data based on a large-scale mobile phone dataset from the US [154]. The experimental evaluation is done across four American cities (Austin, Baltimore, Chicago and Minneapolis) and for multiple types of crimes, because crime patterns might differ across geographies and types of crimes, and the predictive models might have different performance across scenarios. Through the comprehensive experiments and analysis, I aim to

answer the following questions:

3A): Does incorporating mobility features improve or hurt the accuracy of short-term crime prediction?

3B): Does incorporating mobility features improve or hurt the fairness of short-term crime prediction?

3C): If incorporating mobility features makes short-term crime prediction less fair, what are the potential factors that contribute to the less fair prediction?

3D): How can the under-reporting process be modelled into short-term crime predictors and how will the accuracy and fairness be affected by it?

3E): Can transfer learning techniques help cities with limited mobility data improve mobility-based crime prediction by leveraging knowledge from cities with abundant mobility data?

5.2 Data

The main objective of this study is to design and evaluate short-term crime prediction models that use human mobility data, and to analyze their fairness. For that purpose, two types of data are needed: crime incidents and human mobility. Next, I describe these and provide general statistics for the four cities evaluated in this study: Baltimore (Bal), Minneapolis (Min), Austin (Aus) and Chicago (Chi). These four cities were chosen based on the diversity of their demographics, as shown in Table 5.1, with Baltimore having majority Black and African-American population, Minneapolis majority White, Austin has a high White and Latino and Hispanic

	% Not Hispanic or Latino, White Alone	% Black or African-American	% Hispanic or Latino	% Asian
Bal	27.54%	62.46%	5.12%	2.59%
Min	59.80%	19.36%	9.58%	6.13%
Aus	49.08%	7.60%	33.64%	7.34%
Chi	33.61%	29.48%	28.89%	6.40%

Table 5.1: The percentage of population across race and ethnicity for the four cities according to the American Community Survey (2019 ACS 5-year estimates)[1]. The cities are: Baltimore (Bal), Minneapolis (Min), Austin (Aus) and Chicago (Chi).

population and Chicago with a balanced mix of White, Black and African-American and Hispanic and Latino communities. Replicating the short-term crime prediction and fairness analysis across these four cities will provide a robust analysis across geographies.

5.2.1 Crime incident data

The crime incident datasets for the four cities are obtained from their open data portals, covering crimes from January to December, 2020[2]. Each crime incident is associated with the crime category it belongs to and with the time and location where it took place. Crime locations are generally geo-coded to the closest street or block in the city, however, to account for the potential spatial precision inaccuracy, I use a 50-meter buffer to associate crime incidents to urban census tracts (a similar approach has been implemented in prior work *e.g.,* Kadar and Pletikosa [18], De Nadai et al. [20]). Although crime incidents could be associated to smaller spatial units, the choice for spatial units is determined by the availability of human mobility data at the census tract level only. I group the crime incidents

[2]Bal: https://data.baltimorecity.gov/; Min: https://opendata.minneapolismn.gov/; Aus: https://data.austintexas.gov/; Chi: https://data.cityofchicago.org/;

		Jan	Feb	Mar	Apr	May	Jun
Property Crime	Bal	28.0%	27.2%	24.6%	22.4%	23.6%	25.0%
	Min	35.0%	33.4%	34.1%	35.3%	37.6%	34.7%
	Aus	32.9%	31.9%	30.6%	30.5%	31.2%	31.5%
	Chi	23.5%	22.6%	19.7%	16.6%	19.6%	20.4%
Violent Crime	Bal	21.6%	21.1%	21.8%	17.0%	21.6%	23.4%
	Min	9.4%	9.3%	10.7%	8.5%	10.3%	13.0%
	Aus	4.0%	3.7%	4.5%	4.2%	5.0%	5.4%
	Chi	11.5%	11.0%	9.9%	8.3%	10.2%	11.6%

		Jul	Aug	Sep	Oct	Nov	Dec
Property Crime	Bal	24.0%	22.7%	24.7%	25.6%	24.2%	21.3%
	Min	41.6%	43.3%	40.7%	41.3%	37.0%	33.2%
	Aus	31.8%	34.3%	35.0%	33.3%	36.1%	34.4%
	Chi	22.5%	23.5%	22.2%	21.0%	19.7%	18.2%
Violent Crime	Bal	23.2%	23.4%	22.4%	22.5%	21.1%	18.6%
	Min	16.4%	14.6%	13.7%	12.9%	10.4%	8.3%
	Aus	5.7%	5.3%	5.2%	4.7%	5.3%	5.2%
	Chi	12.9%	12.8%	12.4%	11.1%	10.8%	9.3%

Table 5.2: Crime occurrence monthly density for the four cities in 2020: Baltimore (Bal), Minneapolis (Min), Austin (Aus) and Chicago (Chi).

into two types: property and violent crimes, and I will evaluate short-term crime prediction and fairness for each type separately. Property crimes include arson, burglary, larceny-theft, and motor vehicle theft; while violent crimes include aggravated assault, forcible rape, murder, and robbery. Table 5.2 shows the monthly crime density for each city throughout 2020, where monthly crime density is computed as the percentage of census tracts with crime incidents during that month. The table shows that the four cities selected generally suffer from higher volumes of property crimes than violent crimes; and that they represent a diverse group with some cities suffering from higher volumes of violent and property crimes than others.

5.2.2 Human mobility data

The pervasive presence of ubiquitous technologies such as smart phones, has allowed for the collection of large-scale human mobility data. Location intelligence companies like SafeGraph, collect pseudonymized mobile GPS location data using SDKs installed on individuals' mobile phones via mobile apps. SafeGraph offers multiple datasets. For this study, I have used daily origin-to-destination flows at the census tract (CT) level from January to December, 2020. This dataset is publicly available (see [154]). To extract this dataset, SafeGraph assigns to each device a home location at the census block group level based on its night-time activity. Then, it tracks for each device all the trips from its home location to points-of-interest (POIs) in SafeGraphs' large POI database. Origin-destination (OD) flows are finally computed by transforming all the home-to-POIs trips to CT(O)-CT(D) trips and by computing the number of devices associated to each OD across all census tracts in a city. OD flow volumes are computed at a daily granularity. Since the devices in SafeGraph's database account for about 10% of the entire population in the U.S., the OD flow volumes are re-scaled by the census population.

Table 5.3 shows general OD flow volume statistics for the four cities under study for the year 2020. For each measure, the table shows the mean and standard deviation of its daily average values across all census tracts in each city. In-city OD flows refer to flows whose origin and destination census tracts (CT(O) and CT(D)) are within the city; while out-of-city OD flows are flows in which either the origin or the destination census tract is outside the city under study. To characterize mobility

	Bal	Min	Aus	Chi
Number of census tracts	200	116	204	809
Volume of in-city OD flow	4040.1 (1733.9)	4004.3 (1653.7)	8167.2 (3866.3)	5307.3 (2821.6)
Volume of out-of-city OD flow	1413.6 (1149.9)	2055.8 (1749.5)	2102.6 (1651.3)	1198.9 (1646.3)
The number of unique census tracts connected by in-city OD flow	38.7 (14.6)	30.5 (10.7)	66.8 (20.2)	61.0 (28.5)
The number of unique counties connected by out-of-city OD flow	14.5 (11.9)	23.6 (20.8)	29.6 (17.2)	15.1 (20.5)
The number of unique states connected by out-of-city OD flow	5.9 (3.3)	7.0 (4.2)	7.7 (4.0)	6.2 (4.0)

Table 5.3: Human mobility flow statistics for the four cities under study: Baltimore (Bal), Minneapolis (Min), Austin (Aus) and Chicago (Chi). The numbers in each cell represent the mean (standard deviation) of the daily average across all census tracts in a given city in 2020. OD flows outside the city are flows that either start or end in a census tract that is not part of the city of interest.

diversity, the table also shows the number of unique census tracts connected by in-city OD flows and the number of unique counties and states connected by out-of-city OD flows. It can be observed that most of the OD flows identified take place within the cities under study, with smaller volumes being associated to trips to counties outside the city, and even a smaller number to trips to other states. Consequently, there is a higher diversity in the number of distinct areas visited inside than outside the city (counties or states). A more detailed description of the features extracted from this dataset is covered in the next section.

It is important to note that human mobility, especially during March and April in 2020, has been greatly affected by the COVID-19 pandemic and the stay-at-home orders. The median travel distance in the US has been greatly reduced from March 15th to early May, 2020, with regional variance [155]. It would be interesting to explore how the pandemic would impact the short-term crime prediction with mobility features, compared with the pre-pandemic period. However, due to the availability of mobility data, this topic is beyond the scope of this study.

5.3 Short-term Crime Prediction with Mobility Flows

As stated in the Introduction, the first objective of this study is to analyze the effect of using mobility features on the accuracy of short-term crime prediction models when compared to crime predictors solely based on historical crime data. In this section, I describe the problem setting for short-term crime prediction with mobility data, present the models which will be used in the analysis, describe the

experimental and evaluation protocols and finalize discussing model performance.

5.3.1 Problem setting

In this study, I focus on placed-based short-term crime prediction for a given city. For that purpose, a city is divided into N spatial units $\mathbf{S} = \{s_1, s_2, ..., s_N\}$ which for this study are defined as census tracts. Census tracts are chosen as spatial units because the human mobility flow dataset is only available at the census tract level. The short-term crime prediction is framed as determining whether there will be at least one crime the next day at a given census tract using prior crime and mobility data for that tract. Crime occurrences at a census tract s_i on day t are denoted as $h_{i,t}$ and $h_{i,t} = 1$ is referred to as a crime hotspot.

For each census tract s_i, two sets of daily predictive features are computed: 1) historical crimes (C), defined as the daily number of past crime incidents; the input sequence for crime prediction at day t is represented as $\mathbf{C}_{i,t} = \{c_{i,t-T}, c_{i,t-T+1}, ..., c_{i,t-1}\}$ with T being the length of the *look-back* period *i.e.,* the time range used to characterize *history* and $c_{i,t-d}$ being the number of crime incidents d days before day t; and 2) mobility features (M), defined as a set of ten daily features extracted from SafeGraph's daily OD matrices and denoted as $\mathbf{M}_{i,t} = \{\mathbf{M}^j_{i,t} | j \in \{1, 2, ..., 10\}\}$ and $\mathbf{M}^j_{i,t} = \{m^j_{i,t-T}, m^j_{i,t-T+1}, ..., m^j_{i,t-1}\}$, where $m^j_{i,t-d}$ is the value of the j-th mobility feature at d days before day t. The ten features identified characterize mobility volumes and mobility diversity. Mobility volume features characterize the daily total number of people going in (inflow) and out (outflow) of a census tract within or

Types	Features
Crimes	Daily number of crimes
Mobility Volumes	Volumes of in-city inflow Volumes of in-city outflow Volumes of out-of-city inflow Volumes of out-of-city outflow
Mobility Diversity	Number of CT connected by in-city inflow Number of CT connected by in-city outflow Number of counties connected by out-of-city inflow Number of counties connected by out-of-city outflow Number of states connected by out-of-city inflow Number of states connected by out-of-city outflow
day of week	Day of week

Table 5.4: Complete list of predictive (input) features for short-term crime prediction models. For census tract s_i, inflow (outflow) means s_i is the destination (origin) of the OD flow.

outside the city under study, which have been shown to be related with the volumes of crime incidents [18, 19, 153]; while mobility diversity features characterize the regional influence, *i.e.*, the number of unique regions visited by in/outflows, including census tracts, counties and states. Past research has shown that crimes committed by visitors are associated to different patterns (behaviors) than those of residents [156]; and that pass-through traffic information improves crime prediction accuracy [21]. Therefore, mobility diversity features are extracted to reflect the connections between the census tract s_i and other regions. Table 5.4 shows a summary of all the features used in the short-term crime prediction models. Besides crime and human mobility data, I also add *Day of week* to the feature set to capture the difference between crime data and human mobility behaviors during weekdays and weekends.

In order to evaluate the effects of modeling crime with mobility features, I consider 3 combinations of input (predictive) features to the model: 1) C: the

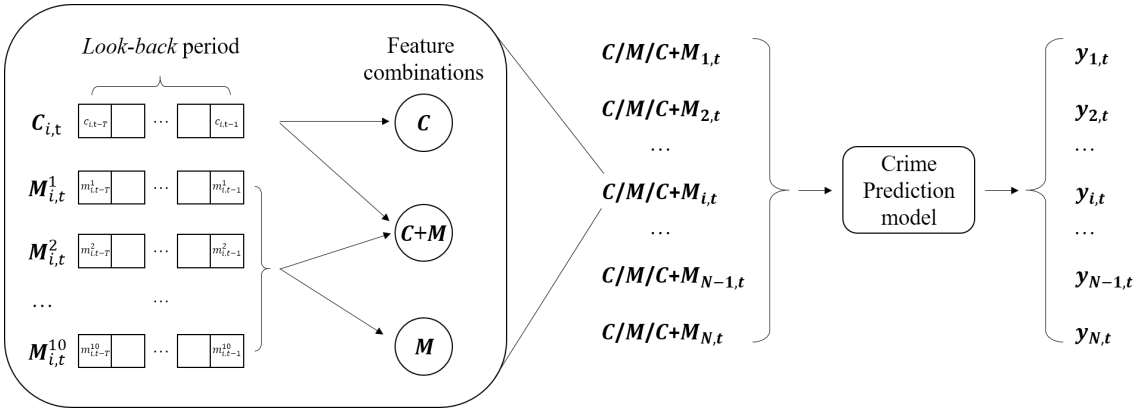


Figure 5.1: Framework of the place-based short-term crime prediction.

input contains only the historical crime features; 2) M: the input contains only the mobility features; 3) $C + M$: the input contains both historical crimes and mobility features.

Problem Statement. Given the temporal sequences of input features ($C/M/C+M$) within the *look-back* period T for all census tracts in a city, predict whether a census tract will be a hotspot in the next day $h_{i,t} = 1, i \in [1, N]$. The framework of the problem setting is shown in Figure 5.1.

5.3.2 Models

I explore a wide variety of state-of-art deep learning models to analyze their predictive power when using crime and/or mobility data as input features.

Historical average logistics regression (HALR). Historical average is a common baseline in crime prediction studies [16, 93]. It predicts the risk score of a spatial unit being a crime hotspot based on the average number of historical crimes for that unit. To incorporate mobility features within this baseline, I add a logistic regression model. The input of the logistic model are $\overline{\mathbf{C}}_{i,t}$ and $\overline{\mathbf{M}}_{i,t}$, which represent

the average of historical crimes and mobility features in the *look-back* period.

Gated recurrent units (GRU). GRU is a variant of recurrent neural networks and is commonly used for modeling sequential data. In this study, multiple layers of GRU are stacked to model the temporal dependency between the probability of being the next-day crime hotspot $h_{i,t}$ and the input temporal feature sequences $\mathbf{C}_{i,t}$ and $\mathbf{M}_{i,t}$ in the *look-back* sequence for census tract s_i.

Attention crime prediction (Attn). Since the success of the Transformer model in natural language processing [157], the attention mechanism has become very popular in modeling sequential data. Here, I use the encoder of the Transformer model with an approach similar to the BERT training setting [158] *i.e.,* a *cls* token is added at the start of the input feature sequences $\mathbf{C}_{i,t}$ and $\mathbf{M}_{i,t}$ in the *look-back* period to predict the probability of crime incidents occurring in census tract s_i in the next day.

Graph convolution network (GCN). By treating all census tracts in a city as nodes in a graph, graph neural networks can be applied to model the spatial dependency of the historical crimes and mobility features among census tracts. In the graph of census tracts, the edges between each pair of census tracts is defined as queen neighbouring (there is an undirected edge between two census tracts if they are queen neighbours, *i.e.*, their boundaries intersect with each other). Graph convolution network (GCN) is one of the earliest neural network architectures for graph structured data [159]. Although more sophisticated graph neural network architectures have been proposed, a simple GCN has been shown to outperform more sophisticated ones if the same hyper-parameter selection and training procedures

Figure 5.2: Arrange the nearest neighbors set for the target census tract s_1 and construct the 2D feature map for historical crimes. In the neighboring set of s_1, s_2 and s_3 is the closest to s_1; s_4 and s_5 are the next closest to s_2 and s_3 respectively; s_6 and s_7 are the next closest to s_4 and s_5; s_8 and s_9 are the next closest to s_6 and s_7. Similar process is applied to each of the ten mobility features.

are used [160]. Therefore, in this study, GCN is adopted for its simplicity and effectiveness for the crime prediction task.

GCN with gated 1D convolution (GGConv). The above deep learning models consider either the temporal or the spatial dependency of the input features for the census tracts. To model the temporal and spatial dependency simultaneously, Yu *et al.* proposed a spatio-temporal convolutional block, which consists of a two gated 1d convolution for the temporal dependency and one GCN layer for the spatial dependency [161]. For this model, the same definition of census tracts graph is used as for the GCN previously described.

Neighbor convolution (NbConv). Neighbor convolution models that account for spatio-temporal dependency have been used for crime prediction using historical data over a spatial grid [93]. To adapt this model to the setting in this study, where the spatial units are census tracts (non-regular division), I extract a fixed-length nearest neighbors set for each census tract for which the model outputs the next-day crime prediction. Specifically, I focus on the eight nearest census tracts for each target census tract. I arrange the target census tract in the middle and sort

the nearest neighboring census tracts from closest to furthest to form a 2D feature map per input feature, as explained in Figure 5.2. Such arrangement allows the kernel of the convolutional layer to model the spatio-temporal dependency through its local receptive field. These 2D feature maps are then input to the full convolution architecture. The original model in [93] contains inception and fractal blocks. In this study, I discuss results for a model with only the first regular convolution blocks because it provided better performance than the full model.

To sum up, HALR is the baseline model that will be used to compare against all the other deep learning approaches. GRU and Attn will be used to test the importance of modeling the temporal dependency of the input features within each census tract, while GCN models will assess the effect of spatial dependencies among neighboring census tracts on short-term crime prediction. Finally, GGConv and NbConv model both the temporal and spatial dependencies of the input features simultaneously, and I will explore whether using such approach is beneficial to improving short-term crime prediction performance when compared to simpler models.

5.3.3 Experiment and Evaluation Protocols

Next, I introduce the experiment and evaluation protocols to evaluate the performance of short-term crime prediction models with mobility features. Given the 1 year of data, I chronologically split the dataset into training (6.5 months), validation (0.5 month), and testing (1 month) sets. The validation set is used to tune the learning rate and early stopping *i.e.,* deciding the maximum number of

epochs for training. Then I re-train the model using the combination of training and validation set (a total of 7 months) and use the testing set to make next-day predictions (5 months). The overall performance of a model is represented by its monthly F1 score, computed comparing the next-day crime prediction with the daily ground truth over all days for each testing month. This experimental protocol with time series data has also been followed in other related work such as Huang et al. [152].

In order to evaluate whether mobility flow features improve short-term crime prediction models, I explore three input feature combinations: 1) Historical crime features only (C); 2) Mobility features only (M) and 3) Historical crimes and Mobility features (C+M). I use the relative change in the F1 score to evaluate the effect of adding mobility features to the short-term crime prediction problem. The F1 score using C combination serves as baseline and the relative change in F1 score using C+M (M) is computed as: $(\frac{F1_{C+M(M)}}{F1_C} - 1) * 100\%$.

5.3.4 Model Implementation and Hyper-parameters

HALR is implemented using its scikit-learn library with the default hyperparameters. All neural network models are implemented with the PyTorch library. The neural networks use Adam as the optimizer with a weight decay of 0.0001 and the learning rate is tuned using the validation set. The dimension of the hidden states for GRU, GCN, GGConv and NbConv is 100. These models have 3 layers of their core blocks *i.e.,* gated recurrent units for GRU, graph convolution for GCN,

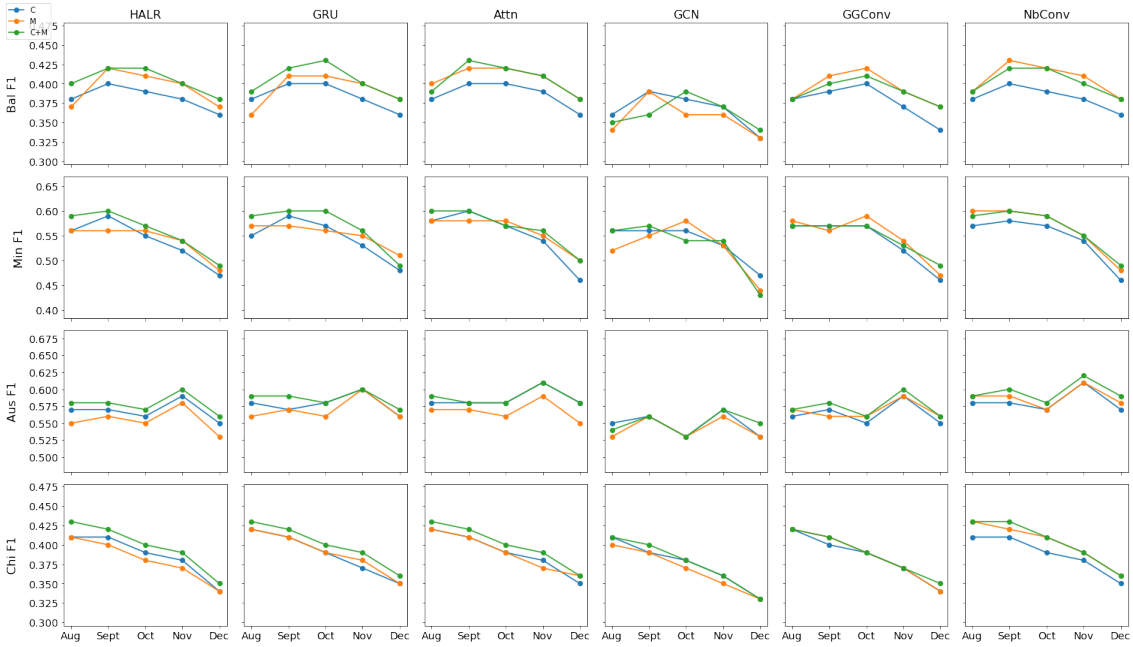


Figure 5.3: Monthly F1 scores for predicting next-day property crime hotspots. Each row represents the F1 scores for one city across all predictive models: Baltimore (Bal), Minneapolis (Min), Austin (Aus) and Chicago (Chi). The blue lines represent F1 scores for models with only crime data (C); the orange lines represent F1 scores for models that use only mobility data (M) and the green line are F1 scores both models that use both (C+M).

spatio-temporal block for GGConv and convolution layer for NbConv. The number of nearest census tracts in NbConv is set as 8. As for Attn, I follow the Mini setting of BERT [3], where the dimensions of the hidden states are 256, the number of attention heads is 4 and the number of layers of attention is 4. The length of the *look-back* period is set to 14 and an analysis of the sensitivity to this parameter is explained in Section 5.3.7.

	Bal	Min	Aus	Chi
HALR	0.403 (0.015)	0.557 (0.039)	0.579 (0.014)	0.398 (0.026)
GRU	0.405 (0.017)	0.567 (0.042)	0.586 (0.010)	0.402 (0.027)
Attn	**0.408** (0.018)	0.564 (0.038)	0.588 (0.011)	0.400 (0.024)
GCN	0.363 (0.016)	0.528 (0.051)	0.551 (0.012)	0.375 (0.030)
GGConv	0.391 (0.014)	0.544 (0.033)	0.576 (0.015)	0.386 (0.027)
NbConv	0.407 (0.013)	**0.571** (0.039)	**0.593** (0.012)	**0.406** (0.026)

Table 5.5: Average (standard deviation) of monthly F1 score using *C+M* for property crime prediction from Aug. to Dec. 2020 for each city.

	Bal	Min	Aus	Chi
HALR	0.390 (0.024)	0.290 (0.056)	0.159 (0.015)	0.269 (0.023)
GRU	0.393 (0.023)	0.284 (0.054)	**0.166** (0.014)	0.267 (0.021)
Attn	0.398 (0.024)	0.284 (0.055)	0.160 (0.012)	0.266 (0.019)
GCN	0.370 (0.025)	0.275 (0.051)	0.149 (0.008)	0.266 (0.020)
GGConv	0.387 (0.028)	0.285 (0.048)	0.152 (0.005)	0.268 (0.021)
NbConv	**0.400** (0.024)	**0.296**(0.047)	0.159 (0.007)	**0.270** (0.022)

Table 5.6: Average (standard deviation) of monthly F1 score using *C+M* for violent crime prediction from Aug. to Dec. 2020 for each city.

5.3.5 Model Performance Analysis

Figure 5.3 shows the monthly F1 scores for predicting property crimes for each model in each city using the three different input combinations: historical crimes only (*C*), mobility features only (*M*) and both (*C+M*). In most cases, the F1 scores using *C+M* are better than using only *C* or *M*; and this observation is true across across cities, test months and models. In other words, adding mobility features improves the predictive accuracy of most of the models explored across all cities. A similar trend was observed for violent crimes.

As *C+M* is the best combination in most cases, I aim to understand what model is giving the best performance. For that purpose, I calculate the average

[3]https://github.com/google-research/bert

Model	Aug	Sept	Oct	Nov	Dec
HALR	3.7%	3.0%	1.4%	2.3%	2.7%
GRU	4.0%	4.3%	2.5%	3.9%	2.4%
Attn	2.9%	2.2%	2.2%	3.2%	3.7%
GCN	1.2%	1.0%	0.3%	0.1%	-1.8%
GGConv	1.2%	0.3%	0.9%	-0.6%	1.3%
NbConv	**4.3%**	**4.9%**	**5.6%**	**4.2%**	**4.5%**

Table 5.7: Relative change in F1 score using $C+M$ for property crime prediction in Chicago in each test month.

monthly F1 score across the five test months for each model and city for both property and violent crimes. Tables 5.5 and 5.6 shows the results. Overall, HALR, GRU, Attn, GGConv, and NbConv have comparable prediction performance and NbConv is the model with best performance in most scenarios, *i.e.*, with the largest F1 scores in three out of four cities for both property crimes and violent crimes. On the other hand, GCN is the model with the worst performance across all scenarios. Since GCN is the only deep learning model in the evaluation that exclusively considers spatial dependency, these results suggest the importance of including temporal dependencies in short-term crime prediction models. I also observe from Figure 5.3 that NbConv is the only model that has the better performance using mobility features only (M) than using historical crimes (C) consistently across different months, cities and types of crimes.

5.3.6 Effects of Mobility Features

To understand the effect of using mobility features in short-term crime prediction models, I compute the relative change in F1 score between using $C+M$ or only M features and the baseline model with only C features, as described in Sec-

Model	Bal	Min	Aus	Chi
HALR	**6.1%**	3.6%	1.6%	2.6%
GRU	5.3%	4.6%	1.3%	3.4%
Attn	5.4%	2.7%	0.5%	2.8%
GCN	-0.6%	-1.2%	0.4%	0.2%
GGConv	4.0%	1.4%	**2.4%**	0.6%
NbConv	5.8%	**4.7%**	2.1%	**4.7%**

Table 5.8: Average relative change in F1 score using *C+M* for property crimes over all test months (Aug-Dec) in each city.

Model	Bal	Min	Aus	Chi
HALR	3.7%	0.2%	-2.9%	-2.3%
GRU	1.9%	2.4%	-1.6%	0.3%
Attn	4.6%	1.7%	-2.9%	0.4%
GCN	-2.8%	-1.7%	-1.5%	-2.0%
GGConv	4.2%	1.8%	0.9%	0.4%
NbConv	**6.0%**	**2.8%**	**1.4%**	**3.3%**

Table 5.9: Average relative change in F1 score using *M* for property crimes over all test months (Aug-Dec) in each city.

tion 5.3.3. This analysis will reveal the effect of using only mobility features or adding mobility features to historical crime features on model performance, when compared to the only crime data baseline. As an example, Table 5.7 shows the relative change in monthly F1 score using *C+M* in Chicago over each test month, from August to December in 2020. I observe that adding mobility features to the models help boost the crime prediction performance in most scenarios (most relative changes are positive for different months and models). However, the improvement of the performance differs across models. NbConv makes the best use of mobility features, *i.e.*, the largest relative improvement in F1 scores in all months in Chicago; while the mobility features sometimes hurt the performance of models with a graph convolution layer: GCN and GGConv have a negative relative change in one month.

Model	Bal	Min	Aus	Chi
HALR	4.1%	2.5%	1.7%	**2.8%**
GRU	3.3%	1.8%	4.0%	1.3%
Attn	3.7%	4.5%	1.9%	2.1%
GCN	0.1%	-0.4%	-3.0%	-1.4%
GGConv	1.9%	2.5%	-0.1%	0.7%
NbConv	**5.0%**	**6.6%**	**7.0%**	2.2%

Table 5.10: Average relative change in F1 score using *C+M* for violent crimes over all test months (Aug-Dec) in each city.

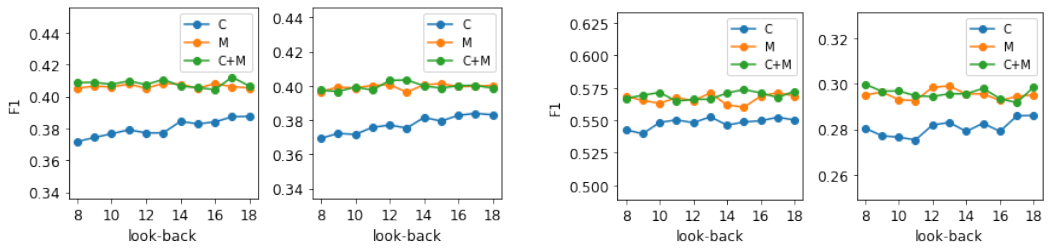
Model	Bal	Min	Aus	Chi
HALR	0.4%	-5.7%	-9.2%	-4.0%
GRU	2.3%	-5.4%	-1.3%	-1.8%
Attn	1.1%	-2.8%	-7.7%	-0.7%
GCN	-1.9%	-1.1%	-2.1%	-4.1%
GGConv	2.4%	-1.4%	0.1%	-0.4%
NbConv	**5.2%**	**5.9%**	**8.2%**	**2.4%**

Table 5.11: Average relative change in F1 score using *M* for violent crimes over all test months (Aug-Dec) in each city.

To be able to analyze the global effect of using mobility features (either *C+M* or *M*) across models, cities and types of crimes, I compute the average relative change over the five test months for each model, city and type of crime and discuss main findings. Tables 5.8 to 5.11 display the results for all combinations described. Based on these average relative changes, I present the following observations:

1) GCN not only has the worst prediction performance but also fails to leverage mobility features, *i.e.*,, the relative changes are mostly negative or small positive values in all cities and types of crimes. In the following observations, GCN is excluded from the analysis.

2) Adding mobility features along with historical crimes as inputs (*C+M*) is consistently beneficial to short-term crime prediction for all cities, types of crimes



(a) Baltimore property crimes (left) and violent crimes (right)

(b) Minneapolis property crimes (left) and violent crimes (right)

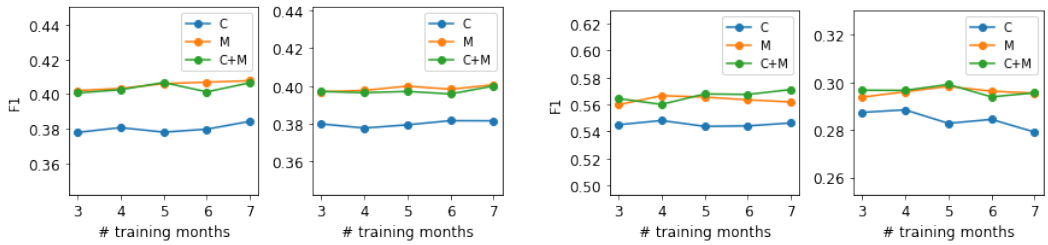
Figure 5.4: Average F1 score in crimes prediction using NbConv across August to December 2020 with different lengths for the look-back period. In the two plots for each city, the one on the left is for property crimes and on the right is for violent crimes.

and models, although the extent of improvement varies (Tables 5.8 and 5.10). NbConv achieves the largest improvement in two cities for property crime and in three cities for violent crimes and the second largest improvement in the rest of the cases.

3) Replacing historical crimes input (C) with mobility features only (M) does not always provide better or comparable crime prediction performance for property crimes (*i.e.*, many relative changes in Table 5.9 are less than 1%) and often hurts prediction performance for violent crimes (*i.e.*, most relative changes in Table 5.11 are negative). The exception is NbConv, whose relative changes using M are consistently positive and improvements are often substantial.

5.3.7 Effect of Length of *look-back* Period and Length of Training Months

In the problem and evaluation setting I have kept two parameters fixed: the length of the *look-back* period is set to 14 and the number of training months (includ-



(a) Baltimore property crimes (left) and violent crimes (right)

(b) Minneapolis property crimes (left) and violent crimes (right)

Figure 5.5: Average F1 score in crimes prediction using NbConv across August to December 2020 with different length of training months. In the two plots for each city, the one on the left is for property crimes and on the right is for violent crimes.

ing the validation set) is set to 7. To investigate the effect of these two parameters on the evaluation, I consider a battery of values for the length of the *look-back* and the number of training months, retrain the best performing model - NbConv - and compute the new F1 scores averaged across all testing months for each of the parameter values considered, city, type of crime and combinations of input features (C+M, M and C). To test the effect of the *look-back*, I consider values ranging from 8 to 18, with the number of training months fixed to 7. To analyze the effect of the number of months, I consider training months varying from 3 to 7, with *look-back* fixed to 14. The results for Baltimore and Minneapolis are shown in Figures 5.4 and 5.5. The results for Austin and Chicago follow similar trends, and thus are not shown in the paper.

It can be observed that the impact of changing the length of the *look-back* period on NbConv models with M and C+M input features is very small, with maximum changes in the F1 score smaller than 1% (see the orange and green lines in the four plots of Figure 5.4). On the other hand, the NbConv model with historical crimes only as input features (C) is slightly more impacted by changes in the length

of the *look-back* period, with F1 scores increasing up to 1.6% as the *look-back* grows until it saturates around *look-back*= 14 (value that is selected for the analysis). These numbers reveal that the improvements in F1 scores for *C+M* and *M* probably represent a lower-bound with potentially larger improvements if the *look-back* period considered was reduced. As for the length of the training months, the impact on F1 scores is also small. I observe maximum F1 score changes of less than 1% and a very slight increase in the F1 score as the length of the training months increases for all input combinations and cities, except for NbConv with input features *C* in Minneapolis. This analysis shows that the F1 scores discussed for NbConv are stable across diverse training lengths.

5.4 Fairness Analysis for Modeling Crimes with Human Mobility

In the previous section, I have shown that the combination of historical crimes and mobility features improves the performance of short-term crime prediction models, with the Neighbor Convolution model (NbConv) being the one with the best F1 scores. In this section, I aim to evaluate the fairness of the proposed model with a focus on the presence of bias that might favor certain race or ethnicity groups, since this has been shown to be an issue for prior crime prediction models [26].

5.4.1 Evaluation Methods for Fairness

The fairness evaluation of short-term crime prediction models can be framed within the field of algorithmic fairness, which is based on the notion of protected

groups. A protected group represents a population sub-group that has historically suffered from discrimination and therefore some form of (approximate) parity or non-discrimination regulation in the predictive algorithm is desired for these groups [82]. Since the discussion around algorithmic fairness for crime prediction has mostly focused on race and ethnicity [26, 162, 163], in this study I evaluate the fairness of the short-term crime prediction results obtained with NbConv with respect to three protected (minority) groups: Black or African-American (BA), Hispanic or Latino (HL), and Asian (A); and one non-protected (non-minority) group comprised of non-Hispanic and non-Latino Whites (W), as defined by the American Community Survey [1].

Fairness is a complex concept and there are different metrics measuring different aspects of fairness. Given the problem setting as a binary classification - positive crime prediction means a census tract is likely to be a crime hotspot in the next day - and my focus on analyzing fairness using protected and non-protected groups, I apply four fairness metrics commonly used in the literature [26, 84, 163]: statistical parity (SP); false positive error rate balance (FPR); false negative error rate balance (FNR); and predicted positive to ground truth positive ratio (the metric used in the study by Lum and Isaac [26], LI). Achieving fairness with respect to these metrics means that the metric value for the protected group should be equal or similar to the metric value for the non-protected group. For example, if $SP_{pg} = SP_{npg}$ then the prediction model is considered to be fair in terms of statistical parity, where *pg* stands for the protected group and *npg* for the non-protected group. Next, I describe each fairness metric in detail and present how I use these metrics to compute

a measure of fairness across protected groups. The four fairness metrics used in this paper are:

Statistical Parity. SP measures the fraction of the population in a (non-)protected group associated to a positive prediction *i.e.,* to a census tract with crime occurring the next day. Fairness in terms of SP suggests that the percentage of the population for a given group associated to positive predictions (crime) should be independent of the group itself, regardless of the ground truth crime data. If the SP_{pg} is larger than SP_{npg}, then the short-term crime prediction model is biased towards protected groups, who would have a higher probability of being associated to crime in next-day predictions than non-protected groups.

The SP metric is computed per (non-)protected group g as follows: $SP_g = \frac{TP+FN}{TP+FP+TN+FN}$ where TP in my setting is defined as the total population of the (non-)protected group g associated to census tracts that were correctly predicted with crime occurring the next day across the whole testing period (August to December 2020); TN is defined as the total population of the (non-)protected group g associated to census tracts that were correctly predicted as non-crime hotspots for the next day across the testing period; FP represents the total population of the (non-)protected group g associated to census tracts that were incorrectly associated to crime occuring the next day across the testing period; and FN refers to the total population of the (non-)protected group g associated to census tracts that were incorrectly predicted as not having crime across the testing period.

False positive error rate balance. FPR measures the fraction of population in a (non-)protected group that is incorrectly associated to a positive predic-

tion *i.e.,* to a census tract with predicted crime occurring the next day, despite the ground truth saying the opposite (no crime). Fairness in terms of FPR suggests that the percentage of errors in the positive prediction should be independent of the population groups. If the FPR_{pg} is larger than FPR_{npg}, the short-term crime prediction model is biased towards incorrectly making larger errors in the prediction of protected groups being involved in crimes. The FPR metric is computed for each (non-)protectd group g as follows: $FPR_g = \frac{FP}{TN+FP}$ with FP and TN defined as explained for the SP metric.

False negative error rate balance. FNR measures the fraction of population in a (non-)protected group that is incorrectly associated to a negative prediction *i.e.,* to a census tract without crimes predicted for the next day, despite the fact that the ground truth points to the presence of crime in that tract. Fairness in terms of FNR suggests that the percentage of errors in the negative prediction should be independent of the population groups. If the FNR_{pg} is smaller than FNR_{npg}, the short-term crime prediction model is biased in incorrectly believing that the non-protected group is less likely to be involved in crimes. The FNR metric is computed for each (non-)protected group g as follows: $FNR_g = \frac{FN}{TP+FN}$ with FN and TP as explained in the SP metric.

Lum and Isaac. LI measures the ratio between (1) the total population of a (non-)protected group associated to predicted crime hotspots by the short-term prediction model and (2) the total population of the same (non-)protected group associated to ground truth crime hotspots i.e., population in census tracts where crime occurrences are predicted versus the population for whom those crime

occurrences are ground truth. Fairness in terms of LI suggests that the (non-)protected groups are represented in the model predictions proportionally to the ground truth crime dataset. If LI_{pg} is larger than LI_{npg}, the protected groups would be over-represented in the predicted hotspots when compared to the non-protected group. The LI metric is computed for each (non-)protected group g as follows: $LI_g = \frac{TP+FP}{TP+FN}$ with TP and FP as explained in the SP metric.

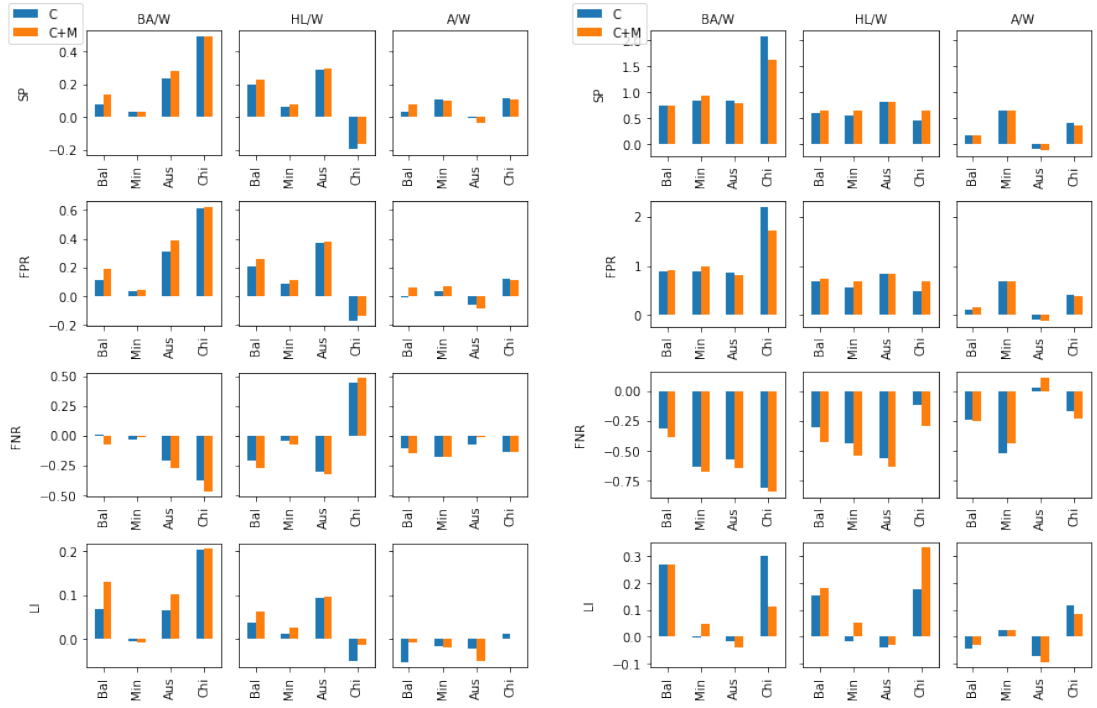
To quantify the degree of unfairness (D) in the short-term crime prediction model, I calculate for each fairness metric described the ratio between each pair of protected and non-protected group and subtract 1. The closer D is to zero, the lower the degree of unfairness associated to the prediction. For SP, FPR and LI, positive D values point to higher degrees of unfairness for the protected groups, while negative D values point to higher degrees of unfairness for the non-protected groups. For example, $D_{BA/W,FPR,C+M} = \frac{FPR_{BA,C+M}}{FPR_{W,C+M}} - 1$ represents the degree of unfairness in crime prediction using historical crimes and mobility features (C+M) in terms of FPR for the protected group BA (Black and African-American) compared to the non-protected group W (non-Hispanic and non-Latino White). If $D_{BA/W,FPR,C+M} > 0$ this reflects a higher degree of unfairness for Black and African-Americans when compared to non-Hispanic, non-Latino Whites. For FNR, positive (negative) D point to higher degrees of unfairness for not-protected (protected) groups. As mentioned earlier, I focus the assessment of the degree of unfairness on the NbConv model since it showed the best performance improvement when using mobility features.

In the next two sections, I first explore the fairness of short-term crime predic-

tion broadly looking into models that use only historical crimes or both historical crimes and mobility data, and analyzing the potential impact of data bias in the findings. In the second section, I will follow with an in-depth analysis of the impact that adding mobility features in short-term crime prediction models has on the fairness of these models, showing the presence of algorithmic bias that exacerbates data bias in some of the models that incorporate mobility data.

5.4.2 Understanding Degree of Unfairness in Short-term Crime Prediction

Figure 5.6 shows the degree of unfairness of the crime prediction task using NbConv for the four cities under study and with respect to different protected race/ethnicity groups and fairness metrics. It can be observed that most of the degrees of unfairness are substantially larger than zero, meaning that unfair crime predictions are commonly observed in the experiments, with protected groups being associated to worse model performances than the non-Hispanic and non-Latino White group. Next I discuss a few examples. I observe that crime predictions tend to have larger FPR values for Black and African-Americans (BA) than for non-Hispanic and non-Latino White Americans (W) which reveals that the short-term crime prediction model is incorrectly predicting more Black and African-American (BA) communities as being involved in crime hotspots. I also observe that crime predictions tend to have lower FNR for non-Hispanic and non-Latino White Americans (W) than for Black and African-American (BA) communities *i.e.,* the short-term



(a) Property crime prediction

(b) Violent crime prediction

Figure 5.6: Degrees of unfairness of crime prediction using NbConv for four cities (Baltimore, Minneapolis, Austin and Chicago) and two types of crime (property and violent). Results are shown for each fairness metric explained in Section 5.4.1 and for each protected group. Crime prediction uses either (i) only historical crime features (C) or (ii) both historical and mobility features (C+M).

crime prediction model is incorrectly predicting less W communities as being involved in crime hotspots. This unfairness issue is consistent with individual level crime risks prediction, where black defendants were far more likely than white defendants to be incorrectly judged to be at a higher risk of recidivism, while white defendants were more likely than black defendants to be incorrectly flagged as low risk by COMPAS, a software used by U.S. courts to assess the likelihood of a defendant becoming a recidivist [164]. The results also show that in terms of statistical parity, minority groups have larger SP than non-Hispanic and non-Latino Whites, revealing that minority communities receive disproportionately larger attention from

the crime prediction model pointing to larger volumes of crime hotspots.

I have shown that short-term crime prediction models that use historical crimes and/or mobility data are associated to worse models performances for protected groups than for non-Hispanic and non-Latino Whites. However, the degree, direction and scale of unfairness varies widely across cities, types of crimes and race/ethnicity. For example, in Figure 5.6(a) for property crimes, $D_{BA/W,FPR,C+M}$ is 0.62 in Chicago *i.e.,* the degree of unfairness for Black and African-Americans in terms of FPR is 62% larger than for non-Hispanic and non-Latino Whites, but is 0.04 in Minneapolis (crime prediction is almost fair for Black and African-Americans in terms of FPR). The direction of unfairness can also be different. For example, $D_{HL/W,FNR,C+M}$ for property crimes in Figure 5.6(a) is positive in Chicago *i.e.,* the model is incorrectly assigning lower risk of being in crime hotspots for Hispanic and Latino communities than the White population; but is negative in the other three cities. The scale of unfairness can be different for different types of crimes. For example, $D_{BA/W,SP,C}$ for violent crimes in Figure 5.6(b) is at least three times larger than for property crimes in Figure 5.6(a) for all four cities.

As pointed out in the algorithmic fairness literature, the bias encoded in the data can be one of the causes of unfairness; the algorithm itself can be another cause [82]. Here, I explore the effect of data bias; the next section will explore the effect of algorithmic bias. To understand if data bias can partially explain the differences in degree of unfairness in crime prediction across types of crimes, cities and race/ethnicity groups, I compute the correlation between the population for each (non-)protected group and the number of crimes in each census tract for each

Race/Ethnicity	Property crime			
	Bal	Min	Aus	Chi
W	0.121	0.287**	-0.020	-0.054
BA	0.267***	0.210*	0.265***	0.483***
HL	0.248***	0.180	0.292***	-0.160***
A	0.181*	0.312***	0.000	-0.021

Table 5.12: Spearman correlation between population in race/ethnicity groups and number of property crimes from August to December 2020 of census tracts in four cities. Significance levels: 0 '***'; 0.001 '**'; 0.01 '*'; 0.05 ' '

Race/Ethnicity	Violent crime			
	Bal	Min	Aus	Chi
W	-0.352***	-0.392***	-0.343***	-0.504***
BA	0.431***	0.734***	0.501***	0.711***
HL	0.013	0.393***	0.557***	-0.230***
A	-0.174*	0.414***	-0.264***	-0.378***

Table 5.13: Spearman correlation between population in race/ethnicity groups and number of violent crimes from August to December 2020 of census tracts in four cities. Significance levels: 0 '***'; 0.001 '**'; 0.01 '*'; 0.05 ' '

city in the study. Specifically, I conduct Spearman's rank correlation coefficients between the population for each race/ethnicity and the number of crimes during the testing period (from August to December, 2020) since the unfairness issues discussed were identified for the testing phase. The correlation results are shown in Table 5.12 and 5.13. Comparing the empirical results in Figure 5.6 and Table 5.12 and 5.13, I observe that the degrees of unfairness mostly align with the direction and coefficient of the correlation *i.e.,* unfairness can be partially explained by the bias in the crime data. A few more in-depth observations:

1) The coefficients of the correlation for non-Hispanic and non-Latino White (W) are significantly negative in the four cities for violent crimes; while they are a mix of positive and negative, and often times at a smaller scale, for the protected

groups (BA/HL/A). For property crimes, the correlation coefficients for different race/ethnicity groups are smaller and closer in range. This means that the ground truth crime incident data is biased towards associating protected groups with higher numbers of violent crimes than of property crimes. This is aligned with my prior observation for the degree of unfairness, where the crime prediction is more unfair for violent crimes than for property crimes (see Figure 5.6).

2) The correlations can also partially explain the difference in degree of unfairness across cities. For example, the coefficients across different race/ethnicity groups for property crimes in Minneapolis are similar, and Minneapolis has most of the smallest degrees of unfairness in property crimes for BA/W, HL/W, and A/W (see Figure 5.6). On the other hand, it can be observed that some large correlation coefficient differences between protected groups and non-Hispanic, non-Latino Whites for property crimes, *e.g.*, 0.483 for Black and African-America versus -0.054 for Whites in Chicago, or 0.292 for Hispanic-Latino versus -0.020 for Whites in Austin. These examples correspond to the largest degrees of unfairness in Figure 5.6.

3) The direction of unfairness also aligns with the direction of the correlation. For example, in property crime prediction, the direction of unfairness for Hispanic-Latino across all fairness metrics is the opposite to that of all the other groups. $D_{HL/W,FNR,C+M}$ is positive which is different from other cities. So is the direction of correlation for Hispanic-Latino in Chicago, which is significant and negative as opposed to all the other protected groups across all four cities.

These observations suggest that the data bias in the crime incident data plays

an important role in the unfairness issues of short-term crime prediction models using historical crimes and mobility features.

5.4.3 Effects of Modeling Crimes with Mobility Features on Fairness

I have shown that short-term crime prediction models that use historical crime and/or mobility features suffer from unfairness issues that might be partially explained by data bias. In this section, I take a deeper look and evaluate the changes in the degree of unfairness brought by adding mobility features to the crime prediction models. Although crime prediction with additional mobility features has consistent better performance for all cities and types of crimes in this study (see Tables 5.8 and 5.10), it is important to understand the effect that mobility features have in the degree of unfairness across cities, types of crimes and races/ethnicity.

To carry out the fairness analysis, I compare the degree of unfairness in crime prediction using historical crimes and mobility features (C+M) to using historical crimes only (C) in order to evaluate how additional mobility features might affect the fairness of the crime prediction; and I use a 5% relative change in the degree of unfairness as a threshold to determine whether using mobility features impacts or not the fairness of crime prediction. For example, $D_{BA/W,FPR,C+M} = \frac{FPR_{BA,C+M}}{FPR_{W,C+M}} - 1$ represents the degree of unfairness in crime prediction using historical crimes and mobility features (C+M) in terms of FPR for the protected group BA (Black and African-American) compared to the non-protected group W (non-Hispanic and non-Latino White). If $\frac{D_{BA/W,FPR,C+M}}{D_{BA/W,FPR,C}} > 1.05$, adding mobility features to the predictive

		Property crime prediction			
		Bal	Min	Aus	Chi
BA/W	SP	- (1.83)	+ (0.88)	- (1.19)	
	FPR	- (1.67)	- (1.12)	- (1.24)	
	FNR	- (16.00)	+ (0.43)	- (1.30)	- (1.25)
	LI	- (1.91)	- (1.85)	- (1.60)	
HL/W	SP	- (1.15)	- (1.24)		+ (0.84)
	FPR	- (1.22)	- (1.29)		+ (0.81)
	FNR	- (1.30)	- (1.69)	- (1.09)	- (1.08)
	LI	- (1.69)	- (2.06)	- (1.06)	+ (0.26)
A/W	SP	- (2.67)		- (4.83)	+ (0.91)
	FPR	- (11.79)	- (1.78)	- (1.54)	+ (0.93)
	FNR	- (1.44)		+ (0.12)	
	LI	+ (0.15)	- (1.26)	- (2.35)	+ (0.19)
C+M improvement over C		5.8%	4.7%	2.1%	4.7%

Table 5.14: Change in degree of unfairness using crimes and mobility features (C+M) compared to using historical crimes only (C) for property crime prediction. For each cell in table: "-" means $\frac{|D_{BA/W,f,C+M}|}{|D_{BA/W,f,C}|} > 1.05$, using M makes crime prediction less fair; "+" means $\frac{|D_{BA/W,f,C+M}|}{|D_{BA/W,f,C}|} < 0.95$, using M makes crime prediction more fair; Blank means $0.95 \leq \frac{|D_{BA/W,f,C+M}|}{|D_{BA/W,f,C}|} \leq 1.05$, using M has little effect on crime prediction fairness. "C+M improvement C" is from Table 5.9.

model increases the degree of unfairness, *i.e.*, the prediction with additional mobility features is less fair than using historical crimes only.

Table 5.14 and 5.15 show the results for the analysis. Negative signs represent settings where adding mobility features increases the degree of unfairness by at least 5%, while positive signs represent settings where adding mobility features to the short-term crime prediction decreases the degree of unfairness by at least 5%. The values between parentheses represent the ratio and the blank values in the table are associated to changes smaller than 5%. For example, the table shows that for Baltimore City, the crime prediction improvement brought in by incorporating mobility features comes at the price of fairness, with all relative changes in the

		Violent crime prediction			
		Bal	Min	Aus	Chi
BA/W	SP		- (1.11)	+ (0.95)	+ (0.78)
	FPR		- (1.12)	+ (0.95)	+ (0.78)
	FNR	- (1.21)	- (1.06)	- (1.12)	
	LI		- (465.8)	- (2.46)	+ (0.37)
HL/W	SP	- (1.06)	- (1.20)		- (1.42)
	FPR	- (1.07)	- (1.20)		- (1.41)
	FNR	- (1.42)	- (1.23)	- (1.13)	- (2.59)
	LI	- (1.17)	- (3.73)	+ (0.84)	- (1.86)
A/W	SP	- (1.11)		- (1.23)	+ (0.90)
	FPR	- (1.46)		- (1.21)	+ (0.89)
	FNR		+ (0.85)	- (3.71)	- (1.41)
	LI	+ (0.66)	+ (0.93)	- (1.31)	+ (0.73)
C+M improvement over C		5.0%	6.6%	7.0%	2.2%

Table 5.15: Change in degree of unfairness using crimes and mobility features (C+M) compared to using historical crimes only (C) for violent crime prediction. For each cell in table, '-', '+' and blank has the same meaning as Table 5.14. "C+M improvement C" is from Table 5.11.

degree of unfairness being negative except for the LI metric for the Asian group. For Minneapolis, using mobility features in crime prediction also causes the prediction to be less fair most of the times, except for property crime predictions for the Black and African-American group. On the other hand, the effect of adding mobility features on prediction fairness for Austin and Chicago is more diverse: 10 and 13 out of the 24 degrees of unfairness are improved by using mobility features in Austin and Chicago respectively. In other words, Chicago is the least likely to provide unfair predictions among the four cities in this study when adding mobility features to the short-term crime prediction.

To explore the reasons behind the diverse effects of mobility features on the degree of unfairness, I design two analyses. The first one focuses on understanding

the effect of mobility data bias *i.e.,* evaluate if the changes in degrees of unfairness can be partially caused by mobility data being biased for certain population groups. The second analysis will evaluate algorithmic bias *i.e.,* analyze whether the NbConv algorithm trained with mobility and crime data might be exacerbating existing data bias and predicting crime hotspots at a higher rate for protected groups than when using only crime data.

To explore the effect of **mobility data bias**, I compute four multivariate linear regressions for each city under study where the dependent variable is the population of each (non-)protected group and the independent variables are the ten mobility features in Table 5.4. Table 5.16 shows the R^2 for the linear regressions for each (non)protected group and each city (all linear regressions are significant at the 0.001 level). R^2 measures the level of variance of population being explained by the mobility features for each (non-)protected group. The larger the R^2 is, the better the mobility features serve as proxies to reconstruct information about population of each race/ethnicity group, and thus the larger potential for *indirect discrimination* [165] through mobility data bias. Any correspondence between the mobility data bias (measured as R^2 of the linear regression in Table 5.16) and the changes in the degree of unfairness brought by additional mobility features (Table 5.14 and 5.15) would point to biased mobility data as one of the causes behind the unfair results reported towards protected groups. However, comparing the two tables it is evident that the role of mobility data bias is minimal on the degrees of unfairness reported. For example, the R^2 for Baltimore and Hispanic-Latinos (HL) is the smallest (0.315) but using the crimes and mobility features (C+M) increases

	Bal	Min	Aus	Chi
W	0.767	0.713	0.802	0.678
BA	0.815	0.510	0.448	0.341
HL	0.315	0.487	0.794	0.456
A	0.446	0.609	0.437	0.389

Table 5.16: R^2 of multivariate linear regression with the population of each (non-)protected group as the dependent variable and all mobility features as the independent variables for each city. All linear regressions are significant at the 0.001 level.

the degree of unfairness of HL/W for both property and violent crime prediction compared to using historical crimes only (C). I also observe that while the R^2 for Chicago and Black-African Americans is the second smallest (0.341), using C+M increases the degree of unfairness of BA/W for property crime prediction (only in terms of FNR), and decreases the degree of unfairness for violent crime prediction. In addition, the R^2 for Minneapolis and Hispanic-Latino (0.487) is similar to the R^2 for Chicago and Hispanic-Latino (0.456). However, the degree of unfairness for HL/W increases for both types of crimes in Minneapolis but decreases for property crimes in Chicago. To sum up, these results suggest that the mobility data bias has little impact on the change in degree of unfairness when using C+M features.

To explore the effect of **algorithmic bias**, I analyze the difference in the racial distribution across the next-day crime hotspots predicted by the NbConv model. Specifically, I calculate the ratio of protected population to non-Hispanic, non-Latino White population involved in the predicted crime hotspots by the NbConv model when using either C+M or only C features. Then, I calculate the relative change between the two ratios to see how adding mobility features to the crime prediction might change the racial distribution in the next-day crime prediction. The ratio

Race/ Ethnicity	Property crime prediction				Violent crime prediction			
	Bal	Min	Aus	Chi	Bal	Min	Aus	Chi
BA/W	5.80%	-0.39%	3.63%	0.14%	-0.05%	4.88%	-2.35%	-14.74%
HL/W	2.52%	1.39%	0.49%	3.96%	2.31%	7.00%	0.63%	12.97%
A/W	4.60%	-0.41%	-2.96%	-0.88%	1.55%	-0.18%	-2.34%	-2.83%

Table 5.17: Change in ratio of minority population to non-Hispanic or Latino White population being involved in crime hotspots predicted by NbConv model with C+M feature combination and with C feature combination.

change is defined as $\frac{TP_{pg,C+M}+FP_{pg,C+M}}{TP_{W,C+M}+FP_{W,C+M}} / \frac{TP_{pg,C}+FP_{pg,C}}{TP_{W,C}+FP_{W,C}} - 1$, where pg is a protected group (BA/HL/A) and TP/FP is explained in Section 5.4.1. The ratio change being positive (negative) suggests that adding mobility features in crime prediction models increases (decreases) the presence of protected groups in the predicted crime hotspots with respect to the non-protected group (the non-Hispanic or Latino White). All ratio changes are shown in Table 5.17.

Looking into the ratio changes in Table 5.17 and the change in the degrees of unfairness that I discussed in Table 5.14 and 5.15, I make three observations: 1) When the ratio change is positive, the degree of unfairness increases when mobility features are added to crime prediction. For example, the ratio change of BA/W in Baltimore is 5.8% for property crime prediction and the degree of unfairness increases for all four fairness metrics. 2) When the ratio change is negative, adding mobility features makes crime prediction more fair. For example, the ratio change of BA/W and A/W in Chicago for violent crimes are -14.74% and -2.83% and the degree of unfairness decreases in terms of SP, FPR and LI. 3) When the ratio change is close to zero, adding mobility features make crime prediction less fair in some fairness metrics. For example, the ratio change of BA/W in Chicago for

property crime is 0.14% and is −0.05% for violent crime in Baltimore. The degree of unfairness only increases in terms of FNR.

There are three exceptions to the pattern described above. The first one is the ratio change of A/W for property crime prediction in Austin. This ratio change is −2.96%, but the degree of unfairness only decreases for FNR and increases for the other three fairness metrics ($SP/FPR/LI$). This suggests that adding mobility features to property crime prediction in Austin mostly balances the false negative rate between Asian and non-Hispanic, non-Latino White population group. The second one is the ratio change of HL/W for property crime prediction in Chicago. The ratio change is 3.96%, but $D_{HL/W}$ decreases for three fairness metrics in Table 5.14. This is because the direction of the unfairness for HL/W in Chicago is different from the other three cities. By increasing the presence of Hispanic and Latino in the predicted next-day crime, the crime prediction model with mobility features improves the fairness. The last exception is the ratio change of A/W for violent crime prediction in Austin. Similar to the second exception, the direction of unfairness for A/W in Austin is different the other three cities. A decreasing presence of Asian population in predicted crime worsens the unfairness of the prediction.

Despite these three exceptions, all the other trends described earlier point to the presence of algorithmic bias that might partially explain the degrees of unfairness *i.e.,* NbConv with mobility and crime data might be exacerbating the crime data bias more than NbConv with only crime data as shown by the increase in the presence of protected groups for the crimes predicted when using both mobility and crime data versus only crime data.

5.5 Improving Fairness in Short-term Crime Prediction with Under-reporting-aware Models

In the previous section, I have shown that unfair predictions are pervasive in short-term crime predictions using either historical crimes only or crimes and mobility features combined. One of the factors contributing to the unfairness is the inherent data bias in the reported crime incidents data. As described in Section 2.3, an important source of data bias of reported crimes is the under-reporting of crimes embedded in the crime reporting process. In this section, I will introduce an under-reporting-aware deep learning model, *i.e.*, models that explicitly take into account the under-reporting issue of crime incident data, with a convolutional gate mechanism to model the under-reporting of crimes and evaluate the effect that this mechanism has on the accuracy and fairness of the short-term crime prediction with mobility features.

5.5.1 Modeling Crime-reporting Process with a Convolutional Gate

As mentioned in Section 2.3, the crime reporting process can be simplified as two stages: 1) a crime incident is reported to the police and 2) the reported incident is recorded in the police database. Inheriting notation from the Study 2, the number of *true* crimes $y_{i,t}$ is defined as the actual number of crimes that will occur regardless of whether they will be reported; and the reporting rate π_i quantifies the under-reporting issue during the two-stage crime reporting process as

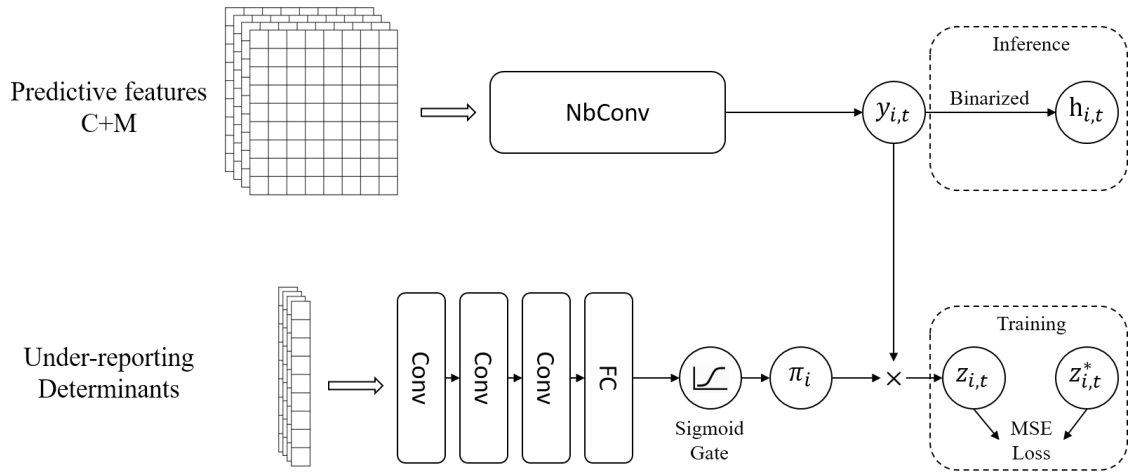


Figure 5.7: Under-reporting-aware short-term crime prediction with crime-reporting convolutional gate. The 2D feature maps for historical crimes, mobility features and under-reporting determinants are constructed based on the neighboring set for census tract s_1 in the same way as shown in Figure 5.2.

the ratio of the number of reported crimes $z_{i,t}$ to the true crimes, where i refers to census tract s_i and t refers to day t. To model the crime-reporting process, $y_{i,t}$ is considered as a function of the predictive features for crimes, *e.g.*, the features extracted from historical crimes and from the human mobility dataset in Section 5.3.1; and π_i is treated as a function of under-reporting determinants based on domain knowledge. Finally, $z_{i,t}$ is the product of the number of true crimes and reporting rate $z_{i,t} = y_{i,t} \times pi_i$.

The under-reporting-aware short-term crime prediction model consists of two neural network branches, as shown in Figure 5.7. The first branch is the true crime predictor which infers the number of true crimes $y_{i,t}$ in the next day t with predictive features for crimes. Since NbConv with historical crimes and mobility features (C+M) is the best performing model in Section 5.3.5, in this section, I apply NbConv as the predictive model for $y_{i,t}$. The second branch is the crime-reporting convolutional gate to infer the reporting rate π_i based on the under-reporting determinants.

As described in Section 4.2.2, the crime-reporting behavior or the reporting rate can be modeled by under-reporting determinants. The convolutional gate proposed here not only models the non-linear relationship between the determinants and reporting rate π_i of the target census tract s_i, but also between π_i and the determinants of the neighboring set of s_i to capture the spatial dependency. The convolutional gate consists of three layers of convolutional blocks (Conv), a fully-connected layer (FC) and a Sigmoid gate.

5.5.2 Experiment Setting

The under-reporting determinants on the census-tract-level are socio-demographic variables obtained from 2019 American Community Survey (ACS) 5-year estimates [4]. As described in Section 4.2.2, the determinants for property crimes are poverty rate (PR) and unemployment rate (UR); and for violent crimes are poverty rate (PR), adult rate (AR), the percentage of people who are never married (never married rate, NMR), male to female ratio (M/F), percentage of female householder with children under 18 years old (FHHR), percentage of people who cannot speak English (linguistic isolation rate, LIR) and the percentage of foreign born population (foreign-born rate, FR). Since ACS provides not only the estimates but also the margin of error of the under-reporting determinants, I include both the estimates and margin of error in the feature maps for the convolutional gate to model the uncertainty in ACS estimates.

The training and inference process for under-reporting-aware short-term crime

[4] https://data.census.gov/

prediction is different from the under-reporting-unaware models in Section 5.3 which are directly trained and make inference on the binary variable $h_{i,t}$, *i.e.*, whether census tract s_i is a hotspot on the next day t. Instead, numeric inferred variables ($y_{i,t}$, π_i and $z_{i,t}$) are introduced for the under-reporting-aware model. To obtain predictions for the binary variable $h_{i,t}$, the under-reporting-aware model follows the training and inference process as shown in Figure 5.7. First, the under-reporting-aware NbConv is trained in the regression setting with mean squared error (MSE) as the loss function $Loss = \frac{1}{N*TD} \sum_{i \in N} \sum_{t \in TD} \left(z^*_{i,t} - z_{i,t}\right)^2$, where $z^*_{i,t}$ is the ground truth and $z_{i,t}$ is the predicted number of reported crimes for census tract s_i on day t, N is the total number of census tract in a city and TD is the number of days in the training period. In the training process, both the branch for true crimes $y_{i,t}$ and for reporting rate π_i of the model are activated to infer the number of reported crimes $z_{i,t}$. Because only the ground truth of the reported crime incidents is available, the goal of the training process is to to minimize the error in predicting the number of reported crimes. Then in the inference phase, only the first branch for $y_{i,t}$ is utilized to predict next-day crime hotspots h_i. $y_{i,t}$ is binarized as $h_{i,t}$ as follows:

$$h_{i,t} = \begin{cases} 1, & y_{i,t} > \overline{y}_t \\ 0, & y_{i,t} \leq \overline{y}_t \end{cases} \tag{5.1}$$

$$\overline{y}_t = \frac{1}{N} \sum_{i \in N} y_{i,t} \tag{5.2}$$

$\overline{y}_t$ represents the average predicted number of true crimes across all census

tracts in the city on day t and Equation 5.1 means that a census tract is considered a hotspot on the next day t if the predicted number of true crimes is larger than the average of all census tracts.

The rest of experiment setting is the same as Section 5.3, including evaluation protocols, implementation and hyper-parameters.

In order to evaluate the effects of modeling under-reporting on accuracy and fairness, I define two baselines. 1) the under-reporting-unaware NbConv model using historical crimes and mobility features (C+M features), denoted as UU. 2) UU with a baseline fairness improvement method proposed in [87] which adds a fairness regularization to the loss function to minimize the per-capita score between protected and non-protected group. For this study, since the score corresponds to the predicted number of reported crimes, we compute the per-capita score as the predicted number of crimes divided by the population of the (non-)protected group. The fairness regularization, named as individual-based fairness gap (IFG) in [87], is computed as:

$$Loss_{IFG,t} = \frac{1}{\sum_{i \in N} z^*_{i,t}} \left| \frac{\sum_{i \in N} z_{i,t} w_i^+}{\sum_{i \in N} p_i w_i^+} - \frac{\sum_{i \in N} z_{i,t} w_i^-}{\sum_{i \in N} p_i w_i^-} \right|, \tag{5.3}$$

where p_i is the total population of s_i and w_i^+ (w_i^-) is the percentage of population in the protected (non-protected) groups. In this study, protected groups refer to Black and African-American, Hispanic and Latino, and Asian population. Non-protected group refers to non-Hispanic and non-Latino White population. The second baseline is denoted as IFG. The under-reporting-aware model proposed in this

	Property crime				Violent crime			
	Bal	Min	Aus	Chi	Bal	Min	Aus	Chi
UU(C)	0.345	0.533	0.533	0.309	0.369	0.299	0.124	0.210
UU	0.405	0.543	0.576	0.362	0.410	0.331	0.201	0.286
IFG	0.403	0.538	0.575	0.315	0.372	0.213	0.090	0.199
TC	0.386	0.444	0.559	0.362	0.366	0.287	0.151	0.215

Table 5.18: Average monthly F1 score for property and violent crime prediction from Aug. to Dec. 2020 for each city. UU(C) means UU model but with historical crimes only as input features.

study is denoted as TC (predicting crime hotspots based on inferred *true crimes*).

We inherit the performance metric for the prediction accuracy defined - in Section 5.3.5 - as the average of the monthly F1 score. The algorithmic fairness is also measured by the degree of unfairness (D) as described in Section 5.4.1. Here I also use a 5% relative change as the threshold to determine whether the TC approach improves fairness compared to the baselines. If the ratio between the degree of unfairness of TC and UU (denoted as TC/UU) or between the degree of unfairness of TC and IFG (denoted as TC/IFG) is smaller than 0.95, then the TC approach improves fairness for a specific metric and protected group. For example, if $\frac{D_{BA/W,FPR,TC}}{D_{BA/W,FPR,UU}} < 0.95$, the TC approach improves fairness for Black and African-American community in terms of false positive rate.

5.5.3 Analysis of Accuracy and Fairness for Under-reporting-aware NbConv Model

Accuracy. Table 5.18 shows the F1 scores for property and violent crime prediction in the testing phase. It can be observed that both the baseline fairness improvement method (IFG) and the proposed under-reporting-aware model (TC)

tend to have lower F1 scores across cities for both types of crimes. This reflects a potential trade-off between accuracy and fairness. In fact, as I later discuss, the two models proposed generally have higher fairness measures at the cost of a decrease in accuracy. For IFG, the trade-off is due to the added fairness regularization, that is, instead of minimizing solely the MSE (the error between predicted and ground truth number of reported crimes), IFG balances between the MSE and the fairness regularization which leads to an increase in the error and a decrease in the accuracy. While for TC, since only reported crime data is available, there is no ground truth to evaluate the actual accuracy of the predicted *true* crime hotspots. The mismatch between the predicted true crime hotspots and the ground truth reported crime hotspots might be the cause of decrease in accuracy. For example, a census tract with low reporting rate could be considered as a hotspot based on the inferred number of true crimes, but not a hotspot based on the reported crime data. In study 2, the Bayesian model for under-reported crimes is observed to improve reported crime prediction accuracy as well as the fairness in long-term crime prediction. However, the TC approach has an non-trivial impact in decreasing accuracy for short-term crime prediction. This suggests that incorporating Bayesian models into deep learning might be helpful in reducing the loss in accuracy. I think this is an important line of inquiry, and I leave this for future research beyond my dissertation work. Although TC has a decrease in accuracy compared with UU, TC still has similar accuracy with UU(C), which means the improvement in accuracy by the additional mobility features has been offset by the convolutional gate mechanism.

Fairness. Figure 5.8 and 5.9 shows the degree of unfairness of crime prediction

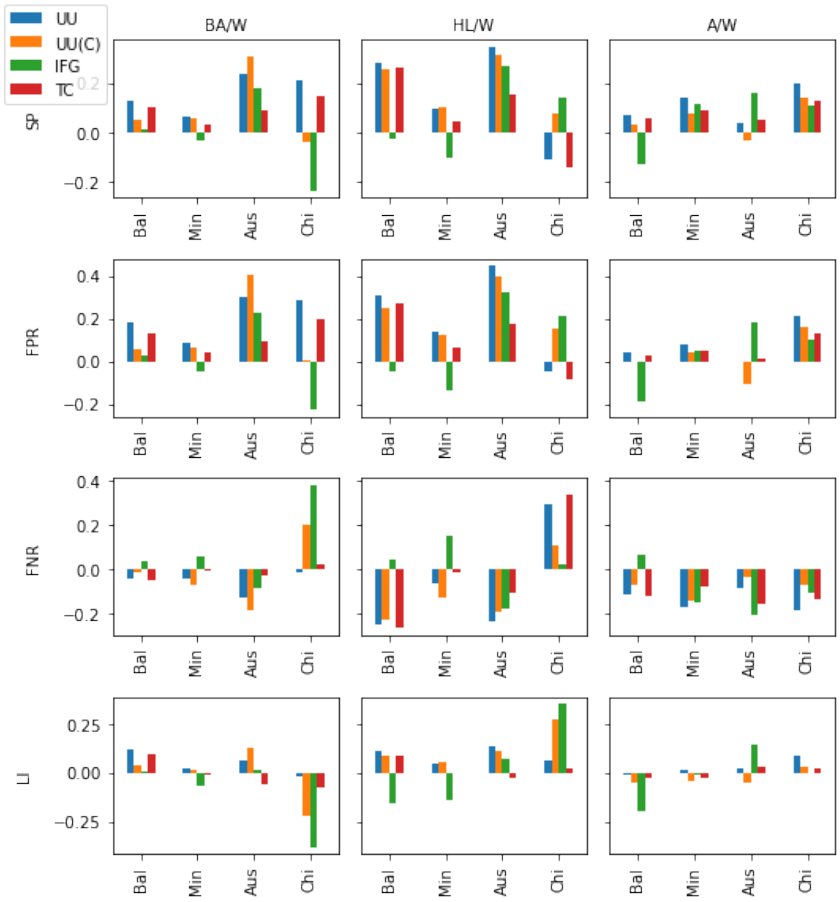


Figure 5.8: Degrees of unfairness of property crime prediction for four cities (Baltimore, Minneapolis, Austin and Chicago). Results are shown for each fairness metric explained in Section 5.4.1 and for each protected group. Crime prediction models include under-reporting-unaware model (UU), UU with historical crimes only (UU(C)), UU with individual-based fairness gap regularization, and the proposed under-reporting-aware model (TC).

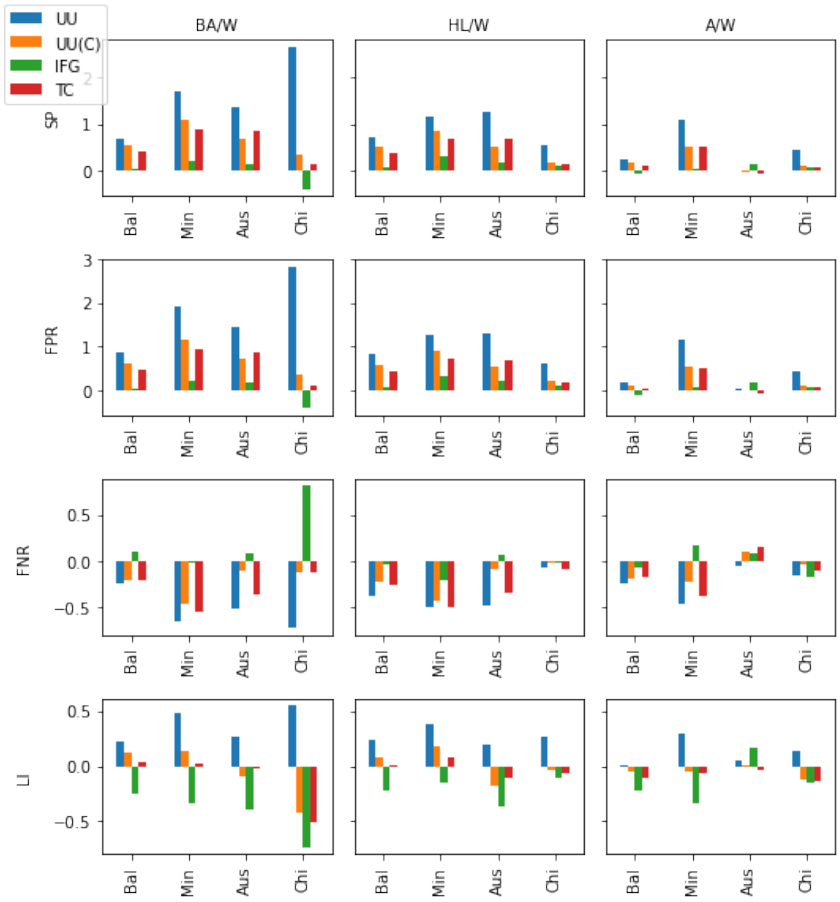


Figure 5.9: Degrees of unfairness of violent crime prediction for four cities (Baltimore, Minneapolis, Austin and Chicago). Results are shown for each fairness metric explained in Section 5.4.1 and for each protected group. Crime prediction models include under-reporting-unaware model (UU), UU with historical crimes only (UU(C)), UU with individual-based fairness gap regularization, and the proposed under-reporting-aware model (TC).

	IFG/UU	TC/UU	TC/IFG	TC/UU(C)
Property crime	58%	69%	65%	58%
Violent crime	71%	85%	40%	54%

Table 5.19: Percentage of settings where applying the convolutional gate for crime-reporting process (TC) improves the baselines.

for the four cities and the two types of crimes in terms of four fairness metrics for different race and ethnicity groups. It shows that the effect on fairness resulting from adding a fairness regularization (IFG) or a convolutional gate for crime-reporting (TC) varies across cities and types of crimes. To be able to summarize the impact of fairness treatments, I calculate the percentage of settings for which the TC approach improves fairness compared to the baselines, with a setting being defined as a combination of city, fairness metric and protected group. In other words, I compute the percentage of settings for which there is at least a 5% relative change in the degree of unfairness of the crime prediction when using TC versus one of the other baselines. If the percentage of settings is larger than 50%, applying the TC approach is beneficial in terms of improving fairness. For example, our results show that when using the convolutional gate for crime reporting (TC), 69% of the times the fairness was improved when compared to an under-reporting-unaware model (UU) suggesting that among the cities, fairness metrics and race/ethnicity groups considered in this study, 69% of the degrees of unfairness decrease more than 5%. Next I discuss the main findings.

The results are shown in Table 5.19. For property crime prediction, the IFG and TC approach both are beneficial to improving fairness (the percentages for IFG/UU and for TC/UU are both larger than 50%) and the proposed TC approach

has a better chance of improving fairness than IFG when compared to the UU baseline (69% versus 58%). In addition, the overall percentage of TC/IFG is 65% suggesting the TC approach can further improve fairness than the baseline IFG approach.

For violent crime prediction, both IFG and TC also are beneficial to improving fairness and similar with property crime prediction, TC has a better chance to improve over baseline UU than IFG (85% for TC/UU versus 71% vs IFG/UU). However, comparing the TC approach with the baseline IFG approach, the chance for improvement (TC/IFG) is 40% suggesting that the scale of reducing degrees of unfairness brought by TC tends to be smaller than the IFG approach in violent crime prediction. This reflects the trade-off between accuracy and fairness, as the IFG has a larger decrease in F1 score for violent crimes than the TC approach, *e.g.*, the decrease can be as large as 50% in violent crime prediction for Austin (Aus) (F1 score using IFG is 0.090 while using UU is 0.201 in Table 5.18.

Comparing property and violent crimes prediction, the percentage to improve fairness is larger for violent crimes (85% for TC/UU) than for property crimes prediction (69% for TC/UU). As discussed in Section 5.4.2, the data bias issue is more severe for violent crime incidents than for property crime incidents. This suggests that the more severe the data bias is, the higher the percentage of improving fairness fairness using the convolutional gate for crime-reporting (TC).

Comparing with UU model with historical crimes only, *i.e.*, UU(C), the percentage of fairness improvement is over 50% for both types of crimes. This means the proposed TC approach can still slightly improve fairness over UU(C) with similar

		Property crime	Violent crime
Metric	SP	83%	92%
	FPR	83%	92%
	FNR	50%	75%
	LI	58%	83%
Race/Ethnicity	BA/W	75%	100%
	HL/W	75%	88%
	A/W	56%	69%
City	Bal	67%	92%
	Min	92%	92%
	Aus	58%	75%
	Chi	58%	83%

Table 5.20: Percentage of settings for which applying the convolutional gate for crime-reporting process (TC) improves fairness when compared with the under-reporting-unaware model (UU) by fairness metrics, race/ethnicity groups and cities.

accuracy with the additional inference on the under-reporting rate for each census tract.

Finally, since TC is the model that improves fairness the most - at the cost of reducing accuracy - I take an in-depth look into the fairness improvement brought about by the TC approach versus the UU baseline by disaggregating the results by metrics, by race/ethnicity groups and by cities as shown in Table 5.20. I highlight three main results. 1) In terms of fairness metrics, adding the convolutional gate for crime-reporting (TC approach) to the under-reporting-unaware model (UU approach) is especially good for improving fairness in terms of statistical parity (SP) and false positive rate (FPR) *e.g.*, the chance of improving fairness is larger than 80% in terms of SP and FPR for both types of crimes. This means that by modeling the data bias generated by the under-reporting issue, the percentage of population in the protected groups being involved in the predicted crime hotspots is reduced

(*i.e.,* the degree of unfairness of SP decreases) due to less false positive prediction. 2) For race/ethnicity groups, the chance to improve fairness is better for the Black and African-American as well as the Hispanic and Latino groups than for the Asian group. Based on Table 5.12 and 5.13, where data bias in reported crimes is quantified by the Spearman correlation between population and number of crimes, the correlation coefficients for the former two protected groups tend to be larger than the coefficients for the Asian group. This suggests that the TC approach mitigates the data bias in reported crimes better for the protected race/ethnicity groups with more severe data bias. 3) The per-city values show the trade-off between accuracy and fairness, that is, larger chances of improving fairness generally correspond to larger decreases in prediction accuracy. For example, the decrease in prediction accuracy for Austin and Chicago in property crime prediction is small (F1 score decreases less than 0.02, see Table 5.18; while the chance of improving fairness is the smallest for Austin and Chicago (58%). On the other hand, although the F1 score for Minneapolis in property crime prediction decreases by almost 0.1, the chance to improve fairness for Minneapolis is as high as 92%.

5.6 Decision-Making Framework

In this study, I have presented a comprehensive analysis of short-term crime prediction with mobility features in terms of prediction accuracy and fairness. By conducting experiments with multiple state-of-the-art deep learning models across different cities and types of crimes, I show that adding mobility features does improve

the accuracy of short-term crime prediction. However, while most studies about short-term crime prediction focus solely on improving the prediction accuracy [13, 92, 93], my proposed detailed fairness analysis suggests that unfair crime prediction is common across cities, types of crimes and fairness metrics; and in some scenarios, the improvement in accuracy comes at the cost of fairness. Thus, improving accuracy should not be the only goal in searching for a better solution for short-term crime prediction, but rather, a measure taken into account together with fairness. As a result, I have also proposed the use of a convolutional gate mechanism to model the crime-reporting process in order to mitigate the data bias in crime incident data due to the issue of under-reporting. Results show that the more severe the data bias issue is, the better the chance that the convolutional gate for crime-reporting improve the fairness in crime hotspot prediction based on the inferred true crimes, albeit at the cost of prediction accuracy.

Using the methods presented in this paper, I propose a decision-making framework to assist in determining whether a new short-term crime prediction model that incorporates mobility features and the under-reporting-aware model should be deployed based on the analysis of crime prediction accuracy and fairness. Figure 5.10 shows the flowchart with the main steps for the decision making framework proposed. The first step identifies whether using mobility features improves the prediction accuracy of the short-term prediction model. My experimental evaluation has shown that Neighbor Convolution models are the best in terms of prediction accuracy. Thus, the proposed framework will first evaluate the prediction accuracy for models that only use historical crime data (C) versus models that incorporate mobility

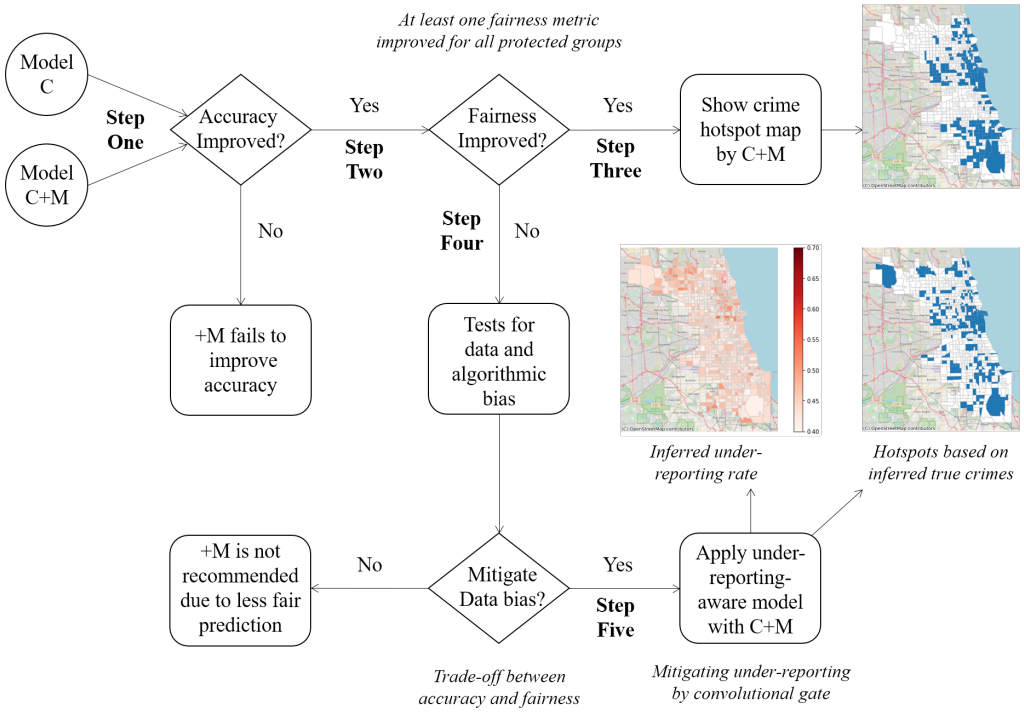


Figure 5.10: Decision-making framework for incorporating mobility features into short-term crime prediction model. Model C (C+M) means crime prediction model (such as NbConv in this study) with historical crimes only (historical crimes and mobility features) as input. +M means adding mobility features to crime prediction model. Blue census tracts are the ones predicted as hotspots on August 9th, 2020 in Chicago.

features ($C + M$). If the prediction accuracy for mobility-based models is worse, the flowchart process ends since adding mobility features does not help; otherwise, move to step two. Step one identifies settings where mobility features improve the prediction. However, decision makers also need to understand the fairness of the algorithms they deploy. Thus, the second step looks for fairness metrics for which the change in the degree of unfairness is positive or neutral across all protected groups *i.e.,* adding mobility features to the predictive model makes the model more fair or maintains the same degree of unfairness. If there exists at least one metric with that condition, move to step three, which recommends incorporating mobil-

ity features to the prediction model and displays a daily map with the predicted census tracts with crime incidents. In this case, the crime predictions are not only more accurate, but also more fair, than predictions that only use historical crime data. On the other hand, if no fairness metric improves or maintains the change in the degree of unfairness, then decision makers are advised to conduct the proposed tests to identify data and algorithmic bias. If the decision makers have no intention to mitigate the data bias at the cost of accuracy, then the flowchart process ends and mobility features are not recommended for crime prediction. Otherwise, the under-reporting-aware model with the convolutional gate for crime-reporting process is recommended and the map of inferred under-reporting rate and daily crime prediction based on inferred true crimes will be displayed.

Finally, I describe a couple of examples of how the decision-making framework could be applied using the results presented in Table 5.14 and 5.15. I have shown that incorporating mobility features in short-term property crime prediction for Chicago improves the accuracy of the prediction by 4.7% (step one), and that the fairness metrics SP, FPR and LI show a positive or neutral change in the degree of unfairness across all protected groups (step two). In this case, the decision making framework would recommend to add mobility features to the prediction model and would display daily maps with crime predictions per census tract (step three). On the other hand, my analysis has also shown that incorporating mobility features in short-term property crime prediction for Baltimore City improves the prediction accuracy by 5.8% (step one). However, there is no fairness metric associated to a positive change in the degree of unfairness across all protected groups (step two).

The analysis of data and algorithmic bias (step four) has shown that short-term crime prediction for Baltimore City suffers from unfairness in part due to data bias in the crime incident dataset and in part due to algorithmic bias, since the predictive model appears to exacerbate the bias in the crime and mobility data. If the decision makers in Baltimore City prioritize fairness over accuracy, they could move to step five and apply the under-reporting-aware model proposed in this study. In this case, they could improve the fairness in 67% of the cases across all fairness metrics and race/ethnicity groups (as shown in Table 5.20).

5.7 A Preliminary Study on Transferring Knowledge from Data-rich Cities

As discussed in the previous section, adding mobility features to short-term crime prediction can improve prediction accuracy for the four cities in this study, while for some cities, the improvement of accuracy brought by the additional mobility features comes at the cost of fairness. The rise of information and communication technologies (ICT) such as mobile phones and wearable devices, and location-based services *e.g.*, geotagged social media and ride sharing, has generated large amounts of human mobility data in cities with well developed infrastructures. However, human mobility data may be scarce for some cities where the infrastructures and services are not ready or even built yet. Therefore, these cities have not accumulated large scale human mobility data to leverage the predictive power of mobility features in short-term crime prediction.

Transfer learning is an effective solution to address the data scarcity problem. The goal of transfer learning is to extract the knowledge from one or more source domains and to apply that knowledge to a target domain [94]. In a cross-city transfer learning setting, cities with abundant data resources correspond with the source domains and are referred to as source cities, while cities with data scarcity issues correspond with the target domains and are referred to as target cities. Studies have shown promising results in cross-city transfer learning, *i.e.,* leveraging knowledge from source cities to improve models built for the target cities [63, 100]. As a preliminary study on the topic of transferring knowledge from cities with large amounts of mobility data to cities with limited mobility data, I will apply a widely used transfer learning technique and evaluate the effect of this technique on the accuracy of short-term crime prediction. I leave more advanced transfer learning techniques and their evaluation on fairness as future work, and discuss it in Chapter 6.

5.7.1 Experiment Setting for Transfer Learning

There are four categories of transfer learning for deep learning models: a) instances-based, which utilizes instances in the source domain by assigning appropriate weights; b) mapping-based, which maps instances from two domains into a new data space with better similarity; c) network-based, which reuses parts of a pre-trained network in the source domain; and d) adversarial-based, which uses adversarial technology to find transferable features suitable for the two domains.

[166]. In this study, I apply the network-based approach to transfer knowledge from a source city to a target city, as shown in Figure 5.11. The advantage of network-based approach is to distill knowledge from the data distribution in the source city in the form of learned parameters of the neural network. The pretrained network can be viewed as a feature extractor learned from the data in the source city. As discussed in Section 5.3, NbConv is the model that can make best use of the mobility features. Therefore, in this section, NbConv is chosen as the base model for transfer learning, which means the crime prediction models for both source and target city share the same network structure. The process of transfer learning is as follows: 1) the base model is pre-trained with the data in the source city, denoted as $Model_s$; 2) the learned parameters (weights) of all layers in the whole architecture of $Model_s$, *i.e.*, θ_s, are used to initialize the parameters of model for the target city $Model_{s,t}$; 3) the parameters of $Model_{s,t}$ are fine-tuned with the limited data in the target city. Fine-tuning refers to the process of updating the transferred parameters θ_s with the local limited data in the target city and the post-fine-tuning parameters are denoted as $\theta_{s,t}$. $Model_{s,t}$ will be also referred as a fine-tuned model in the following discussion. To evaluate the effects of transfer learning in crime prediction accuracy, I conduct experiments in the same four cities as the above sections: Baltimore, Minneapolis, Austin and Chicago. Each of the four cities will be treated as a data-scarce target city with the remaining three being treated as the data-rich source cities. For example, when Baltimore is considered as a target city, Minneapolis, Austin and Chicago are considered as source cities from which knowledge is extracted.

Similar to the experiment and evaluation protocol in Section 5.3.3, the pre-

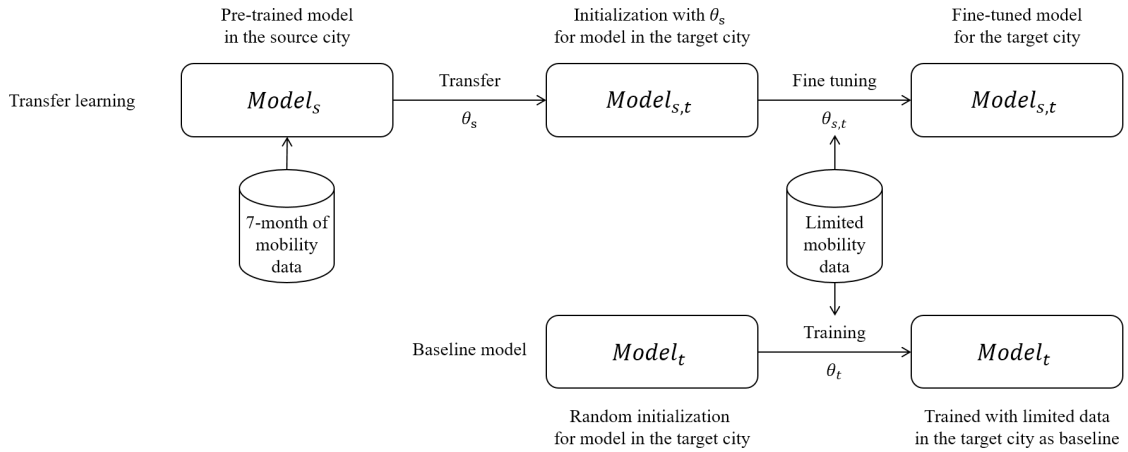


Figure 5.11: The framework of the transfer learning technique applied in this study. $Model_s$ and $Model_{s,t}$ have the same network architecture. The parameters θ_s of the whole architecture of $Model_s$ is transferred to $Model_{s,t}$ as the initialization of $\theta_{s,t}$.

diction accuracy is evaluated as the average monthly F1 score over the five testing months (August to December, 2020). For each testing month, $Model_s$ is pre-trained with training data in the previous 7 months in the source city. To simulate different levels of data scarcity, I vary the number of months of "limited mobility data" in the target city from 1 to 7 months. The shorter the length of data available in the target city is, the more severe the data scarcity issue is for the target city. 7 months of mobility data (which is as long as the training data in the source city) is included in the experiment to evaluate the effects of transfer learning when the target city also has abundant mobility data.

In order to evaluate whether transfer learning can improve crime prediction accuracy in the target city, the baseline model for the target city $Model_t$ for each level of data scarcity (number of months of collected mobility data) is trained as a random-initialized model with the limited mobility data, as shown in Figure 5.11. In order words, the baseline model is trained in the same way as in Section 5.3.3, but with

a shorter length of training data. I use the relative change in the average monthly F1 score to evaluate the effect of transfer learning: $\frac{F1_{model_{s,t}}}{F1_{model_t}} - 1 * 100\%$. Positive (negative) relative change in F1 score suggests that the transferred knowledge from the source city improves (degrades) the prediction accuracy when compared with the model trained only by the limited data in the target city.

To make use of transferred knowledge from multiple source cities, I apply the majority voting to aggregate the next-day crime predictions from multiple fine-tuned models for a target city, *i.e.*, a census tract in the target city is predicted as a crime hotspot in the next day if it is predicted as a hotspot by at least half of the fine-tuned models. For example, when Baltimore is the target city, there are three fine-tuned models with transferred knowledge from Minneapolis, Austin and Chicago. A census tract in Baltimore is considered as a hotspot in the next day if it is a predicted hotspot in at least two of the fine-tuned models.

5.7.2 Analysis of Effects of the Transfer Learning on Prediction Accuracy

Figure 5.12 shows the average monthly F1 score of the fine-tuned crime prediction models with transferred knowledge from different source cities and the baseline model without transfer learning in each target city. Each column shows the results for property and violent crime prediction in each target city. In each plot, x-axis is the number of months of collected mobility data in the target city; *voting* refers to the majority voting aggregating knowledge from multiple source cities, *base* refers

Figure 5.12: Average monthly F1 score for crime prediction by fine-tuned models with transferred knowledge from different source cities.

to the baseline model with training data in the target city only, and each of the three cities refer to the fine-tuned model.

The *base* (purple) lines in all plots suggest that across all cities and types of crimes, the prediction F1 score tends to be smaller when the number of months of collected data in the target cities is ≤ 2. These results show that data scarcity does affect the performance of short-term mobility-based crime prediction. The other lines in the plots suggest that the F1 scores of fine-tuned models tend to be higher than the *base* model in most cases, especially when the number of months of collected data in the target city is ≤ 2. When the number of months is ≥ 3, as the number of months increases, F1 scores of all baseline models and fine-tuned models are mostly stable. For example, the F1 scores of all models for property crime prediction in the target city Minneapolis fall into the range of 0.58 and 0.59.

To further investigate the effects that transfer learning has on the performance of short-term mobility-based crime prediction, I compute the relative change in F1

Target city	Source city	Number of months of collected data in the target city 1	2	3	4	5	6	7
Bal	Voting	**3.17%**	**2.64%**	**1.51%**	**1.86%**	**1.87%**	**1.34%**	1.11%
	Min	2.95%	2.35%	1.37%	1.17%	0.83%	1.29%	**1.15%**
	Aus	1.83%	1.65%	0.74%	0.53%	1.46%	0.43%	0.58%
	Chi	2.90%	1.78%	1.26%	0.99%	1.18%	0.31%	0.37%
Min	Voting	**1.59%**	0.52%	**0.14%**	**0.57%**	0.04%	**0.30%**	**-0.34%**
	Bal	1.09%	-0.45%	-0.99%	-0.33%	-0.96%	-1.06%	-1.10%
	Aus	1.42%	**0.66%**	-0.37%	-0.03%	-0.29%	0.02%	-0.42%
	Chi	1.30%	0.43%	-0.36%	0.09%	**0.13%**	-0.09%	-0.71%
Aus	Voting	**1.50%**	**1.58%**	**1.12%**	**0.51%**	0.22%	**-0.04%**	**0.35%**
	Bal	0.92%	0.69%	0.76%	0.08%	**0.26%**	-0.39%	-0.16%
	Min	0.40%	0.76%	0.17%	-0.42%	-0.43%	-0.45%	-0.35%
	Chi	1.28%	1.25%	0.55%	0.19%	-0.03%	-0.40%	0.13%
Chi	Voting	**0.97%**	**0.67%**	**0.73%**	**1.27%**	**0.26%**	**0.43%**	**0.02%**
	Bal	0.03%	0.24%	0.30%	0.78%	-0.26%	-0.21%	-0.38%
	Min	0.45%	-0.21%	-0.04%	0.72%	-0.36%	-0.17%	-0.10%
	Aus	0.54%	0.11%	0.23%	0.75%	-0.19%	-0.08%	-0.32%

Table 5.21: Relative change in average monthly F1 score using transfer learning over all test months (Aug-Dec) for property crime prediction.

score between the fine-tuned model $Model_{s,t}$ and the baseline model $Model_t$ as described in the previous section. Table 5.21 and 5.22 show the results of relative change for all target cities in property and violent crime prediction.

Based on the relative change in F1 score for property crime prediction in Table 5.21, I highlight the following main observations:

1) The transfer learning is beneficial for target cities with data scarcity, especially when the number of months of available mobility data is small. As the level of data scarcity alleviates, *i.e.*, the number of fine-tuning months increases, the improvement in F1 score brought by transfer learning decreases and, in many cases, could hurt the F1 score in the crime prediction for the target cities. For example, the relative changes in F1 score are all positive when the number of fine-tuning

months is 1 and eight out of twelve of the fine-tuned models (not including *Voting*) have worse F1 scores compared with the baseline when the number of fine-tuning months is 7. This suggests that knowledge extracted from mobility data in a single source city in the form of network parameter initialization conflicts with the local knowledge in the target city. This could be due to a data distribution difference between the source city and the target city. But through majority voting, where the knowledge from multiple source cities is aggregated, the downside of different data distribution is mitigated and the relative change in F1 scores are mostly positive in all cities and all levels of data scarcity.

2) The effects of the transfer learning differ across different target cities. For example, the relative changes in F1 score for Baltimore (Bal) are all positive for any source city considered in this study. Also, the scale of relative changes in F1 score for Baltimore are the largest among all the four target cities across all levels of data scarcity; while Chicago benefits the least from the transfer learning of the three source cities considered. This could be because Chicago is a much larger city than the source cities and the knowledge extracted from a single small city is not enough to provide a good initialization for the model in a big target city. As shown in Table 5.3, the number of census tracts in Chicago is 809, while the number for Baltimore, Minneapolis and Austin is 200, 116 and 204 respectively. In the next chapter, I will suggest a few future work avenues to explore city similarities that could be exploited to inform better the source city selection for a specific target city.

3) The effects of transfer learning also varies by different source cities. For example, knowledge extracted from Baltimore (Bal) has little effect (0.03%) on

Target city	Source city	Number of months of collected data in the target city 1	2	3	4	5	6	7
Bal	Voting	**4.09%**	0.82%	**1.16%**	**-0.04%**	**0.63%**	**0.38%**	**0.96%**
	Min	2.79%	**0.93%**	0.82%	-0.65%	-0.17%	-0.37%	0.19%
	Aus	1.69%	-1.04%	0.06%	-1.29%	-1.05%	-0.63%	0.02%
	Chi	3.91%	0.77%	0.53%	-0.35%	0.56%	-0.05%	0.42%
Min	Voting	**4.15%**	**-0.18%**	**-0.24%**	**0.18%**	**-0.54%**	**0.03%**	**1.54%**
	Bal	3.92%	-0.65%	-1.45%	-1.17%	-1.31%	-0.82%	-0.95%
	Aus	-0.22%	-2.83%	-2.81%	-1.75%	-1.30%	-1.61%	-0.34%
	Chi	3.64%	-1.83%	-0.86%	-0.83%	-0.75%	-1.40%	0.60%
Aus	Voting	7.42%	**4.88%**	**1.88%**	**3.28%**	**3.19%**	**3.06%**	**2.01%**
	Bal	2.51%	1.82%	0.32%	0.19%	0.68%	1.70%	0.41%
	Min	3.68%	2.54%	-0.13%	2.39%	2.29%	0.78%	0.36%
	Chi	**9.72%**	3.51%	-0.28%	1.57%	2.49%	1.65%	1.30%
Chi	Voting	0.38%	0.53%	**0.01%**	-0.18%	**-0.04%**	-0.41%	0.08%
	Bal	-0.93%	-0.31%	-1.25%	-1.25%	-0.84%	-1.45%	-0.91%
	Min	**0.91%**	**1.23%**	-0.23%	**0.11%**	-0.08%	**0.00%**	**0.44%**
	Aus	-1.68%	-1.17%	-1.19%	-0.98%	-0.98%	-0.90%	-0.80%

Table 5.22: Relative change in average monthly F1 score using transfer learning over all test months (Aug-Dec) for violent crime prediction.

the property crime prediction for Chicago (Chi) when the number of fine-tuning months is 1 while knowledge from other source cities help slightly increase the F1 score (0.45% and 0.54%). As an approach making use of knowledge from multiple source cities, the majority voting often has the best improvement in F1 score for all four target cities, as highlighted as bold in the table.

Table 5.22 shows the results for transfer learning in violent crime prediction. Most of the observations for violent crime prediction are similar with the property crime prediction, including 1) transfer learning improves the prediction accuracy when the fine-tuning data in the target cities is limited; 2) the effects of transfer learning varies across different target and source cities and Chicago is the target city which benefits the least from the transferred knowledge from the source cities

considered in this study; 3) and the majority voting tends to provide the best performance compared with transfer learning from a single source city.

However, there are some findings for violent crime prediction that are different when compared with for property crime prediction:

1) The improvement in F1 score tends to be larger for violent crime prediction. For example, the largest improvement in F1 score (9.72%) is observed for Austin (Aus) with knowledge transferred from Chicago (Chi) when the number of fine-tuning months is 1. While the largest improvement in F1 score for property crime prediction is 3.71%.

2) The variance in relative change of F1 score among different source cities is larger for violent crime prediction than for property crime prediction. For example, the relative changes in F1 score for property crime prediction in Minneapolis (Min) as the target city are 1.09%, 1.42% and 1.30% when the source cities are Baltimore, Austin and Chicago. But the corresponding relative changes for violent crime prediction are 3.92%, -0.22% and 3.64%. This suggests that for violent crime prediction, it is important to choose an appropriate source city or develop a good mechanism to incorporate knowledge from multiple source cities.

The results shown in this section show promising results for applying transfer learning techniques to improve prediction accuracy in mobility-based short-term crime prediction when the mobility data is scarce. In Chapter 6 I will discuss some potential directions for more advanced transfer learning techniques as future work.

Chapter 6: Conclusions and Future Directions

6.1 Conclusions

With the rise of mobile phones and other information communication technologies, in this dissertation, I identify three challenges in leveraging large scale human mobility data in urban crime prediction. Three empirical studies are conducted to address these three challenges with the aim to improve the robustness, accuracy, fairness, and applicability of mobility-based crime prediction techniques. Next, I summarize these studies and answer the research questions raised in the Introduction section.

6.1.1 Spatial Sensitivity Analysis for Urban Hotspots using Cell Phone Traces

The rich spatio-temporal information provided by CDR data has provided great potential for studying human mobility dynamics in urban environments. A bulk of the literature has used CDR data to study the relationship between human dynamics - modelled through hotspot areas in cities - and various urban characteristics, such as spatial structure, transportation efficiency and energy consumption.

However, most of these studies are based on ad-hoc selections of city boundaries and spatial units. In this study, I provide a novel interpretable approach to carry out a systematic analysis of the stability of various hotspot indices in both static (inter-city) and dynamic (intra-city) settings.

I have found that at the inter-city level, the urban municipality boundary is the best setting to obtain stable and robust city ranking results. Indices for scale of hotspots and degree of urban sprawl are strongly stable across all city boundary settings. Therefore, when a particular city boundary setting is desired, NHS, AHS, COMP, and MCOMP are good indices to work with. If the compactness of the family of indices (COHE, PROX, NMI, and NMMI) is of interest, it is better to use the municipalities with urban and rural areas (PerMuni-UR setting). If other city boundaries are required, we recommend using the (G, Pop) interpolating method as this method tends to be most correlated with other methods in all settings.

For intra-city level, the stability of indices is mostly weakly stable. Only the degree of urban sprawl (COMP and MCOMP) in Metro-based settings has moderate stability. The stability of indices in different cities has large variation, meaning some indices can be very strongly stable across interpolation methods in some cities but not stable at all in other cities. I have not found any index or boundary setting that would work well across cities. Thus, it is vital for researchers to use a consistent type of spatial units and interpolation methods across analyses.

Therefore, to conclude this study with respect to the questions in Research Objective 1:

1A) Mobility-based hotspot features are sensitive to the methodology choices

that construct such features;

1B) For inter-city level analysis, there are some methodology choices which could provide stable results given different methodology priorities; while for intra-city level analysis, no combination of methodology considered in this study could provide stable results across cities which emphasizes the importance of consistent methodology for cross-city comparison.

6.1.2 Addressing Under-Reporting to Enhance Fairness and Accuracy in Mobility-based Long-term Crime Prediction

With the increasingly available human mobility data, various empirical studies have demonstrated the relationship between human mobility and reported crimes in the context of long-term crime estimation and prediction analysis. However, these studies have overlooked the data bias in the reported crime data, which can affect the quality of the final predictions in terms of accuracy and fairness. To mitigate the data bias issue, in this study, I propose a Bayesian hierarchical model (BURC) to identify and mitigate the under-reporting of crimes, which is an important sources of bias in reported crimes. This model utilizes human mobility data as predictive features to generate long-term crime prediction and leverages the domain knowledge about possible determinants for the under-reporting of crimes, *e.g.*, poverty rate has influence on under-reporting of property and violent crimes, to enhance both fairness and accuracy of the prediction.

The experiments and evaluations show that the proposed model not only im-

prove substantially the accuracy in reported crimes prediction, but also the fairness in crime prediction. In terms of accuracy, compared with the baseline machine learning models, the BURC model can reduce the RMSE and MAE by 21.2% and 9% for property crimes prediction and by 10.4% and 14.4% for violent crimes. As for fairness, the BURC model improves the fairness in crime prediction for all income groups and for municipalities with large indigenous population, which is measured by mean difference and group errors.

Another advantage of the proposed BURC model is the interpretability. In addition to the predicted number of crimes, BURC also infers the crime reporting rates for each municipalities, which is consistent with the ENVIPE survey in Mexico. The coefficients in the BURC model contains insights about the influence of the mobility features on the true crime generating process and the influence of under-reporting determinants on the under-reporting process.

To conclude this study with respect to the questions in Research Objective 2:

2A) Modeling the under-reporting process improves the performance of predicting the number of reporting crimes;

2B) Modeling the under-reporting process improve the fairness of long-term crime prediction;

2C) In terms of the influence of mobility-based features on the true crime generating process, more hotspots detected, *i.e.*, people actively moving around the municipalities, lower degree of urban sprawl, and less compact of the detected hotspots are associated with more future crimes;

2D) In terms of the influence of determinants on the reporting rate, poverty

rate have a large influence on the reporting rate of property crime prediction and the higher the poverty rate of a municipality, the lower the reporting rate is. As for violent crime prediction, the more presence of adults, never married people, and people born in other municipalities increase the willingness to report violent crimes, while more males and male-headed households as well as larger poverty rate are associate with lower reporting rates.

6.1.3 Enhancing Short-term Crime Prediction with Human Mobility Flows: An Analysis of Accuracy and Fairness

In this study, I leverage large-scale human mobility flows for short-term place-based crime prediction with neural networks. To robustly analyze the effect of adding mobility features to next-day crime prediction in terms of prediction accuracy, I conduct comprehensive experiments with a wide range of neural network architectures on cities with diverse demographic characteristics and different types of crimes. Experiments show that adding human mobility flow features to historical crimes can improve the F1 scores for a variety of neural short-term crime prediction models across cities and types of crimes. The improvement in F1 scores varies across models. Neighbor convolution architectures (NbConv) that model the spatio-temporal patterns of the input features simultaneously produce the best prediction accuracy when adding mobility features with relative improvements from 2% to 7%.

I also perform an algorithmic fairness analysis of short-term crime prediction models. I frame the fairness analysis as an evaluation of the crime prediction per-

formance across race and ethnicity, and select a set of metrics that measure the differences across population groups. Results show that unfair predictions are pervasive in short-term crime predictions using either crime data only or both crime and mobility data; and that one of the factors contributing to the unfair prediction is the inherent data bias in the crime incident data. I also show that short-term crime prediction models that add mobility features on top of crime data have diverse performance in terms of degree of unfairness, and that the unfairness differs across cities and types of crimes. The analyses suggest that there exists almost no relationship between mobility data bias and degrees of unfairness, pointing to potential algorithmic bias whereby NbConv appears to exacerbate the crime incident data bias as shown by the increase in the presence of protected groups for the crimes predicted when using both mobility and crime data versus only crime data.

To mitigate the data bias in reported crimes, I propose a convolutional gating mechanism to model the crime reporting process. Compared with the under-reporting-unaware model and a fairness regularization baselines, the proposed convolutional gate can improve the fairness in crime prediction in many scenarios, especially in terms of statistical parity and false positive rate. Based on the comprehensive evaluation, I propose a decision making framework for incorporating mobility features in short-term crime prediction balancing between accuracy and fairness.

Lastly, I conduct a preliminary study on transferring knowledge from data-rich cities to data-scarce cities. Results show that by transferring the parameters of the pre-trained model in a source city with long period training data to the model for the

target city with limited training data, the prediction accuracy can be improved by 1% to 3% for property crime prediction and improved by 1% to 7% for violent crime prediction. This shows promising results for applying transfer learning techniques to short-term mobility-based crime prediction.

To conclude this study with respect to the questions in Research Objective 3:

3A): Incorporating mobility features improve the accuracy of short-term crime prediction for a variety of neural network architectures;

3B): The effects of incorporating mobility features on fairness is diverse across cities and types of crimes;

3C): The unfairness in crime prediction is likely associated with the data bias in reported crimes. The bias and unfairness appears to be exacerbated when additional mobility features are incorporated in the short-term crime prediction model;

3D): Modeling the under-reporting process with a convolutional gating mechanism can improve the fairness in mobility-based crime prediction, but at the cost of accuracy;

3E): Transfer learning techniques show promising results in helping mitigate the data-scarcity issue in mobility data for short-term crime prediction.

6.2 Implications

This dissertation has two important implications: 1) Human mobility features have predictive power for both long- and short-term crime prediction, and therefore, it is worthy to further leverage the massive human mobility data to improve crime

prediction model; 2) Unfairness issues has been observed for both long- and short-term crime prediction and one important source of unfairness is the under-reporting of crimes, which can be mitigated by modeling the crime reporting process based on the domain knowledge from the criminology literature.

For decision-makers such as city planners or police departments, the findings from this dissertation suggest that when applying crime prediction models to allocate crime-prevention resources, it is necessary to consider the fairness of the predictive models and their potential negative impacts on communities. The decision making framework in Section 5.6 provides an example for how to carry out analyses that look at the balance between accuracy and fairness for crime prediction. In addition to the prediction outputs, the inferred crime-reporting rates are also valuable for the decision-makers to identify regions with serious under-reporting issues and conduct further investigation on the causes of these issues, such as the community-police relationship. Such efforts can potentially mitigate the data bias in the training data and allow the predictive models perform better in terms of fairness.

For researchers, this dissertation highlights the importance of evaluating place-based crime prediction in the context of algorithmic fairness since the unfairness in the predictions are quite pervasive across cities and types of crimes; while most of the fairness discussion around predictive systems related with crimes are at the individual level, such as the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) risk tool [163, 165, 167]. As a contribution to the algorithmic fairness techniques with the goal to improve fairness in predictive systems, my analysis shows the value of modeling the bias encoded in data based on domain

knowledge about the generation of data bias without imposing a explicit fairness regularization in the model training process.

6.3 Future Directions

In this section, I will describe the potential directions to extend this dissertation for future studies.

In this dissertation, only historical crimes and mobility data have been leveraged as input or predictive features for future crimes. Environmental factors, such as built-in environment, point-of-interests, demographic and socio-economic data, and meteorological data, have also been explored in the crime prediction literature. Therefore, it would be interesting to incorporate more predictive features and evaluate the effects of mobility features. Beyond enriching the input features, there are other directions that are worthy of explore:

Extending Spatial Sensitivity Analysis. Study 1 (Chapter 3) does not provide a comprehensive analysis of all the combinations of spatial units and interpolation methods that exist in the literature, but rather an analysis of the most common approaches currently used in the field. In addition, the analysis presented focuses only in one country. Further, Study 1 only evaluates the spatial sensitivity of call detailed records (CDR), while there are other commonly used mobility datasets such as GPS and social media data. These datasets usually are more fine-grained than CDR data and do not require interpolation methods, however, they often are aggregated to grids or census tracts for downstream analysis.

As future work, it would be interesting to expand and generalize our spatial sensitivity analysis to more methodological choices, such as different grid sizes, or interpolation methods that consider terrain and land use patterns; and to verify the applicability of the results in this study to other countries and other kinds of mobility datasets.

Extension of the Bayesian Hierarchical Model for Long-term Crime Prediction. In Study 2 (Chapter 4), I propose a Bayesian model for under-reported crimes, where the true crime generation process is modeled by the mobility features of each individual municipalities. In the literature, various empirical studies have shown the predictive power of historical crimes, point-of-interests, and built-in environment characteristics. Also, the proposed model assumes independence among different municipalities while spatial auto-correlation is common is geographical data as well as there are mobility flow among municipalities.

Therefore, as future work, the following directions are worth pursuing: 1) The mobility connections among municipalities can be used to compute features similar to the ones described in Table 5.4. These features can also be incorporated into Equation 4.4 to model the true crime occurring rate; 2) The spatial dependency among municipalities, such as spatial autocorrelation, can be modeled by adding spatially structured effects into Equation 4.4 and 4.5. The spatially structured effects can be estimated through the intrinsic Gaussian conditional autoregressive (ICAR) model. The ICAR estimates the co-variance matrix to capture the spatial dependency among all municipalities.

Deep Bayesian Model for Under-reporting-aware Short-term Crime

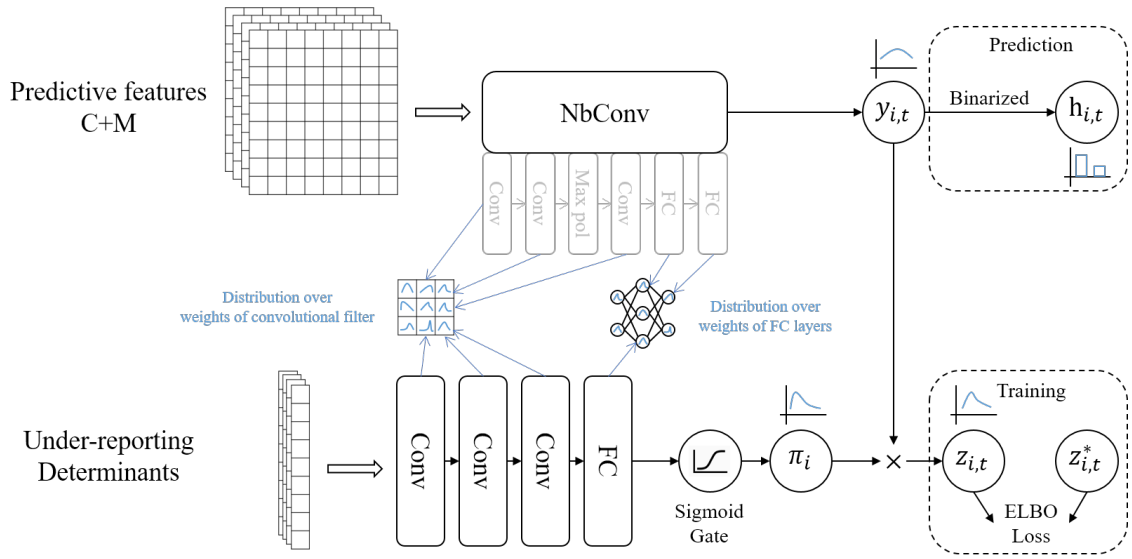


Figure 6.1: Bayesian version of the under-reporting-aware NbConv model.

Prediction. In Study 3 (Chapter 5), I have shown that by modeling the crime reporting process with a convolutional gate can improve the fairness of short-term crime prediction in most scenarios but at a substantial cost of accuracy. This is different from Study 2 where the Bayesian model (BURC) can improve both the accuracy and fairness of crime prediction. Although Study 2 focuses on long-term crime prediction while Study 3 focuses on short-term prediction, this suggests the potential of combining the Bayesian inference of the reporting rate with the deep learning neural network for modeling the spatiotemporal dependency of historical crimes and mobility features. The advantage of Bayesian inference is that it provides principled uncertainty estimates of the reporting rate. This could be beneficial since there also is uncertainty in the under-reporting determinants provided by the census data, *e.g.*, the margin of error in the American Community Survey.

As a future direction to explore, the under-reporting-aware model in Figure 5.7 can be transformed into a Bayesian neural network (BNN) by adding prior

distributions over all the parameters (weights and biases, denoted as θ) in the architecture, as shown in Figure 6.1. In this way, all parameters are considered as random variables. The goal of Bayesian inference (or training the BNN) is to compute the posterior distribution of the parameters given the observations Z (the reported crime data): $P(\theta|Z)$. But computing the exact posterior distribution can be intractable. One approach for inference of BNN is variational inference, which turns the inference problem into an optimization problem. Variation inference approximates the true posterior by finding a variational distribution over the parameters $q^*(\theta)$ from a family of candidate distributions Q. The candidate distribution should be simple for efficient computation yet expressive enough to approximate the true posterior. $q^*(\theta)$ can be found by minimizing the Kullback-Leibler (KL) divergence from the true posterior:

$$q^*(\theta) = \arg\max_{q(\theta)\in Q} KL(q(\theta)||p(\theta|Z)), \tag{6.1}$$

which is equivalent to maximizing the evidence lower bound (ELBO):

$$\text{ELBO} \equiv \mathbb{E}_{q(\theta)}[\log p(Z,\theta) - \log q(\theta)], \tag{6.2}$$

and ELBO can be maximized by gradient ascent. Therefore, training BNN with variational inference is also called *Bayes by Backprop* [168].

In the training phase, the ground truth reported crimes $z^*_{i,t}$ are treated as the observations/evidence to train the model; while in the prediction/inference phase, the reporting rate $\pi_{i,t}$ and number of true crimes $y_{i,t}$ would be sampled from the

learned posterior distribution. For example, $y_{i,t}$ will be sampled multiple times from the posterior distribution so that we can measure the uncertainty in predicting $y_{i,t}$ and calculate the point estimate for $y_{i,t}$ as the average over the posterior samples. With Bayesian version of the under-reporting-aware model, we can explore the relationship between uncertainty and fairness of crime prediction.

Multi-city Transfer Learning for Short-term Mobility-based Crime Prediction. In the preliminary study of transfer learning (Section 5.7), I have shown that transferring knowledge in the form of learned parameters of pre-trained model in the data-rich source city can improve the prediction accuracy in the target city with data scarcity issue, *e.g.*, there are limited short-period of mobility data collected in the target city. The improvement brought by transfer learning varies by different source cities, which might be due to different data distribution and different characteristics between each source city and the target city. This unstable results can be mitigated by a simple majority voting mechanism, which tends to provide better accuracy improvement than any single-source-city fine-tuned model.

These results highlight the importance of measuring the similarity between source cities and the target city as well as the fusion of multi-city knowledge. The similarity can be measured at the spatial unit level. For example, for a given spatial unit in the target city, only leverage knowledge from similar spatial units in the source cities. It can also be measured at the city level, *e.g.*, use the similarity between the source city and target city as a weight to control the influence of the transfer knowledge in the fine-tuning process. The similarity can be computed with multiple sources of data, such as the human mobility features, socio-demographics,

distribution of historical crimes and built-in environments features.

Bibliography

[1] US Census Bureau. Acs demographic and housing estimates, 2019. URL `https://data.census.gov/cedsci/table?q=demographic`. Online; accessed 20 November 2020.

[2] Thomas Louail, Maxime Lenormand, Oliva G Cantu Ros, Miguel Picornell, Ricardo Herranz, Enrique Frias-Martinez, José J Ramasco, and Marc Barthelemy. From mobile phone data to the spatial structure of cities. *Sci. Rep.*, 4:5276, June 2014. ISSN 2045-2322. doi: 10.1038/srep05276.

[3] Carlo Ratti, Dennis Frenchman, Riccardo Maria Pulselli, and Sarah Williams. Mobile landscapes: Using location data from cell phones for urban analysis. *Environ. Plann. B Plann. Des.*, 33(5):727–748, October 2006. ISSN 0265-8135. doi: 10.1068/b32047.

[4] Rein Ahas, Anto Aasa, Y Yuan, Martin Raubal, Zbigniew Smoreda, Yu Liu, Cezary Ziemlicki, Margus Tiru, and Matthew Zook. Everyday space–time geographies: using mobile phone-based sensor data to monitor urban activity in harbin, paris, and tallinn. *Int. J. Geogr. Inf. Sci.*, 29(11):2017–2039, November 2015. ISSN 1365-8816. doi: 10.1080/13658816.2015.1063151.

[5] Jonathan Reades, Francesco Calabrese, Andres Sevtsuk, and Carlo Ratti. Cellular census: Explorations in urban data collection. *IEEE Pervasive Computing*, 6(3):30–38, July 2007. ISSN 1536-1268. doi: 10.1109/MPRV.2007.53.

[6] Danya Bachir, Vincent Gauthier, Mounim El Yacoubi, and Ghazaleh Khodabandelou. Using mobile phone data analysis for the estimation of daily urban dynamics. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pages 626–632, October 2017. doi: 10.1109/ITSC.2017.8317956.

[7] Lingzi Hong, Cheng Fu, Jiahui Wu, and Vanessa Frias-Martinez. Information needs and communication gaps between citizens and local governments online during natural disasters. *Information Systems Frontiers*, 20(5):1027–1039, 2018.

[8] Marcus Felson and Lawrence E Cohen. Human ecology and crime: A routine activity approach. *Hum. Ecol.*, 8(4):389–406, December 1980. ISSN 0046-8169, 1572-9915. doi: 10.1007/BF01561001.

[9] Carlos Caminha, Vasco Furtado, Tarcisio H C Pequeno, Caio Ponte, Hygor P M Melo, Erneson A Oliveira, and José S Andrade, Jr. Human mobility in large cities as a proxy for crime. *PLoS One*, 12(2):e0171609, February 2017. ISSN 1932-6203. doi: 10.1371/journal.pone.0171609.

[10] R Norman Davidson. *Crime and Environment*. St. Martin's Press, New York, 1981. doi: 10.4324/9780429026997.

[11] Marcus Felson and Ronald V Clarke. *Opportunity makes the thief: Practical theory for crime prevention*. Police Research Series Paper 98. Home Office, London, 1998.

[12] George O Mohler. Marked point process hotspot maps for homicide and gun crime prediction in chicago. *Int. J. Forecast.*, 30(3):491–497, July 2014. ISSN 0169-2070. doi: 10.1016/j.ijforecast.2014.01.004.

[13] Xian Wu, Chao Huang, Chuxu Zhang, and Nitesh V Chawla. Hierarchically structured transformer networks for Fine-Grained spatial event forecasting. In *Proceedings of The Web Conference 2020*, WWW '20, pages 2320–2330, New York, NY, USA, April 2020. Association for Computing Machinery. ISBN 9781450370233. doi: 10.1145/3366423.3380296.

[14] John A Eterno, Arvind Verma, and Eli B Silverman. Police manipulations of crime reporting: Insiders' revelations. *Justice Q.*, 33(5):811–835, July 2016. ISSN 0741-8825. doi: 10.1080/07418825.2014.980838.

[15] Albert Meijer and Martijn Wessels. Predictive policing: Review of benefits and drawbacks. *International Journal of Public Administration*, 42(12):1031–1039, September 2019. ISSN 0190-0692. doi: 10.1080/01900692.2019.1575664.

[16] Spencer Chainey, Lisa Tompson, and Sebastian Uhlig. The utility of hotspot mapping for predicting spatial patterns of crime. *Secur. J.*, 21(1):4–28, February 2008. ISSN 0955-1662, 1743-4645. doi: 10.1057/palgrave.sj.8350066.

[17] Xiangyu Zhao and Jiliang Tang. Modeling Temporal-Spatial correlations for crime prediction. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, pages 497–506, New York, NY, USA, 2017. ACM. ISBN 9781450349185. doi: 10.1145/3132847.3133024.

[18] Cristina Kadar and Irena Pletikosa. Mining large-scale human mobility data for long-term crime prediction. *EPJ Data Science*, 7(1):26, July 2018. ISSN 2193-1127. doi: 10.1140/epjds/s13688-018-0150-z.

[19] Andrey Bogomolov, Bruno Lepri, Jacopo Staiano, Emmanuel Letouzé, Nuria Oliver, Fabio Pianesi, and Alex Pentland. Moves on the street: Classifying crime hotspots using aggregated anonymized data on people dynamics. *Big Data*, 3(3):148–158, September 2015. ISSN 2167-647X, 2167-6461. doi: 10.1089/big.2014.0054.

[20] Marco De Nadai, Yanyan Xu, Emmanuel Letouzé, Marta C González, and Bruno Lepri. Socio-economic, built environment, and mobility conditions associated with crime: a study of multiple cities. *Sci. Rep.*, 10(1):13871, August 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-70808-2.

[21] Cristina Kadar, Stefan Feuerriegel, Anastasios Noulas, and Cecilia Mascolo. Leveraging mobility flows from location technology platforms to test crime pattern theory in large cities. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 339–350. aaai.org, 2020.

[22] Stan Openshaw. *The modifiable areal unit problem*. Norwich: Geo Abstracts Univ. of East Anglia, 1984.

[23] Roger Tarling and Katie Morris. Reporting crime to the police. *Br. J. Criminol.*, 50(3):474–490, May 2010. ISSN 0007-0955. doi: 10.1093/bjc/azq011.

[24] Gail Mason and Rachael Stanic. Reporting and recording bias crime in new south wales. *Current Issues in Criminal Justice*, 31(2):164–180, April 2019. ISSN 1034-5329. doi: 10.1080/10345329.2019.1594920.

[25] Amalia R Miller and Carmit Segal. Do female officers improve law enforcement quality? effects on crime reporting and domestic violence. *Rev. Econ. Stud.*, 86(5):2220–2247, October 2019. ISSN 0034-6527. doi: 10.1093/restud/rdy051.

[26] Kristian Lum and William Isaac. To predict and serve? *Significance*, 13(5): 14–19, October 2016. ISSN 1740-9705, 1740-9713. doi: 10.1111/j.1740-9713.2016.00960.x.

[27] Man Wang, Rachel Garshick Kleit, Jane Cover, and Christopher S Fowler. Spatial variations in US poverty: beyond metropolitan and non-metropolitan. *Urban Stud.*, 49(3):563–585, 2012. ISSN 0042-0980. doi: 10.1177/0042098011404932.

[28] Genevieve Giuliano and Kenneth A Small. Subcenters in the los angeles region. *Reg. Sci. Urban Econ.*, 21(2):163–182, July 1991. ISSN 0166-0462. doi: 10.1016/0166-0462(91)90032-I.

[29] Elizabeth Burton. The compact city: Just or just compact? a preliminary analysis. *Urban Stud.*, 37(11):1969–2006, October 2000. ISSN 0042-0980. doi: 10.1080/00420980050162184.

[30] Tim Schwanen, Frans M Dieleman, and Martin Dijst. Travel behaviour in dutch monocentric and policentric urban systems. *J. Transp. Geogr.*, 9(3):173–186, September 2001. ISSN 0966-6923. doi: 10.1016/S0966-6923(01)00009-6.

[31] Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang. Urban computing: Concepts, methodologies, and applications. *ACM Trans. Intell. Syst. Technol.*, 5(3):1–55, September 2014. ISSN 2157-6904. doi: 10.1145/2629592.

[32] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *nature*, 453(7196):779–782, 2008.

[33] Shao-Meng Qin, Hannu Verkasalo, Mikael Mohtaschemi, Tuomo Hartonen, and Mikko Alava. Patterns, entropy, and predictability of human mobility and life. *PLoS One*, 7(12):e51353, December 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0051353.

[34] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, February 2010. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1177170.

[35] Trinh Minh Tri Do and Daniel Gatica-Perez. Where and what: Using smartphones to predict next locations and applications in daily life. *Pervasive Mob. Comput.*, 12:79–91, June 2014. ISSN 1574-1192. doi: 10.1016/j.pmcj.2013.03.006.

[36] Basmah Altaf, Lu Yu, and Xiangliang Zhang. Spatio-Temporal attention based recurrent neural network for next location prediction. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 937–942, December 2018. doi: 10.1109/BigData.2018.8622218.

[37] Vanessa Frias-Martinez, Jesus Virseda-Jerez, and Enrique Frias-Martinez. On the relation between socio-economic status and physical mobility. *Information Technology for Development*, 18(2):91–106, 2012.

[38] Lingzi Hong, Jiahui Wu, Enrique Frias-Martinez, Andrés Villarreal, and Vanessa Frias-Martinez. Characterization of internal migrant behavior in the immediate post-migration period using cell phone traces. In *Proceedings of the Tenth International Conference on Information and Communication Technologies and Development*, pages 1–12, 2019.

[39] Pierre Deville, Catherine Linard, Samuel Martin, Marius Gilbert, Forrest R Stevens, Andrea E Gaughan, Vincent D Blondel, and Andrew J Tatem. Dynamic population mapping using mobile phone data. *Proc. Natl. Acad. Sci. U. S. A.*, 111(45):15888–15893, November 2014. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1408439111.

[40] Marco Hernandez, Lingzi Hong, Vanessa Frias-Martinez, and Enrique Frias-Martinez. *Estimating poverty using cell phone data: evidence from Guatemala*. The World Bank, 2017.

[41] Joshua Blumenstock, Gabriel Cadamuro, and Robert On. Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264):1073–1076, November 2015. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aac4420.

[42] Nathan Eagle, Michael Macy, and Rob Claxton. Network diversity and economic development. *Science*, 328(5981):1029–1031, 2010.

[43] Vanessa Frias-Martinez, Victor Soto, Jesus Virseda, and Enrique Frias-Martinez. Computing cost-effective census maps from cell phone traces. In *Workshop on pervasive urban applications*, 2012.

[44] Lingzi Hong, Enrique Frias-Martinez, and Vanessa Frias-Martinez. Topic models to infer socio-economic maps. In *Thirtieth AAAI Conference on Artificial Intelligence*. aaai.org, 2016.

[45] Florent Le Néchet. Urban spatial structure, daily mobility and energy consumption: a study of 34 european cities. *Cybergeo*, January 2012. ISSN 1278-3366. doi: 10.4000/cybergeo.24966.

[46] Reid H Ewing. Characteristics, causes, and effects of sprawl: A literature review. In John M Marzluff, Eric Shulenberger, Wilfried Endlicher, Marina Alberti, Gordon Bradley, Clare Ryan, Ute Simon, and Craig ZumBrunnen, editors, *Urban Ecology: An International Perspective on the Interaction Between Humans and Nature*, pages 519–535. Springer US, Boston, MA, 2008. ISBN 9780387734125.

[47] Weipan Xu, Haohui Chen, Enrique Frias-Martinez, Manuel Cebrian, and Xun Li. The inverted u-shaped effect of urban hotspots spatial compactness on urban economic growth. *R Soc Open Sci*, 6(11):181640, November 2019. ISSN 2054-5703.

[48] Jingnan Huang, Xi X Lu, and Jefferey M Sellers. A global comparative analysis of urban form: Applying spatial metrics and remote sensing. *Landsc. Urban Plan.*, 82(4):184–197, October 2007. ISSN 0169-2046. doi: 10.1016/j.landurbplan.2007.02.010.

[49] Sebastián A Ríos and Ricardo Muñoz. Land use detection with cell phone data using topic models: Case santiago, chile. *Comput. Environ. Urban Syst.*, 61:39–48, January 2017. ISSN 0198-9715. doi: 10.1016/j.compenvurbsys.2016.08.007.

[50] Fei Wu, Zhenhui Li, Wang-Chien Lee, Hongjian Wang, and Zhuojie Huang. Semantic annotation of mobility data using social media. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1253–1263, 2015.

[51] Carmen Vaca, Daniele Quercia, Francesco Bonchi, and Piero Fraternali. Taxonomy-based discovery and annotation of functional areas in the city. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 9, 2015.

[52] Di Chai, Leye Wang, and Qiang Yang. Bike flow prediction with multi-graph convolutional networks. In *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '18, pages 397–400, New York, NY, USA, November 2018. Association for Computing Machinery. ISBN 9781450358897. doi: 10.1145/3274895.3274896.

[53] Chao Zhang, Keyang Zhang, Quan Yuan, Haoruo Peng, Yu Zheng, Tim Hanratty, Shaowen Wang, and Jiawei Han. Regions, periods, activities: Uncovering urban dynamics via cross-modal representation learning. In *Proceedings of the 26th International Conference on World Wide Web*, pages 361–370. dl.acm.org, 2017.

[54] Kun Ouyang, Yuxuan Liang, Ye Liu, Zekun Tong, Sijie Ruan, David Rosenblum, and Yu Zheng. Fine-grained urban flow inference. *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[55] Robin Wilson, Elisabeth zu Erbach-Schoenberg, Maximilian Albert, Daniel Power, Simon Tudge, Miguel Gonzalez, Sam Guthrie, Heather Chamberlain, Christopher Brooks, Christopher Hughes, et al. Rapid and near real-time assessments of population displacement using mobile phone data following disasters: The 2015 nepal earthquake. *PLoS currents*, 8, 2016.

[56] Boyeong Hong, Bartosz J Bonczak, Arpit Gupta, and Constantine E Kontokosta. Measuring inequality in community resilience to natural disasters using large-scale mobility data. *Nature communications*, 12(1):1–9, 2021.

[57] David Weisburd, Elizabeth R Groff, and Sue-Ming Yang. *The Criminology of Place: Street Segments and Our Understanding of the Crime Problem*. Oxford University Press, October 2012. ISBN 9780199709106.

[58] Shane D Johnson and Kate J Bowers. Near repeats and crime forecasting. In Gerben Bruinsma and David Weisburd, editors, *Encyclopedia of Criminology and Criminal Justice*, pages 3242–3254. Springer New York, New York, NY, 2014. ISBN 9781461456902. doi: 10.1007/978-1-4614-5690-2_210.

[59] David Weisburd, Shawn Bushway, Cynthia Lum, and Sue-Ming Yang. Trajectories of crime at places: A longitudinal study of street segments in the city of seattle. *Criminology*, 2004. ISSN 0011-1384.

[60] Kate J Bowers, Shane D Johnson, and Ken Pease. Prospective Hot-Spotting: The future of crime mapping? *Br. J. Criminol.*, 44(5):641–658, September 2004. ISSN 0007-0955. doi: 10.1093/bjc/azh036.

[61] George O Mohler, Martin B Short, P Jeffrey Brantingham, Frederic Paik Schoenberg, and George E Tita. Self-Exciting point process modeling of crime. *J. Am. Stat. Assoc.*, 106(493):100–108, March 2011. ISSN 0162-1459. doi: 10.1198/jasa.2011.ap09546.

[62] Bao Wang, Duo Zhang, Duanhao Zhang, P Jeffery Brantingham, and Andrea L Bertozzi. Deep learning for real time crime forecasting. July 2017.

[63] Huaxiu Yao, Yiding Liu, Ying Wei, Xianfeng Tang, and Zhenhui Li. Learning from multiple cities: A Meta-Learning approach for Spatial-Temporal prediction. In *The World Wide Web Conference*, WWW '19, pages 2181–2191, New York, NY, USA, May 2019. Association for Computing Machinery. ISBN 9781450366748. doi: 10.1145/3308558.3313577.

[64] Kate Bowers. Risky facilities: Crime radiators or crime absorbers? a comparison of internal and external levels of theft. *J. Quant. Criminol.*, 30(3):389–414, September 2014. ISSN 0748-4518, 1573-7799. doi: 10.1007/s10940-013-9208-z.

[65] Dennis M Gorman, Paul W Speer, Paul J Gruenewald, and Erich W Labouvie. Spatial dynamics of alcohol availability, neighborhood structure and violent crime. *J. Stud. Alcohol*, 62(5):628–636, September 2001. ISSN 0096-882X. doi: 10.15288/jsa.2001.62.628.

[66] Matthew Quick, Jane Law, and Guangquan Li. Time-varying relationships between land use and crime: A spatio-temporal analysis of small-area seasonal property crime trends. *Environment and Planning B: Urban Analytics and City Science*, 46(6):1018–1035, July 2019. ISSN 2399-8083. doi: 10.1177/2399808317744779.

[67] Matthew Ranson. Crime, weather, and climate change. *J. Environ. Econ. Manage.*, 67(3):274–302, May 2014. ISSN 0095-0696. doi: 10.1016/j.jeem.2013.11.008.

[68] Christopher R Browning, Nicolo P Pinchak, and Catherine A Calder. Human mobility and crime: Theoretical approaches and novel data collection strategies. *Annu. Rev. Criminol.*, January 2021. ISSN 2572-4568. doi: 10.1146/annurev-criminol-061020-021551.

[69] Ronald V Clarke. Opportunity makes the thief. really? and so what? *Crime Science*, 1(1):3, December 2012. ISSN 2193-7680. doi: 10.1186/2193-7680-1-3.

[70] Marcus Felson and Rémi Boivin. Daily crime flows within a city. *Crime Science*, 4(1):31, October 2015. ISSN 2193-7680. doi: 10.1186/s40163-015-0039-0.

[71] Steven D Levitt. The relationship between crime reporting and police: Implications for the use of uniform crime reports. *J. Quant. Criminol.*, 14(1):61–81, March 1998. ISSN 0748-4518, 1573-7799. doi: 10.1023/A:1023096425367.

[72] Timothy C Hart and Callie Marie Rennison. Reporting crime to the police, 1992-2000. Technical report, 2003.

[73] A Keith Bottomley and Clive Coleman. *Understanding crime rates: Police and public roles in the production of official statistics*. Gower Publishing, UK, 1981.

[74] Jennifer Cole and Alexandra Stickings. The future of crime reporting. *The RUSI Journal*, 162(1):68–78, January 2017. ISSN 0307-1847. doi: 10.1080/03071847.2017.1301640.

[75] Kevin J Strom and Erica L Smith. The future of crime data: The case for the national Incident-Based reporting system (NIBRS) as a primary data source for policy evaluation and crime analysis. *Criminol. Public Policy*, 16(4):1027–1048, November 2017. ISSN 1538-6473. doi: 10.1111/1745-9133.12336.

[76] Bureau of Justice Statistics. Data collection: National crime victimization survey (ncvs). `https://www.inegi.org.mx/programas/envipe/2014/`, 2020. Accessed: 2020-4-17.

[77] Jiahui Wu and Vanessa Frias-Martinez. An analysis of the relationship between crime incidents and 911 calls. *Proc. Assoc. Info. Sci. Tech.*, 55(1):933–935, January 2018. ISSN 2373-9231, 2373-9231. doi: 10.1002/pra2.2018.14505501182.

[78] Sean P Varano, Joseph A Schafer, Jeffrey Michael Cancino, and Marc L Swatt. Constructing crime: Neighborhood characteristics and police recording behavior. *J. Crim. Justice*, 37(6):553–563, November 2009. ISSN 0047-2352. doi: 10.1016/j.jcrimjus.2009.09.004.

[79] ENVIPE. National survey of victimization and perception of public security. `https://www.inegi.org.mx/programas/envipe/2014/`, 2014. Accessed: 2020-4-17.

[80] Barbara D Warner. Community characteristics and the recording of crime: Police recording of citizens' complaints of burglary and assault. *Justice Q.*, 14(4):631–650, December 1997. ISSN 0741-8825. doi: 10.1080/07418829700093531.

[81] Ziggy MacDonald. The impact of under-reporting on the relationship between unemployment and property crime. *Appl. Econ. Lett.*, 7(10):659–663, October 2000. ISSN 1350-4851. doi: 10.1080/135048500415978.

[82] Alexandra Chouldechova and Aaron Roth. The frontiers of fairness in machine learning. October 2018.

[83] Indre Zliobaite. A survey on measuring indirect discrimination in machine learning. October 2015.

[84] Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pages 1–7. ieeexplore.ieee.org, May 2018. doi: 10.23919/FAIRWARE.2018.8452913.

[85] Toon Calders, Asim Karim, Faisal Kamiran, Wasif Ali, and Xiangliang Zhang. Controlling attribute effect in linear regression. In *2013 IEEE 13th International Conference on Data Mining*, pages 71–80, December 2013. doi: 10.1109/ICDM.2013.114.

[86] Jack Fitzsimons, Abdulrahman Al Ali, Michael Osborne, and Stephen Roberts. A general framework for fair regression. *Entropy*, 21(8):741, July 2019. doi: 10.3390/e21080741.

[87] An Yan and Bill Howe. FairST: Equitable spatial and temporal demand prediction for new mobility systems. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '19, pages 552–555, New York, NY, USA, November 2019. Association for Computing Machinery. ISBN 9781450369091. doi: 10.1145/3347146.3359380.

[88] Junpei Komiyama, Akiko Takeda, Junya Honda, and Hajime Shimao. Nonconvex optimization for regression with fairness constraints. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2737–2746, Stockholmsmässan, Stockholm Sweden, 2018. PMLR.

[89] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. A convex framework for fair regression. June 2017.

[90] Danielle Ensign, Sorelle A Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. Runaway feedback loops in predictive policing. In Sorelle A Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 160–171, New York, NY, USA, 2018. PMLR.

[91] P Jeffrey Brantingham, Matthew Valasik, and George O Mohler. Does predictive policing lead to biased arrests? results from a randomized controlled trial. *Statistics and Public Policy*, 5(1):1–6, January 2018. doi: 10.1080/2330443X.2018.1438940.

[92] George O Mohler, Martin B Short, Sean Malinowski, Mark Johnson, George E Tita, Andrea L Bertozzi, and P Jeffrey Brantingham. Randomized controlled field trials of predictive policing. *J. Am. Stat. Assoc.*, 110(512):1399–1411, October 2015. ISSN 0162-1459. doi: 10.1080/01621459.2015.1077710.

[93] Lian Duan, Tao Hu, En Cheng, Jianfeng Zhu, and Chao Gao. Deep convolutional neural networks for spatiotemporal crime prediction. In *Proceedings of the International Conference on Information and Knowledge Engineering (IKE)*, pages 61–67. csce.ucmss.com, 2017.

[94] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.*, 22(10):1345–1359, October 2010. ISSN 1041-4347, 1558-2191. doi: 10.1109/TKDE.2009.191.

[95] Tao Xu, Yutao Ma, and Qian Wang. Cross-Urban Point-of-Interest recommendation for Non-Natives. *Int. J. Web Serv. Res.*, 2018.

[96] Jingtao Ding, Guanghui Yu, Yong Li, Depeng Jin, and Hui Gao. Learning from hometown and current city: Cross-city POI recommendation via interest drift and transfer learning. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 3(4):1–28, December 2019. doi: 10.1145/3369822.

[97] Zipei Fan, Xuan Song, Ryosuke Shibasaki, Tao Li, and Hodaka Kaneda. CityCoupling: bridging intercity human mobility. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '16, pages 718–728, New York, NY, USA, September 2016. Association for Computing Machinery. ISBN 9781450344616. doi: 10.1145/2971648.2971737.

[98] Tianfu He, Jie Bao, Ruiyuan Li, Sijie Ruan, Yanhua Li, Li Song, Hui He, and Yu Zheng. What is the human mobility in a new city: Transfer mobility knowledge across cities. In *Proceedings of The Web Conference 2020*, WWW '20, pages 1355–1365, New York, NY, USA, April 2020. Association for Computing Machinery. ISBN 9781450370233. doi: 10.1145/3366423.3380210.

[99] Zhaoyang Liu, Yanyan Shen, and Yanmin Zhu. Inferring dockless shared bike distribution in new cities. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, pages 378–386, New York, NY, USA, February 2018. Association for Computing Machinery. ISBN 9781450355810. doi: 10.1145/3159652.3159708.

[100] Leye Wang, Xu Geng, Xiaojuan Ma, Feng Liu, and Qiang Yang. Cross-city transfer learning for deep spatio-temporal prediction. In *IJCAI International Joint Conference on Artificial Intelligence*, page 1893, 2019.

[101] Tanwi Mallick, Prasanna Balaprakash, E Rask, and J F MacFarlane. Transfer learning with graph neural networks for Short-Term highway traffic forecasting. *ArXiv*, 2020.

[102] Bill Y Lin, Frank F Xu, Eve Q Liao, and Kenny Q Zhu. Transfer learning for traffic speed prediction: A preliminary study. In *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*. aaai.org, June 2018.

[103] Marcos R Vieira, Vanessa Frias-Martinez, Nuria Oliver, and Enrique Frias-Martinez. Characterizing dense urban areas from mobile Phone-Call data: Discovery and social dynamics. In *2010 IEEE Second International Conference on Social Computing*, pages 241–248, August 2010. doi: 10.1109/SocialCom.2010.41.

[104] Sahar Hoteit, Stefano Secci, Stanislav Sobolevsky, Carlo Ratti, and Guy Pujolle. Estimating human trajectories and hotspots through mobile phone data. *Computer Networks*, 64:296–307, May 2014. ISSN 1389-1286.

[105] Xiaoqing Zuo and Yongchuan Zhang. Detection and analysis of urban area hotspots based on cell phone traffic. *JCP*, 7(7):1753–1760, 2012.

[106] Mohammadhossein Ghahramani, MengChu Zhou, and Chi Tin Hon. Mobile phone data analysis: A spatial exploration toward hotspot detection. *IEEE Transactions on Automation Science and Engineering*, 16(1):351–362, 2018.

[107] Andrey Bogomolov, Bruno Lepri, Jacopo Staiano, Emmanuel Letouzé, Nuria Oliver, Fabio Pianesi, and Alex Pentland. Moves on the street: Classifying crime hotspots using aggregated anonymized data on people dynamics. *Big data*, 3(3):148–158, 2015.

[108] Martin Traunmueller, Giovanni Quattrone, and Licia Capra. Mining mobile phone data to investigate urban crime theories at scale. In *International Conference on Social Informatics*, pages 396–411. Springer, 2014.

[109] Sibren Isaacman, Richard Becker, Ramón Cáceres, Margaret Martonosi, James Rowland, Alexander Varshavsky, and Walter Willinger. Human mobility modeling at metropolitan scales. In *Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services*, MobiSys '12, pages 239–252, New York, NY, USA, 2012. ACM. ISBN 9781450313018. doi: 10.1145/2307636.2307659.

[110] Richard A Becker, Ramon Caceres, Karrie Hanson, Ji Meng Loh, Simon Urbanek, Alexander Varshavsky, and Chris Volinsky. A tale of one city: Using cellular network data for urban planning. *IEEE Pervasive Comput.*, 10(4): 18–26, April 2011. ISSN 1536-1268. doi: 10.1109/MPRV.2011.44.

[111] Demographia. Definition of urban terms. `http://demographia.com/db-define.pdf`, 2020. Accessed: 2020-01-17.

[112] Roberto Trasarti, Ana-Maria Olteanu-Raimond, Mirco Nanni, Thomas Couronné, Barbara Furletti, Fosca Giannotti, Zbigniew Smoreda, and Cezary Ziemlicki. Discovering urban and country dynamics from mobile phone data with spatial correlation patterns. *Telecomm. Policy*, 39(3):347–362, May 2015. ISSN 0308-5961.

[113] Song Gao. Spatio-Temporal analytics for exploring human mobility patterns and urban dynamics in the mobile age. *Spat. Cogn. Comput.*, 15(2):86–114, April 2015. ISSN 1387-5868. doi: 10.1080/13875868.2014.984300.

[114] Rex W Douglass, David A Meyer, Megha Ram, David Rideout, and Dongjin Song. High resolution population estimates from telecommunications data. *EPJ Data Science*, 4(1):4, May 2015. ISSN 2193-1127.

[115] Jonathan Reades, Francesco Calabrese, and Carlo Ratti. Eigenplaces: Analysing cities using the Space–Time structure of the mobile phone network. *Environ. Plann. B Plann. Des.*, 36(5):824–836, October 2009. ISSN 0265-8135. doi: 10.1068/b34133t.

[116] Wei Tu, Jinzhou Cao, Yang Yue, Shih-Lung Shaw, Meng Zhou, Zhensheng Wang, Xiaomeng Chang, Yang Xu, and Qingquan Li. Coupling mobile phone and social media data: a new approach to understanding urban functions and diurnal patterns. *Int. J. Geogr. Inf. Sci.*, 31(12):2331–2358, December 2017. ISSN 1365-8816.

[117] John Doyle, Peter Hung, Ronan Farrell, and Seán McLoone. Population mobility dynamics estimated from mobile telephony data. *Journal of Urban Technology*, 21(2):109–132, April 2014. ISSN 1063-0732. doi: 10.1080/10630732.2014.888904.

[118] Oscar F Peredo, José A García, Ricardo Stuven, and Julián M Ortiz. Urban dynamic estimation using mobile phone logs and locally varying anisotropy. In J Jaime Gómez-Hernández, Javier Rodrigo-Ilarri, María Elena Rodrigo-Clavero, Eduardo Cassiraga, and José Antonio Vargas-Guzmán, editors, *Geostatistics Valencia 2016*, pages 949–964. Springer International Publishing, Cham, 2017. ISBN 9783319468198. doi: 10.1007/978-3-319-46819-8_66.

[119] Petr Kubíček, Milan Konečný, Zdeněk Stachoň, Jie Shen, Lukáš Herman, Tomáš Řezník, Karel Staněk, Radim Štampach, and Šimon Leitgeb. Population distribution modelling at fine spatio-temporal scale based on mobile phone data. *International Journal of Digital Earth*, 12(11):1319–1340, November 2019. ISSN 1753-8947.

[120] Shlomo Angel, Jason Parent, and Daniel L Civco. Ten compactness properties of circles: measuring shape in geography. *The Canadian Geographer / Le Géographe canadien*, 54(4):441–461, December 2010. ISSN 0008-3658. doi: 10.1111/j.1541-0064.2009.00304.x.

[121] Martin Ouředníček, Jiří Nemeškal, Petra Špačková, Martin Hampl, and Jakub Novák. A synthetic approach to the delimitation of the prague metropolitan area. *J. Maps*, 14(1):26–33, January 2018.

[122] Claudio Gariazzo, Armando Pelliccioni, and Maria Paola Bogliolo. Spatiotemporal analysis of urban mobility using aggregate mobile phone derived presence and demographic data: A case study in the city of rome, italy. *Brown Univ. Dig. Addict. Theory Appl.*, 4(1):8, January 2019. ISSN 1040-6328.

[123] Jie Chen, Tao Pei, Shih-Lung Shaw, Feng Lu, Mingxiao Li, Shifen Cheng, Xiliang Liu, and Hengcai Zhang. Fine-grained prediction of urban population using mobile phone location data. *Int. J. Geogr. Inf. Sci.*, 32(9):1770–1786, September 2018. ISSN 1365-8816.

[124] Nina Schwarz. Urban form revisited—selecting indicators for characterising european cities. *Landsc. Urban Plan.*, 96(1):29–47, May 2010. ISSN 0169-2046.

[125] Chaogui Kang, Yu Liu, Xiujun Ma, and Lun Wu. Towards estimating urban population distributions from mobile call data. *Journal of Urban Technology*, 19(4):3–21, October 2012. ISSN 1063-0732.

[126] Alberto Rubio, Angel Sanchez, and Enrique Frias-Martinez. Adaptive non-parametric identification of dense areas using cell phone records for urban analysis. *Eng. Appl. Artif. Intell.*, 26(1):551–563, January 2013. ISSN 0952-1976.

[127] Johannes Scholz, Michael Andorfer, and Manfred Mittlboeck. Spatial accuracy evaluation of population density grid disaggregations with corine landcover. In Danny Vandenbroucke, Bénédicte Bucher, and Joep Crompvoets, editors, *Geographic Information Science at the Heart of Europe*, pages 267–283. Springer International Publishing, Cham, 2013. ISBN 9783319006154.

[128] Robert W Wassmer. Urban sprawl in a us metropolitan area: ways to measure and a comparison of the sacramento area to similar metropolitan areas in california and the us. *CSUS Public Policy and Administration Working Paper*, (2000-03), 2000.

[129] Wenwen Li, Michael F Goodchild, and Richard Church. An efficient measure of compactness for two-dimensional shapes and its application in regionalization problems. *Int. J. Geogr. Inf. Sci.*, 27(6):1227–1250, June 2013. ISSN 1365-8816. doi: 10.1080/13658816.2012.752093.

[130] Wenwen Li, Tingyong Chen, Elizabeth A Wentz, and Chao Fan. NMMI: A mass compactness measure for spatial pattern analysis of areal features. *Ann. Assoc. Am. Geogr.*, 104(6):1116–1133, November 2014. ISSN 0004-5608. doi: 10.1080/00045608.2014.941732.

[131] Statstutor. Spearman's correlation. `http://www.statstutor.ac.uk/resources/uploaded/spearmans.pdf`, 2020. Accessed: 2020-1-17.

[132] CONAPO. Delimitación de zonas metropolitanas. `http://www.conapo.gob.mx/es/CONAPO/Datos_Abiertos_Delimitacion_de_Zonas_Metropolitanas`, 2015. Accessed: 2020-01-17.

[133] INEGI. Colección: Cartografía geoestadística urbana, cierre del censo de población y vivienda 2010. `https://www.inegi.org.mx/app/mapas/?t=0710000000000000&tg=3604`, 2010. Accessed: 2020-1-17.

[134] Charlie Catlett, Eugenio Cesario, Domenico Talia, and Andrea Vinci. Spatio-temporal crime predictions in smart cities: A data-driven approach and experiments. *Pervasive Mob. Comput.*, 53:62–74, February 2019. ISSN 1574-1192. doi: 10.1016/j.pmcj.2019.01.003.

[135] Chao Huang, Junbo Zhang, Yu Zheng, and Nitesh V Chawla. DeepCrime: Attentive hierarchical recurrent networks for crime prediction. In *Proceedings*

of the 27th ACM International Conference on Information and Knowledge Management, CIKM '18, pages 1423–1432, New York, NY, USA, October 2018. Association for Computing Machinery. ISBN 9781450360142. doi: 10.1145/3269206.3271793.

[136] Hongjian Wang, Daniel Kifer, Corina Graif, and Zhenhui Li. Crime rate inference with big data. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 635–644, New York, NY, USA, 2016. ACM. ISBN 9781450342322. doi: 10.1145/2939672.2939736.

[137] Sarah White, Tobin Yehle, Hugo Serrano, Marcos Oliveira, and Ronaldo Menezes. The spatial structure of crime in urban environments. In *International Conference on Social Informatics*, pages 102–111. Springer, 2014.

[138] Keith Kirkpatrick. It's not the algorithm, it's the data. *Commun. ACM*, 60 (2):21–23, January 2017. ISSN 0001-0782. doi: 10.1145/3022181.

[139] Elías Moreno and Javier Girón. Estimating with incomplete count data a bayesian approach. *J. Stat. Plan. Inference*, 66(1):147–159, January 1998. ISSN 0378-3758. doi: 10.1016/S0378-3758(97)00073-6.

[140] Oliver Stoner, Theo Economou, and Gabriela Drummond Marques da Silva. A hierarchical framework for correcting Under-Reporting in count data. *J. Am. Stat. Assoc.*, pages 1–17, March 2019. ISSN 0162-1459. doi: 10.1080/01621459.2019.1573732.

[141] Gabriel Ferreyra-Orozco. Race, ethnicity, crime and criminal justice in mexico. In Anita Kalunta-Crumpton, editor, *Race, Ethnicity, Crime and Criminal Justice in the Americas*, pages 169–191. Palgrave Macmillan UK, London, 2012. ISBN 9780230355866. doi: 10.1057/9780230355866_8.

[142] SESNSP. Datos abiertos de incidencia delictiva. `https://www.gob.mx/sesnsp/acciones-y-programas/datos-abiertos-de-incidencia-delictiva`, 2011. Accessed: 2020-4-17.

[143] CONEVAL. MediciÓn de la pobreza. `https://www.coneval.org.mx/Medicion/MP/Paginas/Medicion-de-la-pobreza-municipal-2010.aspx`, 2010. Accessed: 2020-4-17.

[144] INEGI. 2010 census of population and housing units. `https://www.coneval.org.mx/Medicion/MP/Paginas/Medicion-de-la-pobreza-municipal-2010.aspx`, 2010. Accessed: 2020-4-17.

[145] Perry de Valpine, Daniel Turek, Christopher J Paciorek, Clifford Anderson-Bergman, Duncan Temple Lang, and Rastislav Bodik. Programming with

models: Writing statistical algorithms for general model structures with NIMBLE. *J. Comput. Graph. Stat.*, 26(2):403–413, April 2017. ISSN 1061-8600. doi: 10.1080/10618600.2016.1172487.

[146] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[147] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785. URL `http://doi.acm.org/10.1145/2939672.2939785`.

[148] Stephen P Brooks and Andrew Gelman. General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, 7(4):434–455, 1998.

[149] Shay Cohen. *Bayesian Analysis in Natural Language Processing*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool, California, second edition edition, 2019. ISBN 9781681735276. doi: 10.2200/S00905ED2V01Y201903HLT041.

[150] David Evans, Nicholas Fyfe, and David Herbert. *Crime, Policing and Place: Essays in Environmental Criminology*. Taylor & Francis, January 2002. ISBN 9780203007860. doi: 10.4324/9780203007860.

[151] Patricia Brantingham and Paul Brantingham. Criminality of place: Crime generators and crime attractors. *European Journal on Criminal Policy and Research*, 3(3):5–26, September 1995. ISSN 1572-9869. doi: 10.1007/BF02242925.

[152] Chao Huang, Chuxu Zhang, Jiashu Zhao, Xian Wu, Dawei Yin, and Nitesh Chawla. MiST: A multiview and multimodal Spatial-Temporal learning framework for citywide abnormal event forecasting. In *The World Wide Web Conference*, WWW '19, pages 717–728, New York, NY, USA, May 2019. Association for Computing Machinery. ISBN 9781450366748. doi: 10.1145/3308558.3313730.

[153] Jiahui Wu, Enrique Frias-Martinez, and Vanessa Frias-Martinez. Addressing Under-Reporting to enhance fairness and accuracy in mobility-based crime prediction. In *Proceedings of the 28th International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '20, pages 325–336, New York, NY, USA, November 2020. Association for Computing Machinery. ISBN 9781450380195. doi: 10.1145/3397536.3422205.

[154] Yuhao Kang, Song Gao, Yunlei Liang, Mingxiao Li, and Jake Kruse. Multiscale dynamic human mobility flow dataset in the u.s. during the covid-19 epidemic. *Scientific Data*, pages 1–13, 2020.

[155] Song Gao, Jinmeng Rao, Yuhao Kang, Yunlei Liang, and Jake Kruse. Mapping county-level mobility pattern changes in the united states in response to covid-19. *SIGSpatial Special*, 12(1):16–26, 2020.

[156] Rémi Boivin and Marcus Felson. Crimes by visitors versus crimes by residents: The influence of visitor inflows. *J. Quant. Criminol.*, 34(2):465–480, June 2018. ISSN 0748-4518, 1573-7799. doi: 10.1007/s10940-017-9341-1.

[157] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł Ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017.

[158] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186, 2019. doi: 10.18653/v1/N19-1423.

[159] Thomas N Kipf and Max Welling. Semi-Supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

[160] Oleksandr Shchur, Maximilian Mumme, Aleksandar Bojchevski, and Stephan Günnemann. Pitfalls of graph neural network evaluation. In *Relational Representation Learning Workshop*, 2018.

[161] Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-Temporal graph convolutional networks: A deep learning framework for traffic forecasting. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 3634–3640, California, July 2018. International Joint Conferences on Artificial Intelligence Organization. ISBN 9780999241127. doi: 10.24963/ijcai.2018/505.

[162] P Jeffrey Brantingham. The logic of data bias and its impact on place-based predictive policing. *Ohio St. J. Crim. L.*, 15:473, 2017.

[163] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent Trade-Offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, Leibniz International Proceedings in Informatics (LIPIcs), pages 1–23. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2017. doi: 10.4230/LIPIcs.ITCS.2017.43.

[164] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How we analyzed the compas recidivism algorithm, May 2016. URL https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

[165] Niki Kilbertus, Adrià Gascón, Matt J Kusner, Michael Veale, Krishna P Gummadi, and Adrian Weller. Blind justice: Fairness with encrypted sensitive attributes. June 2018.

[166] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A survey on deep transfer learning. In *Artificial Neural Networks and Machine Learning – ICANN 2018*, pages 270–279. Springer International Publishing, 2018. doi: 10.1007/978-3-030-01424-7_27.

[167] Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Sci Adv*, 4(1):eaao5580, January 2018. ISSN 2375-2548. doi: 10.1126/sciadv.aao5580.

[168] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1613–1622, Lille, France, 2015. PMLR.